MedLinker

Medical Entity Linking with Neural Representations and Dictionary Matching

*Daniel Loureiro and Alípio Mário Jorge

LIAAD INESCTEC LABORATÓRIO ASSOCIADO

U.PORTO

FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

RECAP preterm

ECIR 2020

Medical Entity Linking

- Medical literature is growing rapidly.
- This information is extremely important, but also hard to parse.
- Current SOTA of NLP can help with Entity Linking.
- Prior methods, such as dictionary matching, are still relevant.

We’ll show how SOTA NLP can benefit from dictionary matching, in this important task.

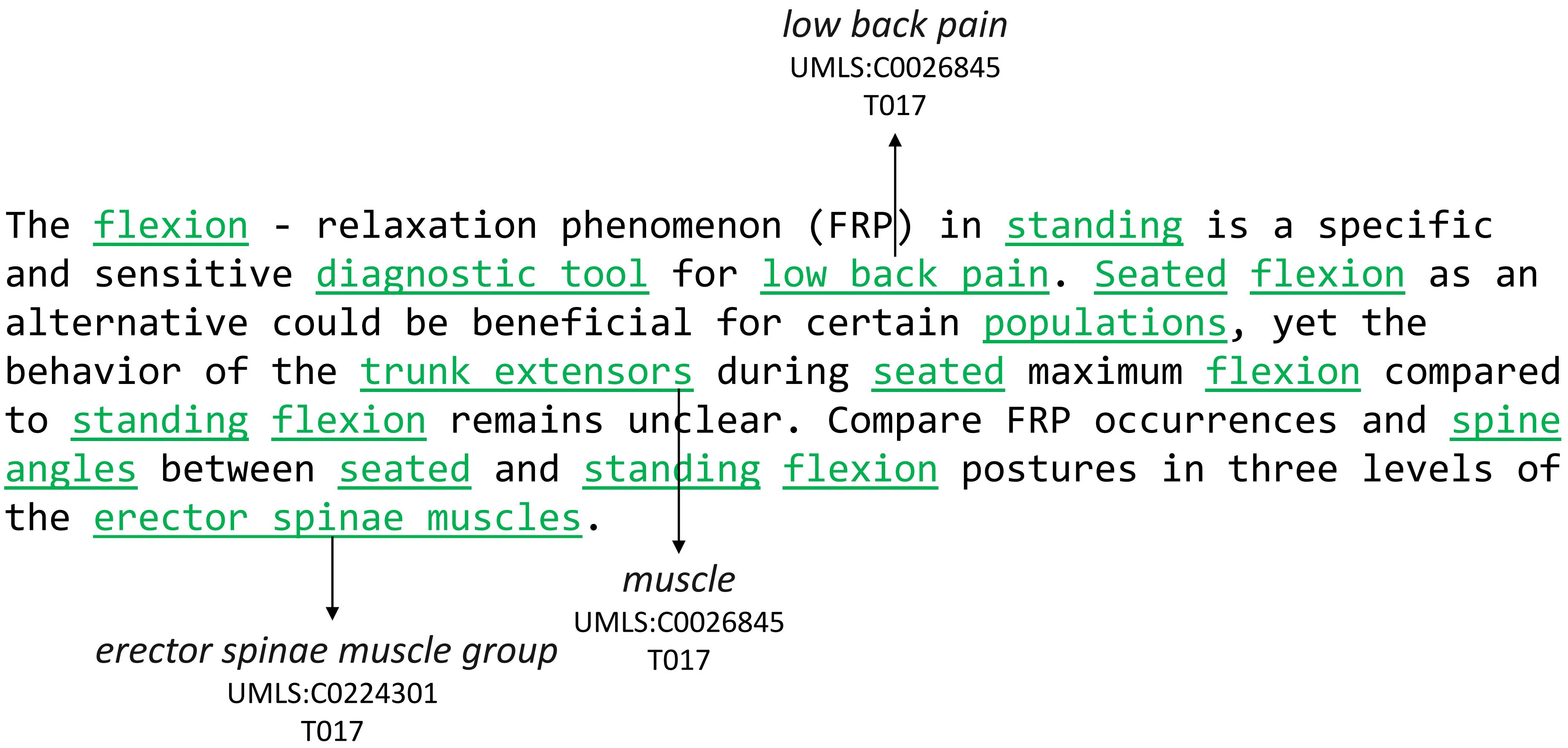
Defining Task

The flexion - relaxation phenomenon (FRP) in standing is a specific and sensitive diagnostic tool for low back pain. Seated flexion as an alternative could be beneficial for certain populations, yet the behavior of the trunk extensors during seated maximum flexion compared to standing flexion remains unclear. Compare FRP occurrences and spine angles between seated and standing flexion postures in three levels of the erector spinae muscles.

Defining Task

The flexion - relaxation phenomenon (FRP) in standing is a specific and sensitive diagnostic tool for low back pain. Seated flexion as an alternative could be beneficial for certain populations, yet the behavior of the trunk extensors during seated maximum flexion compared to standing flexion remains unclear. Compare FRP occurrences and spine angles between seated and standing flexion postures in three levels of the erector spinae muscles.

