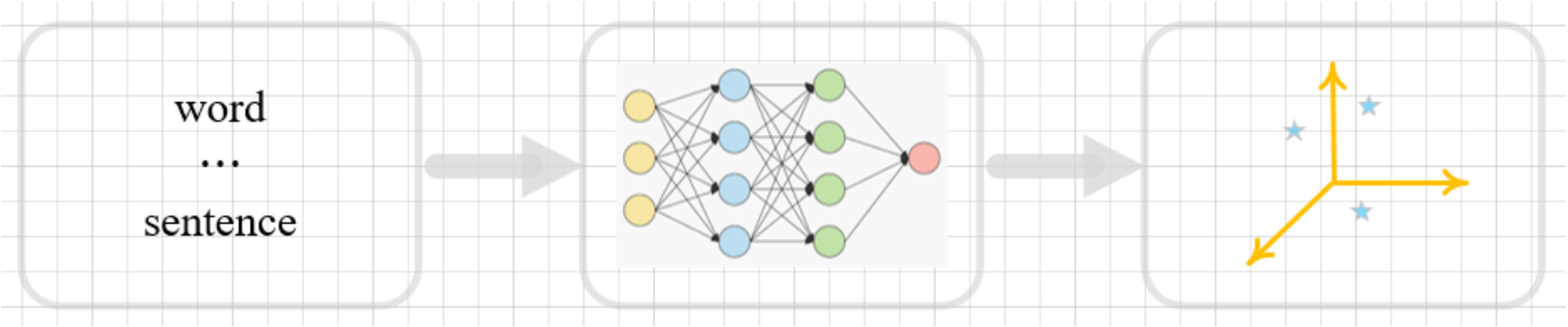
Neural Representation Learning in Natural Language Processing

