Neural Disease Normalization with Graph Embeddings

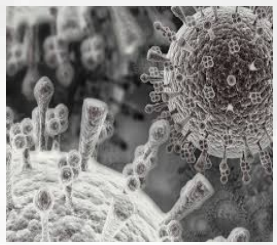
Dhruba Pujary[1]* Camilo Thorne[2] Wilker Aziz[1]

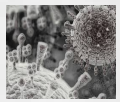
[1] UvA (Amsterdam, Netherlands)

[2] Elsevier (Frankfurt, Germany)

* Email: druv022@gmail.com (main author)

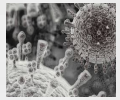
AI Meetup, Frankfurt, 6.11.2019





Motivation

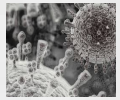
1. To semantically enrich biomedical text we need to detect entities (NER), and resolve them (EL) to a **canonical** name
2. Canonical names, a.k.a. **concepts**, are defined in gold repositories such as:
   $\Rightarrow$ thesauri, databases, taxonomies, knowledge graphs
3. Systems used in practice are often based on dictionaries/TFIDF:
   $\Rightarrow$ high precision, low recall
4. Big advances in NER and EL for other domains

Motivation

1. To semantically enrich biomedical text we need to detect entities (NER), and resolve them (EL) to a **canonical** name
2. Canonical names, a.k.a. **concepts**, are defined in gold repositories such as:
   $\Rightarrow$ thesauri, databases, taxonomies, knowledge graphs
3. Systems used in practice are often based on dictionaries/TFIDF:
   $\Rightarrow$ high precision, low recall
4. Big advances in NER and EL for other domains

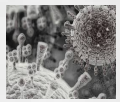
Q: Can we use DL and **graph embeddings** to detect and resolve **diseases**?

Example – NCBI Corpus [2] & MeSH® Taxonomy [7]

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 ...
	⋮

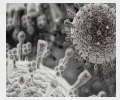


Dataset Statistics

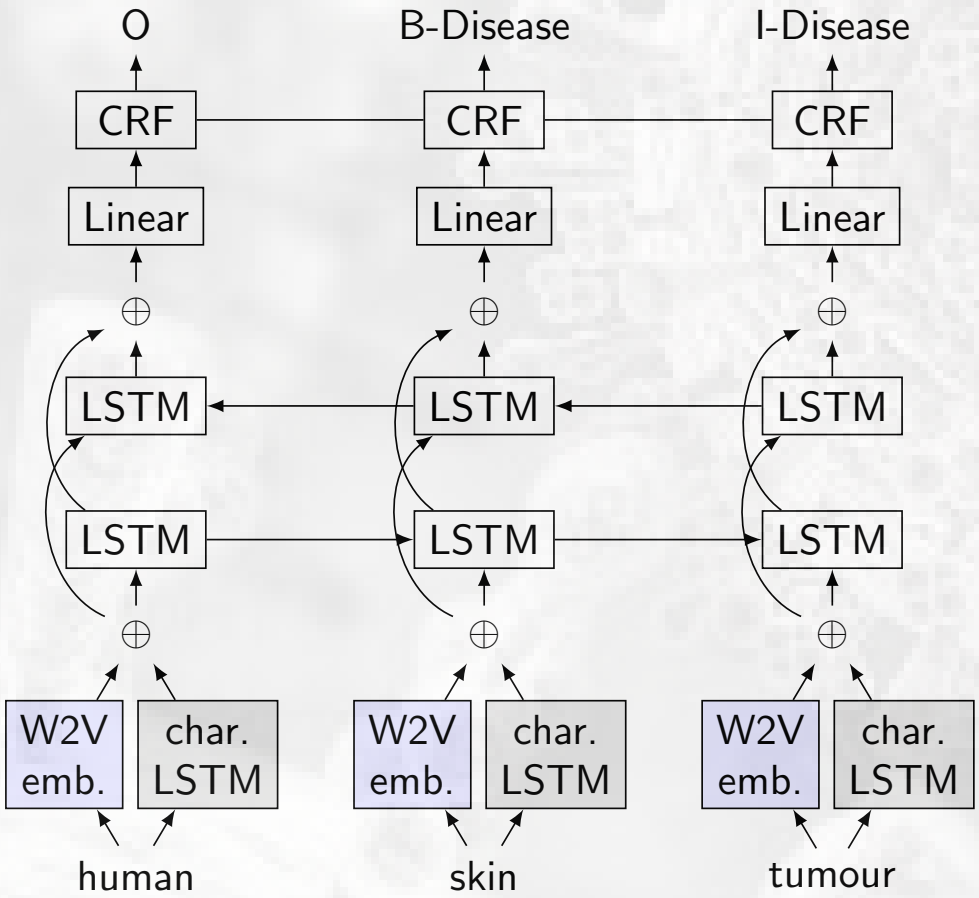
- NCBI corpus:

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)
  - ▷ approx. 10,000 scope notes comprising a 10-20 tokens
  - ▷ approx. 100,000 – 200,000 tokens

