Disease Normalization with Graph Embeddings

Dhruba Pujary[1 2] Camilo Thorne[1] Wilker Aziz[2]

[1] Elsevier (Amsterdam/Frankfurt, Netherlands/Germany)

[2] UvA (Amsterdam, Netherlands)

IntelliSys 2020

Motivation

1. To semantically enrich clinical texts we need to **detect** (NER) and **resolve** (EL) diseases to canonical names
2. Canonical disease names, a.k.a. disease **concepts**, are defined in repositories such as.:

   $\Rightarrow$ thesauri, databases, taxonomies, knowledge graphs
3. Systems used in practice often ignore the (graphical) structure of such resources

Motivation

1. To semantically enrich clinical texts we need to **detect** (NER) and **resolve** (EL) diseases to canonical names
2. Canonical disease names, a.k.a. disease **concepts**, are defined in repositories such as.:

   $\Rightarrow$ thesauri, databases, taxonomies, knowledge graphs
3. Systems used in practice often ignore the (graphical) structure of such resources

Question

Can we use **neural graph embeddings** to detect and resolve diseases?

NCBI Corpus [DL12] & MeSH® Taxonomy [Lip00]

- NCBI corpus:

Split	PubMed® abstracts	Total mentions	Unique mentions	Unique concept IDs	Tokens
Training	592	5,134	1,691	657	136,088
Validation	100	787	363	173	23,969
Test	100	960	424	201	24,497

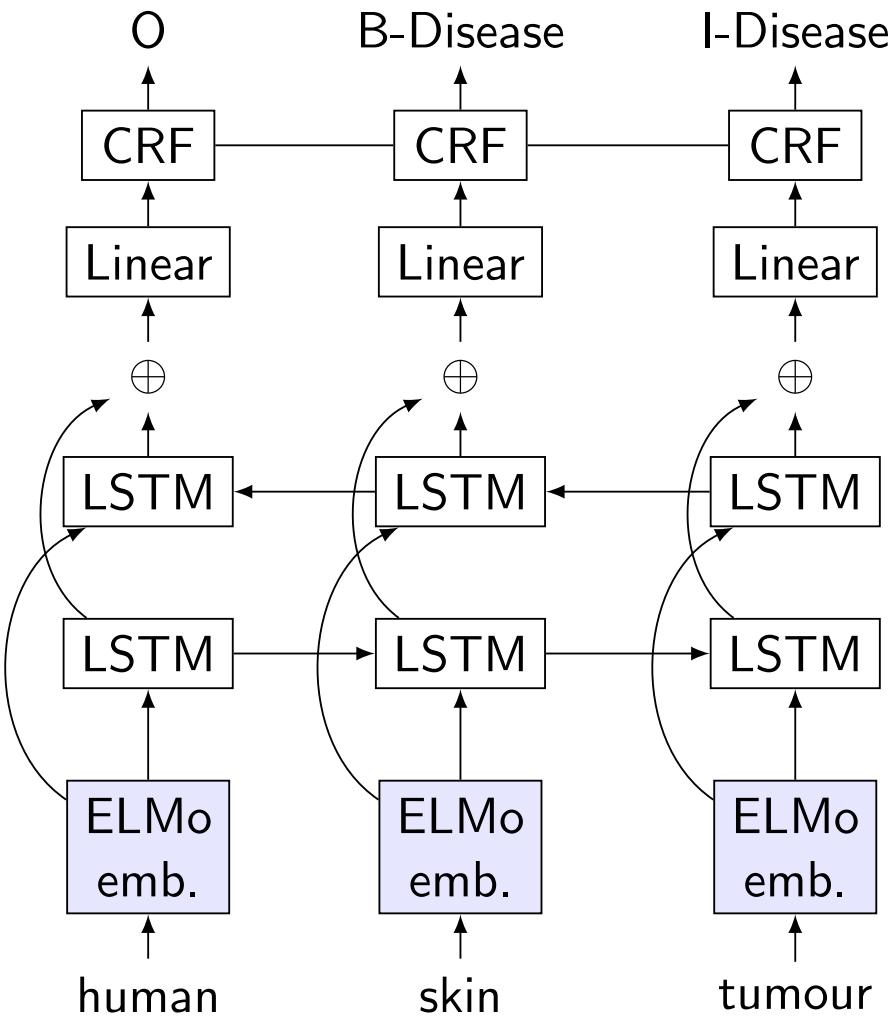
- MeSH® taxonomy (disease branch):
  - ▷ 10,932 diseases/conditions (tree nodes)
  - ▷ rich disease metadata
  - ▷ approx. 10,000 scope notes of 10-20 tokens each (approx. 100,000 – 200,000 tokens)

NCBI Corpus [DL12] & MeSH® Taxonomy [Lip00]

Identification of APC2, a homologue of the [*adenomatous polyposis coli tumour*]$_{\text{D011125}}$ suppressor.