Pengfei Liu
pfliu.com

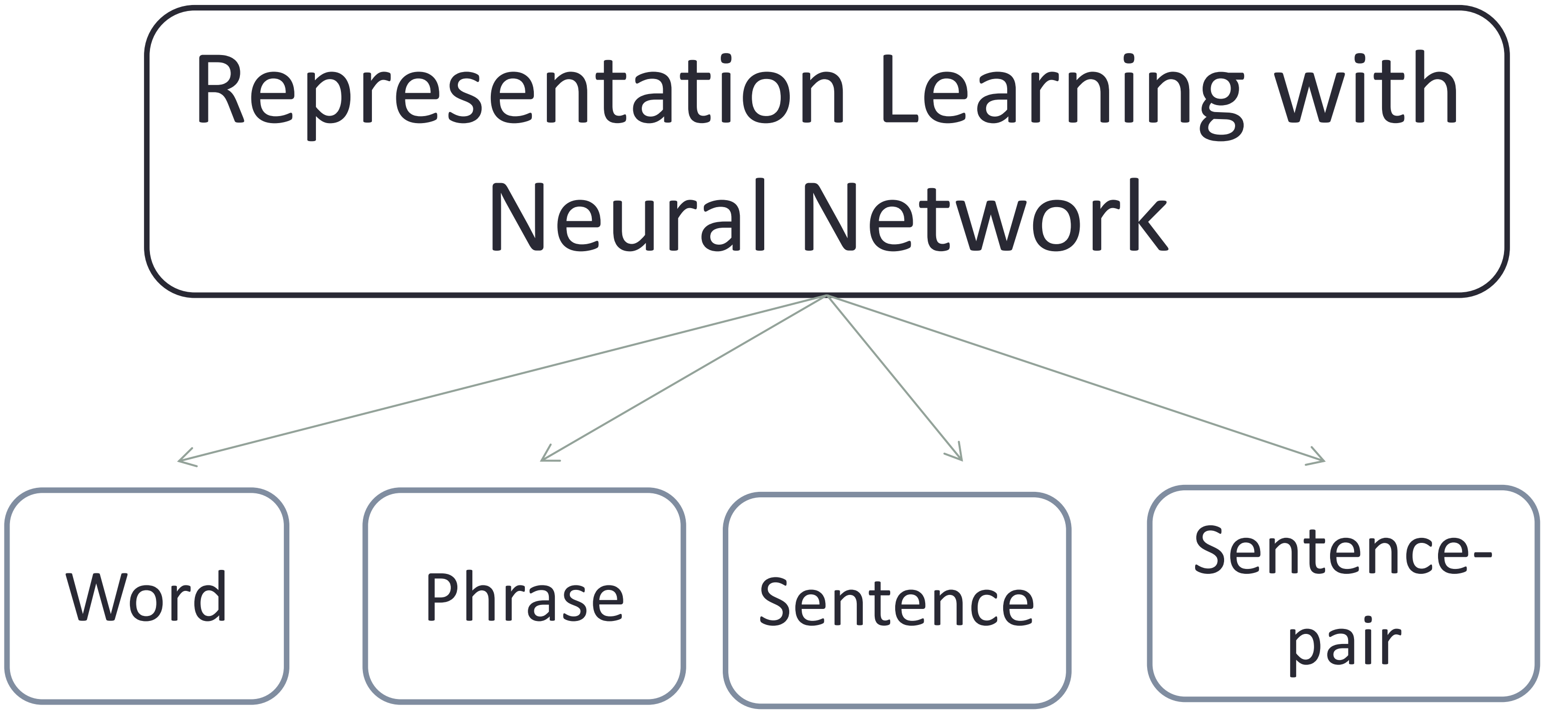
Neural Representation Learning for NLP



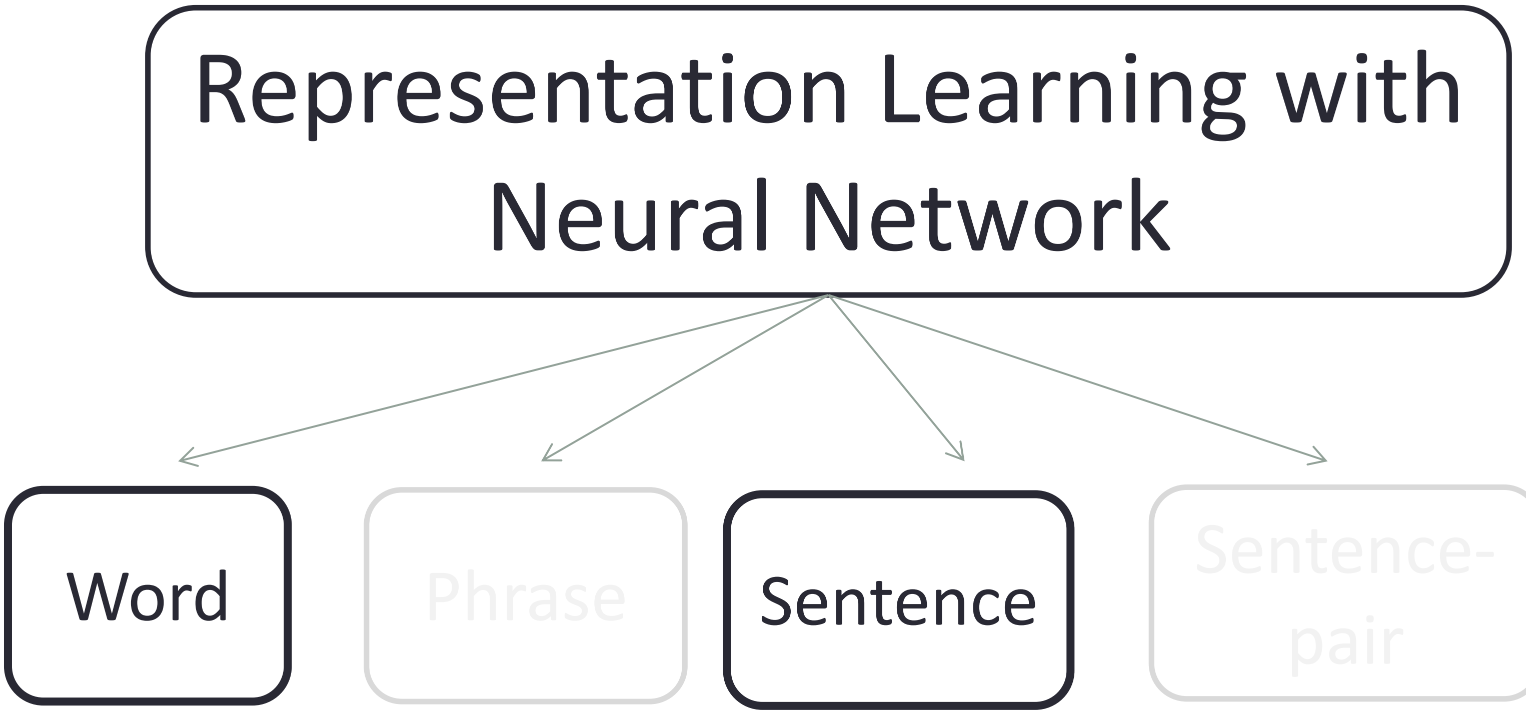
Neural networks

text → vector

Neural Representation Learning for NLP



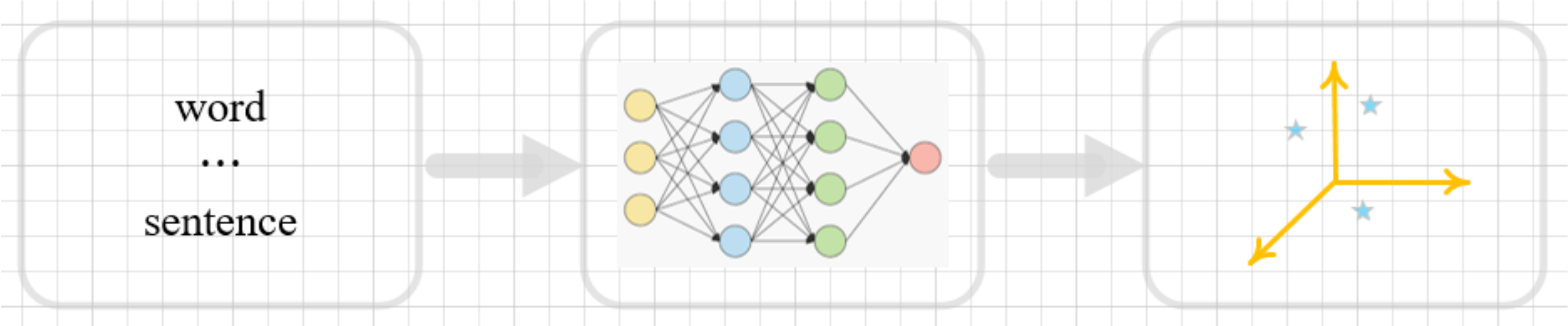
Neural Representation Learning for NLP



Part-I: Word Representation

What is the "word representation"?

a vector



apple ⇨ [0.1,0.3.0,4]

Why should we learn "word representation"?

- Easy to make mathematical calculation
  - What if you want to know the meaning of "red apple"

red apple ?= red + apple

$$\begin{bmatrix} 0.3 \\ 0.9 \\ 0.9 \end{bmatrix} = f\left\{ \begin{bmatrix} 0.1 \\ 0.2 \\ 0.1 \end{bmatrix}, \begin{bmatrix} 0.2 \\ 0.7 \\ 0.8 \end{bmatrix} \right\}$$

Why should we learn "word representation"?

- Easy to make mathematical calculation

Vectorizing discrete signals makes things easier.

How can we get word representations?

One-hot

LSA

PLSA

MT-emb

CoVe

NNLM

C&W

SentiEMB

word2vec

GloVe

ELMo

BERT

BART

A lot of approaches!

Let’s try to cluster them!

Symbolic or Distributed?