Defining Task



low back pain
UMLS:C0026845
T017

The flexion - relaxation phenomenon (FRP) in standing is a specific and sensitive diagnostic tool for low back pain. Seated flexion as an alternative could be beneficial for certain populations, yet the behavior of the trunk extensors during seated maximum flexion compared to standing flexion remains unclear. Compare FRP occurrences and spine angles between seated and standing flexion postures in three levels of the erector spinae muscles.

muscle
UMLS:C0026845
T017

erector spinae muscle group
UMLS:C0224301
T017

Challenges

- We want to use UMLS, the most comprehensive medical ontology.
  - 3M concepts compiled from mutliple sources (SNOMED, NCI, etc.)
  - Very broad, from medical occupations to biological molecules.

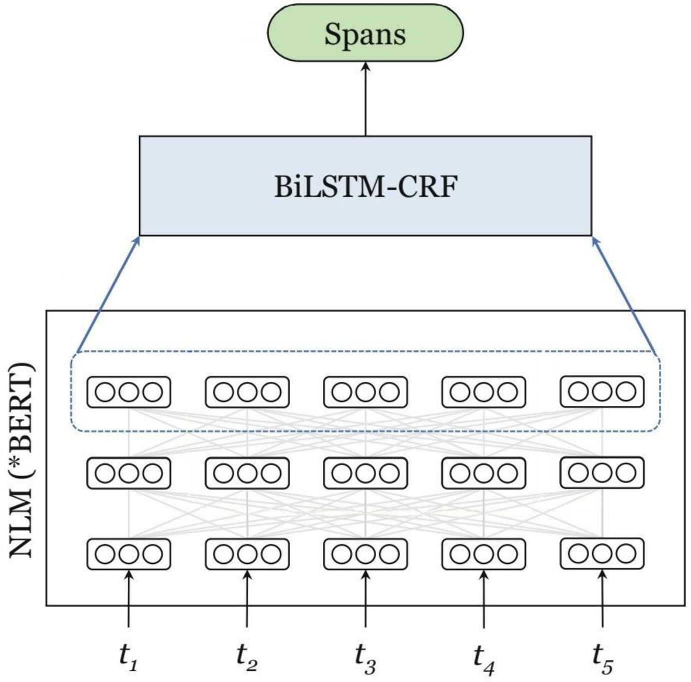
- The largest corpus with UMLS annotations is MedMentions [Mohan et Li, 2019].
  - 4,392 abstracts with 203k annotations (st21pv subset).
  - Covers 1% of concepts in UMLS.
  - Low overlap of concepts between train and test sets (57.5%).

Recognize Relevant Spans

The flexion - relaxation phenomenon (FRP) in standing is a specific and sensitive diagnostic tool for low back pain. Seated flexion as an alternative could be beneficial for certain populations, yet the behavior of the trunk extensors during seated maximum flexion compared to standing flexion remains unclear. Compare FRP occurrences and spine angles between seated and standing flexion postures in three levels of the erector spinae muscles.

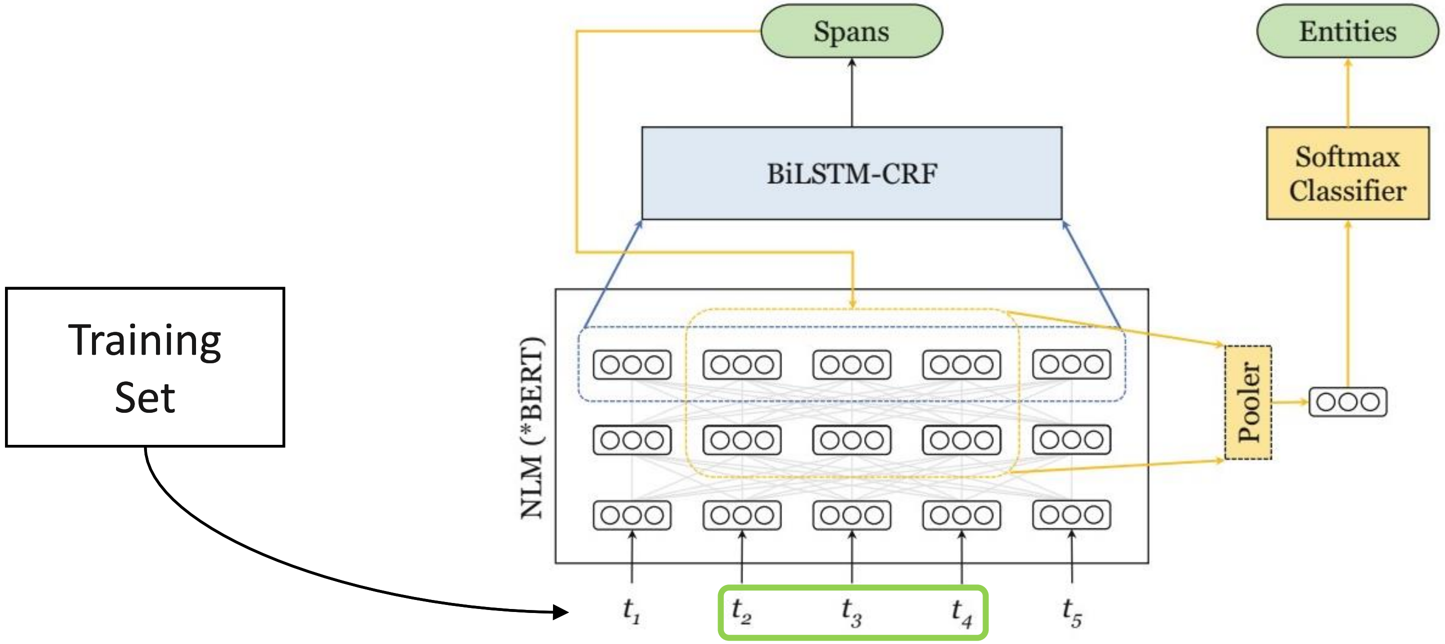
Named Entity Recognition (NER)

- Standard NER architecture, but using SOTA Neural Language Models (NLMs) trained on the medical domain.