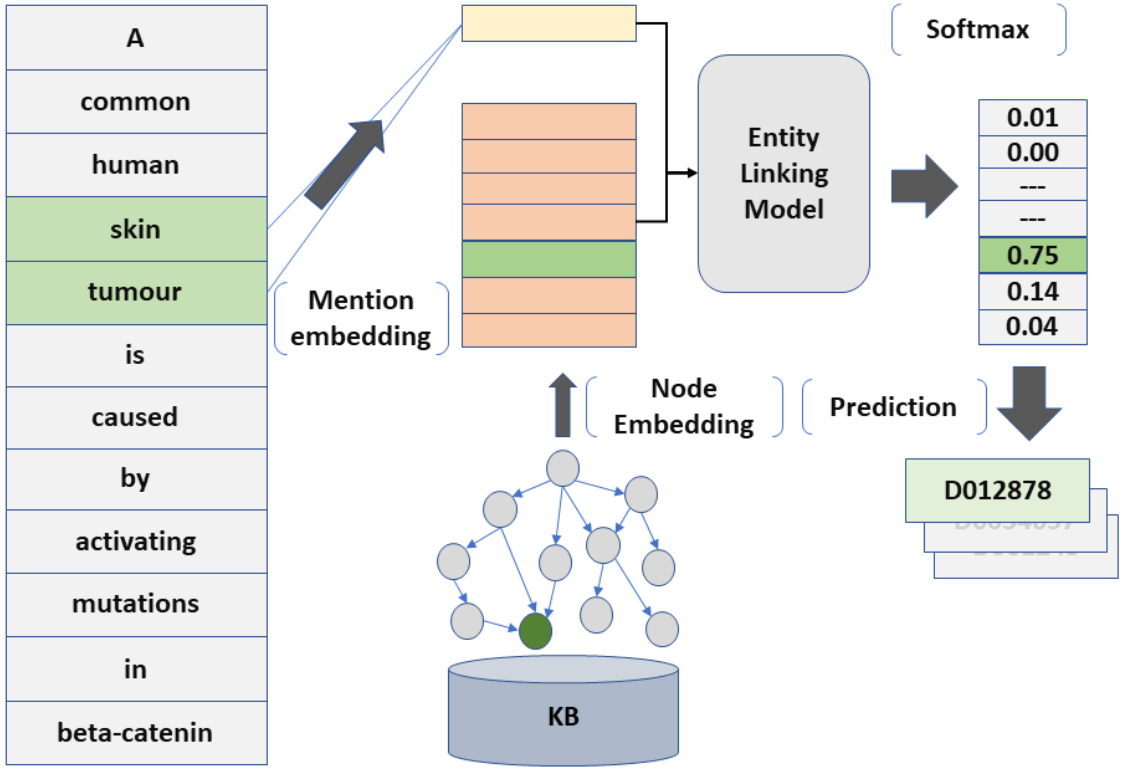
MeSH® Heading	Adenomatous Polyposis Coli
Scope Note	A polyposis syndrome due to an autosomal dominant mutation of the APC genes (GENES, APC) on CHROMOSOME 5. ...
Tree Numbers	C04.557.470.035.215.100 ...
Entry Terms	Polyposis Syndrome, Familial ...
	⋮