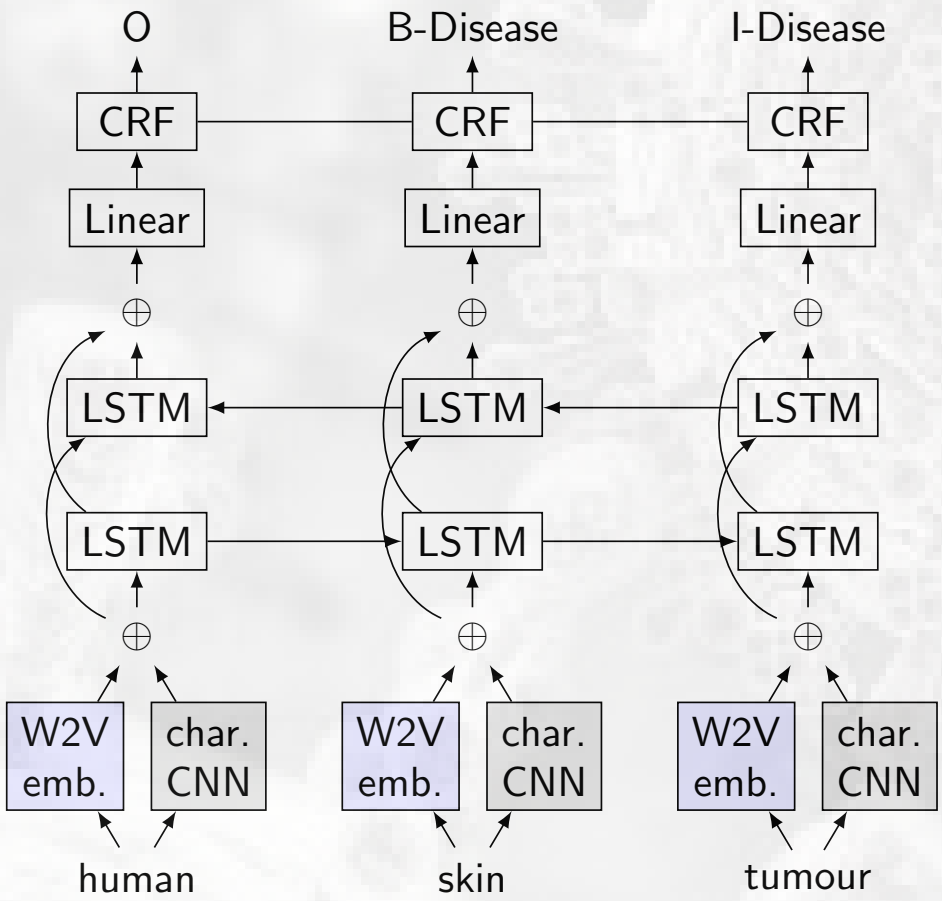



Disease NER – biLSTM-CRF Model I



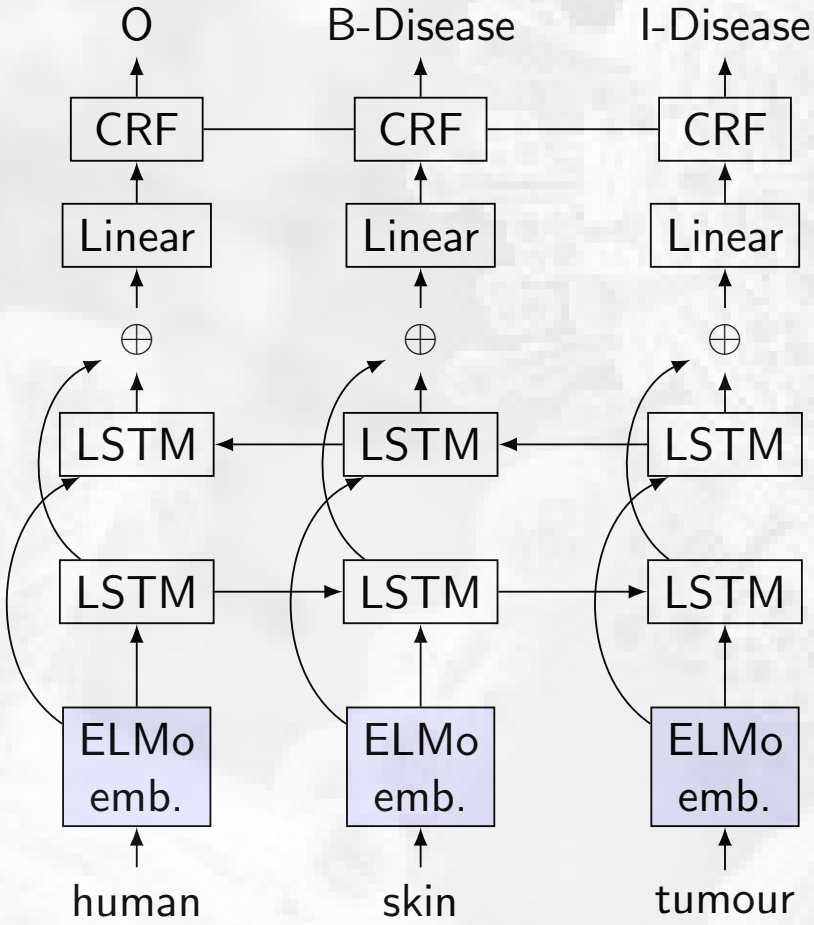
Lample et al. [5, 4, 10]

Disease NER – biLSTM-CRF Model II

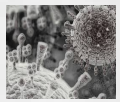


Ma and Hovy [8, 10]