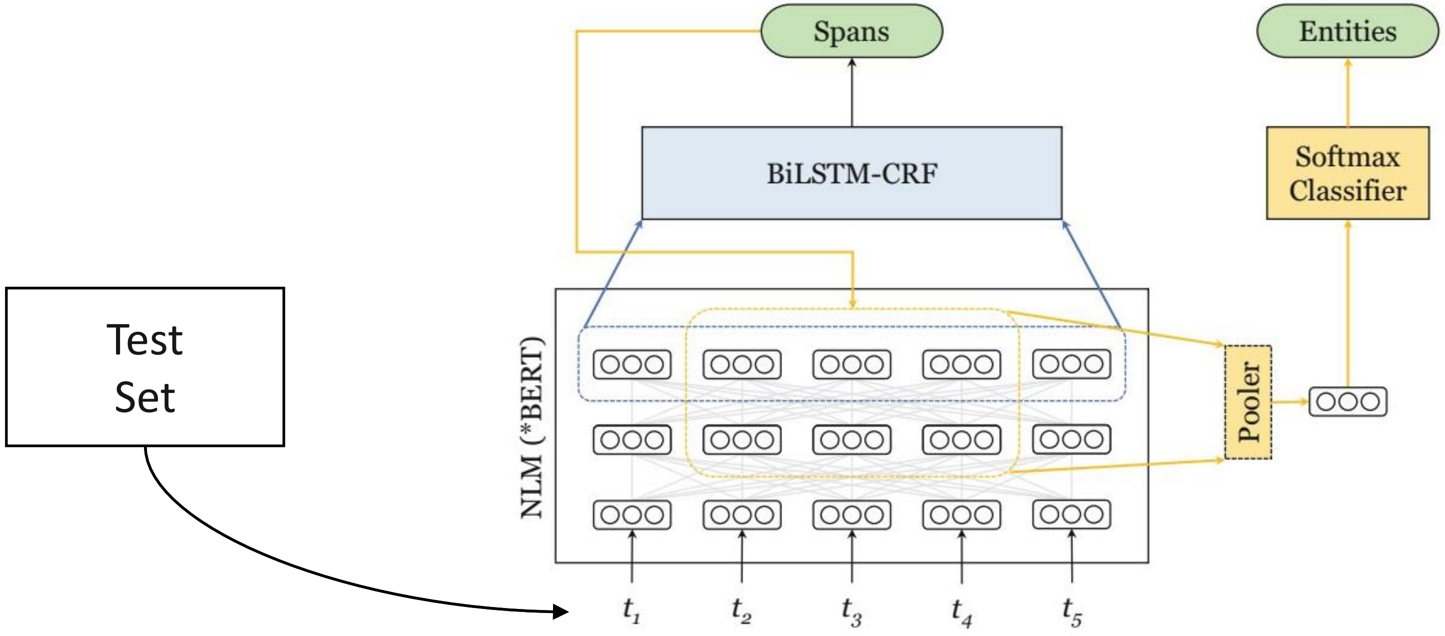
Contextual Embeddings

- Train a minimal Softmax classifier based on pooled internal states of a NLM. Also experimented with kNN, but less effective.



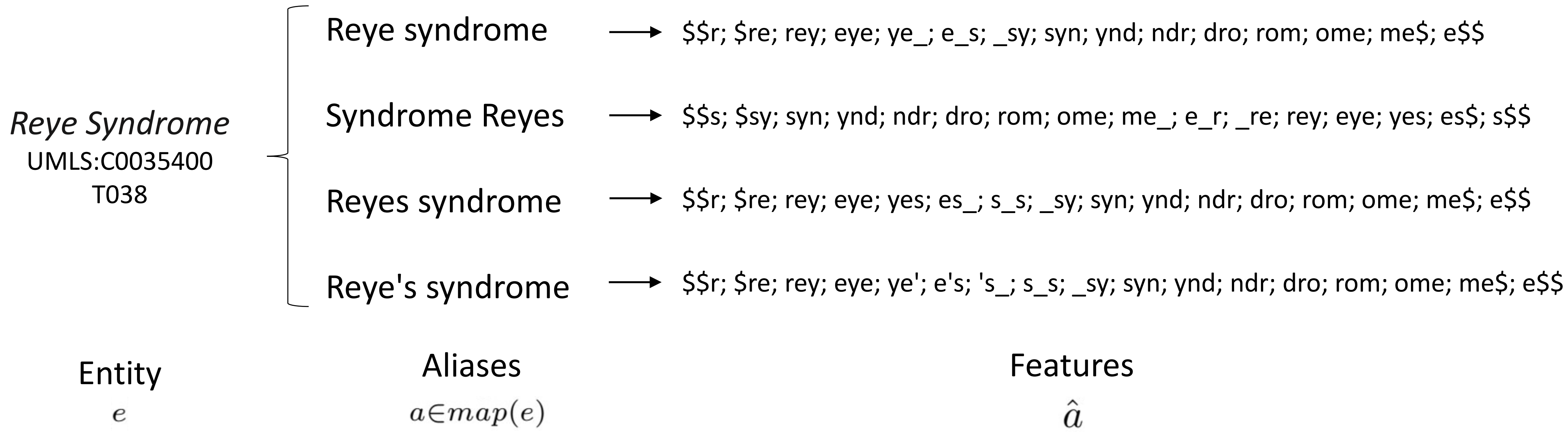
Matching Embeddings

- Inference is performed in three steps, re-using the same NLM.
  1. Predict Spans; 2. Obtain Contextual Embedding; 3. Classify Embedding



Character Ngrams

- UMLS provides aliases (alt. names) for every concept (5M).
- SimString [Okazaki et Tsujii, 2010] breaks words into character n-grams for approximate dictionary matching.