NER Model



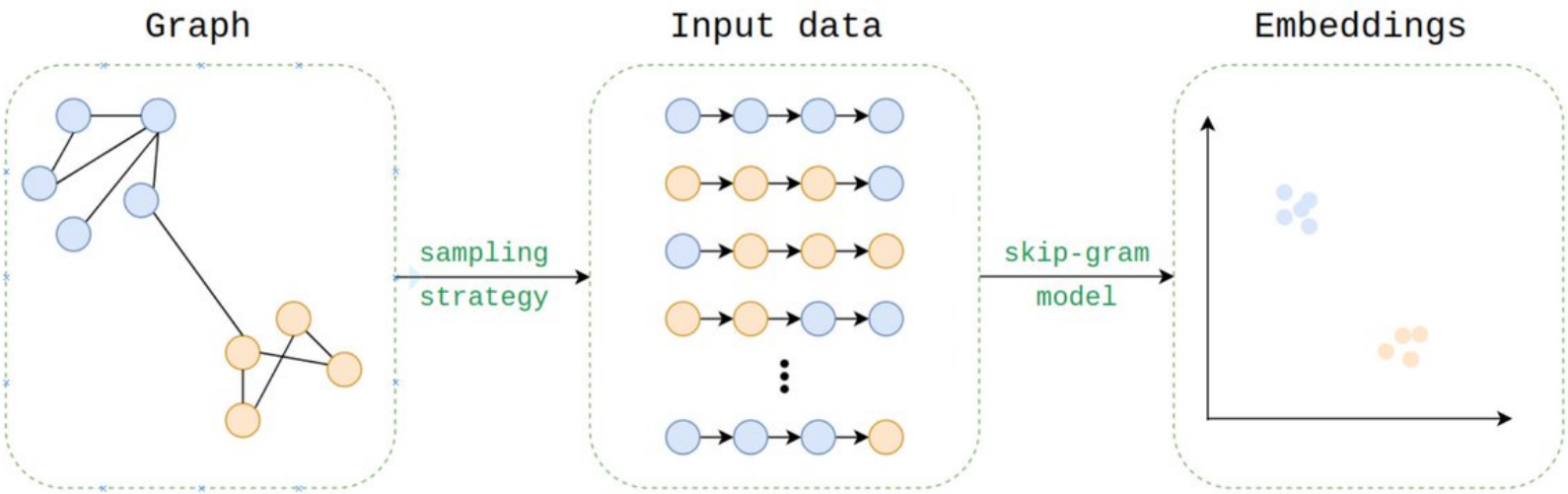
bioELMo [PNI+18]

EL Model



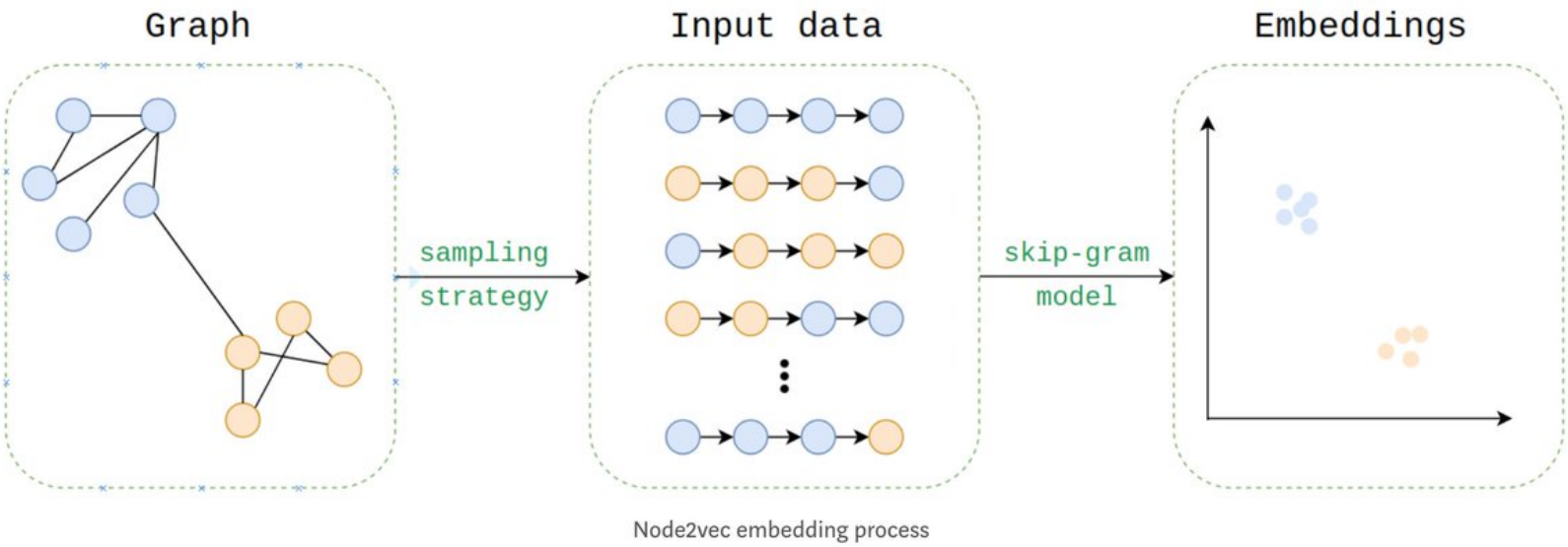
embed mention + embed MeSH® concept + softmax layer

