A Lightweight Neural Model for Biomedical Entity Linking

Lihu Chen[1], Gaël Varoquaux[2], Fabian Suchanek[1]

[1] LTCI & Télécom Paris & Institut Polytechnique de Paris, France

[2] Inria & CEA & Université Paris-Saclay, France

AAAI21

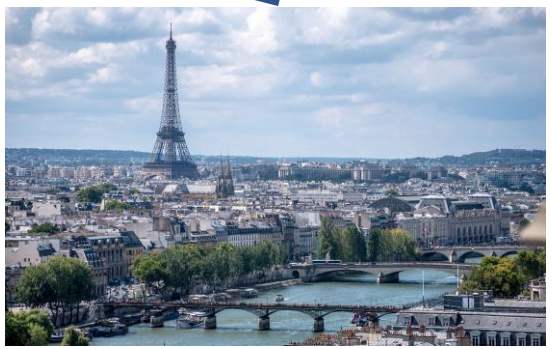
What is entity linking?

Entity linking (Entity Normalization) is the task of mapping entity mentions in text documents to standard entities in a given knowledge base.

Paris is the son of King Priam

?

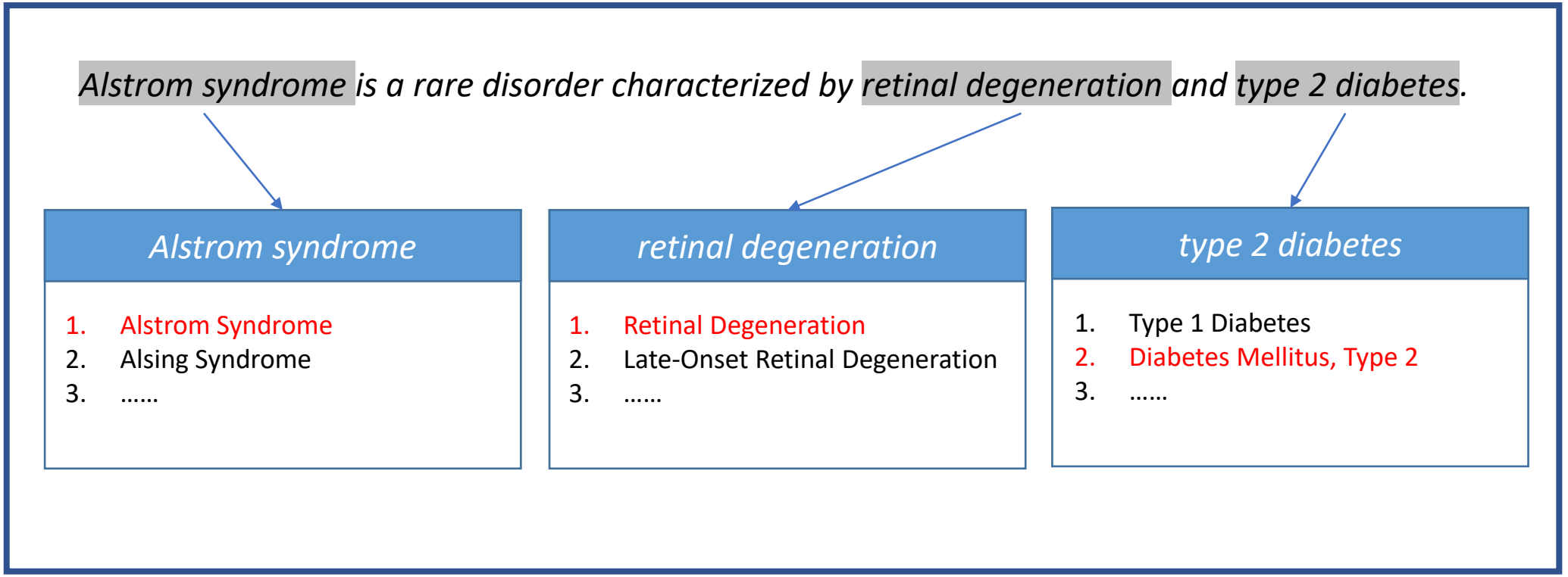
Paris (mythology), a prince of Troy in Greek mythology



Paris is the capital of France.

Biomedical Entity Linking

In the biomedical domain, entity linking maps mentions of diseases, drugs, and measures to normalized entities in standard vocabularies



Applications & Challenges

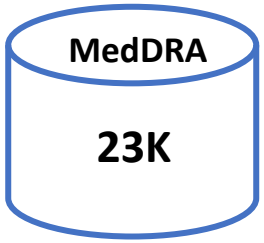
❑ Application

- Biomedical Information Extraction
- Data Integration of Medical Information System

❑ Challenge

- Surface forms vary markedly
- Canonical forms look alike
- Biomedical KBs contain only surface forms of entities

decreases in hemoglobin ?