- Symbolic
  - One-hot vector
- Distributed
  - Real-valued vector

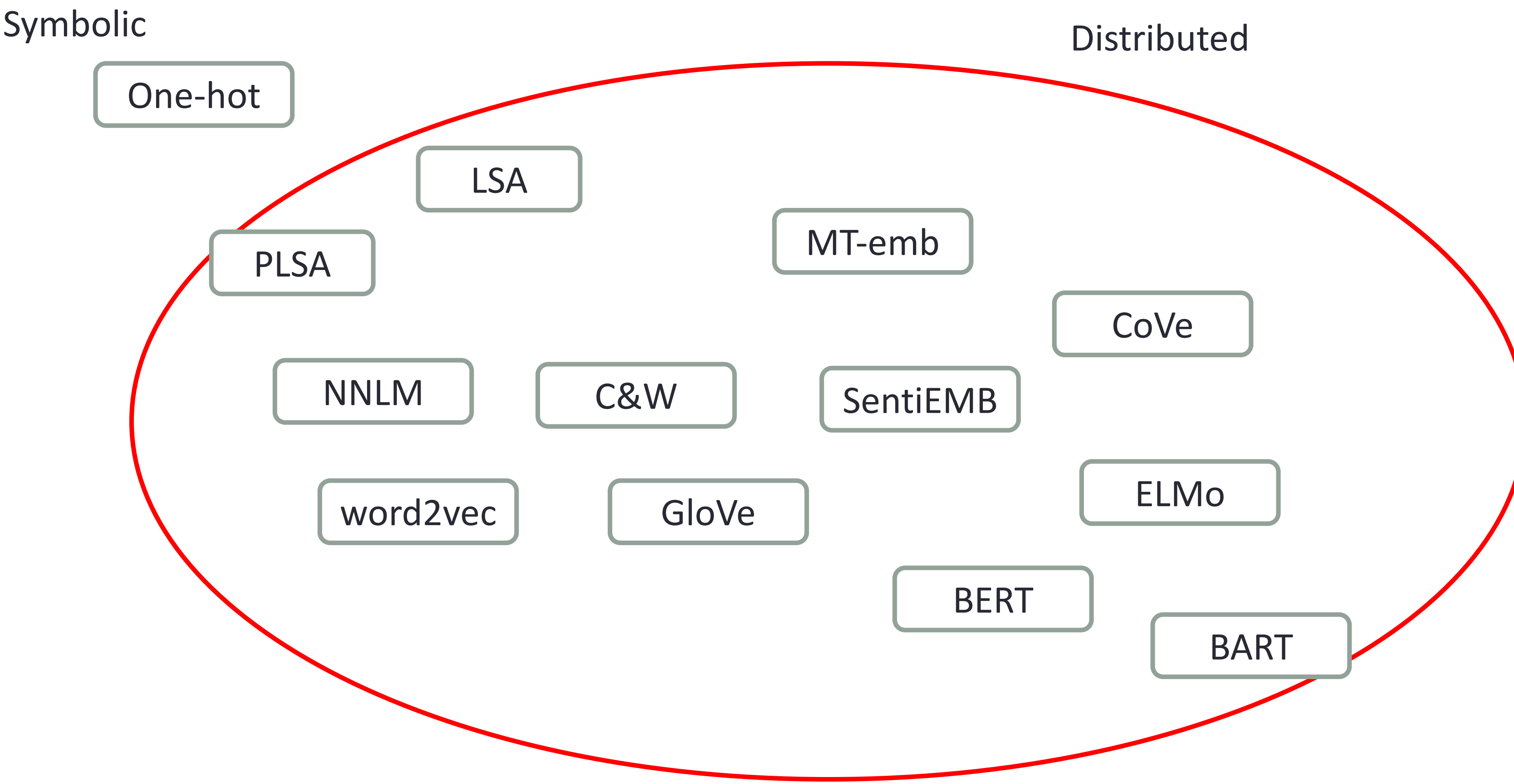
0
1
0
0
0

explainable

0.1
0.3
0.5
0.1
0.9

explainable

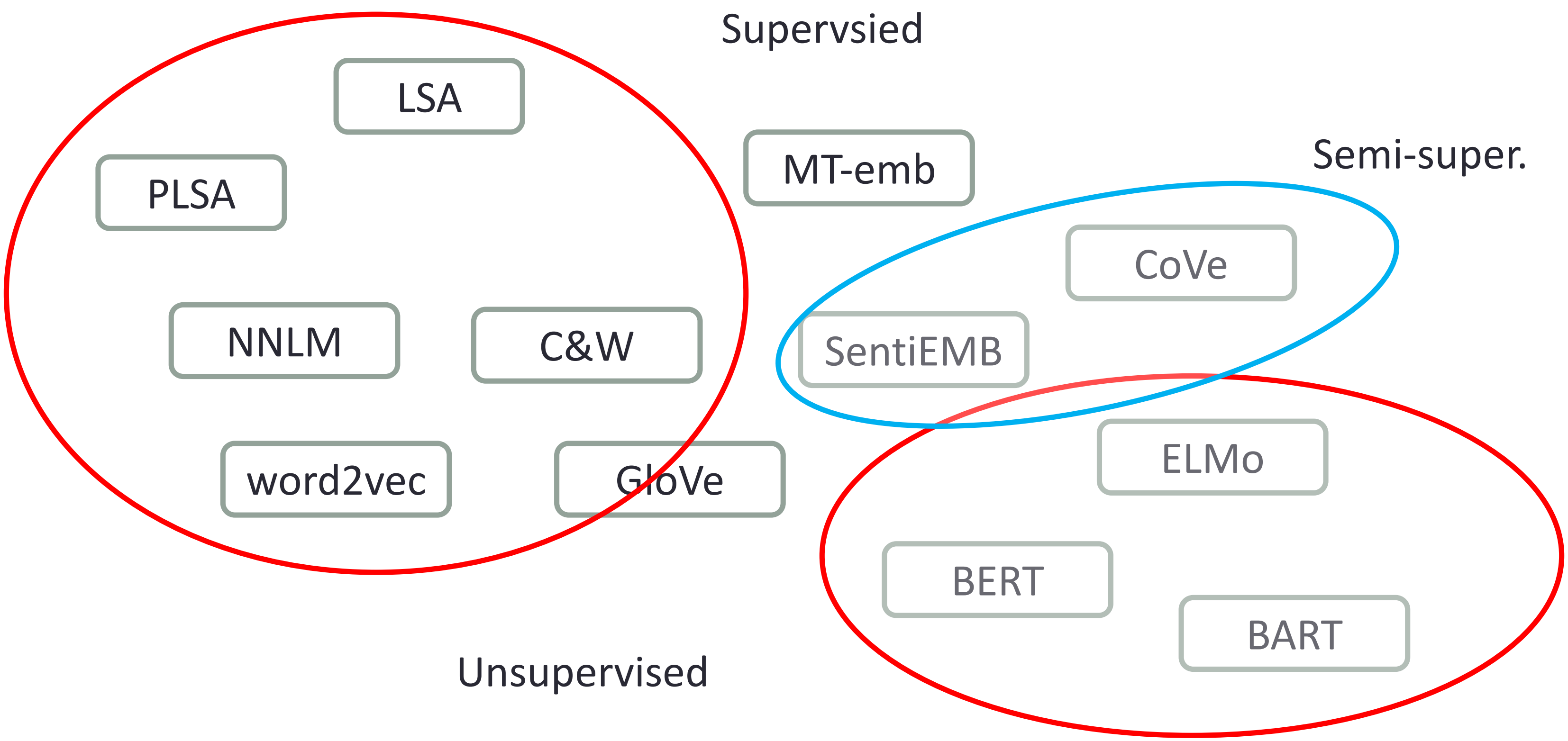
Clusters of Approaches



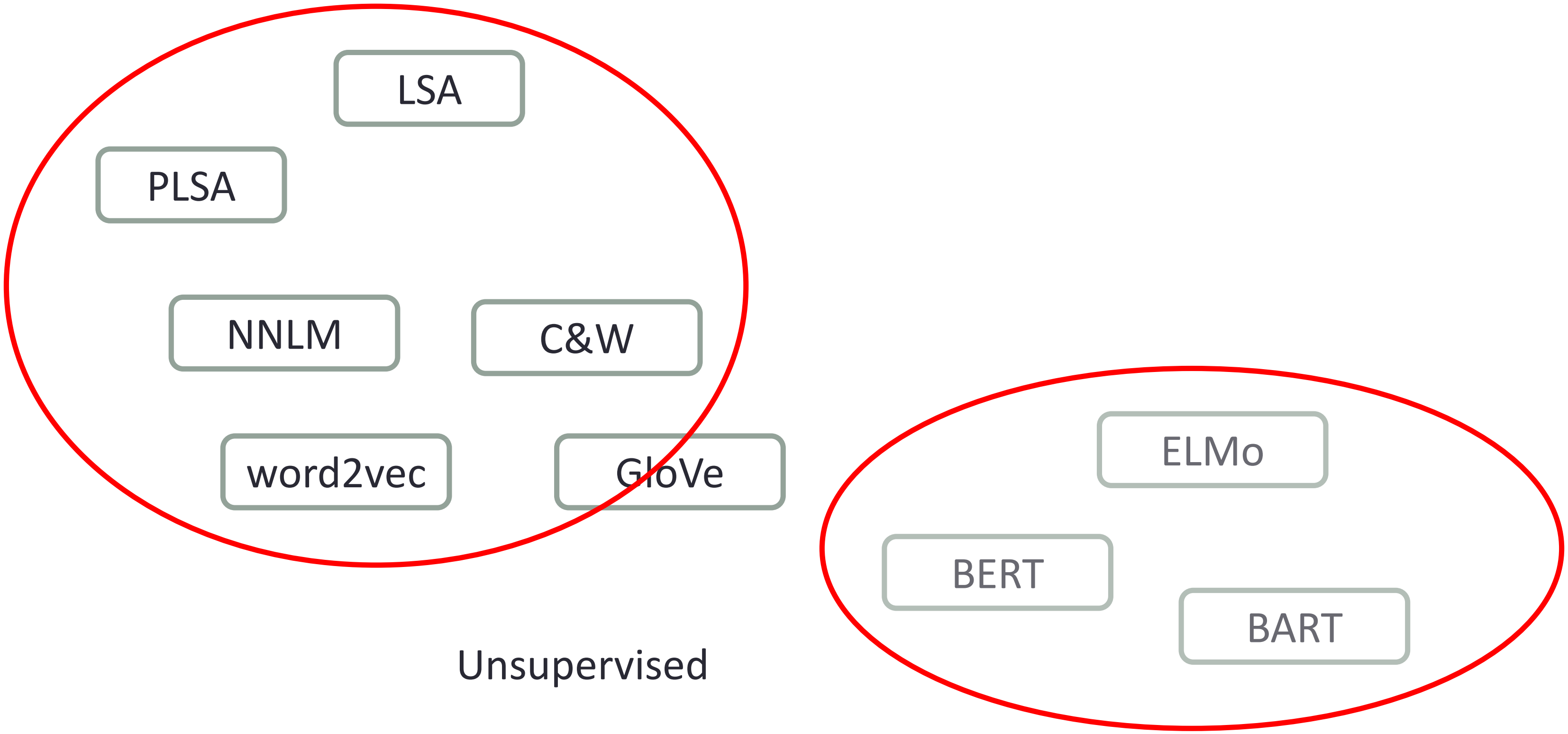
Supervised or Unsupervised?