node2vec [GL16] Embeddings



Node2vec embedding process

node2vec [GL16] Embeddings

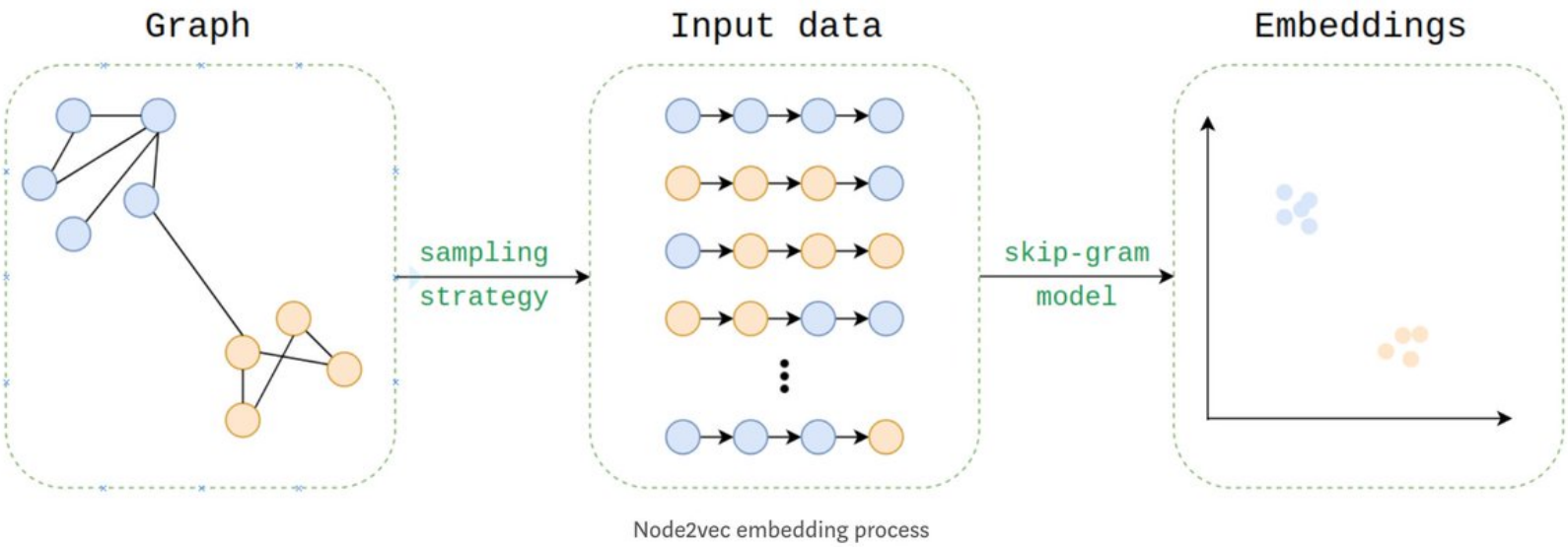


Node2vec embedding process

type I tree structure only

type II tree structure + lexicalization (scope note embedding)

node2vec [GL16] Embeddings



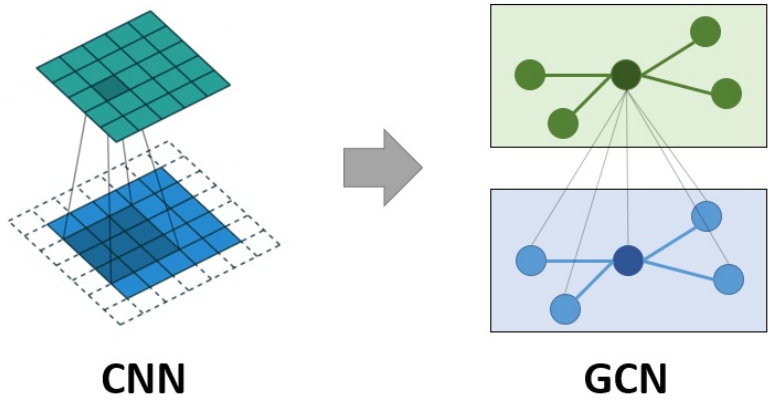
Node2vec embedding process

type I tree structure only

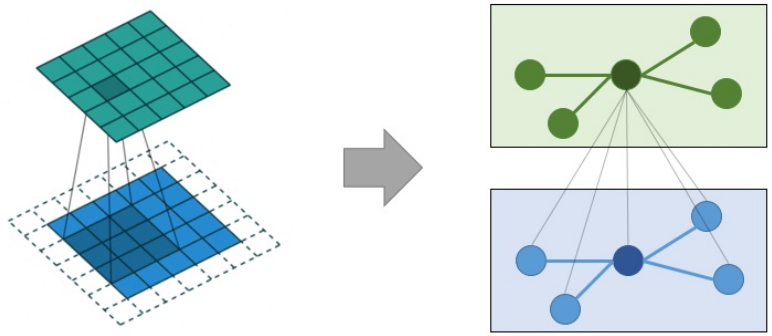
type II tree structure + lexicalization (scope note embedding)