1. *increase in hematocrit*
2. *changes in hemoglobin*
3. *haemoglobin decreased*
4. *decreases in platelets*
5. *……*

Related Work

❑ Rule-based Approach

- Pre-define rules manually to measure a string similarity. [1-3]

need to define rules manually

❑ Machine Learning Approach

- Dnorm[4]
- TaggerOne[5]
- Learning to Rank[6]

cannot recognize semantically related words well

❑ Deep Learning Approach

- CNN-based Ranking[7]
- RNN Model[8]

context-independent representation of each word

❑ BERT-based Approach

- BERT[9]
- Clinical BERT[10]
- BioBERT[11]

large amounts of parameters

Contribution

BERT-based methods have advanced the state-of-the-art.

However, they often have hundreds of millions of parameters and require

heavy computing resources, which limits their applications in resource-limited scenarios.

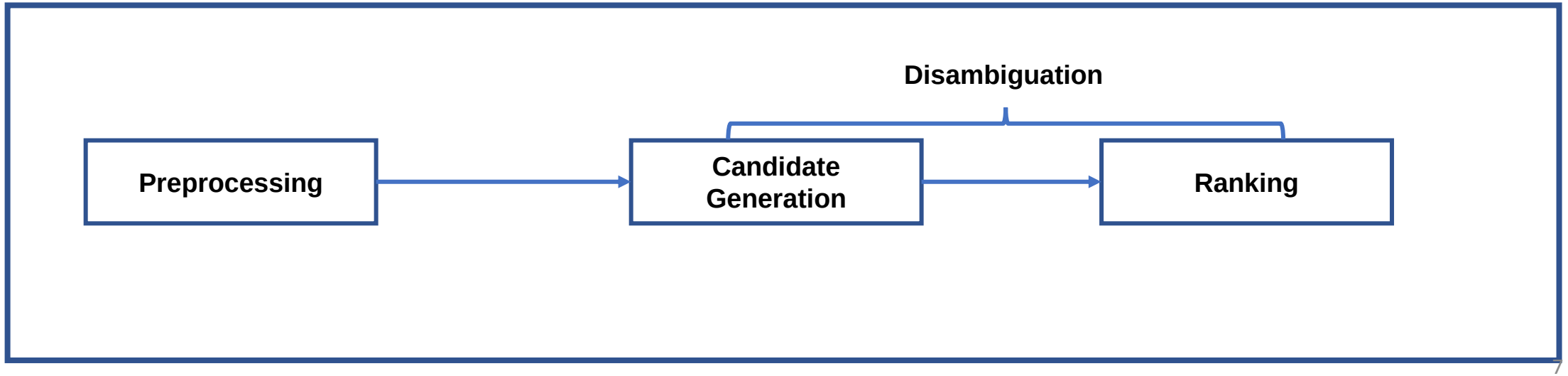
❑ Contribution

- Propose a **simple** and **lightweight** neural model for biomedical entity linking
- Achieve a competitive performance with BERT-based models
- Explore how to add **prior**, **context** and **coherence** features for this task
- **23x smaller** and **6.4x faster** than BERT-based models