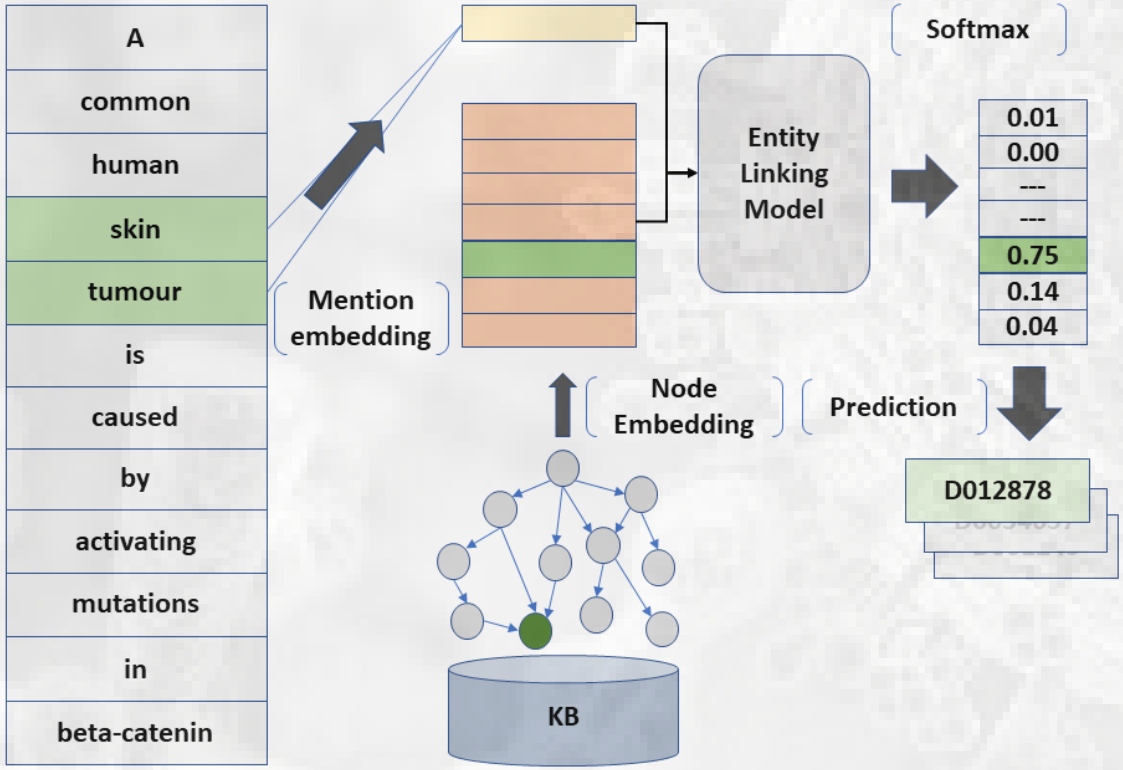
Disease NER – biLSTM-CRF Model III



bioELMo [9]

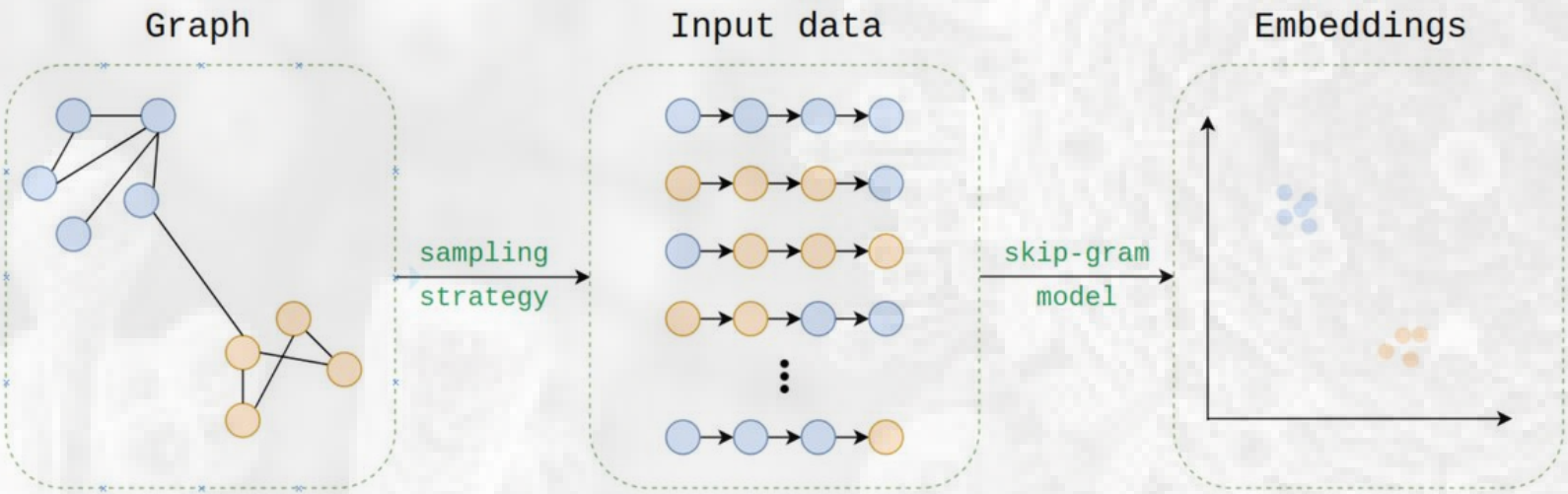
EL Model



- in a nutshell:

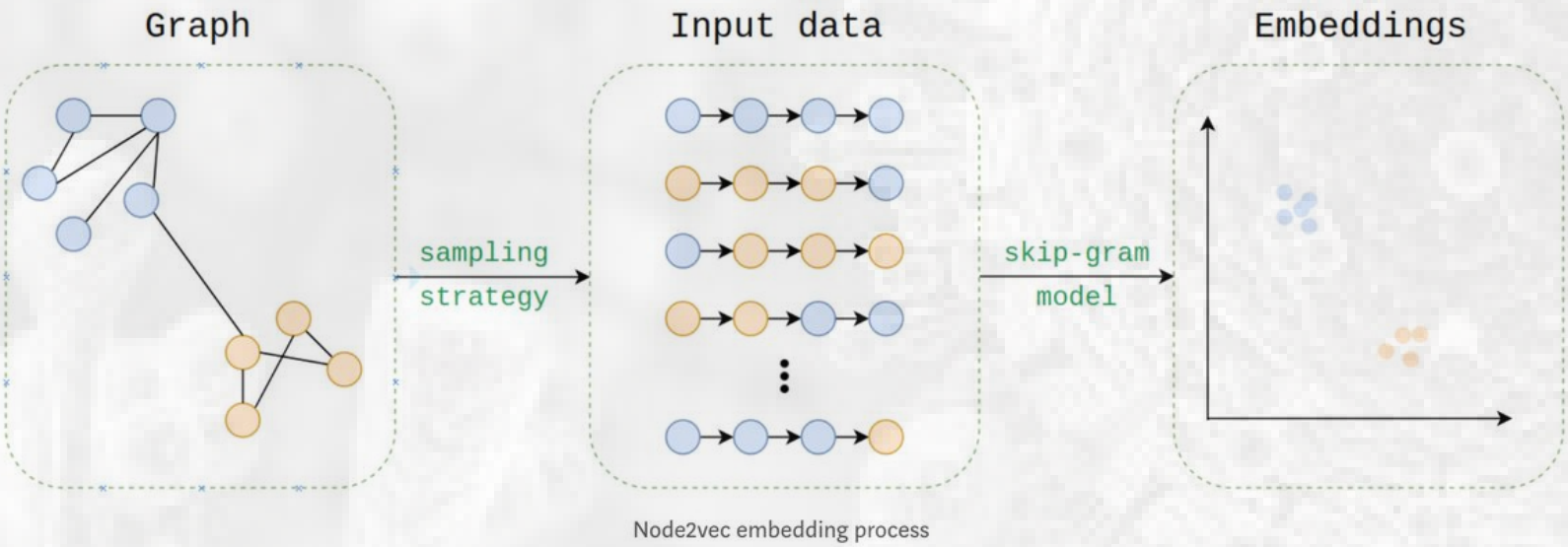
  embed mention + embed MeSH® concept + softmax layer

EL – *node2vec* [3] Embeddings



Node2vec embedding process

EL – *node2vec* [3] Embeddings

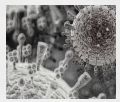


Node2vec embedding process

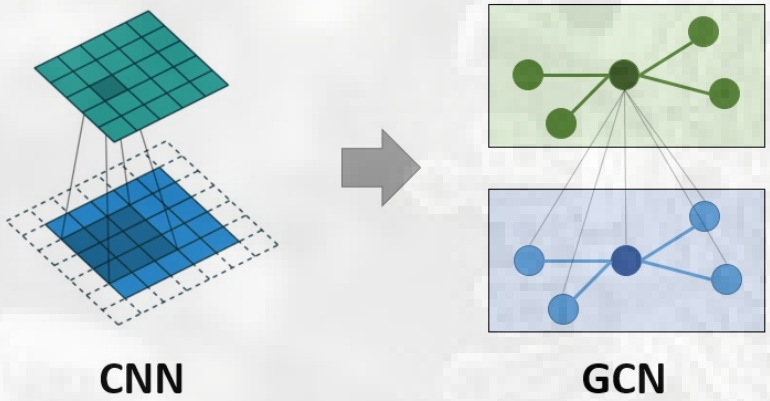
estimate: $p(d_i|\mathbf{m}, \mathbf{d_i}) \propto \exp(\mathbf{m^T W d}_i))$

type I tree structure only

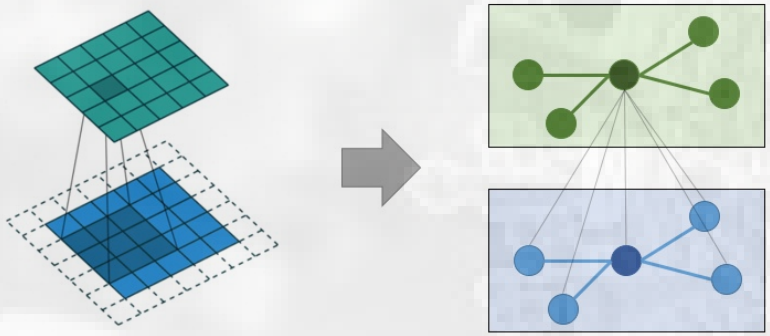
type II tree structure + lexicalization (scope note embedding)



