

A Journey of Deep Neural Networks for Short and Long Text Understanding

Yuan Luo

Assistant Professor

Department of Preventive Medicine

Departments of IEMS and EECS (Courtesy)

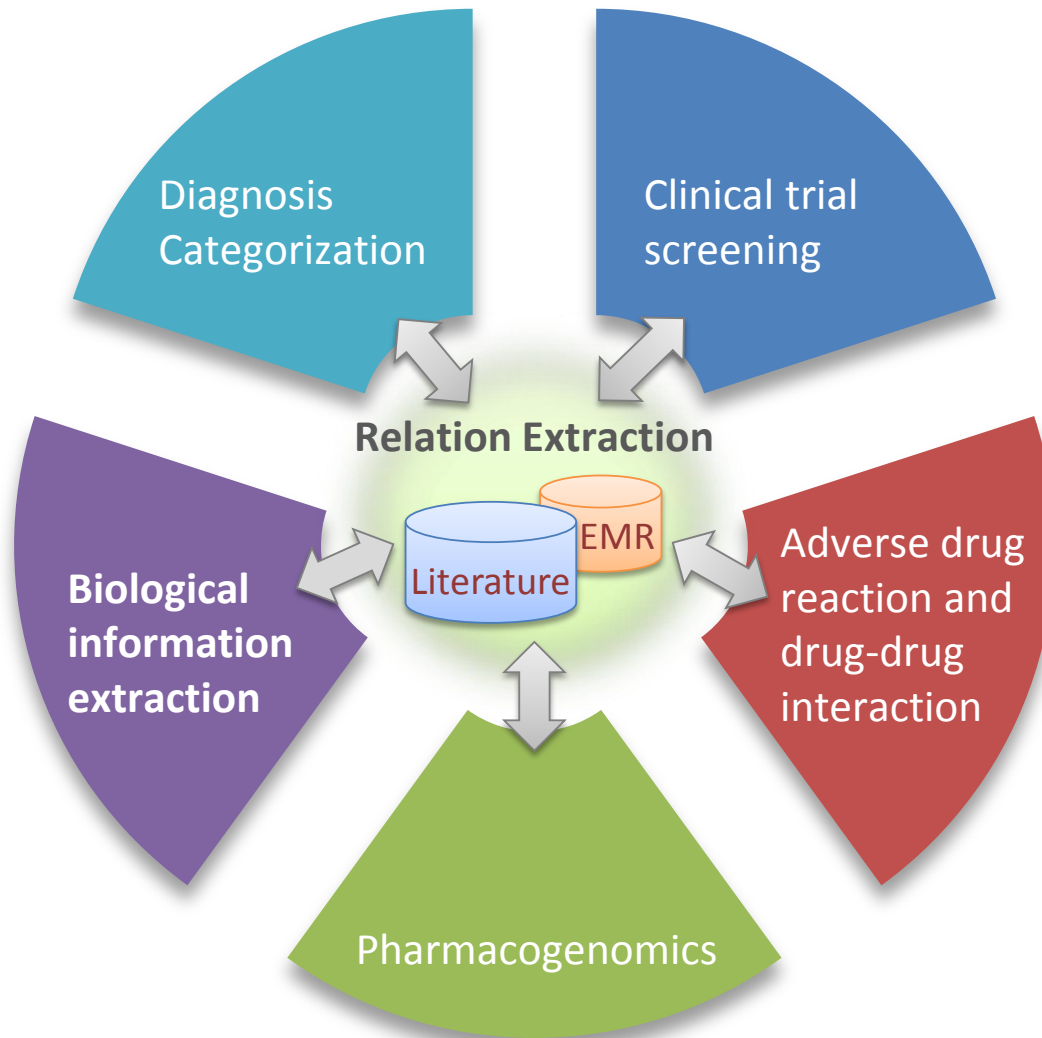
Northwestern University

yuan.luo@northwestern.edu

 @yuanhypnosluo

© Yuan Luo. All Rights Reserved

DNNs for Short Biomedical Text Understanding



Y Luo, Ö Uzuner, P Szolovits. Bridging Semantics and Syntax with Graph Algorithms - State-of-the-Art of Extracting Biomedical Relations. *Briefings in Bioinformatics* 2016 18 (1), 160-178. *PMCID*: 5221425

Biomedical Relations in Diagnosis Categorization

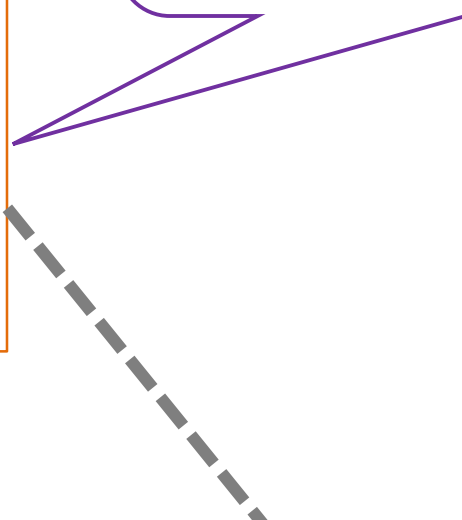
Interpretability

Panel of immunophenotypic test result
 Neoplastic cells express CD19
 Neoplastic cells express CD20
 Neoplastic cells express CD22
 Neoplastic cells express CD79a

WHO guideline: The neoplastic cells express pan B-cell markers such as CD19, CD20, CD22 and CD79a, but may lack one or more of these.

Unsupervised learning

Expert annotation expensive
 Evolving guideline



BCL2	BCL6	CD10	large cells	positive	negative	...	Class label
1	1	0	1	1	0		y
0	0	1	1	0	1		n
1	1	1	1	1	1		n

i2b2/VA Relation Classification Challenge

Treatment Problem Relations

- TrAP (Treatment administered for problem)
 - he was given **Entresto** to treat his **high blood pressure**
- TrNAP (Treatment not administered because of the problem)
 - **Relafen** which is contraindicated because of **ulcers**
- TrIP (Treatment improves problem)
 - **infection** resolved with a full course of **cephalexin**
- TrCP (Treatment causes problem)
 - the patient took **amoxicillin** for two days, which caused **diarrhea**
- TrWP (problem has deteriorated or worsened because of or in spite of a treatment)
 - the **tumor** was growing despite the **drain**
- TrP None (no relation between treatment and problem)

Ö Uzuner, BR South, S Shen, SL DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*. 2011 Jun 16;18(5):552-6.

i2b2/VA Relation Classification Challenge

Test Problem Relations

- TeRP (Test has revealed some medical problem)
 - an echocardiogram revealed a pericardial effusion
- TeCP (Test was performed to investigate a medical problem)
 - chest x-ray done to rule out pneumonia
- TrP None (no relation between test and problem)

Problem Problem Relations

- PIP (Two problems are related to each other)
 - azotemia presumed secondary to sepsis
- PP None (no relation between two problems)