Reye Syndrome
UMLS:C0035400
T038

Reye syndrome → $$r; $re; rey; eye; ye_; e_s; _sy; syn; ynd; ndr; dro; rom; ome; me$; e$$

Syndrome Reyes → $$s; $sy; syn; ynd; ndr; dro; rom; ome; me_; e_r; _re; rey; eye; yes; es$; s$$

Reyes syndrome → $$r; $re; rey; eye; yes; es_; s_s; _sy; syn; ynd; ndr; dro; rom; ome; me$; e$$

Reye's syndrome → $$r; $re; rey; eye; ye'; e's; 's_; s_s; _sy; syn; ynd; ndr; dro; rom; ome; me$; e$$

Entity
e

Aliases
$a \in map(e)$

Features
$\hat{a}$

Approximate Dictionary Matching

- Each word’s n-grams represent features that can be matched using cosine similarity.
- During inference, ngrams of recognized spans are represented as query features.

‘Reye's syndrome (RS)’

$$scoreSTR(\overset{\uparrow}{s}, \underset{\downarrow}{e}) = \max_{a \in map(e)} \cos(\hat{s}, \hat{a})$$

Reye Syndrome

UMLS:C0035400

T038

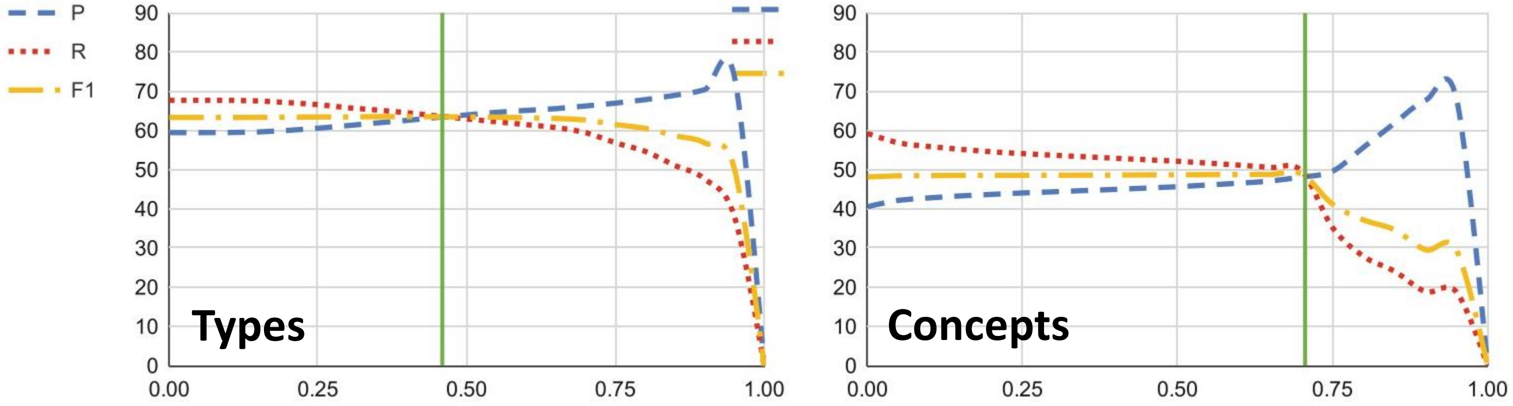
$$\cos(\hat{s}, \hat{a}) = \frac{|\hat{s} \cap \hat{a}|}{\sqrt{|\hat{s}|\,|\hat{a}|}}$$

Combining Matches

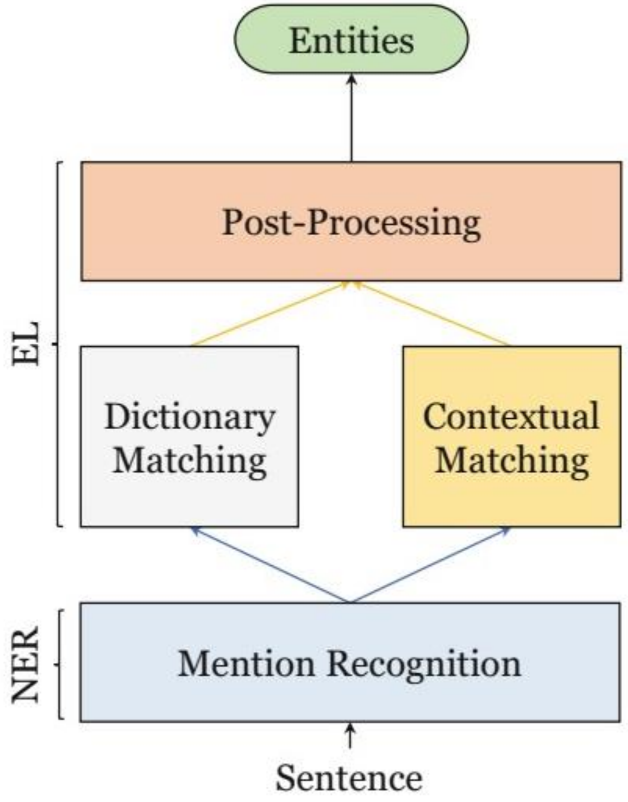
- A simple max-based solution works well.