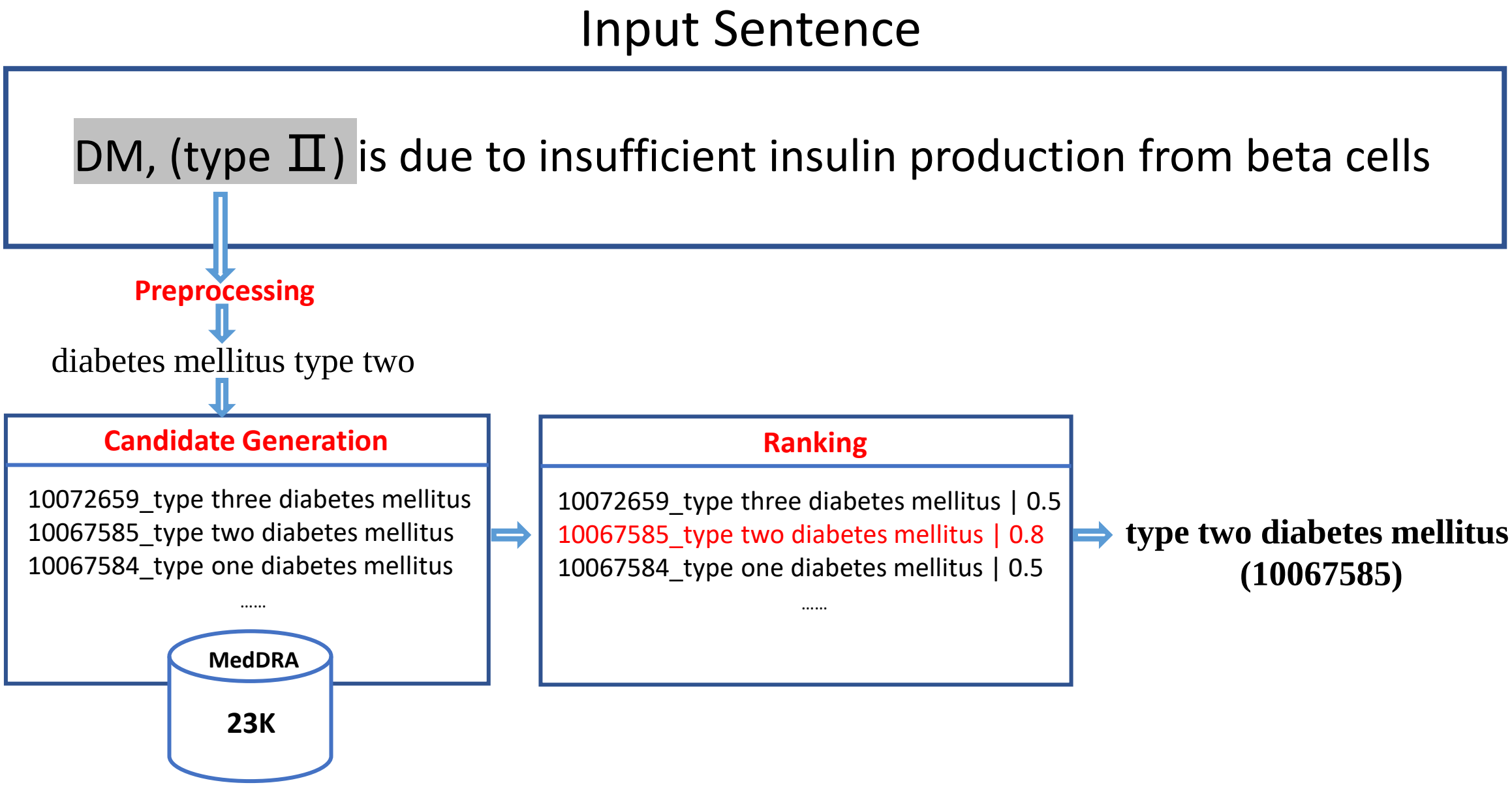
Problem Definition & Framework

- **Input (1):** a knowledge base (KB), i.e., a list of entities, each with one or more names
- **Input (2):** a corpus, i.e., a set of text documents in which certain text spans have been tagged as entity mentions
- **Goal:** to link each entity mention to the correct entity in the KB

A Framework of Biomedical Entity Linking



Framework


Input Sentence
DM, (type Ⅱ) is due to insufficient insulin production from beta cells
Preprocessing
diabetes mellitus type two
Candidate Generation
10072659_type three diabetes mellitus
10067585_type two diabetes mellitus
10067584_type one diabetes mellitus
.....
MedDRA
23K
Ranking
10072659_type three diabetes mellitus | 0.5
10067585_type two diabetes mellitus | 0.8
10067584_type one diabetes mellitus | 0.5
.....
type two diabetes mellitus (10067585)

Our Approach: Preprocessing

❑ Abbreviation Expansion

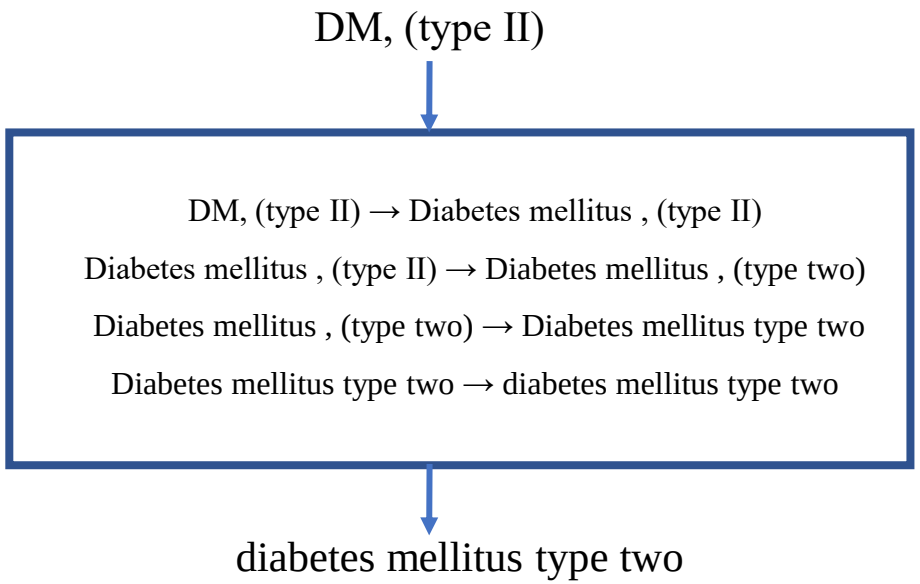
- Ab3p Toolkit[12]
- Abbreviation dictionary[3]

❑ Numeral Replacement

- Replace all forms (e.g., Arabic, Roman…) with spelled-out English numerals.

❑ KB Augmentation

- Augment the KB by adding all names from the training set



Our Approach: Candidate Generation

❑ Input

- a mention M and a knowledge base (KB)
- each entity E in the KB has a certain number of names $\{S_1^E, S_2^E, \dots, S_n^E\}$
- a pre-trained word embedding matrix $\boldsymbol{V} \in \mathbb{R}^{d \times |V|}$

❑ Output

- For the mention M, we generate a set C_M of candidate entities from the KB.