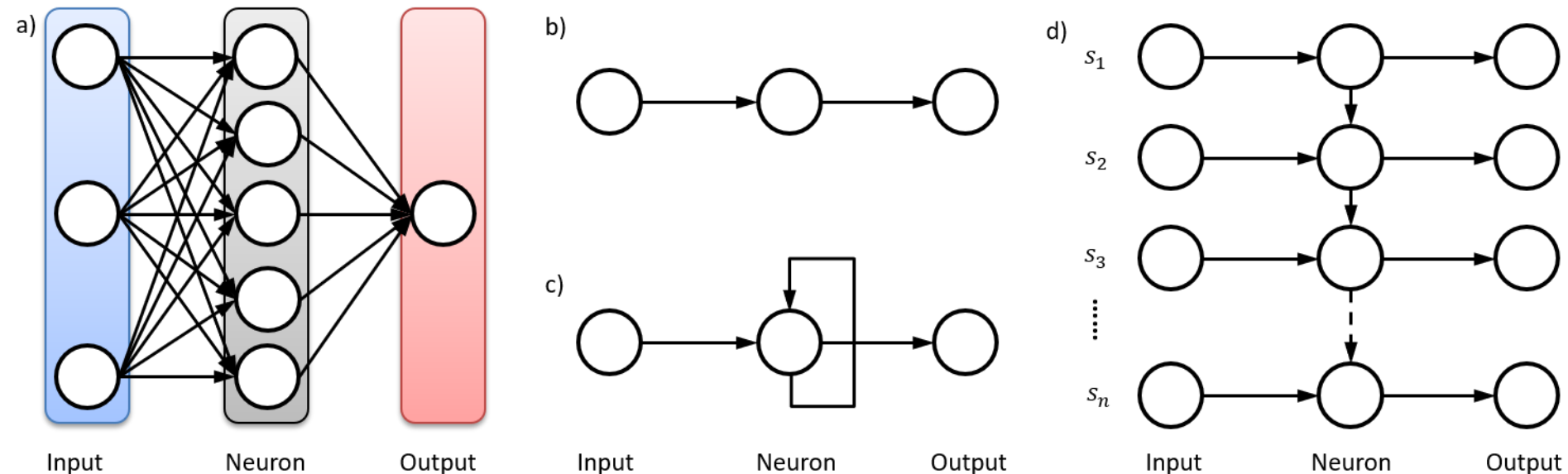
Distribution of i2b2/VA Relation Classes

Relation Type	Training	Training %	Test	Test %	Effective Training*
PIP	1239	38.4%	1986	61.6%	1123
PP None	7349	39.64%	11190	60.36%	4453
TeCP	303	34.0%	588	66.0%	271
TeRP	1734	36.4%	3033	63.6%	1564
TeP None	1535	38.50%	2452	61.50%	1379
TrAP	1423	36.4%	2487	63.6%	1284
TrCP	296	40.0%	444	60.0%	270
TrIP	107	35.1%	198	64.9%	100
TrNAP	106	35.7%	191	64.3%	101
TrWP	56	28.1%	143	71.9%	48
TrP None	2329	40.05%	3486	59.95%	2081

*Effective Training denotes the number of samples used to train each class. It is less than the number of samples in the training dataset due to random allocation of 10% training dataset as validation set for all relations, and down-sampling for PP relations.

Basics on Neural Networks

- From multi-layer neural networks to recurrent neural networks



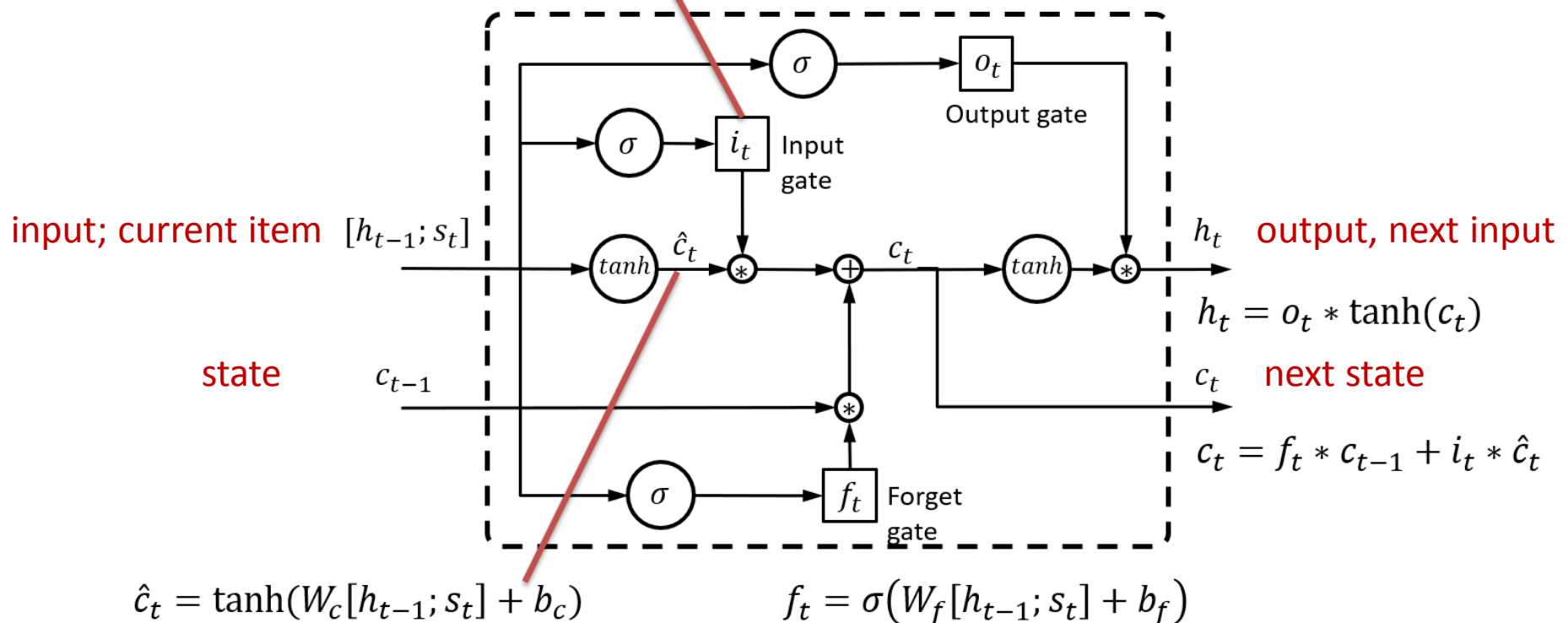
Recurrent Neural Network

The building blocks – LSTM memory cell

given a text sequence $[s_1; s_2; \dots; s_n]$, at each step $t = 1, \dots, n$.