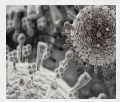
EL – GCN [1] Embeddings



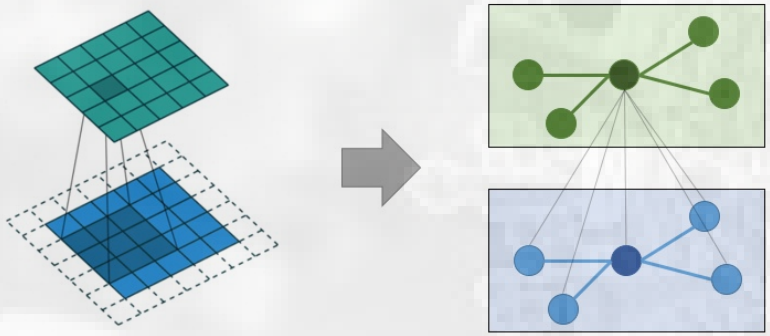
EL – GCN [1] Embeddings



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

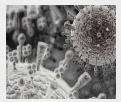



EL – GCN [1] Embeddings



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

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

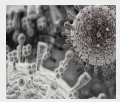



NER & EL – Results

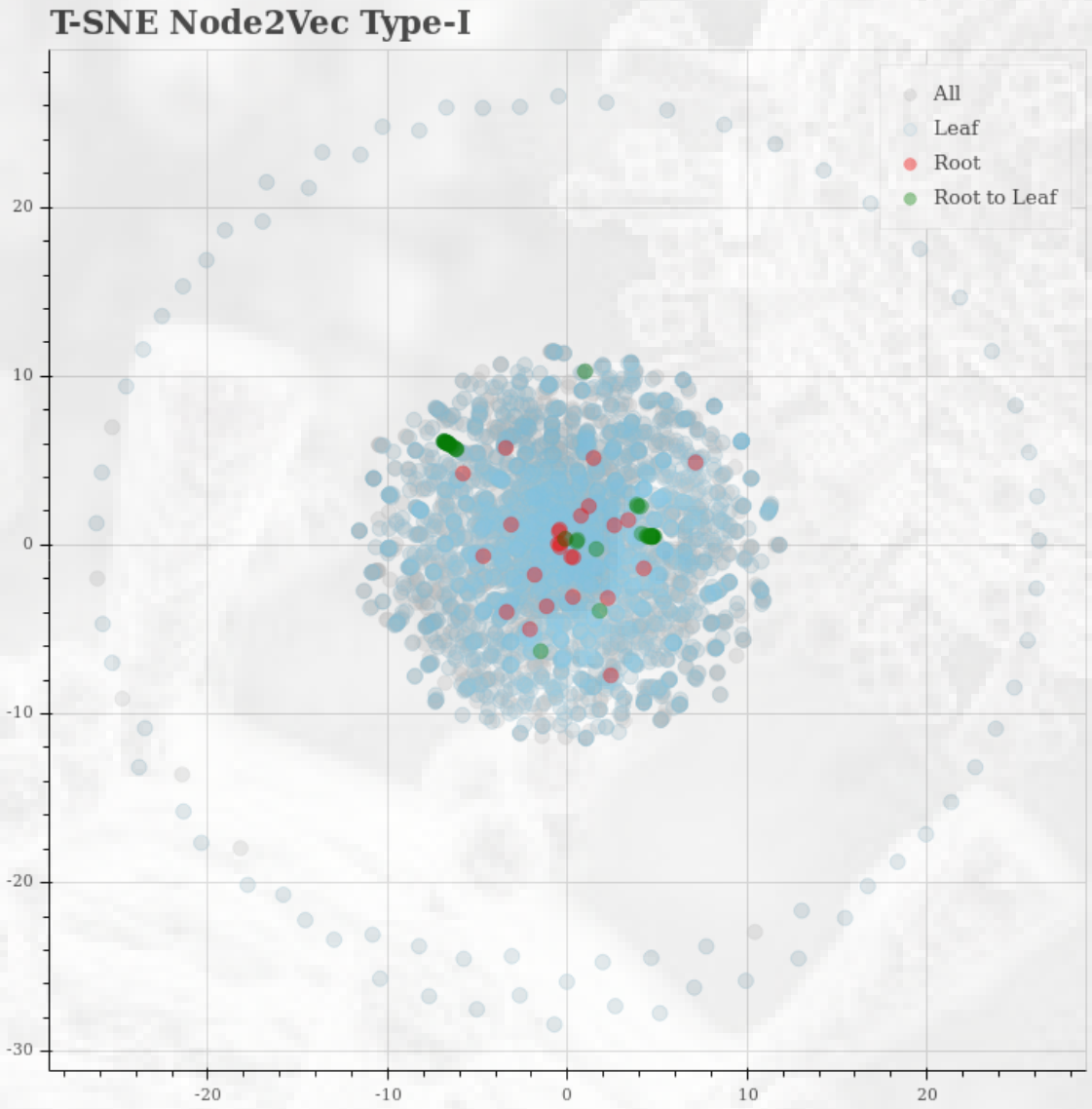
Model	Pre	Rec	F1	Val. F1
Lample et al.	0.824 ± 0.022	0.742 ± 0.019	0.781 ± 0.003	0.805 ± 0.007
Ma and Hovy	0.823 ± 0.011	0.776 ± 0.023	0.799 ± 0.012	0.792 ± 0.005
bioELMo	**0.878 ± 0.003**	**0.856 ± 0.005**	**0.867 ± 0.002**	**0.884 ± 0.001**
bioELMo + 2-layer biLSTM	0.857 ± 0.006	0.873 ± 0.005	0.865 ± 0.005	0.884 ± 0.001
Lample et al.** [4]	0.875**	0.836**	0.844**	-

Model	MRR	F1	Pre	Pre@30	Val. MRR
bioELMo (S.N.)	0.748 ± 0.002	0.715 ± 0.004	0.715 ± 0.002	0.844 ± 0.004	0.791 ± 0.001
node2vec I	0.749 ± 0.002	0.718 ± 0.004	0.720 ± 0.004	0.819 ± 0.006	0.800 ± 0.003
node2vec II	**0.757 ± 0.001**	**0.721 ± 0.004**	**0.724 ± 0.001**	**0.842 ± 0.004**	**0.804 ± 0.006**
GCN	0.744 ± 0.006	0.710 ± 0.008	0.710 ± 0.007	0.831 ± 0.005	0.803 ± 0.007
DNorm** [6]	-	0.782**	-	-	-
NormCo** [11]	-	0.840**	0.878**	-	-