That is, $C_M = \{< E_1, S_1 >, < E_2, S_2 >, \dots, < E_k, S_k >\}$

❑ Algorithm

use word embeddings to represent each tokens in a name, $\mathrm{M} = \{m_1, m_2, \dots, m_{|M|}\}, \mathrm{S} = \{s_1, s_2, \dots, s_{|S|}\}$

foreach entity E in KB **do**

foreach name S of E **do**

foreach token in M calculate $Acos(m_i, S) = \max\{\cos(m_i, s_j) \mid s_j \in S\}$

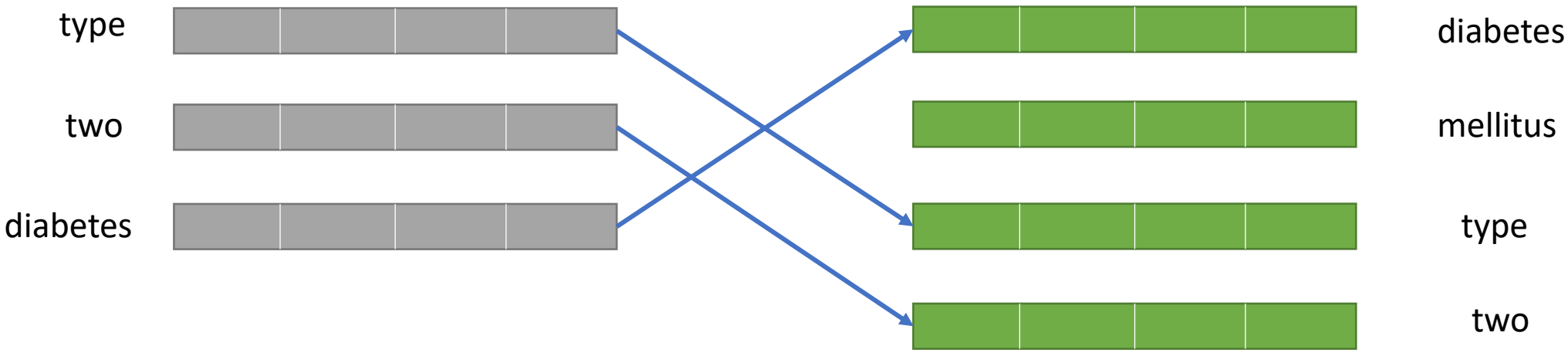
foreach token in S calculate $Acos(s_j, M) = \max\{\cos(s_j, m_i) \mid m_i \in M\}$

calculate similarity: $Sim(M, S) = 1/(|M| + |S|) \left(\sum_{m_i \in M} Acos(m_i, S) + \sum_{s_j \in S} Acos(s_j, M)\right)$

use the maximum $Sim(M, S)$ as the similarity between M and E

return top-k entities with the highest $Sim(M, E)$, so $C_M = \{< E_1, S_1 >, < E_2, S_2 >, \dots, < E_k, S_k >\}$