- Supervised
  - labeled data
- Unsupervised
  - unlabeled data
- Semi-supervised
  - Pre-trained + fine-tuned

Clusters of Approaches



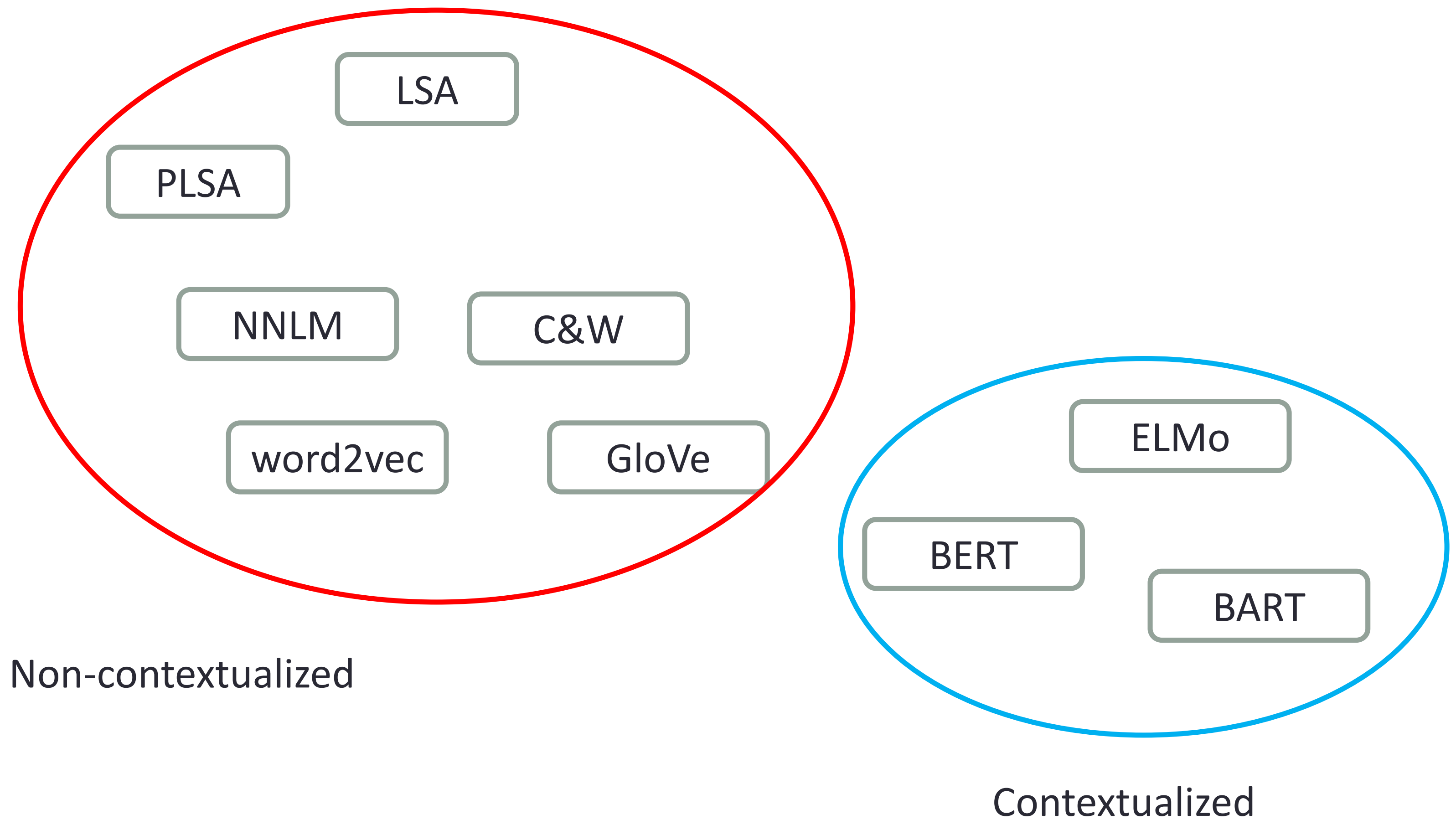
Clusters of Approaches



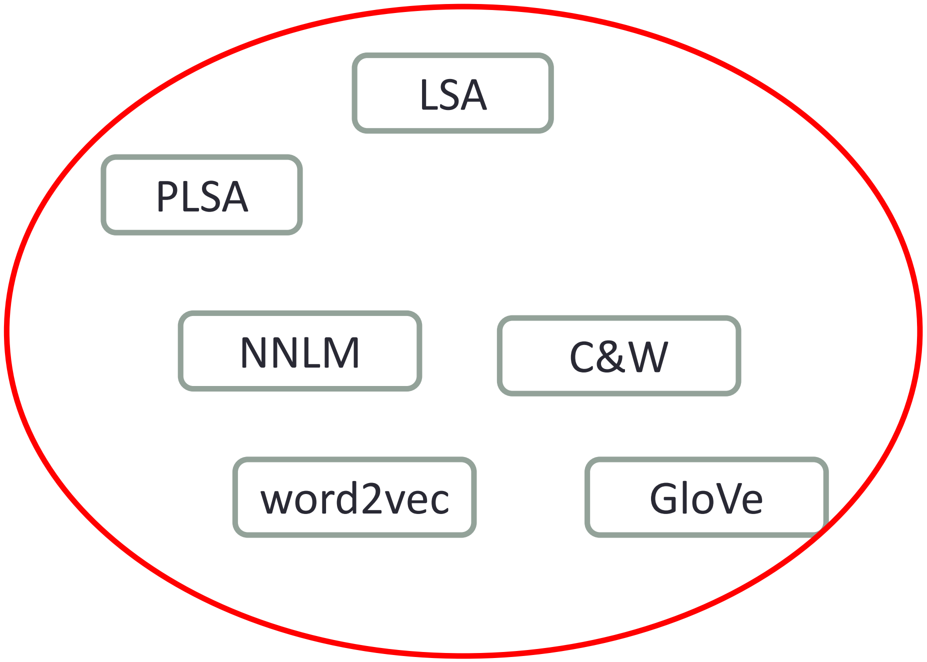
Contextualized or not?

- Non-contextualized
  - Word vector is context-independent
- Contextualized
  - Word vector is context-dependent