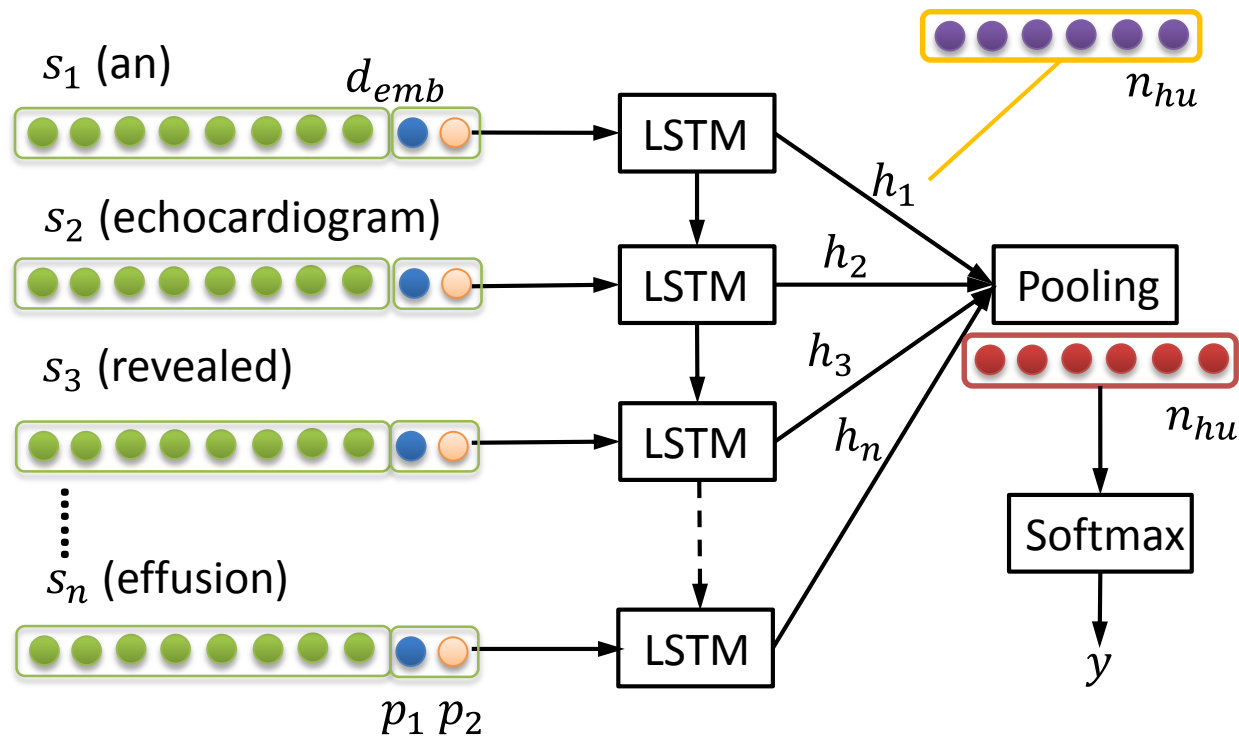
$$i_t = \sigma(W_i[h_{t-1}; s_t] + b_i)$$

$$o_t = \sigma(W_o[h_{t-1}; s_t] + b_o)$$

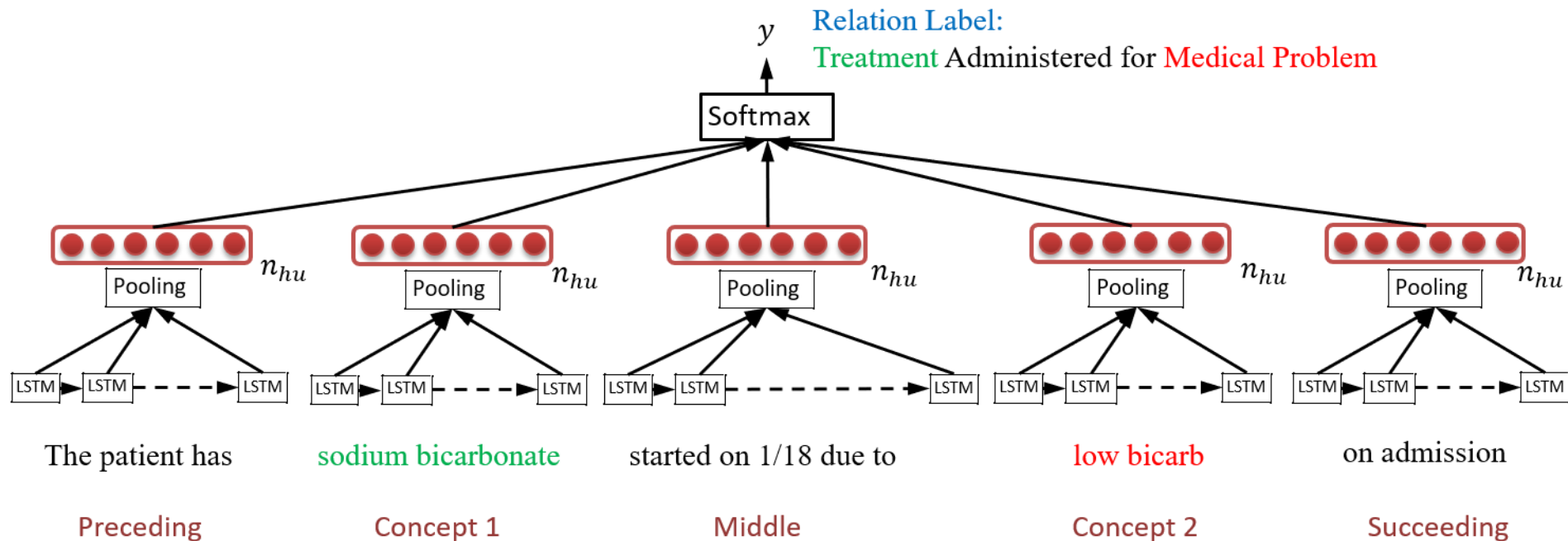


Recurrent Neural Network

The sentence level LSTM model



Segment-LSTM Model



n_{hu} : Number of hidden unit; LSTM: Long short-term memory;
 Preceding, Concept 1, Middle, Concept 2, Succeeding: Five sentence segments partitioned according to concept pairs in the relation

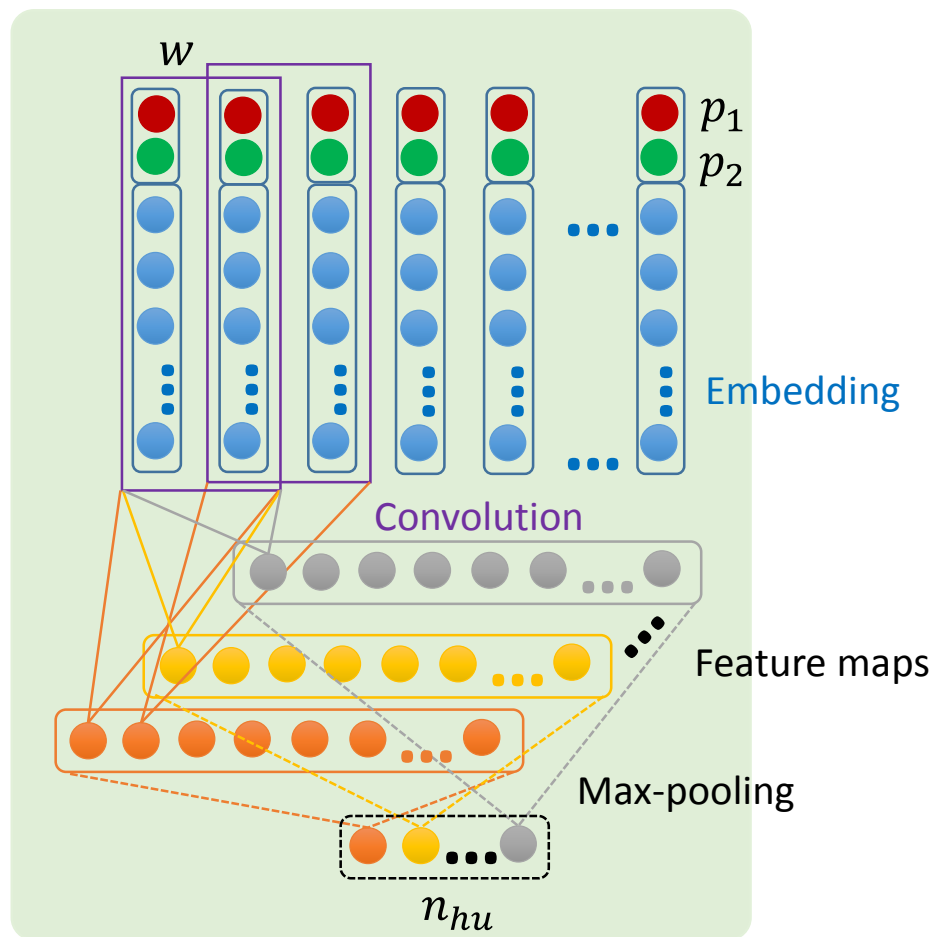
Y Luo. Recurrent neural networks for classifying relations in clinical notes. *Journal of Biomedical Informatics*, 2017, 72: 85-95.