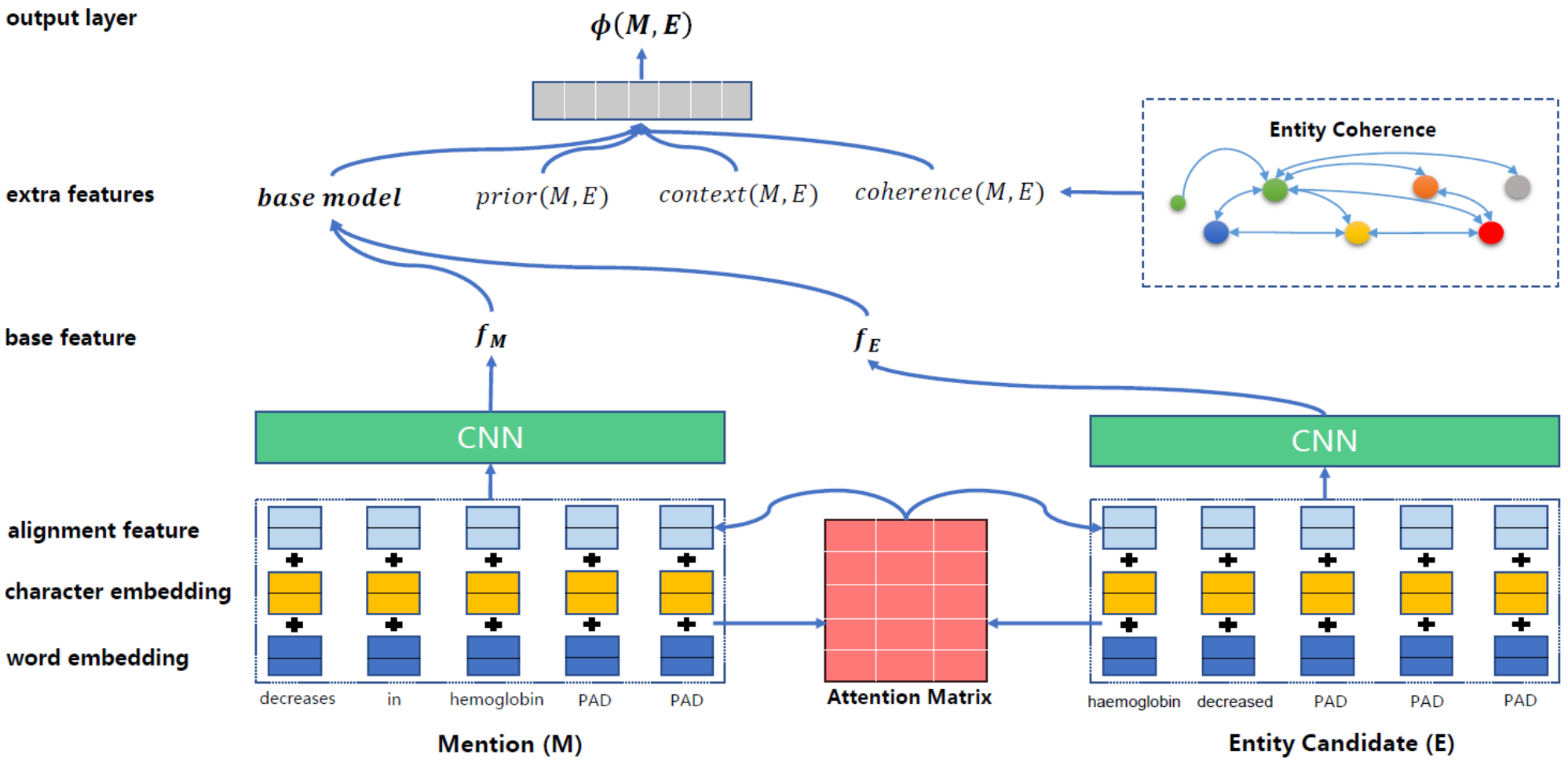
Our Approach: Candidate Generation



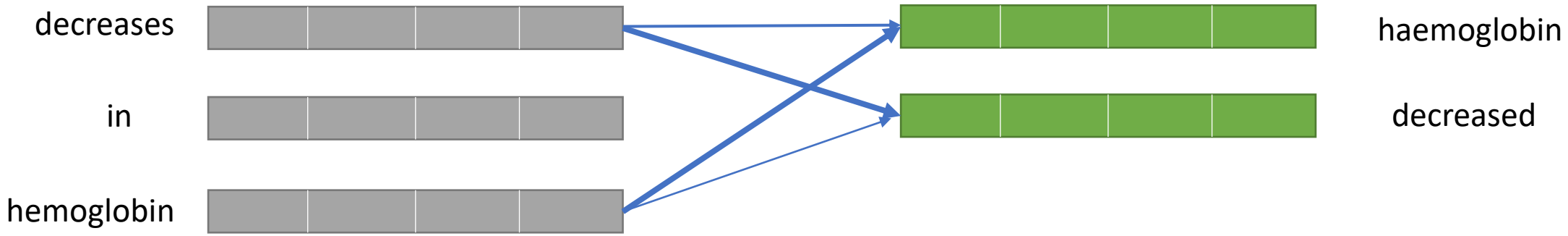
Dataset	ShARe/CLEF	NCBI	ADR
Recall	97.79%	94.27%	96.66%

We generate 20 candidates for each mention across these three datasets.
The table shows the recall score of our method of candidate generation