$$scoreSTR\_CTX(s,e) = \max(scoreSTR(s,e), scoreCTX(s,e))$$

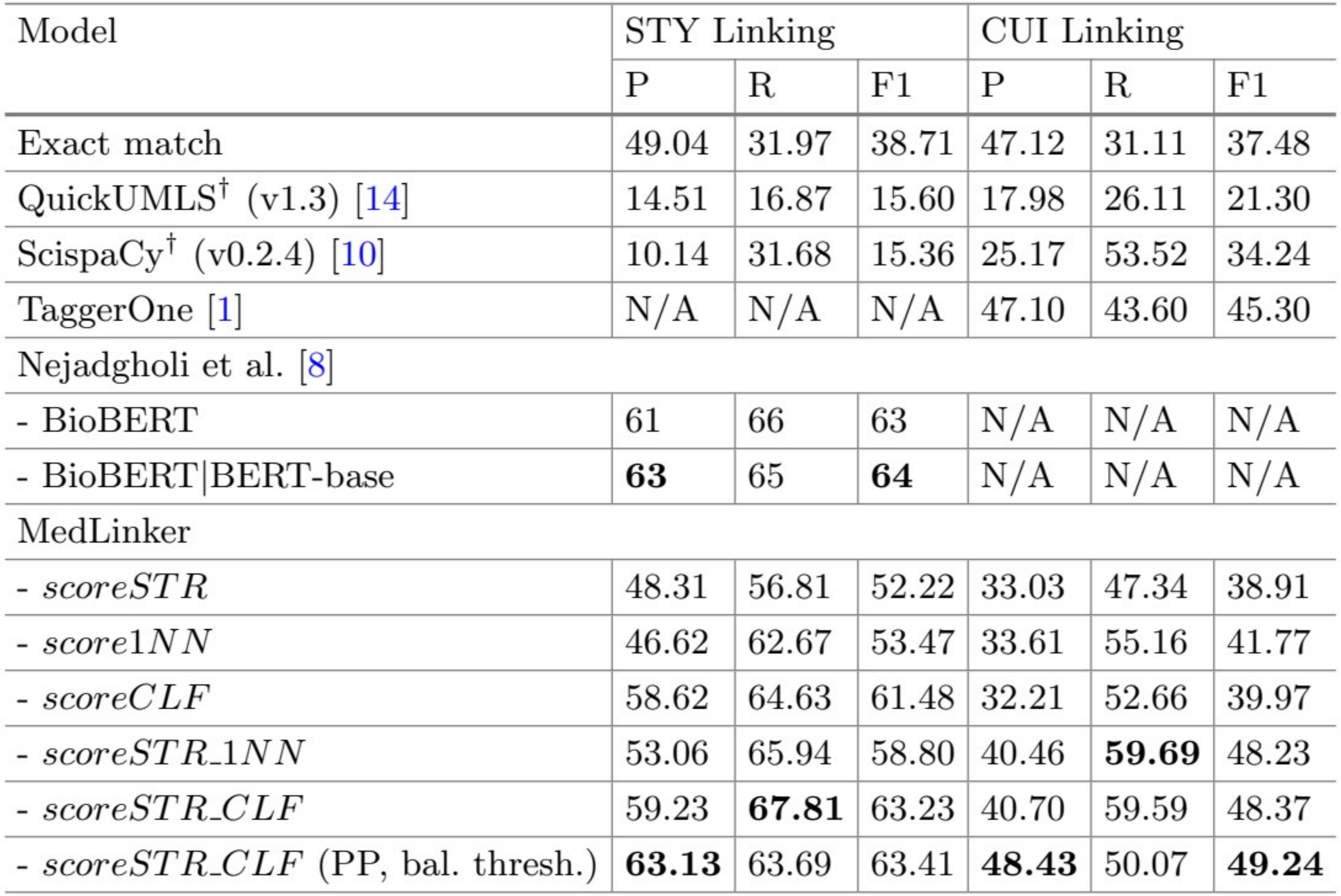
- This allows for many false-positives. We achieve higher Precision by training LR with scores as features, and finding a threshold.