Seg-LSTM Evaluation Results

System	Problem-Treatment (TrP) Relations			Problem-Test (TeP) Relations			Problem-Problem (PP) Relations		
	R	P	F	R	P	F	R	P	F
Segment LSTM mean	0.641	0.683	0.661	0.766	0.838	0.800	0.731	0.640	0.683
Sentence LSTM mean	0.623	0.658	0.640	0.758	0.794	0.775	0.728	0.681	0.704
Roberts et al.	0.686	0.672	0.679	0.833	0.798	0.815	0.726	0.664	0.694
deBruijn et al.	0.583	0.750	0.656	0.789	0.843	0.815	0.712	0.691	0.701
Grouin et al.	0.646	0.647	0.647	0.801	0.792	0.797	0.645	0.670	0.657
Patrick et al.	0.599	0.671	0.633	0.774	0.813	0.793	0.627	0.677	0.651
Jonnalagadda et al.	0.679	0.581	0.626	0.828	0.765	0.795	0.730	0.586	0.650
Divita et al.									0.610
Solt et al.									0.655
Demner-Fushman et al.									0.691
Anick et al.	0.619	0.596	0.608	0.787	0.744	0.765	0.502	0.631	0.559
Cohen et al.	0.578	0.606	0.591	0.781	0.750	0.765	0.492	0.627	0.552

Seg-LSTM comes close to state-of-the-art baselines, without using manual feature engineering

Sentence-CNN Model



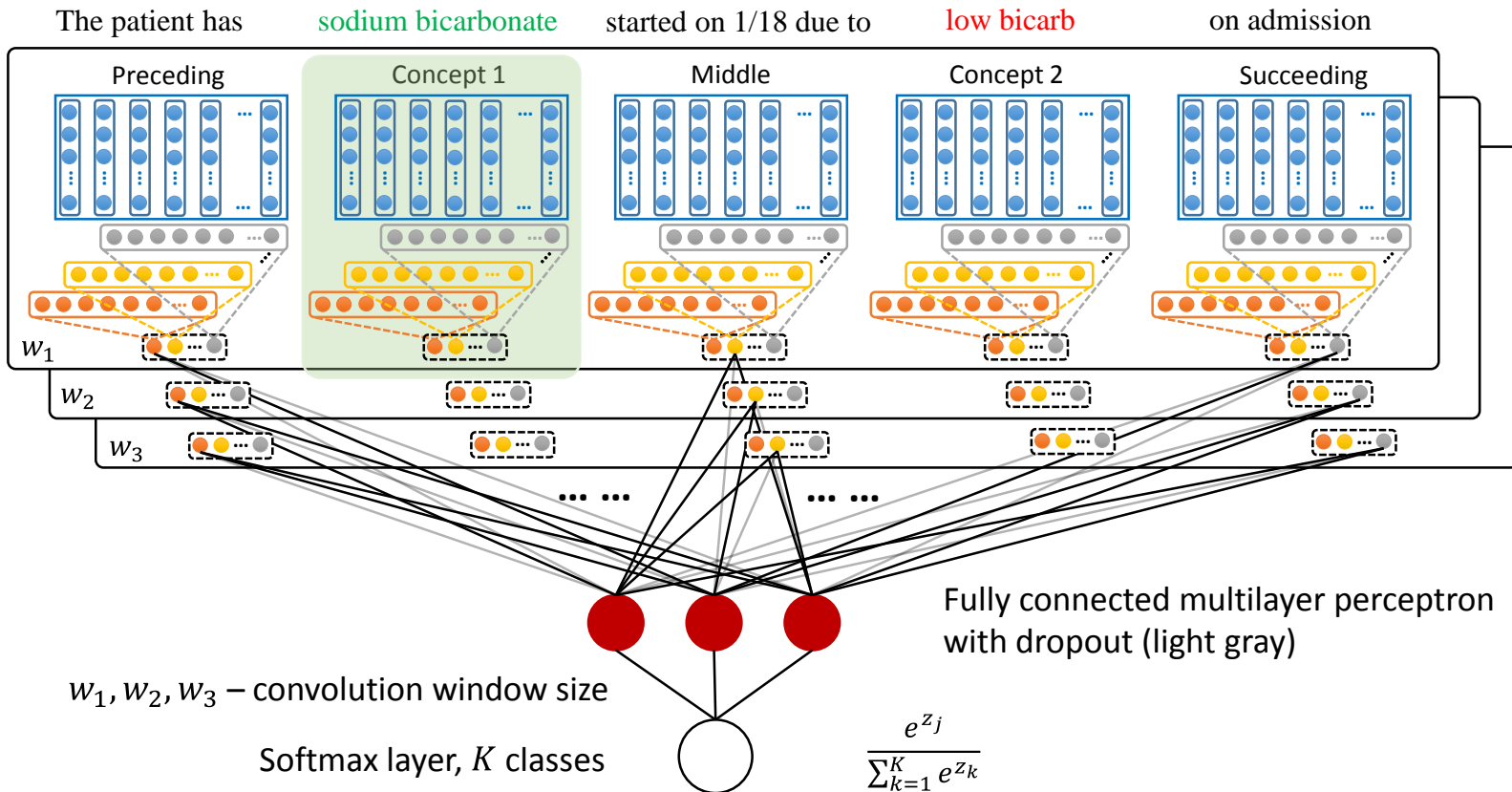
w – convolution window size
 n_{hu} – number of hidden units,
 also the number of different
 feature maps

Segment Convolutional Neural Networks

https://github.com/yuanluo/seg_cnn

Relation Label:

Treatment Administered for Medical Problem



Y Luo, Y Cheng, Ö Uzuner, P Szolovits, J Starren. Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. *JAMIA 2017 Aug 31*;25(1):93-8.