Our Approach: Ranking Model



Our Approach: Alignment Layer



Alignment Layer

$$Attention(M, S) = softmax(MS^T)M$$

m_i 

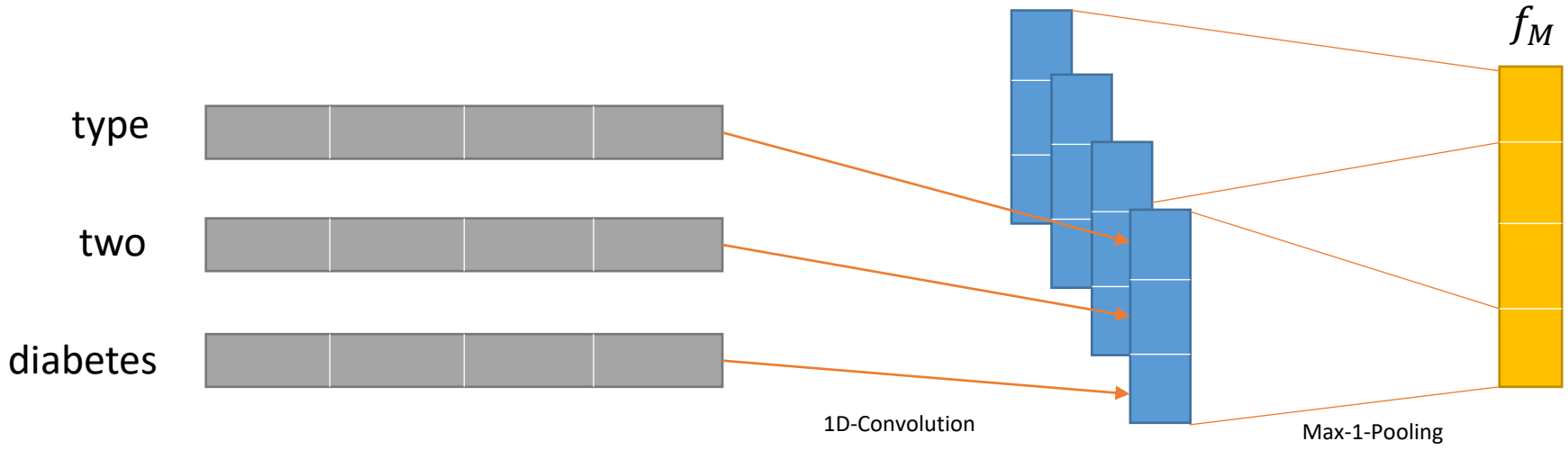
$\bar{m}_i$ 

$$sub(m_i, \bar{m}_i) = (m_i - \bar{m}_i) \odot (m_i - \bar{m}_i)$$

$$mul(m_i, \bar{m}_i) = m_i \odot \bar{m}_i$$

$$\hat{m}_i = [m_i, \bar{m}_i, sub(m_i, \bar{m}_i), mul(m_i, \bar{m}_i)]$$

Our Approach: CNN Layer



$$\hat{m}_i = [m_i, \overline{m}_i, sub(m_i, \overline{m}_i), mul(m_i, \overline{m}_i)]$$

$$f_M = CNN(\hat{m}_1, \hat{m}_2, \dots, \hat{m}_M)$$

$$f_{\text{out}} = [f_M, f_E]$$

Our Approach: Extra Features

Discharge Instructions
*You should shower, let water flow over **wounds**, pat dry with a towel*

surgical wound **gunshot wound**

Mention-Entity Prior

$$prior(M,E) = \log count(M,E)$$

$$\mathrm{f}_{out} = [f_M, f_E, prior(M,E)]$$

Long-term cumulative radiation exposure is associated with an increased risk of cancer

radiation exposure → *cancer*

Context

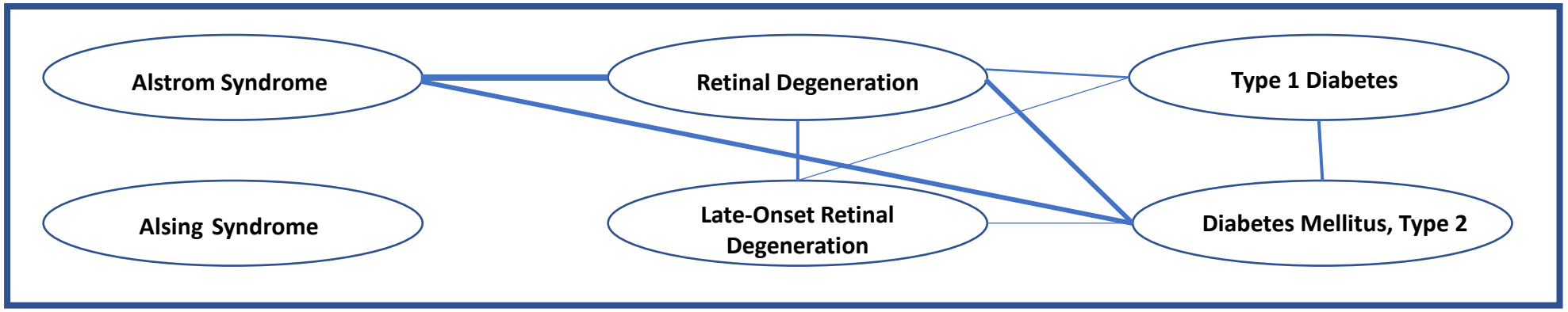
$$context(M,E) = \cos(cxt_M, cxt_E)$$

$$\mathrm{f}_{out} = [f_M, f_E, context(M,E)]$$

Our Approach: Coherence Feature

Alstrom syndrome is a rare disorder characterized by retinal degeneration and type 2 diabetes.