Clusters of Approaches



Clusters of Approaches

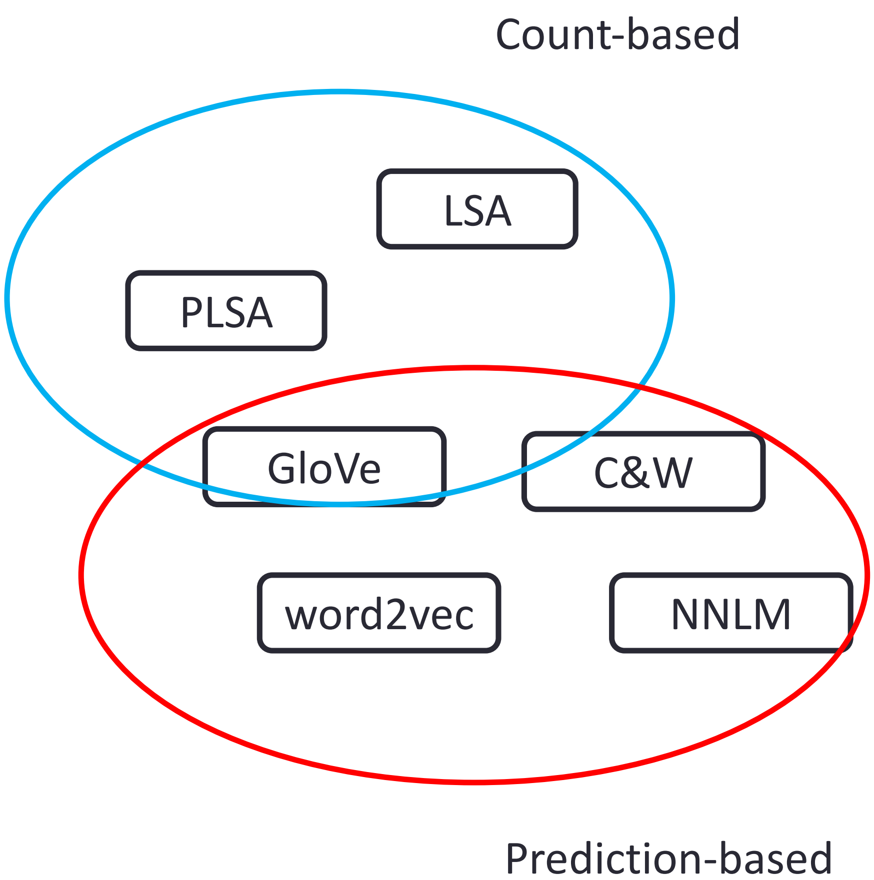


Non-contextualized

Count-based or Prediction-based?

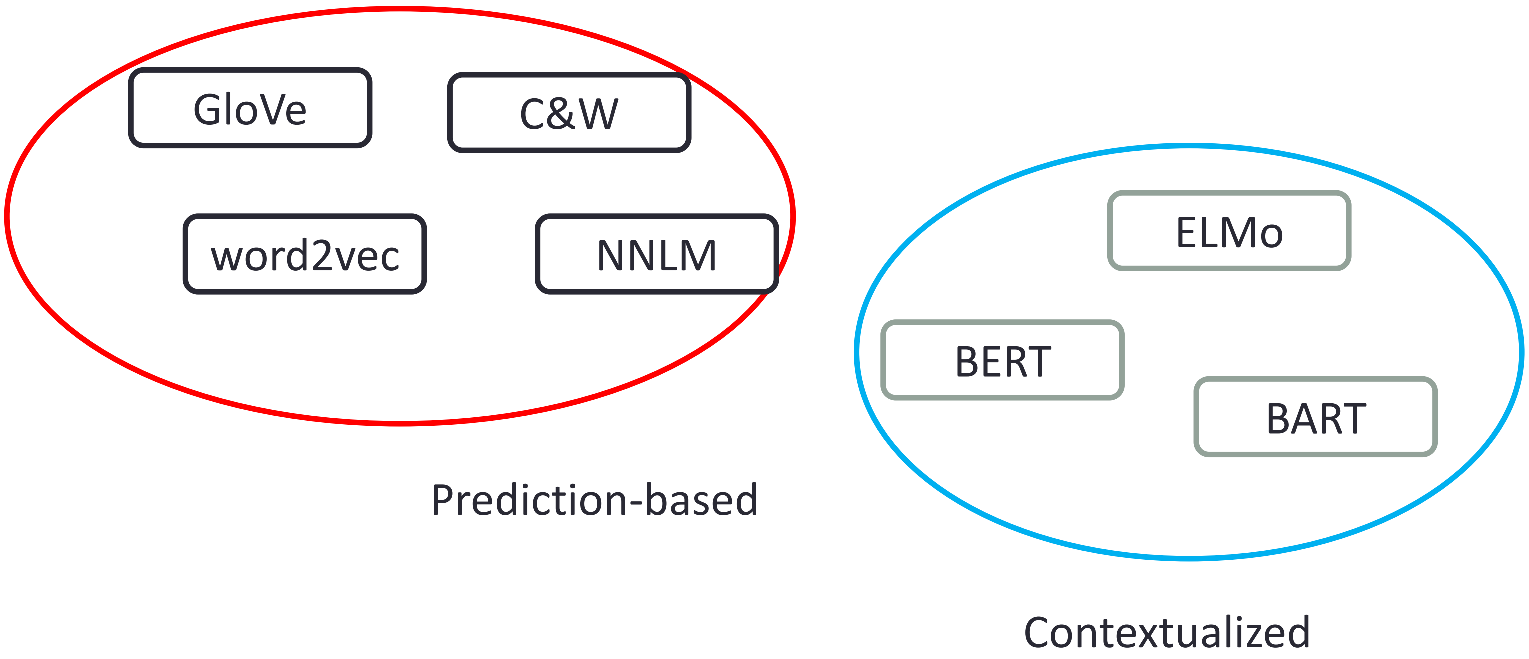
- Count-based
  - **Count** the number of co-occurrences of word/context, with rows as word, columns as contexts
  - Maybe **weight** with pointwise mutual information
  - Maybe **reduce dimensions** using SVD
- Prediction-based
  - try to **predict** the words within a neural network

Clusters of Approaches



Strong connection between count-based methods and prediction-based methods (Levy and Goldberg 2014)

Clusters of Approaches

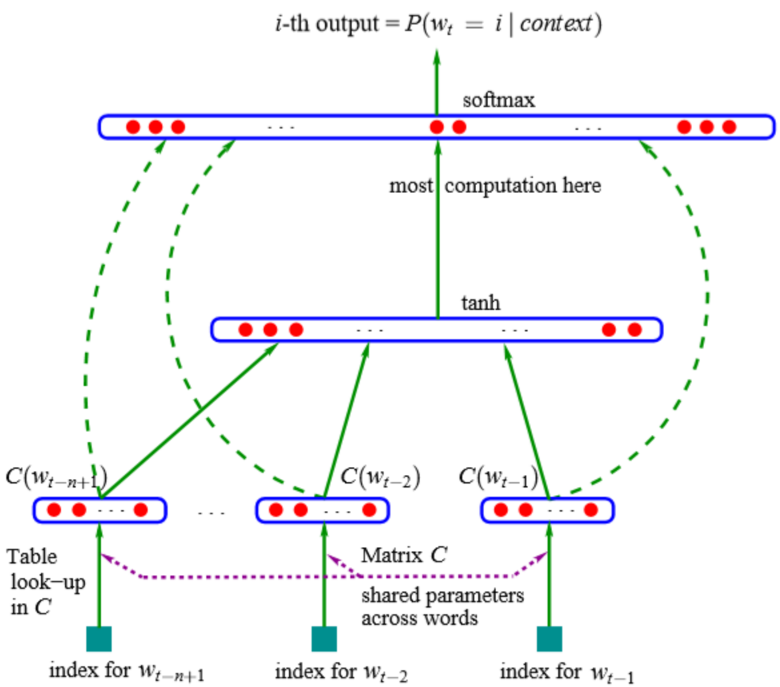


Clusters of Approaches

Year	Conf.	Concept	Cited	Paper	
2014	nips	none	19365	Distributed Representations of Words and Phrases and their Compositionality Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean	word2vec
2013	arxiv	none	15383	Efficient Estimation of Word Representations in Vector Space Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean	word2vec
2014	emnlp	none	13069	Glove: Global Vectors for Word Representation Jeffrey Pennington, Richard Socher, Christopher Manning	GloVe
2003	jmlr	none	6074	A Neural probabilistic language model Yoshua Bengio, Rejean Ducharme, Pascal Vincent	NNLM
2019	naacl	none	5292	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova	BERT
2018	naacl	none	2913	Deep Contextualized Word Representations Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton	ELMo
2013	naacl	none	2578	Linguistic Regularities in Continuous Space Word Representations Tomas Mikolov, Wen-tau Yih, Geoffrey Zweig	
2012	acl	none	1079	Improving Word Representations via Global Context and Multiple Word Prototypes Eric Huang, Richard Socher, Christopher Manning, Andrew Ng	
2014	arxiv	none	971	word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method Yoav Goldberg and Omer Levy	
2015	**tacl**	**none**	903	**Improving Distributional Similarity with Lessons Learned from Word Embeddings** **Omer Levy, Yoav Goldberg, Ido Dagan**	