Seg-CNN Evaluation Results

System	Problem—Treatment Relations			Problem—Test Relations			Problem—Problem Relations		
	R	P	F	R	P	F	R	P	F
Seg-CNN (MIMIC)	0.685	0.687	0.686	0.804	0.836	0.820	0.704	0.700	0.702
Sentence CNN	0.642	0.641	0.641	0.760	0.812	0.785	0.679	0.693	0.686
Embedding max	0.636	0.645	0.641	0.770	0.816	0.791	0.741	0.554	0.634
Embedding mean	0.632	0.618	0.625	0.770	0.825	0.796	0.786	0.533	0.635
Seg-CNN (NYT)	0.641	0.690	0.665	0.790	0.835	0.812	0.708	0.681	0.694
Seg-CNN (NYT+MIMIC)	0.653	0.706	0.678	0.788	0.848	0.817	0.710	0.689	0.700
Roberts et al.	0.686	0.672	0.679	0.833	0.798	0.815	0.726	0.664	0.694
deBruijn et al.	0.583	0.750	0.656	0.789	0.843	0.815	0.712	0.691	0.701
Grouin et al.	0.646	0.647	0.647	0.801	0.792	0.797	0.645	0.670	0.657
Patrick et al.	0.599	0.671	0.633	0.774	0.813	0.793	0.627	0.677	0.651
Jonnalagadda et al.	0.679	0.581	0.626	0.828	0.765	0.795	0.730	0.586	0.650
Divita et al.									0.610
Solt et al.									0.565
Demner-Fushman et al.									0.591
Anick et al.	0.619	0.596	0.608	0.787	0.744	0.765	0.502	0.631	0.559
Cohen et al.	0.578	0.606	0.591	0.781	0.750	0.765	0.492	0.627	0.552

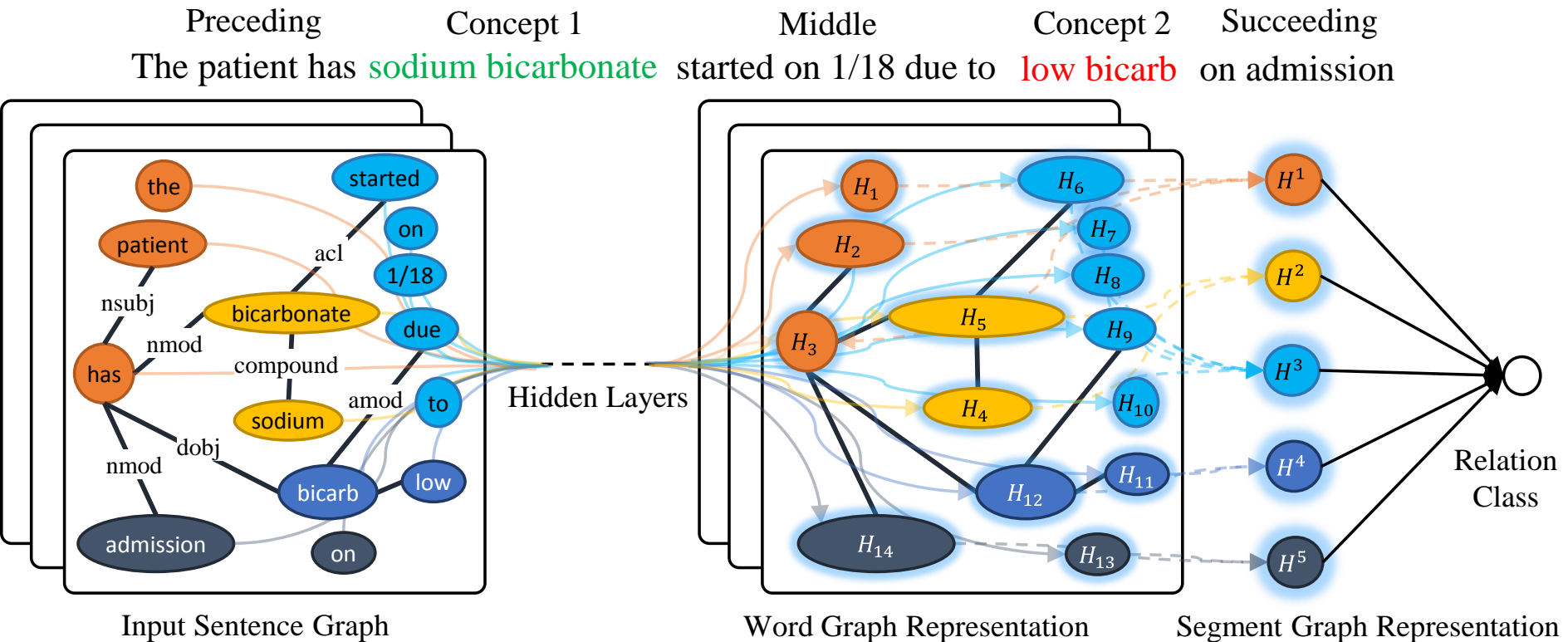
Seg-CNN outperform all state-of-the-art baselines, without using manual feature engineering

Segment-GCRN Model

Seg-GCRN: https://github.com/yuanluo/seg_gcn

Relation Label:

Treatment Administered for Medical Problem



Y Li, R Jin, **Y Luo***. Classifying relations in clinical narratives using Segment Graph Convolutional and Recurrent Neural Networks (Seg-GCRNs). *JAMIA 2018 Accepted*.

Segment-GCRN Model: One GCN Layer

- Denote $A \in \mathbb{R}^{n \times n}$ as the dependency graph G 's adjacency matrix
- Graph Laplacian is $L = D - A$, where $D_{ii} = \sum_j A_{ij}$
- Let $U \in \mathbb{R}^{n \times n}$ be the matrix of eigenvectors of the normalized graph Laplacian $L_n = I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}} = U\Lambda U^T$
- Let $g_w \in \mathbb{R}^{n \times n}$ be a **Fourier domain** filter matrix parametrized with a scalar w as its diagonal elements (**recall signal processing theory**)
- The graph convolution for the 1-dimensional embedding $x \in \mathbb{R}^n$ (for n words) is

$$h = U g_w U^T x = U \text{diag}([w, \dots, w]) U^T x = U U^T x \text{diag}([w, \dots, w])$$
- Simplify using Chebyshev polynomial approximation
- $h = (I + D^{-\frac{1}{2}}AD^{-\frac{1}{2}})xw$, after renormalization $h = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}xw$
- Extend the embedding and convolved signal to d -dimensional $X \in \mathbb{R}^{n \times d}$ and $H \in \mathbb{R}^{n \times d}$