Alstrom syndrome	*retinal degeneration*	*type 2 diabetes*
1. Alstrom Syndrome	1. Retinal Degeneration	1. Type 1 Diabetes
2. Alsing Syndrome	2. Late-Onset Retinal Degeneration	2. Diabetes Mellitus, Type 2
3. ……	3. ……	3. ……



$$coherence(M,E) = \frac{1}{k}\sum_{i=1}^{k}\cos(p_i, p_E)$$

$$\mathrm{f}_{out} = [f_M, f_E, coherence(M,E)]$$

Our Approach: Training

❑ Loss Function

- $\phi(M, E) = sigmoid(W_2 ReLU(W_1 f_{\text{out}} + b_1) + b_2)$
- $\theta^* = \underset{\theta}{\text{argmin}} \sum_{D \in \mathcal{D}} \sum_{M \in D} \sum_{E \in C} \max(0, \gamma + \phi(M, E^+) - \phi(M, E^-))$

❑ The NIL Problem

- When a mention does not correspond to any entity in the KB, we adopt a traditional threshold method

Experiments

❑ Dataset

- ShARe/CLEF[13], NCBI[14], ADR

❑ Metric

- Top-1 Accuracy
- Binomial confidence interval (confidence level=0.05)

❑ Competitors

- Dnorm [4]
- TaggerOne [5]
- Sieve-based Model[3]
- Learning to Rank[6]
- CNN-based Ranking[7]
- BERT-based Ranking[15]

	ShARe/CLEF		NCBI		ADR	
	train	test	train	test	train	test
documents	199	99	692	100	101	99
mentions	5816	5351	5921	964	7038	6343
NIL	1641	1750	0	0	47	18
concepts	88140		9656		23668	
synonyms	42929		59280		0	

Dataset Statistics

Results: Overall Performance

Our simple model is just as good as a BERT-based model

Model	ShARe/CLEF	NCBI	ADR
DNorm (Leaman, Islamaj Doğan, and Lu 2013)	-	82.20±4.05	-
UWM (Ghiasvand and Kate 2014)	89.50±1.38	-	-
Sieve-based Model (D'Souza and Ng 2015)	90.75±1.31	84.65±3.84	-
TaggerOne (Leaman and Lu 2016)	-	88.80±3.32	-
Learning to Rank (Xu et al. 2017)	-	-	92.05±1.12
CNN-based Ranking (Li et al. 2017)	90.30±1.33	86.10±3.63	-
BERT-based Ranking (Ji, Wei, and Xu 2020)	**91.06±1.29**	89.06±3.32	**93.22±1.04**
Our Base Model	90.10±1.35	89.07±3.32	92.89±1.06
Our Base Model + Extra Features	90.43±1.33	**89.59±3.22**	93.00±1.06

We compute a binomial confidence interval for each model (at a confidence level of 0.05), based on the total number of mentions and the number of correctly mapped mentions. Results in gray are not statistically different from the top result.
In other words, the available data cannot demonstrate that sampling a new test set on the same task would not lead to different order.

Results: Ablation study

- **The removal of the Alignment layer causes the biggest drop**
- **Adding extra features is not necessary**

Model	ShARe/CLEF	NCBI	ADR
- Character Feature	-1.21	-0.31	-0.30
- Alignment Layer	-3.80	-4.06	-3.17
- CNN Layer	-1.87	-0.93	-0.35
Our Base Method	90.10	89.07	92.89
+ Mention-Entity Prior	+0.33	+0.04	+0.03
+ Context	-0.09	+0.21	-0.24
+ Coherence	-0.02	+0.27	+0.11

The gray row is the accuracy of our base model
The above is effect of the removal of each component of our base model
The below is the effect of addition of extra features

Results: Inference Time

❑ Parameters and Inference Time

- **Our model has 4.6M parameters, which is 1.6x to 72.9x smaller than the other models**
- **On average, our model is 6.4x faster than other BERT models, and our model is much lighter on the CPU.**

Model	Parameters	ShARe/CLEF		NCBI		ADR		Avg	Speedup
		CPU	GPU	CPU	GPU	CPU	GPU		
BERT (large)	340M	2230s	1551s	353s	285s	2736s	1968s	1521s	12.3x
BERT (base)	110M	1847s	446s	443s	83s	1666s	605s	848s	6.4x
TinyBERT$_6$	67M	1618s	255s	344s	42s	2192s	322s	796s	6.0x
MobileBERT (base)	25.3M	1202s	330s	322s	58s	1562s	419s	649s	4.7x
ALBERT (base)	12M	836s	**129s**	101s	24s	1192s	170s	409s	2.6x
Our Base Model	4.6M	**181s**	131s	**38s**	**22s**	**196s**	**116s**	**114s**	-

A Lightweight Neural Model for Biomedical Entity Linking

❑ Contribution

- Propose a **simple** and **lightweight** neural model for biomedical entity linking
- Achieve a performance that is **statistically indistinguishable** from the BERT-based model
- **23x smaller** and **6.4x faster** than BERT-based models

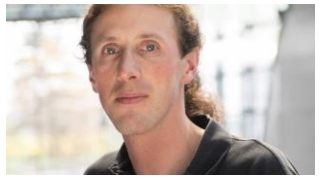
❑ Future Work

- How to Automatically assigning a weight for each word in the mentions and entity names
- Introduce GCN to capture the coherence among entities

❑ Paper and code

- https://arxiv.org/pdf/2012.08844.pdf
- https://github.com/tigerchen52/Biomedical-Entity-Linking