Full Pipeline



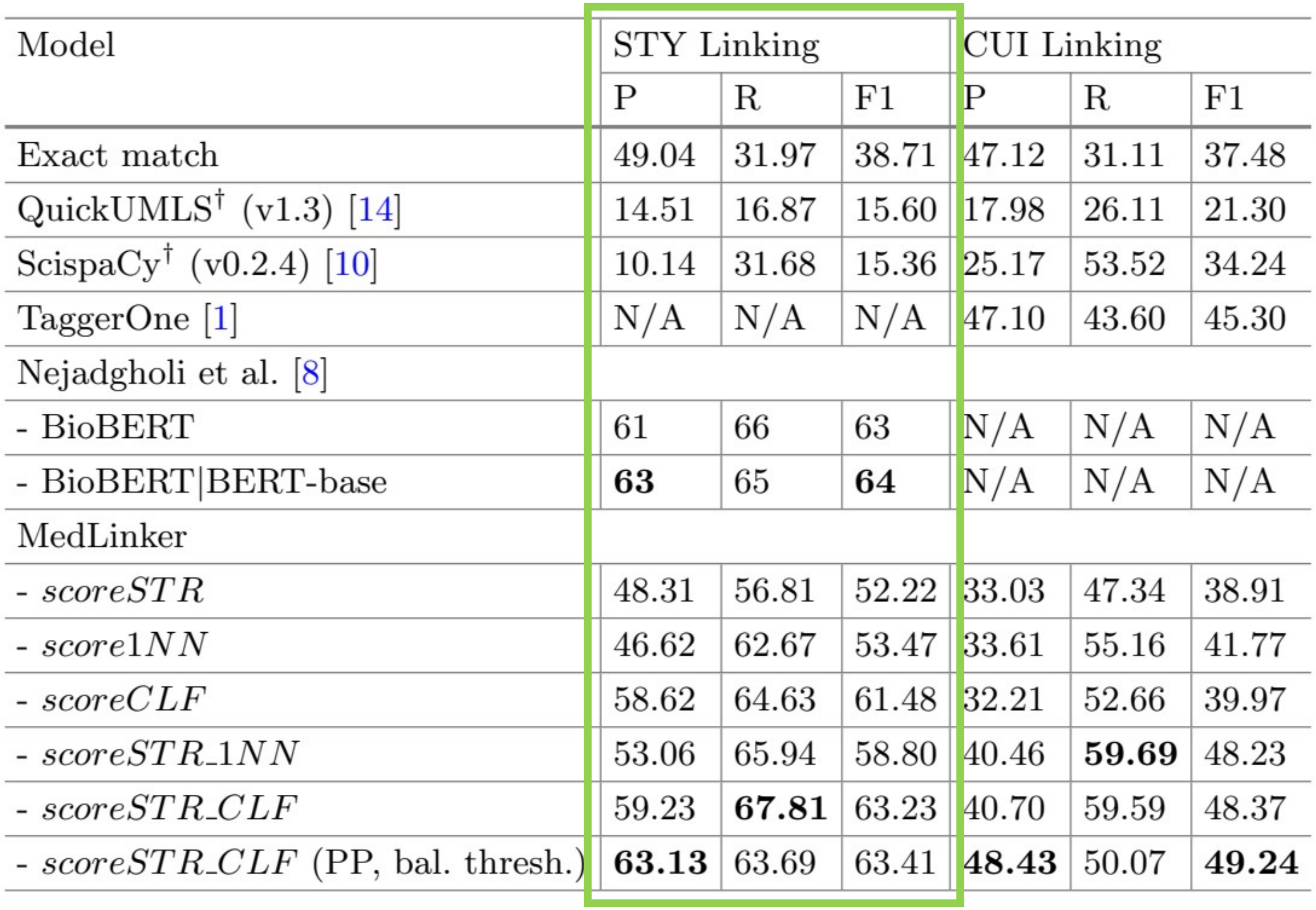
Results



Model	STY Linking			CUI Linking		
	P	R	F1	P	R	F1
Exact match	49.04	31.97	38.71	47.12	31.11	37.48
QuickUMLS† (v1.3) [14]	14.51	16.87	15.60	17.98	26.11	21.30
ScispaCy† (v0.2.4) [10]	10.14	31.68	15.36	25.17	53.52	34.24
TaggerOne [1]	N/A	N/A	N/A	47.10	43.60	45.30
Nejadgholi et al. [8]						
- BioBERT	61	66	63	N/A	N/A	N/A
- BioBERT|BERT-base	**63**	65	**64**	N/A	N/A	N/A
MedLinker						
- *scoreSTR*	48.31	56.81	52.22	33.03	47.34	38.91
- *score1NN*	46.62	62.67	53.47	33.61	55.16	41.77
- *scoreCLF*	58.62	64.63	61.48	32.21	52.66	39.97
- *scoreSTR_1NN*	53.06	65.94	58.80	40.46	**59.69**	48.23
- *scoreSTR_CLF*	59.23	**67.81**	63.23	40.70	59.59	48.37
- *scoreSTR_CLF* (PP, bal. thresh.)	**63.13**	63.69	63.41	**48.43**	50.07	**49.24**