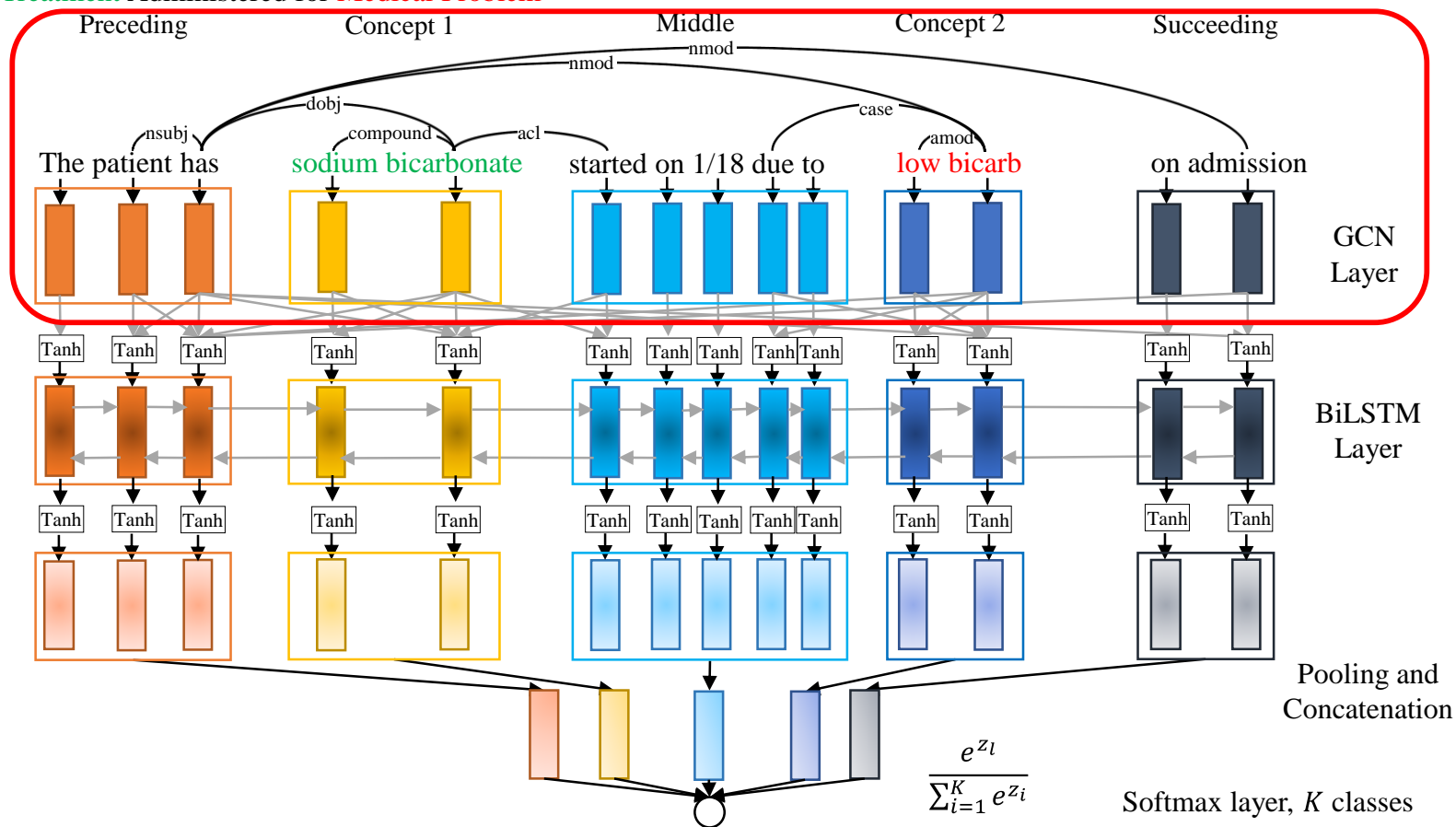
$$H = \text{Tanh}(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}XW)$$

Segment-GCRN Model

Segment Graph Convolutional and Recurrent Neural Networks (Seg-GCRNs)

Relation Label:

Treatment Administered for Medical Problem



Seg-GCRN Evaluation Results

System	Medical treatment– problem relations			Medical test– problem relations			Medical problem– problem relations		
	P	R	F	P	R	F	P	R	F
Seg-GCRN (GENIA+PubMed)	0.703	0.682	0.692	0.833	0.821	0.827	0.762	0.722	0.741
Seg-GCRN (WSJ)	0.684	0.683	0.683	0.842	0.802	0.821	0.787	0.702	0.742
Seg-GCN	0.673	0.684	0.679	0.818	0.795	0.807	0.641	0.717	0.677
Seg-CNN	0.687	0.685	0.686	0.836	0.804	0.820	0.700	0.704	0.702
Seg-CNN (NYT)	0.641	0.690	0.665	0.790	0.835	0.812	0.708	0.681	0.694
Seg-LSTM	0.641	0.683	0.661	0.766	0.838	0.800	0.728	0.681	0.704
Roberts et al.	0.672	0.686	0.679	0.798	0.833	0.815	0.664	0.726	0.694
deBruijn et al.	0.750	0.583	0.656	0.843	0.789	0.815	0.691	0.712	0.701
Grouin et al.	0.647	0.646	0.647	0.792	0.801	0.797	0.670	0.645	0.657
Patrick et al.	0.671	0.599	0.633	0.813	0.774	0.793	0.677	0.627	0.651
Jonnalagadda et al.	0.581	0.679	0.626	0.765	0.828	0.795	0.586	0.730	0.650
Divita et al.	0.704	0.582	0.637	0.794	0.782	0.788	0.710	0.534	0.610
Solt et al.									0.65
Demner-Fushman et al.									0.91
Anick et al.									0.69
Cohen et al.	0.606	0.578	0.591	0.750	0.781	0.765	0.627	0.492	0.552

Seg-GCRN outperform all state-of-the-art baselines by a margin, , without using manual feature engineering

Top Similar Words by Cosine Distance of Word Embedding

insulin

Word	Cosine Dist
hyperglycemia-on	0.688355
sliding	0.673756
insuling	0.673033
humalog	0.639635
nph	0.636425

cancer

Word	Cosine Dist
cancer-	0.727164
melanoma	0.703527
neoplasm	0.678477
metastatic	0.663101
malignant	0.643923

cephalexin

Word	Cosine Dist
keflex	0.754806
monohydrate	0.69552
clarithromycin	0.682608
augmentin	0.626386
amoxicillin	0.624552

infection

Word	Cosine Dist
infxn	0.72388
infections	0.713478
infection-	0.704144
infeciton	0.680275
sources	0.632514

Domain Really Matters!



nph

All

News

Images

Videos

Books

More

Settings

Tools

About 11,100,000 results (0.68 seconds)

Normal pressure hydrocephalus - Wikipedia

https://en.wikipedia.org/wiki/Normal_pressure_hydrocephalus

Normal pressure hydrocephalus (NPH), also termed Hakim's syndrome and symptomatic hydrocephalus, is a type of brain malfunction caused by expansion of ...

People also ask

What is NPH diagnosis?

What is NPH in medical terms?

What are the symptoms of normal pressure hydrocephalus?

What are the symptoms of fluid on the brain?

Feedback

Neil Patrick Harris (@ActuallyNPH) · Twitter

<https://twitter.com/ActuallyNPH>

The word "stat" must only be used for important matters. Like in @Cigna's #ad to get you to your annual check-up, STAT! #GoKnowTakeControl pic.twitter.com/j9UTJAV...

2 days ago · Twitter

My #CarpoolKaraoke with @tylerperry is live now. Check it out! apple.co/_ck pic.twitter.com/RZi0dwr...

4 days ago · Twitter

I just watched the #lastjedi trailer and now I just, you know, really want to be in a Star Wars movie. I feel like a kid again. So stoked.

5 days ago · Twitter

What Is Normal Pressure Hydrocephalus? - WebMD

<https://www.webmd.com/brain/normal-pressure-hydrocephalus>

Sep 11, 2016 - **NPH** is different than other types of hydrocephalus in that it develops slowly over time.