Lihu Chen



Gaël Varoquaux

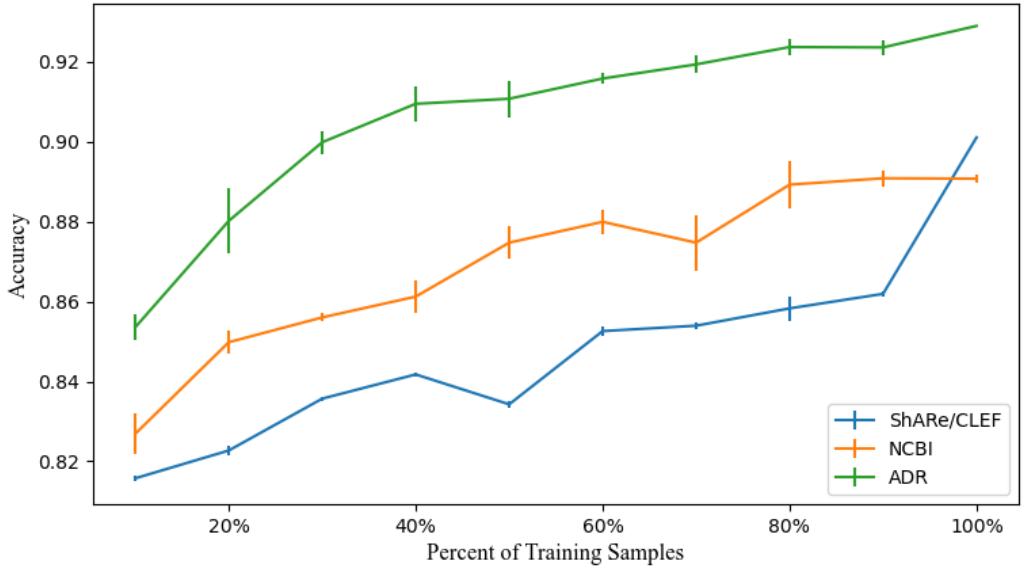
Fabian Suchanek

Appendix

Performance in the face of typos

Model	Original ADR	10%	30%	50%	70%	90%
+ Ordering Change	92.89	92.46	92.44	92.23	92.57	92.31
+ Typo	92.89	92.29	91.87	91.64	91.67	91.39

Model Performance as Data grows



The model is not limited by it's simplicity

Reference

[1] Dogan, R. I.; and Lu, Z. 2012. An inference method for disease name normalization. In 2012 AAAI Fall Symposium Series.

[2] Kang, N.; Singh, B.; Afzal, Z.; van Mulligen, E. M.; and Kors, J. A. 2013. Using rule-based natural language processing to improve disease normalization in biomedical text. Journal of the American Medical Informatics Association 20(5): 876–881.

[3] D'Souza, J.; and Ng, V. 2015. Sieve-based entity linking for the biomedical domain. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 297–302.

[4] Leaman, R.; Islamaj Doˇgan, R.; and Lu, Z. 2013. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics 29(22): 2909–2917.

[5] Leaman, R.; and Lu, Z. 2016. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. Bioinformatics 32(18): 2839–2846.

[6] Xu, J.; Lee, H.-J.; Ji, Z.;Wang, J.;Wei, Q.; and Xu, H. 2017. UTH CCB System for Adverse Drug Reaction Extraction from Drug Labels at TAC-ADR 2017. In TAC.

[7] Li, H.; Chen, Q.; Tang, B.;Wang, X.; Xu, H.;Wang, B.; and Huang, D. 2017. CNN-based ranking for biomedical entity normalization. BMC bioinformatics 18(11): 79–86.

[8] Wright, D. 2019. NormCo: Deep disease normalization for biomedical knowledge base construction. Ph.D. thesis, UC San Diego.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, andKristina Toutanova. 2018. Bert: Pre-training of deepbidirectional transformers for language understand-ing.arXiv preprint arXiv:1810.04805

[10] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath.2019.Clinicalbert: Modeling clinical notes andpredicting hospital readmission.arXiv preprintarXiv:1904.05342.

[11] JinhyukLee,WonjinYoon,SungdongKim,Donghyeon Kim, Sunkyu Kim, Chan Ho So, andJaewoo Kang. 2020. Biobert: a pre-trained biomed-ical language representation model for biomedicaltext mining.Bioinformatics, 36(4):1234–1240

[12] Sohn, S.; Comeau, D. C.; Kim, W.; and Wilbur, W. J. 2008. Abbreviation definition identification based on automatic precision estimates. BMC bioinformatics 9(1): 402.

[13] Pradhan, S.; Elhadad, N.; South, B. R.; Martinez, D.; Christensen, L. M.; Vogel, A.; Suominen, H.; Chapman, W. W.; and Savova, G. K. 2013. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. In CLEF (Working Notes), 212–31.

[14] Doˇgan, R. I.; Leaman, R.; and Lu, Z. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. Journal of biomedical informatics 47: 1–10.

[15] Ji, Z.; Wei, Q.; and Xu, H. 2020. Bert-based ranking for biomedical entity normalization. AMIA Summits on Translational Science Proceedings 2020: 269.