Case Study: NNLM

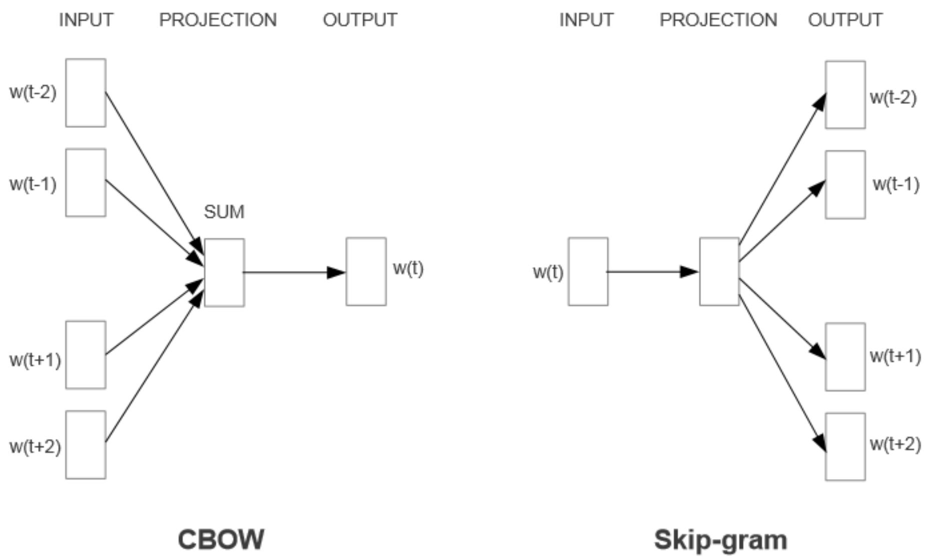
2003	jmlr	none	6074	A Neural probabilistic language model Yoshua Bengio, Rejean Ducharme, Pascal Vincent



- One of the earliest work on neural word representation
- How to neutralize language model task
- Word embedding is a by-product
- Slow!!!

Case Study: word2vec

2014	nips	none	19365	Distributed Representations of Words and Phrases and their Compositionality Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean
2013	arxiv	none	15383	Efficient Estimation of Word Representations in Vector Space Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean



- Two specific models
- Word embedding is our focus
- Efficient to train

Case Study: GloVe

2014	emnlp	none	13069	Glove: Global Vectors for Word Representation Jeffrey Pennington, Richard Socher, Christopher Manning

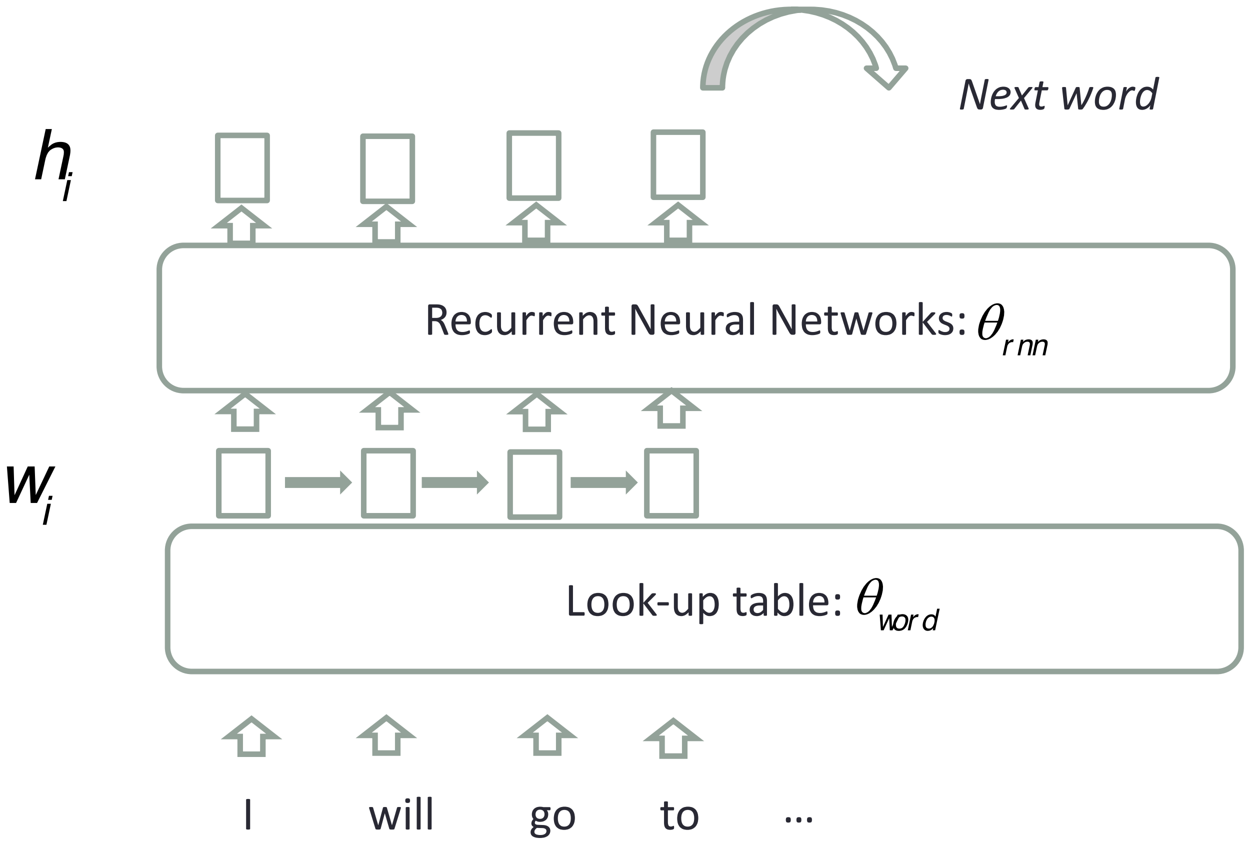
$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

$$f(x) = \begin{cases} (x/\overset{100}{x_{\max}})^{\overset{3/4}{\alpha}} & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases}$$

- The dot product of two word embeddings <-> co-ocurrence
- https://nlp.stanford.edu/projects/glove/