Neil Patrick Harris

American actor


twitter.com/actuallynph

Neil Patrick Harris is an American actor, comedian, magician, and singer, known primarily for his comedy roles on television and his dramatic and musical stage roles. [Wikipedia](#)

Born: June 15, 1973 (age 44), Albuquerque, NM

Height: 6' 0"

Spouse: David Burtka (m. 2014)

Children: Harper Grace Burtka-Harris, Gideon Scott Burtka-Harris

Siblings: Brian Harris

Movies and TV shows

View 45+ more



How I Met Your Mother
2005 – 2014



A Series of Unfortunate Events
Since 2017



Doogie Howser, M.D.
1989 – 1993



Best Time Ever with Neil Patri...
2015

Using GCN for Long Text Understanding

- Challenges of Long Text Understanding
 - CNN and RNN prioritize locality and sequentiality.
 - They can model local consecutive word sequences well
 - They may ignore global word co-occurrence in a corpus
- GCN can
 - Generalizing well-established neural network models like CNN that apply to regular grid structure (2-d mesh or 1-d sequence) to work on arbitrarily structured graphs
 - Can preserve global structure information of a graph in graph embeddings (node, edge, subgraph and whole graph embeddings)

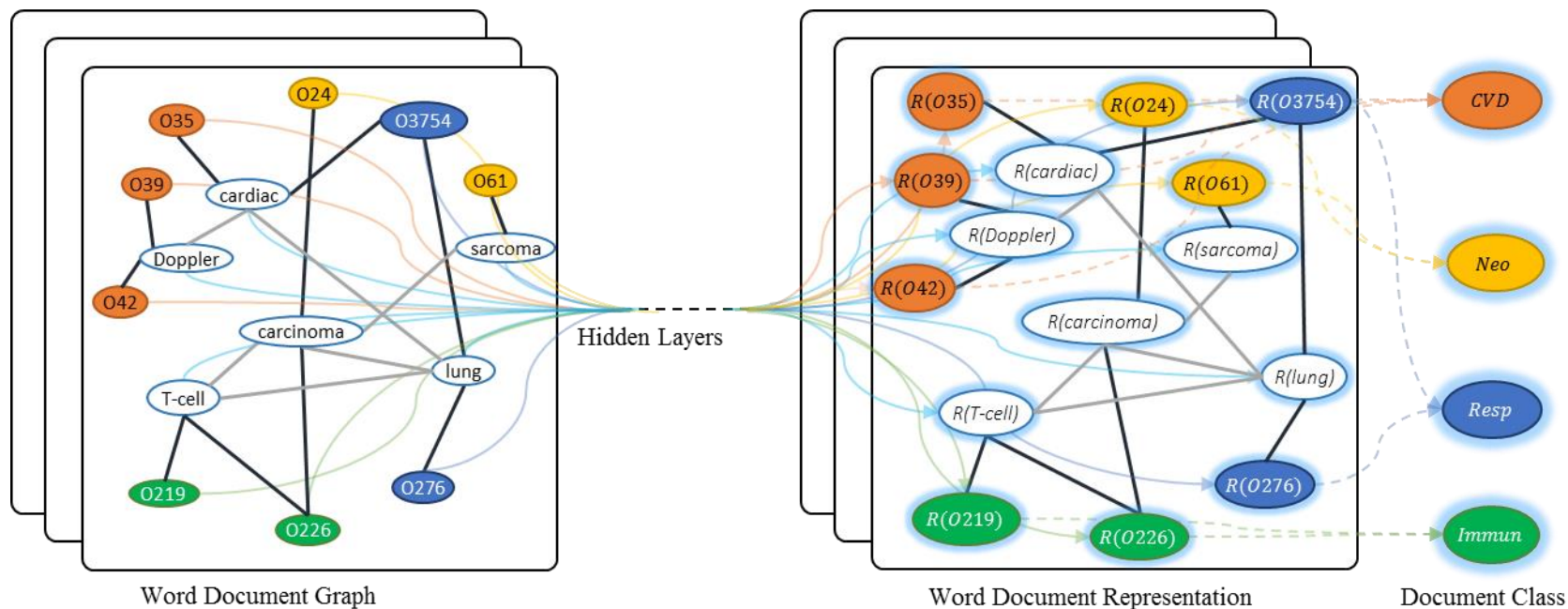
Graph Convolutional Networks (GCN)

- A graph $G = (V, E)$:
 - $(v, v) \in E$ for any v
 - $X \in R^{n \times m}$: node features matrix
 - A : adjacency matrix, degree matrix $D_{ii} = \sum_j A_{jj}$
 - $\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$: normalized symmetric adjacency matrix
 - W_j : weight matrix, trained via SGD
- One layer GCN:
- $L^{(1)} = \rho(\tilde{A} X W_0)$
- Stacking multiple GCN layers:
- $L^{(j+1)} = \rho(\tilde{A} L^{(j)} W_j)$

Graph Convolutional Networks (GCN)

- GCN can capture information only about immediate neighbors with one layer
- When multiple GCN layers are stacked, one can incorporate higher order neighborhoods information
 - e.g., a two-layers GCN can allow message passing among nodes that are at maximum two steps away.
- A special form of Laplacian smoothing:
 - computes the new features of a node as the weighted average of itself and its neighbors (second order neighbors for a two-layer GCN).

Text Graph Convolutional Networks (Text GCN)



L Yao, C Mao, **Y Luo***. Graph Convolutional Networks for Text Classification. *Proceedings of AAAI Conference on Artificial Intelligence 2019 Full paper*.

Text Graph Convolutional Networks (Text GCN)

- Document content and global word co-occurrence
 - Document-word edges: TF-IDF
 - Word-word edges: point-wise mutual information (PMI)

$$\text{PMI}(i, j) = \log \frac{p(i, j)}{p(i)p(j)}$$

$$A_{ij} = \begin{cases} \text{PMI}(i, j) & i, j \text{ are words, } \text{PMI}(i, j) > 0 \\ \text{TF-IDF}_{ij} & i \text{ is document, } j \text{ is word} \\ 1 & i = j \\ 0 & \text{otherwise} \end{cases}$$

Text Graph Convolutional Networks (Text GCN)

- A simple two-layer GCN:
 - one-hot feature matrix for words and documents: $X = I$
 - 1st layer document and word embeddings: $\tilde{A}XW_0$
 - 2nd layer document and word embeddings: $\tilde{A}\text{ReLU}(\tilde{A}XW_0)W_1$
 - \mathcal{Y}_D is the set of document indices that have labels and F is the dimension of the output features, which is equal to the number of classes, Y is the label indicator matrix
 - Loss function

$$Z = \text{softmax}(\tilde{A} \text{ReLU}(\tilde{A}XW_0)W_1)$$

$$\mathcal{L} = - \sum_{d \in \mathcal{Y}_D} \sum_{f=1}^F Y_{df} \ln Z_{df}$$

Datasets

Dataset	# Docs	# Training	# Test	# Words	# Nodes	# Classes	Average Length
20NG	18,846	11,314	7,532	42,757	61,603	20	221.26
R8	7,674	5,485	2,189	7,688	15,362	8	65.72
R52	9,100	6,532	2,568	8,892	17,992	52	69.82
Ohsumed	7,400	3,357	4,043	14,157	21,557	23	135.82
MR	10,662	7,108	3,554	18,764	29,426	2	20.39