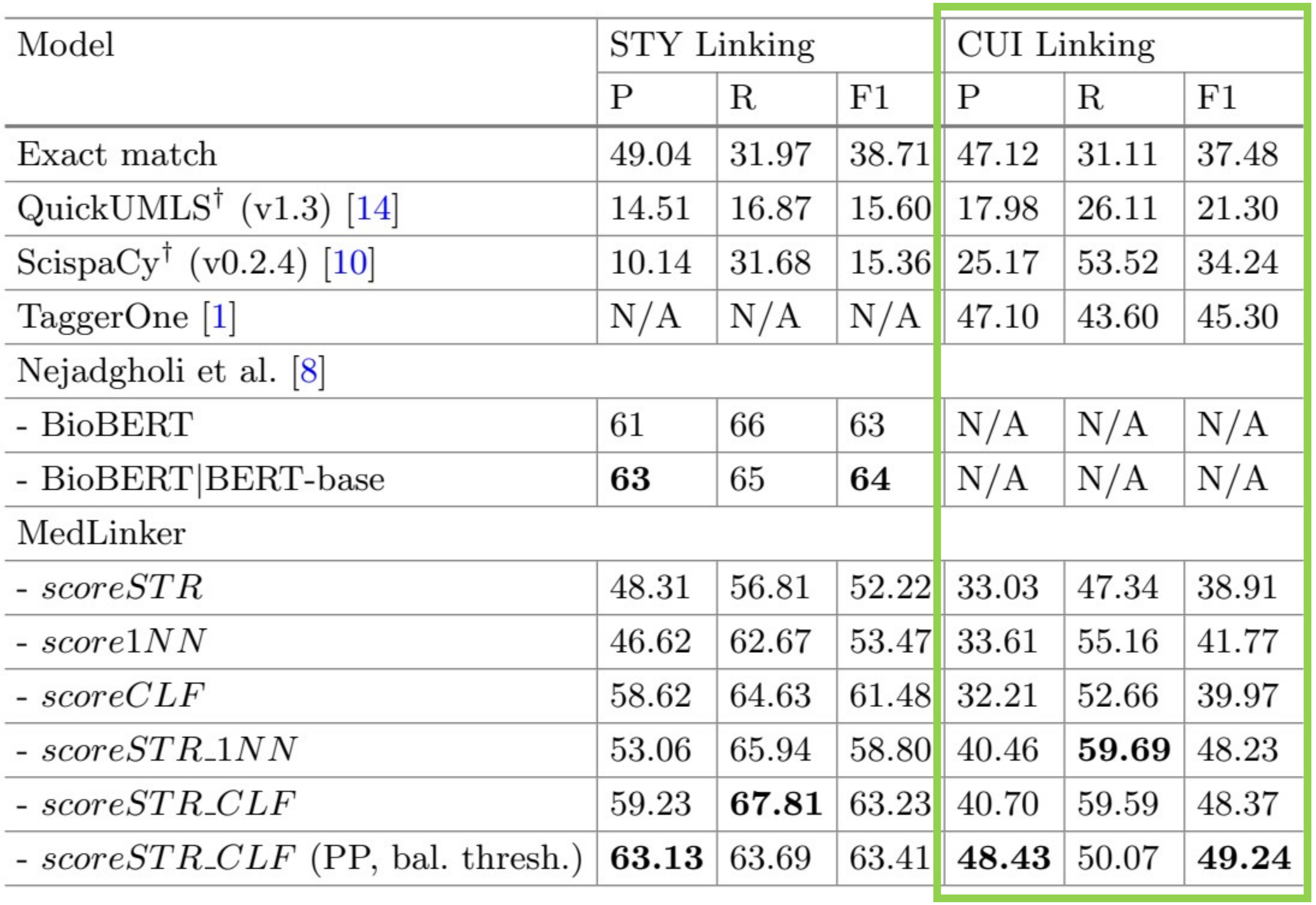
Results

21 labels



Model	STY Linking			CUI Linking		
	P	R	F1	P	R	F1
Exact match	49.04	31.97	38.71	47.12	31.11	37.48
QuickUMLS† (v1.3) [14]	14.51	16.87	15.60	17.98	26.11	21.30
ScispaCy† (v0.2.4) [10]	10.14	31.68	15.36	25.17	53.52	34.24
TaggerOne [1]	N/A	N/A	N/A	47.10	43.60	45.30
Nejadgholi et al. [8]						
- BioBERT	61	66	63	N/A	N/A	N/A
- BioBERT|BERT-base	**63**	65	**64**	N/A	N/A	N/A
MedLinker						
- *scoreSTR*	48.31	56.81	52.22	33.03	47.34	38.91
- *score1NN*	46.62	62.67	53.47	33.61	55.16	41.77
- *scoreCLF*	58.62	64.63	61.48	32.21	52.66	39.97
- *scoreSTR_1NN*	53.06	65.94	58.80	40.46	**59.69**	48.23
- *scoreSTR_CLF*	59.23	**67.81**	63.23	40.70	59.59	48.37
- *scoreSTR_CLF* (PP, bal. thresh.)	**63.13**	63.69	63.41	**48.43**	50.07	**49.24**