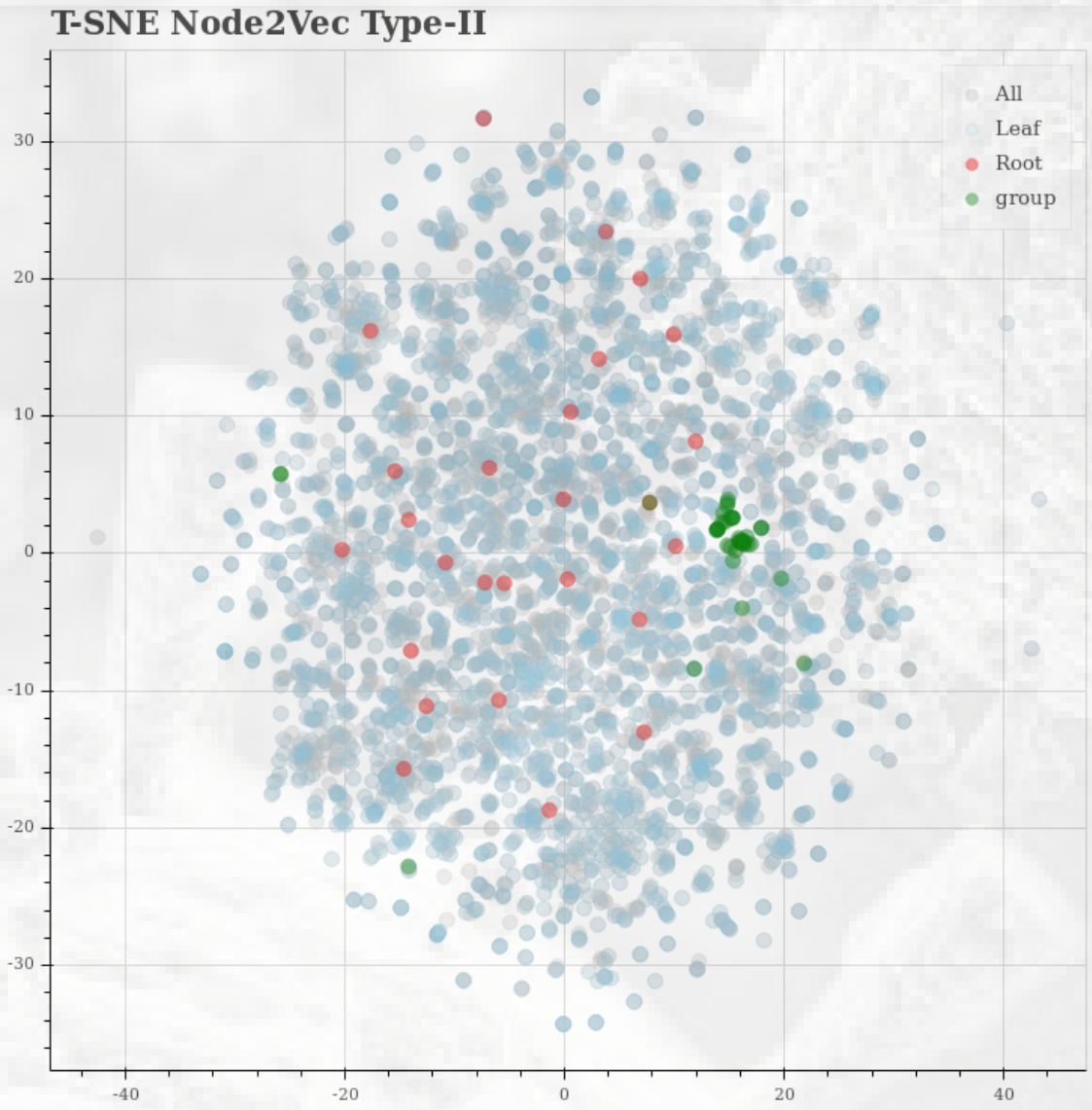
$$(\text{where: MRR} = \frac{1}{|E|}\sum_{i=1}^{|E|}\frac{1}{\text{rank}_i})$$