estimate: $p(d_i|\mathbf{m}, \mathbf{d_i}) \propto \exp(\mathbf{m^T W d}_i))$

GCN [BTA+17] Embeddings

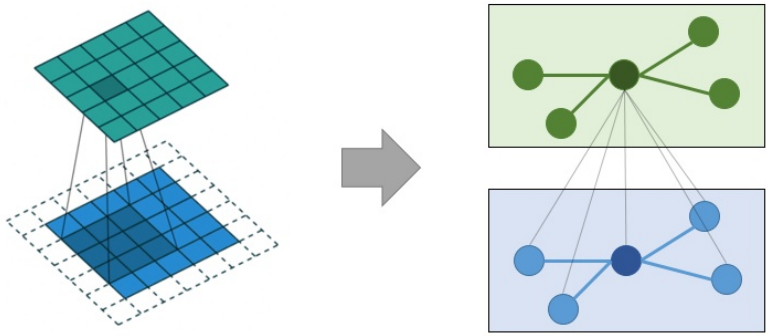


GCN [BTA+17] Embeddings



$$\mathbf{h}_d^{(j+1)} = \sigma\left(\sum_{d' \in \Omega(d)} \mathbf{W}^{(j)} \mathbf{h}_{d'}^{(j)} + \mathbf{b}^{(j)} \right)$$

GCN [BTA+17] Embeddings



$$\mathbf{h}_d^{(j+1)} = \sigma\left(\sum_{d' \in \Omega(d)} \mathbf{W}^{(j)} \mathbf{h}_{d'}^{(j)} + \mathbf{b}^{(j)} \right)$$

estimate: $p(d_i|\mathbf{m}) \propto \exp(\mathbf{m}^{\mathbf{T}} g(d_i; \theta))$

NER, EL & MTL – Results

Model	Pre	Rec	F1	Val. F1
bioELMo [PNI+18]	**0.878 ± 0.003**	**0.856 ± 0.005**	**0.867 ± 0.002**	**0.884 ± 0.001**
Lample et al.** [HWN+17]	0.875**	0.836**	0.844**	-

Model	MRR	F1	Pre@30	Val. MRR
node2vec I	0.749 ± 0.002	0.718 ± 0.004	0.819 ± 0.006	0.800 ± 0.003
node2vec II	**0.757 ± 0.001**	**0.721 ± 0.004**	**0.842 ± 0.004**	**0.804 ± 0.006**
GCN	0.744 ± 0.006	0.710 ± 0.008	0.831 ± 0.005	0.803 ± 0.007
NormCo** [WKMH19]	-	0.840**	-	-

Model	NER			EL	
	Pre	Rec	F1	MRR	Pre@30
NER & EL	**0.880 ± 0.003**	**0.872 ± 0.008**	**0.876 ± 0.003**	0.747 ± 0.003	0.816 ± 0.006

Common Errors

NER Multi-token entities partially detected: in sporadic T-PLL only the head T-PLL is detected

EL Diseases are confused with their MeSH® neighbors: in

> *"Occasional missense mutations in ATM were also found in tumour DNA from patients with [B-cell non-Hodgkins lymphomas]$_{D016393}$ ([B-NHL]$_{D008228}$) and a [B-NHL]$_{D018239}$ cell line."*

D016393 is correct but B-NHL is mapped to D008228 (its child) and then to D018239 (another form of cancer)

Conclusions

1. Adapted graph embeddings that exploit both MeSH®'s hierarchical structure and the description of diseases (scope notes) for EL
2. Combined NER and EL models in a MTL setting

Conclusions

1. Adapted graph embeddings that exploit both MeSH®'s hierarchical structure and the description of diseases (scope notes) for EL
2. Combined NER and EL models in a MTL setting
3. Main findings:

EL	node lexicalization improves over structural or lexical embeddings
NER	bioELMO leads to large gains
MTL	leads to better performance for NER

Conclusions

1. Adapted graph embeddings that exploit both MeSH®'s hierarchical structure and the description of diseases (scope notes) for EL
2. Combined NER and EL models in a MTL setting
3. Main findings:

EL	node lexicalization improves over structural or lexical embeddings
NER	bioELMO leads to large gains
MTL	leads to better performance for NER

4. Further work:
   - incorporate EL optimizations studied in [WKMH19]
   - enlarge target taxonomy by linking it to large knowledge graphs such as DBpedia

Thank you!