Implementation Details

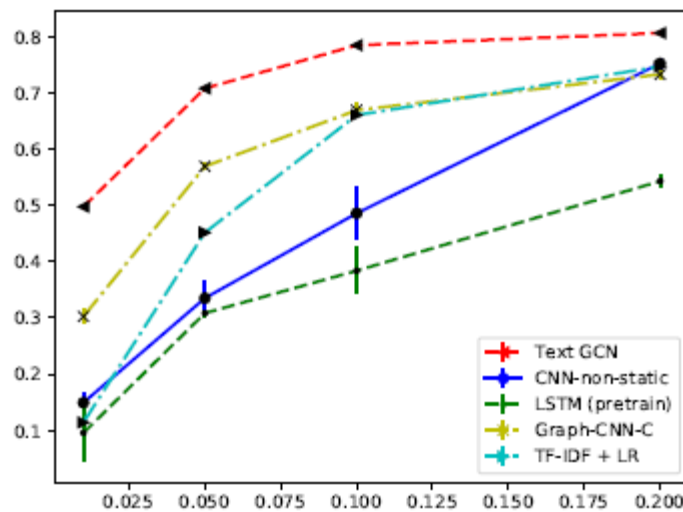
- Parameters:
 - first embedding size: 200
 - window size: 20
 - dropout rate: 0.5
 - learning rate: 0.02
 - validation set: 10% of training set
 - number of epochs: 200
- We also tried other parameters but do not find much difference
- Adam algorithm as the optimizer

Results

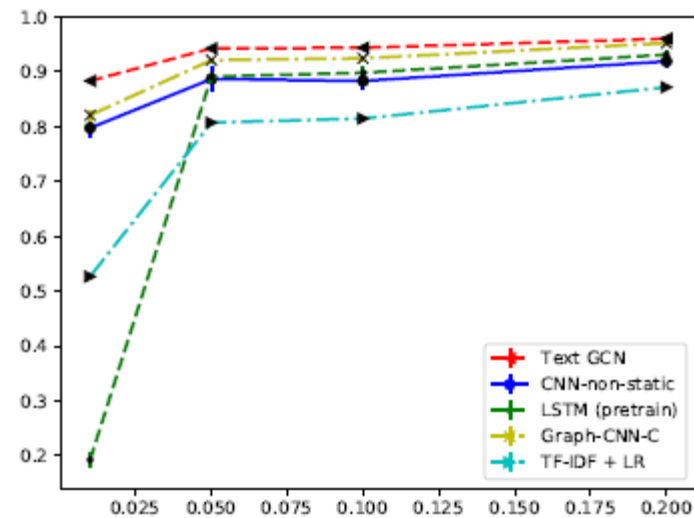
Table 2: Test Accuracy on document classification task. We run all models 10 times and report mean \pm standard deviation. Text GCN significantly outperforms baselines on 20NG, R8, R52 and Ohsumed based on student t -test ($p < 0.05$).

Model	20NG	R8	R52	Ohsumed	MR
TF-IDF + LR	0.8319 \pm 0.0000	0.9374 \pm 0.0000	0.8695 \pm 0.0000	0.5466 \pm 0.0000	0.7459 \pm 0.0000
CNN-rand	0.7693 \pm 0.0061	0.9402 \pm 0.0057	0.8537 \pm 0.0047	0.4387 \pm 0.0100	0.7498 \pm 0.0070
CNN-non-static	0.8215 \pm 0.0052	0.9571 \pm 0.0052	0.8759 \pm 0.0048	0.5844 \pm 0.0106	0.7775 \pm 0.0072
LSTM	0.6571 \pm 0.0152	0.9368 \pm 0.0082	0.8554 \pm 0.0113	0.4113 \pm 0.0117	0.7506 \pm 0.0044
LSTM (pretrain)	0.7543 \pm 0.0172	0.9609 \pm 0.0019	0.9048 \pm 0.0086	0.5110 \pm 0.0150	0.7733 \pm 0.0089
Bi-LSTM	0.7318 \pm 0.0185	0.9631 \pm 0.0033	0.9054 \pm 0.0091	0.4927 \pm 0.0107	0.7768 \pm 0.0086
PV-DBOW	0.7436 \pm 0.0018	0.8587 \pm 0.0010	0.7829 \pm 0.0011	0.4665 \pm 0.0019	0.6109 \pm 0.0010
PV-DM	0.5114 \pm 0.0022	0.5207 \pm 0.0004	0.4492 \pm 0.0005	0.2950 \pm 0.0007	0.5947 \pm 0.0038
PTE	0.7674 \pm 0.0029	0.9669 \pm 0.0013	0.9071 \pm 0.0014	0.5358 \pm 0.0029	0.7023 \pm 0.0036
fastText	0.7938 \pm 0.0030	0.9613 \pm 0.0021	0.9281 \pm 0.0009	0.5770 \pm 0.0049	0.7514 \pm 0.0020
fastText (bigrams)	0.7967 \pm 0.0029	0.9474 \pm 0.0011	0.9099 \pm 0.0005	0.5569 \pm 0.0039	0.7624 \pm 0.0012
SWEM	0.8516 \pm 0.0029	0.9532 \pm 0.0026	0.9294 \pm 0.0024	0.6312 \pm 0.0055	0.7665 \pm 0.0063
LEAM	0.8191 \pm 0.0024	0.9331 \pm 0.0024	0.9184 \pm 0.0023	0.5858 \pm 0.0079	0.7695 \pm 0.0045
Graph-CNN-C	0.8142 \pm 0.0032	0.9699 \pm 0.0012	0.9275 \pm 0.0022	0.6386 \pm 0.0053	0.7722 \pm 0.0027
Graph-CNN-S	–	0.9680 \pm 0.0020	0.9274 \pm 0.0024	0.6282 \pm 0.0037	0.7699 \pm 0.0014
Graph-CNN-F	–	0.9689 \pm 0.0006	0.9320 \pm 0.0004	0.6304 \pm 0.0077	0.7674 \pm 0.0021
Text GCN	0.8634 \pm 0.0009	0.9707 \pm 0.0010	0.9356 \pm 0.0018	0.6836 \pm 0.0056	0.7674 \pm 0.0020

Results



(a) 20NG




(b) R8

Figure 4: Test accuracy by varying training data proportions.

Contributions to Short and Long Text Understanding

- A Trilogy of segment LSTM, CNN, GCRN for classifying three different types of relations from clinical notes
- Comparable Seg-LSTM, Seg-CNN performance to state-of-the-art, with minimal feature engineering
- Superior Seg-GCRN performance to state-of-the-art, with minimal feature engineering
- Seg-GCRN integrates both syntactic dependency information (GCN) and word sequence information (RNN)
- Developing heterogeneous document-word GCN for long text understanding
- Domain really matters: in word embeddings, in dependency parsing

Thank you

- Collaboration welcome, my email: yuan.luo@northwestern.edu
-  @yuanhypnosluo
- IEEE ICHI Missing data challenge (data, 3D-MICE+Evaluation code available)
 - <http://www.ieee-ichi.org/challenge.html>
 - Select papers publish in a JBI supplement issue
 - Remote/recorded presentation possible

The 7th IEEE International Conference on Healthcare Informatics (ICHI 2019)
Beijing, China, June 10-13, 2019