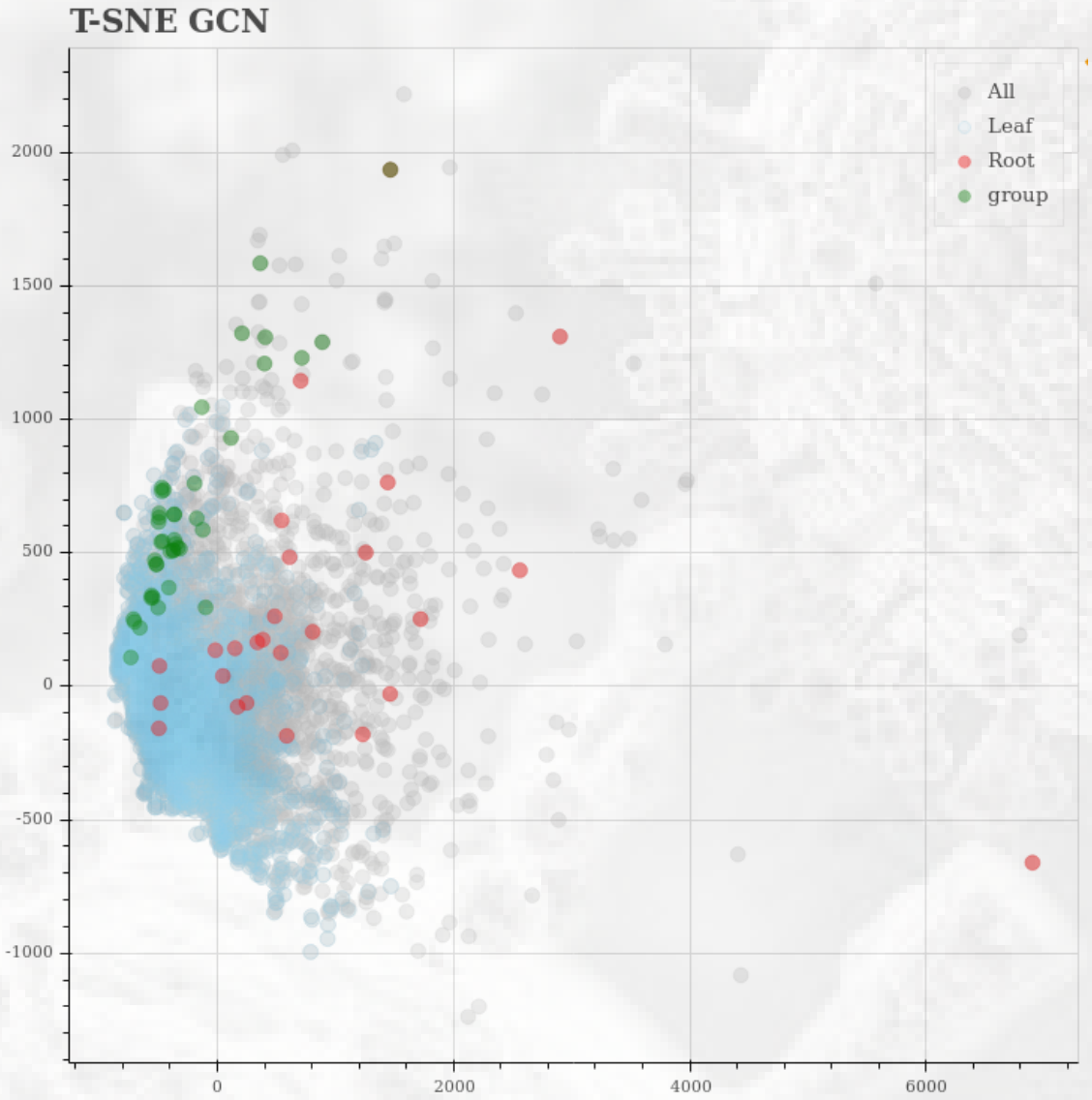
EL – Embedding Visualization I



EL – Embedding Visualization II



EL – Embedding Visualization III

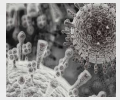


Main Errors

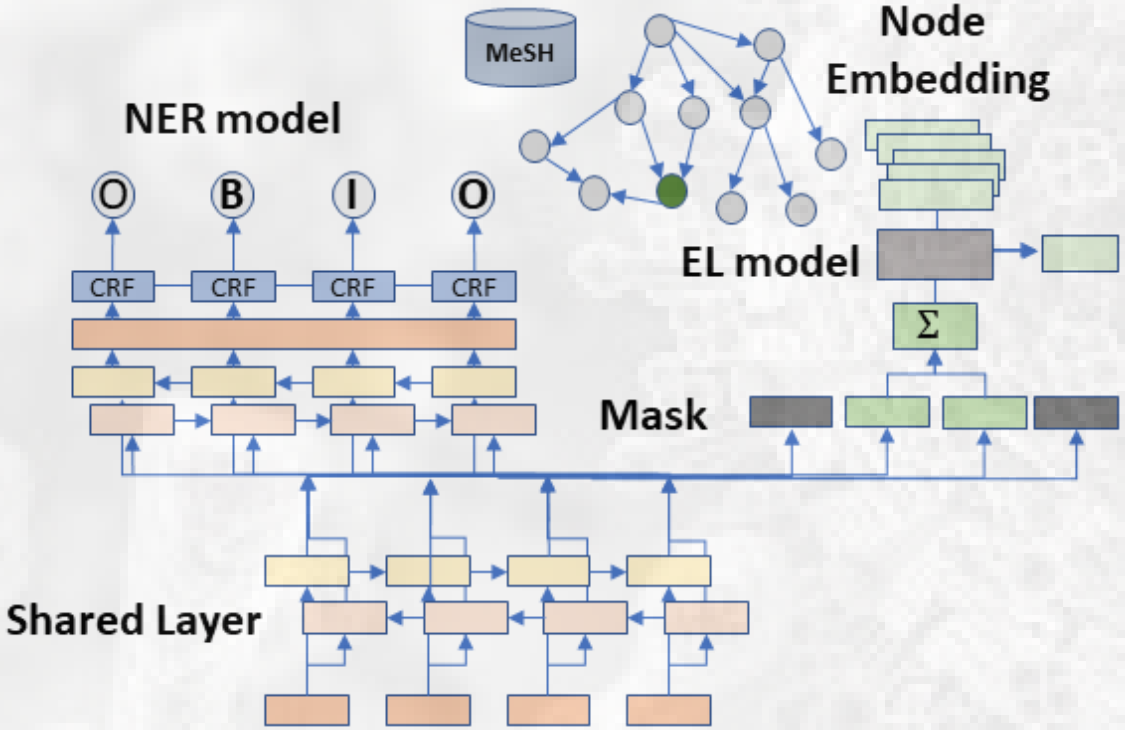
- Multi-token entities only partially detected: in sporadic T-PLL only the head T-PLL is detected
- EL models confuse diseases with their MeSH® neighbors: D016399 - Lymphoma, T-Cell is confused for D015458 - Leukemia, T-Cell, with which it shares an ancestor
- EL models resolve correctly the first instance, but return a neighbour after: 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.*