References I

[BTA+17] Joost Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. Graph convolutional encoders for syntax-aware neural machine translation. *CoRR*, abs/1704.04675, 2017.

[DL12] Rezarta Islamaj Dogan and Zhiyong Lu. An inference method for disease name normalization. In *AAAI Fall Symposium: Information Retrieval and Knowledge Discovery in Biomedical Text*, 2012.

[GL16] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. *CoRR*, abs/1607.00653, 2016.

[HWN+17] Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 2017.

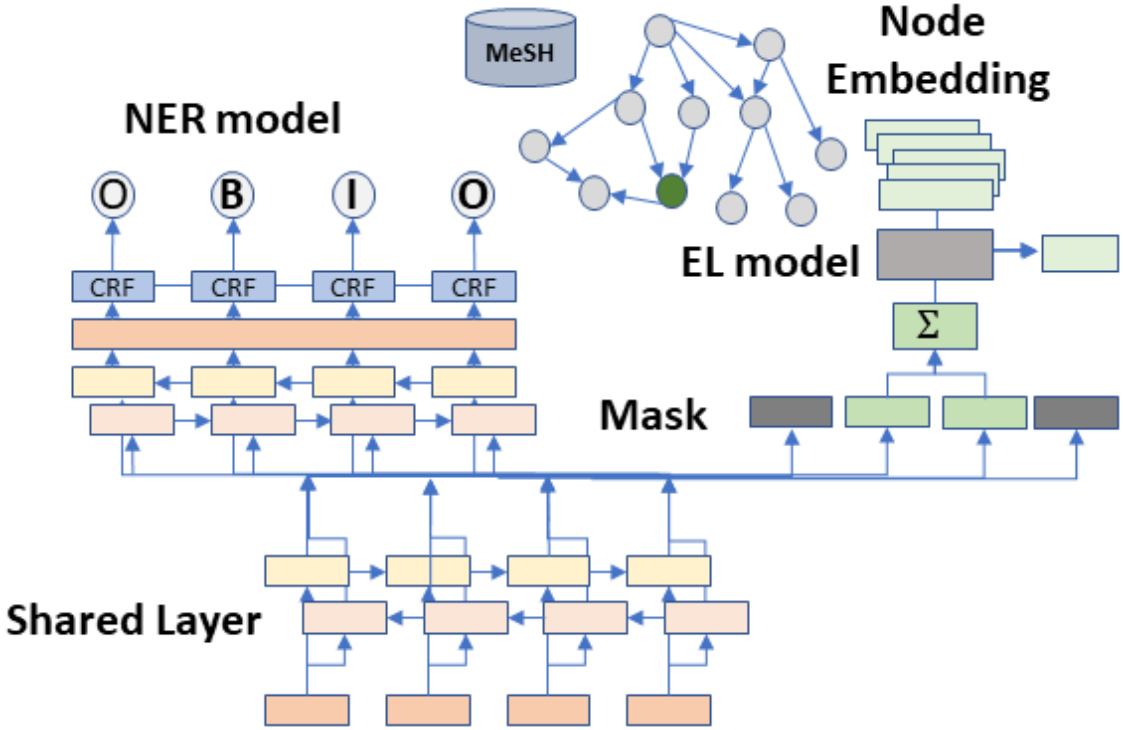
References II

[Lip00] C.E. Lipscomb. Medical subject headings (MeSH). *Bull Medical Library Association*, 88(3):265–266, 2000.

[PNI+18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL 2018*, 2018.

[WKMH19] Dustin Wright, Yannis Katsis, Raghav Mehta, and Chun-Nan Hsu. NormCo: Deep disease normalization for biomedical knowledge base construction. In *Proceedings of AKBC 2019*, 2019.