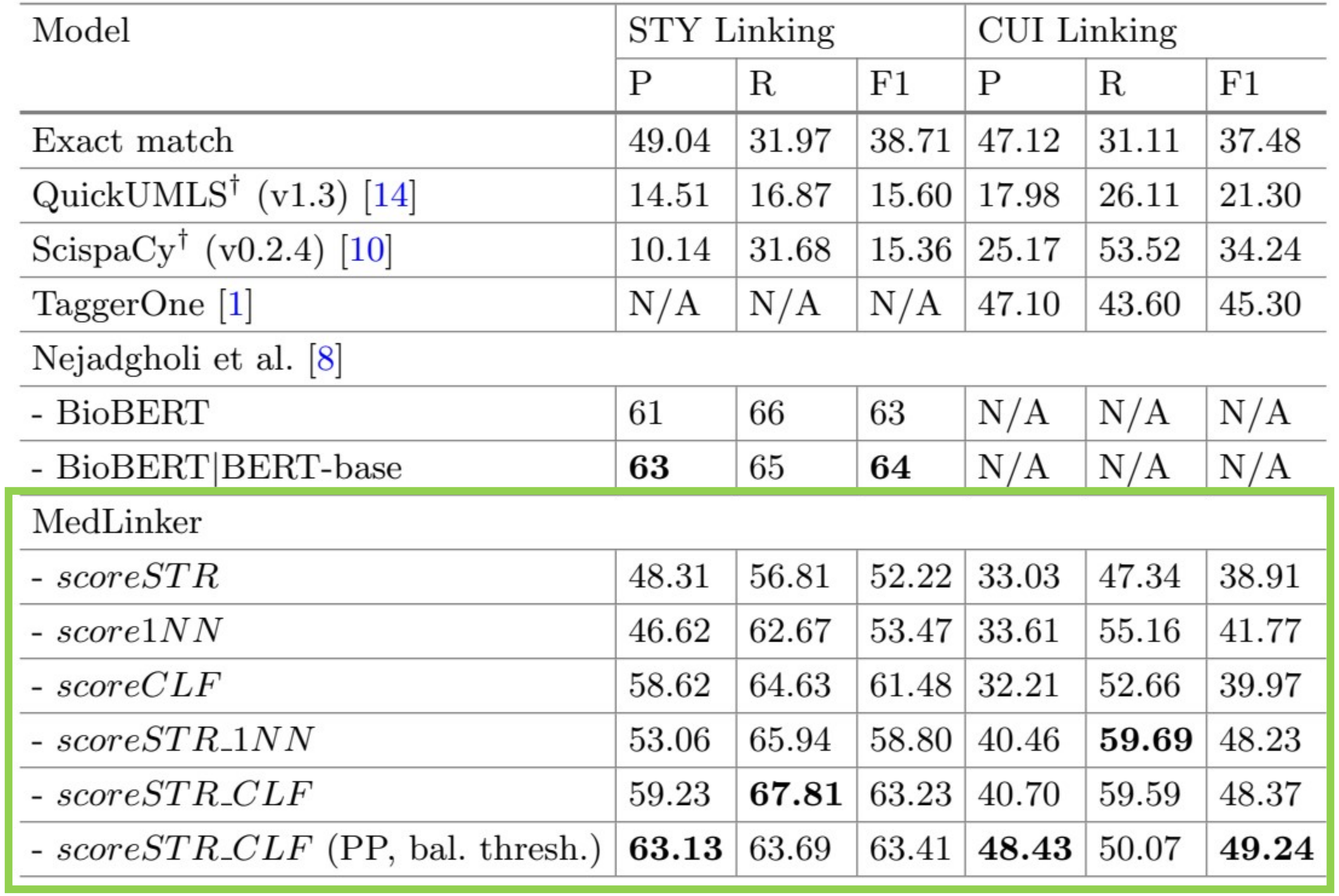
Results



2M labels

Model	STY Linking			CUI Linking		
	P	R	F1	P	R	F1
Exact match	49.04	31.97	38.71	47.12	31.11	37.48
QuickUMLS† (v1.3) [14]	14.51	16.87	15.60	17.98	26.11	21.30
ScispaCy† (v0.2.4) [10]	10.14	31.68	15.36	25.17	53.52	34.24
TaggerOne [1]	N/A	N/A	N/A	47.10	43.60	45.30
Nejadgholi et al. [8]						
- BioBERT	61	66	63	N/A	N/A	N/A
- BioBERT|BERT-base	**63**	65	**64**	N/A	N/A	N/A
MedLinker						
- *scoreSTR*	48.31	56.81	52.22	33.03	47.34	38.91
- *score1NN*	46.62	62.67	53.47	33.61	55.16	41.77
- *scoreCLF*	58.62	64.63	61.48	32.21	52.66	39.97
- *scoreSTR_1NN*	53.06	65.94	58.80	40.46	**59.69**	48.23
- *scoreSTR_CLF*	59.23	**67.81**	63.23	40.70	59.59	48.37
- *scoreSTR_CLF* (PP, bal. thresh.)	**63.13**	63.69	63.41	**48.43**	50.07	**49.24**

Results



Model	STY Linking			CUI Linking		
	P	R	F1	P	R	F1
Exact match	49.04	31.97	38.71	47.12	31.11	37.48
QuickUMLS† (v1.3) [14]	14.51	16.87	15.60	17.98	26.11	21.30
ScispaCy† (v0.2.4) [10]	10.14	31.68	15.36	25.17	53.52	34.24
TaggerOne [1]	N/A	N/A	N/A	47.10	43.60	45.30
Nejadgholi et al. [8]						
- BioBERT	61	66	63	N/A	N/A	N/A
- BioBERT|BERT-base	**63**	65	**64**	N/A	N/A	N/A
MedLinker						
- *scoreSTR*	48.31	56.81	52.22	33.03	47.34	38.91
- *score1NN*	46.62	62.67	53.47	33.61	55.16	41.77
- *scoreCLF*	58.62	64.63	61.48	32.21	52.66	39.97
- *scoreSTR_1NN*	53.06	65.94	58.80	40.46	**59.69**	48.23
- *scoreSTR_CLF*	59.23	**67.81**	63.23	40.70	59.59	48.37
- *scoreSTR_CLF* (PP, bal. thresh.)	**63.13**	63.69	63.41	**48.43**	50.07	**49.24**

Conclusion

- Medical Entity Linking is still a very challenging task, requiring new approaches that make up for lack of annotations.
- Neural Language Models can be effectively combined with Dictionary Matching using lightweight methods.
- Code and supplementary material available at:
  - https://github.com/danlou/medlinker