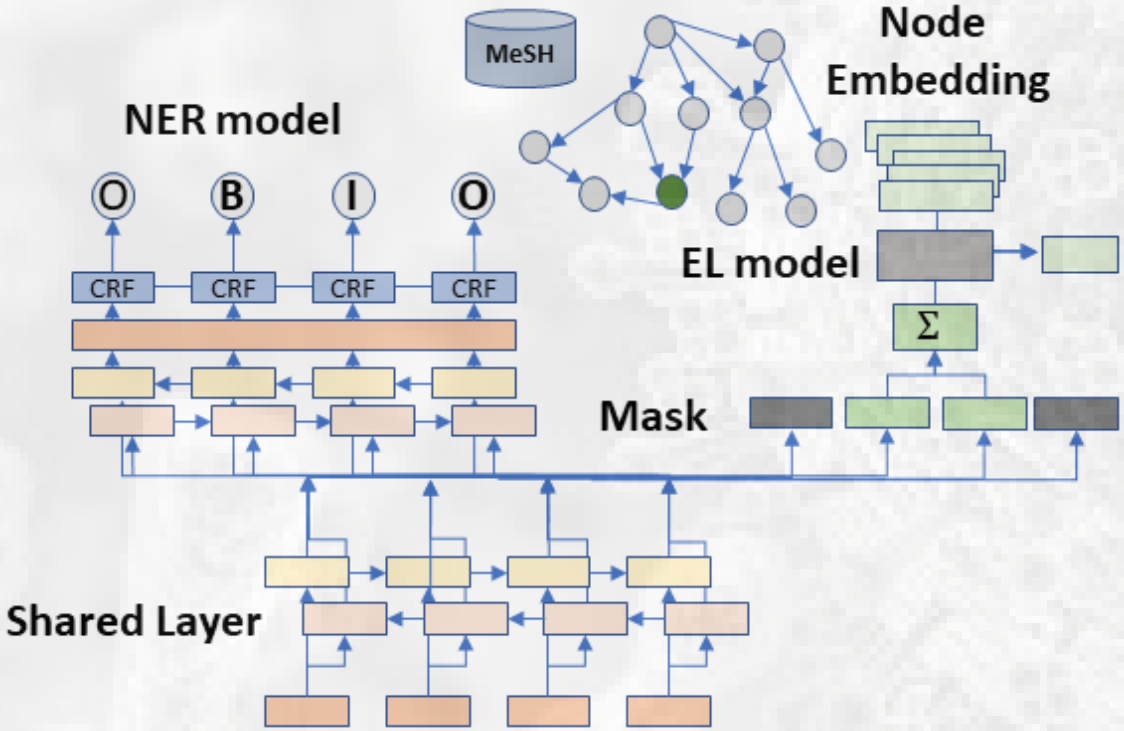
  D016393 is correct but B-NHL is mapped to D008228 (its child) and then to D018239 (another form of cancer)

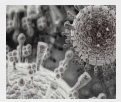
MTL – Model & Results



MTL – Model & Results



Model	NER			EL	
	Pre	Rec	F1	MRR	Pre@30
NER & GCN	**0.880 ± 0.003**	**0.872 ± 0.008**	**0.876 ± 0.003**	**0.747 ± 0.003**	**0.816 ± 0.006**
NER	0.875 ± 0.006	0.869 ± 0.001	0.872 ± 0.003	-	-
GCN	-	-	-	0.745 ± 0.001	0.816 ± 0.001



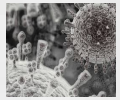
Conclusions

1. Adapted biLSTM-CRFs to the NCBI corpus (NER)
2. Adapted graph embeddings (GCN and *node2vec*) that exploit both MeSH®'s hierarchical structure and the description of diseases (EL)
3. Combined NER and EL models in a MTL setting
4. Main findings:

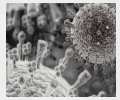
   EL node lexicalization improves over structural or lexical embeddings

   NER bioELMO leads to large gains for NER

   MTL leads to state-of-the-art performance for NER
5. Further work:
   - incorporate EL optimizations studied in [11] and [6]
   - enlarge target taxonomy by linking it to large knowledge graphs such as DBpedia

Thank you!



References I

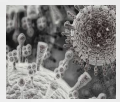
[1] Joost Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. Graph convolutional encoders for syntax-aware neural machine translation. *CoRR*, abs/1704.04675, 2017.

[2] 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.

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

[4] 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.

[5] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *CoRR*, abs/1603.01360, 2016.



References II

[6] Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917, 2013.

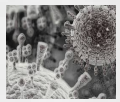
[7] C.E. Lipscomb. Medical subject headings (MeSH). *Bull Medical Library Association*, 88(3):265–266, 2000.

[8] Xuezhe Ma and Eduard H. Hovy. End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF. *CoRR*, abs/1603.01354, 2016.

[9] 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.

[10] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou. Distributional semantics resources for biomedical text processing. In *Proceedings of LBM 2013*, 2013.

[11] 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.



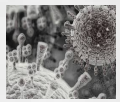
Appendix – CRFs vs. biLSTM-CRFs

- 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