Appendix – MTL Architecture



biLTSM-CRF + mention embedding + MeSH® concept embedding + softmax layer

Appendix – CRFs vs. biLSTM-CRFs

- Traditional linear-chain CRFs estimate:

$$p(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}) \propto \prod_{i=1}^{n} \exp(\sum_{k=1}^{K} \theta_k f_k(y_i, y_{i+1}, \mathbf{x}_{1:n}))$$

- biLSTM-CRFs estimate:

$$p(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}) \propto \exp(s(\mathbf{x}_{1:n}, \mathbf{y}_{1:n})) = \exp(\sum_{i=0}^{n} A_{y_i,y_{i+1}} + \sum_{i=1}^{n} P_{x_i,y_i})$$

Appendix – Hyperparameters/Training

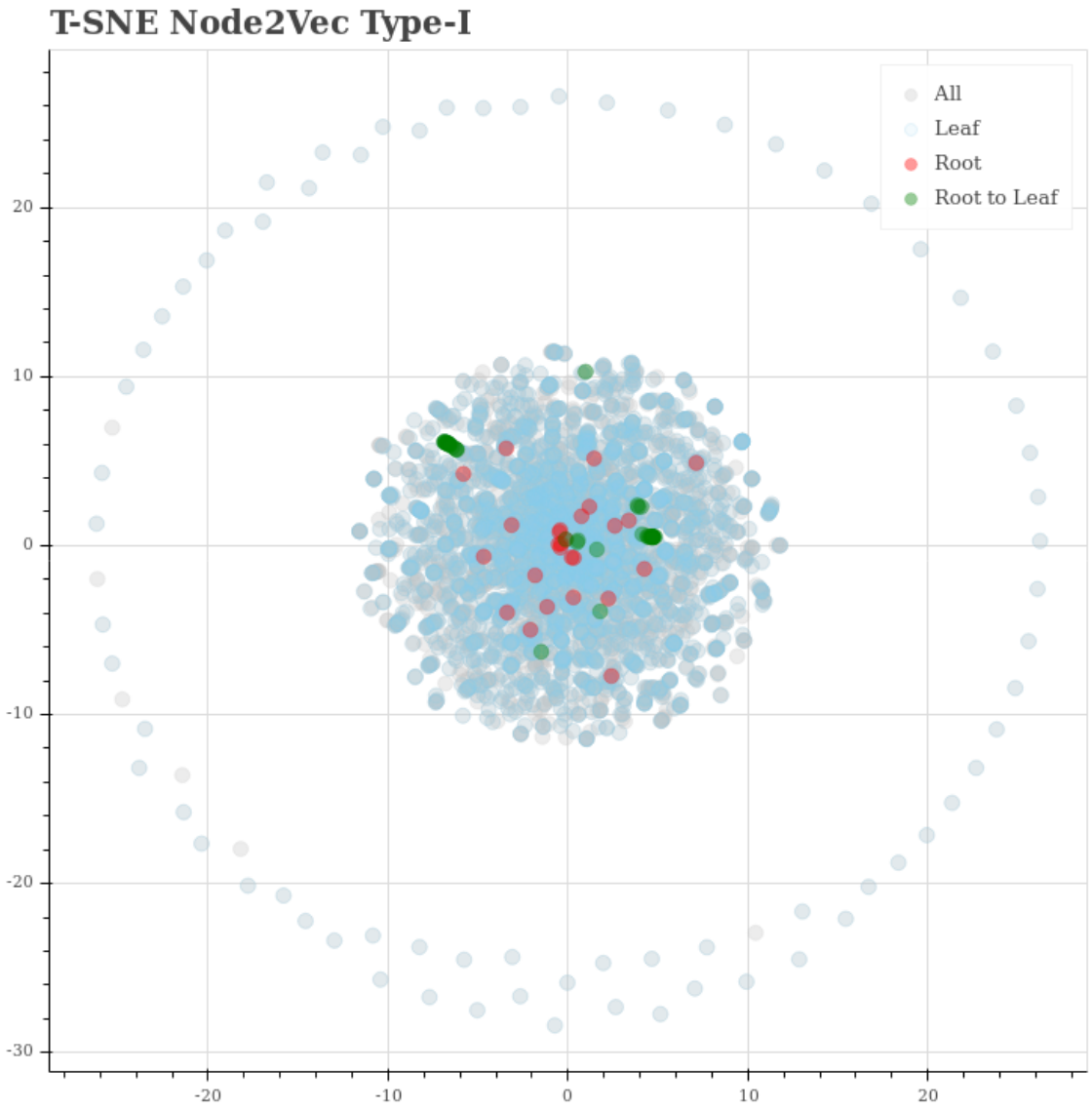
Common:
- ADAM with 10^{-3} learning rate
- 0.5 dropout regularization
- split the abstracts into sentences using NLTK (https://www.nltk.org/)
- used whitespace tokenization
- 200-dim *word2vec* embeddings and 1024-dim bioELMo embeddings

NER:
- learnt 60-dim character embeddings

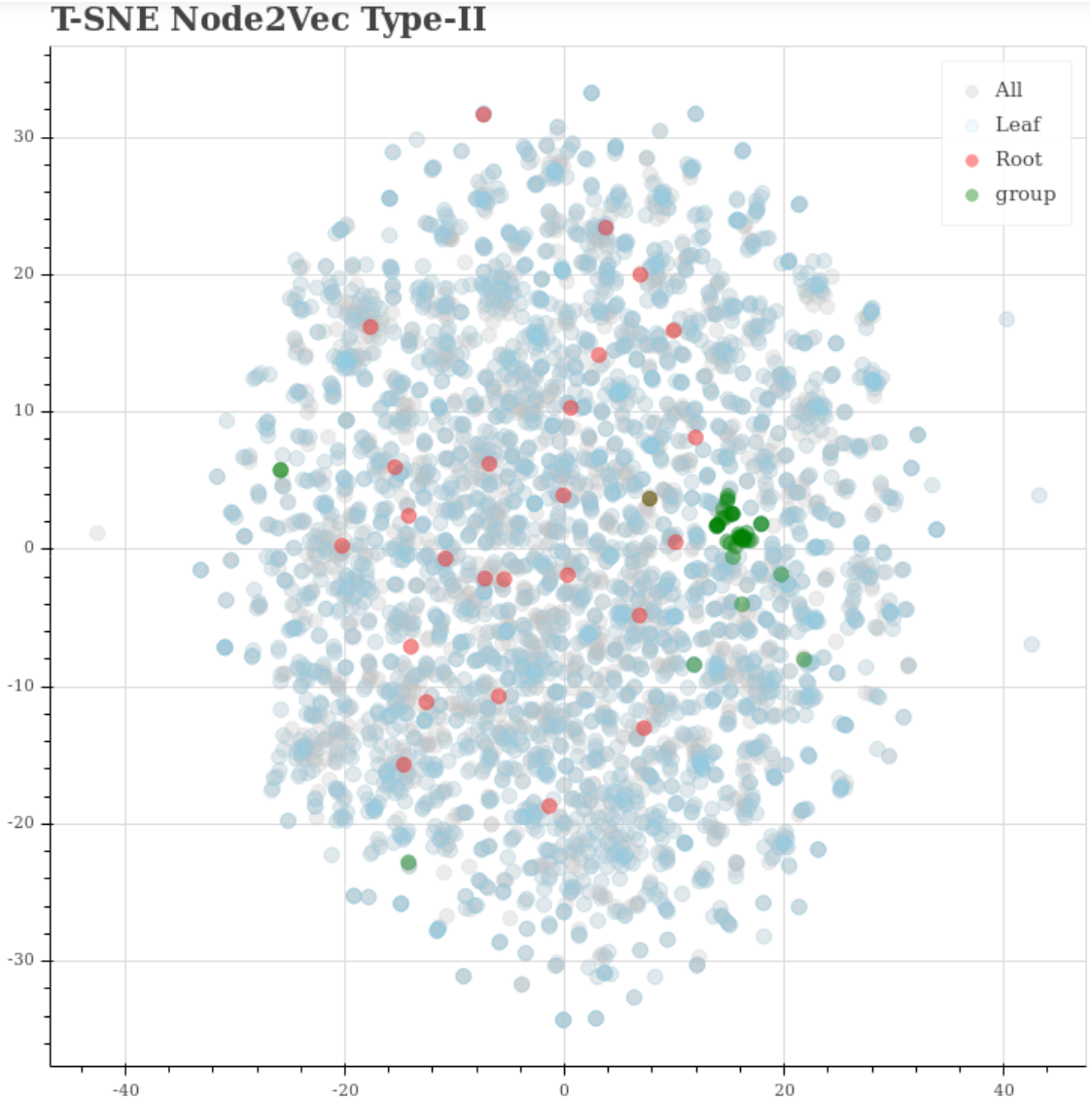
node2vec:
- 1024-dim MeSH® node embeddings trained using *node2vec* for 100 epochs

GCN:
- 1024-dim EL models trained for 500 epochs
- stacked 2 GCN layers with 2048 hidden units and 1024 output units

Appendix – Embedding Visualization I



Appendix – Embedding Visualization II



Appendix – Embedding Visualization III