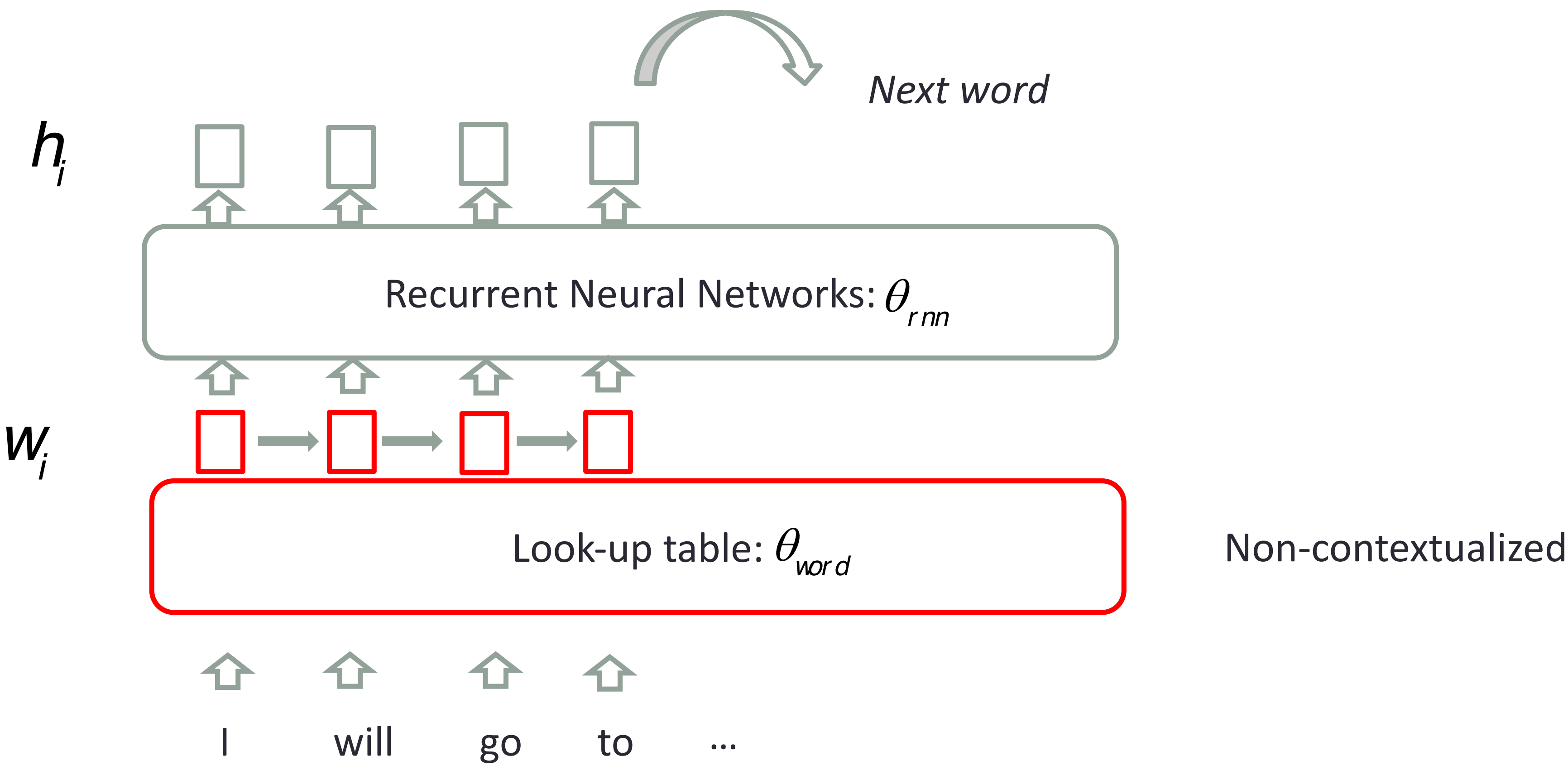
Case Study: ELMo

2018	naacl	none	2913	Deep Contextualized Word Representations Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer



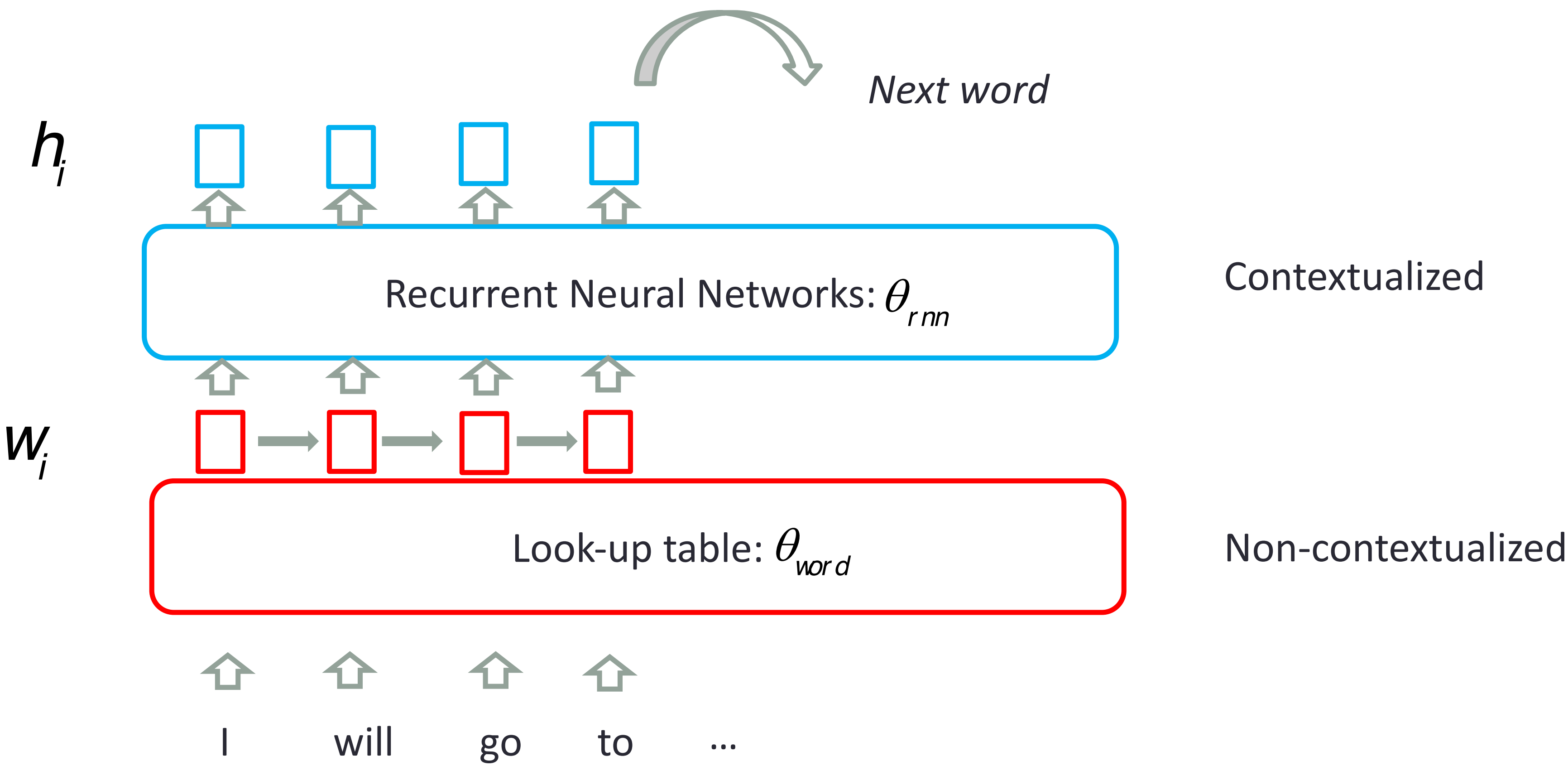
Case Study: ELMo

2018	naacl	none	2913	Deep Contextualized Word Representations Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer



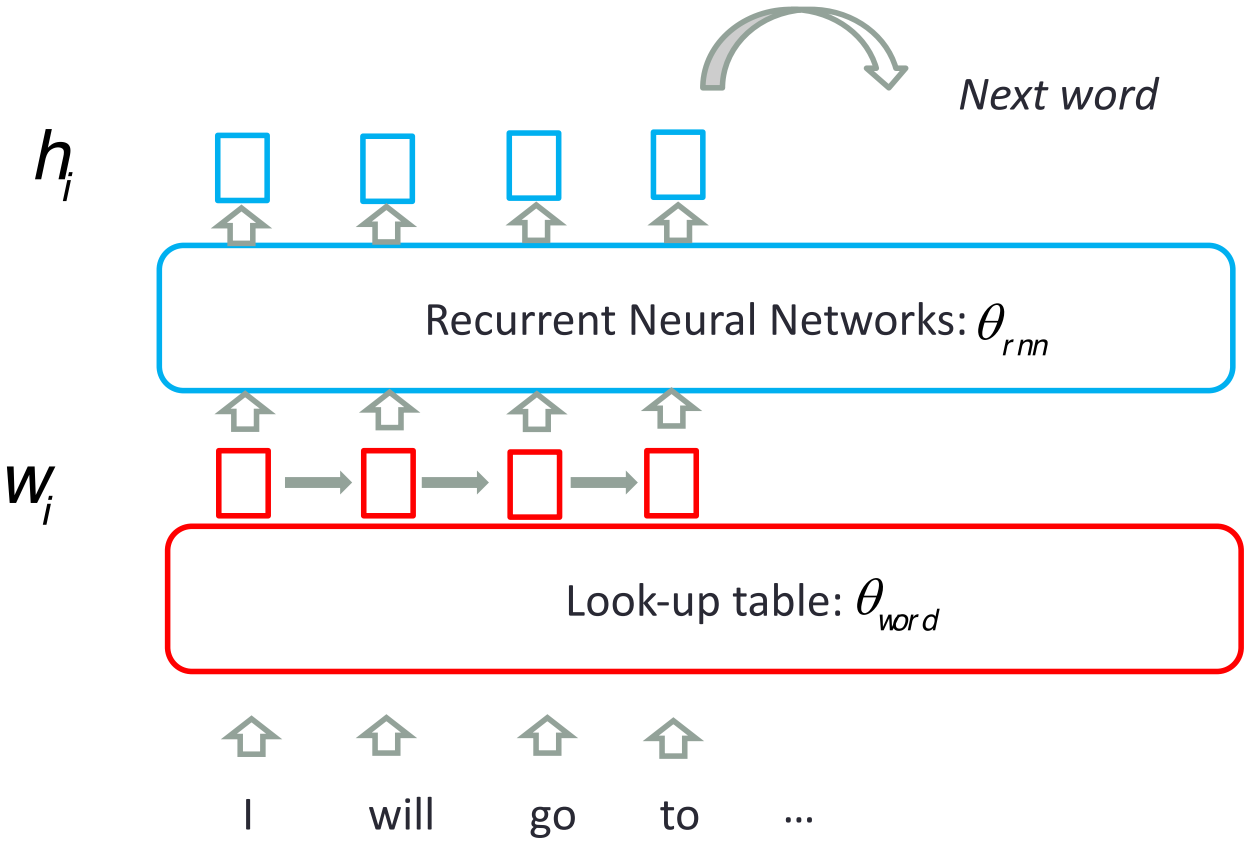
Case Study: ELMo

2018	naacl	none	2913	Deep Contextualized Word Representations Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer



Case Study: ELMo

2018	naacl	none	2913	Deep Contextualized Word Representations Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer



Contextualized

Challenges

1) More parameters
2) Slow

Advantages

1) Disambiguity

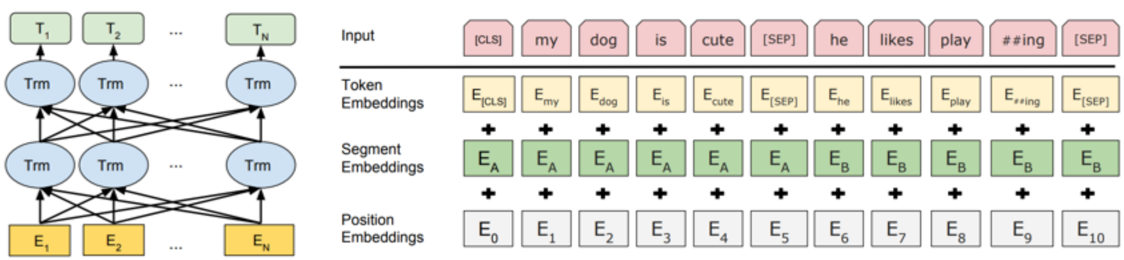
Case Study: BERT

2019	naacl	none	5292	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

Masked Word Prediction (BERT)

(Devlin et al. 2018)

- **Model:** Multi-layer self-attention. Input sentence or pair, w/ [CLS] token, subword representation



- **Objective:** Masked word prediction + next-sentence prediction
- **Data:** BooksCorpus + English Wikipedia

Credit to Graham's slide

Software?   Training corpus?

Software, Model, Corpus

- **GloVe**
  - Code & Off-the-shelf model: https://github.com/stanfordnlp/GloVe
  - Training corpus:
    - Wikipedia-2014: https://en.wikipedia.org/wiki/
    - Gigaword 5: https://catalog.ldc.upenn.edu/LDC2011T07
- **Word2vec**
  - Code: https://www.tensorflow.org/tutorials/text/word_embeddings
  - Off-the-shelf model: https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit
  - Training corpus:
    - Google News dataset
- **BERT**
  - Code & Off-the-shelf model: https://github.com/google-research/bert
  - Training corpus:
    - Wikipedia: https://en.wikipedia.org/wiki/
    - BookCorpus: https://yknzhu.wixsite.com/mbweb

Which one should I choose?

No pretraining when …

language modeling,
machine translation

Using non-contextualized when …

(e.g, word2vec, glove)

- ✓ Limited computation resources
- ✓ Fast training/quickly evaluate your models
- ✓ No off-the-shelf BERT models
- ✓ Hugh Domain shift
- ✓ Best of both worlds

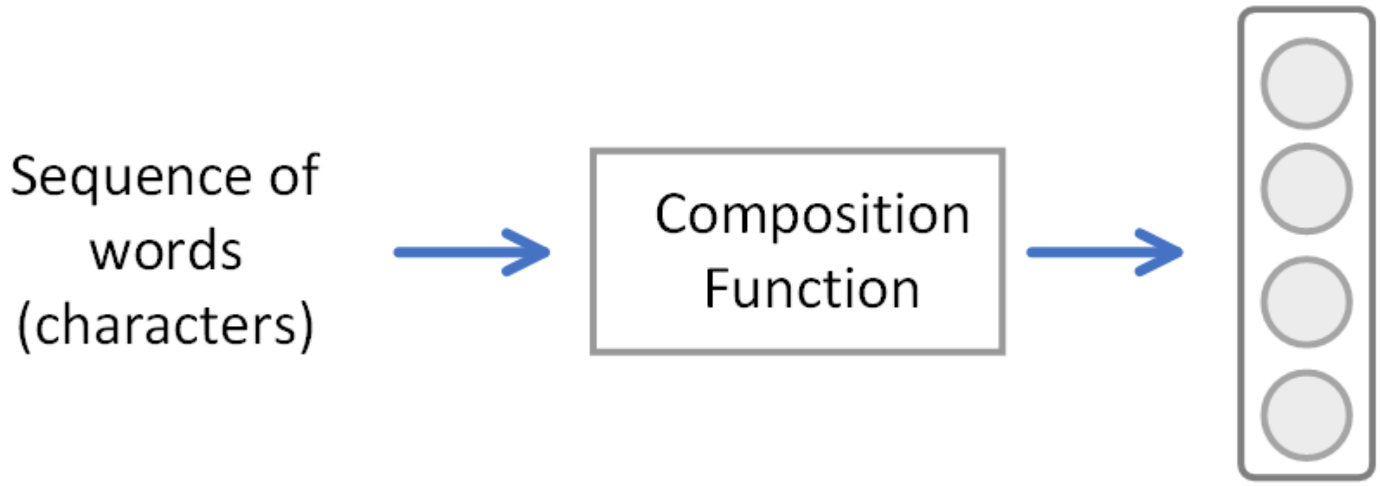
Using contextualized when …

(e.g, BERT, BART)

- ✓ Rich in GPUs
- ✓ Care the SOTA result
- ✓ Don’t care the training time
- ✓ Off-the-shelf BERT models
- ✓ Few training samples/Low-resource

Part-II: Sentence Representation

What is the "sentence representation"?



Why do we need "sentence representation"?

- It is a fundamental step!
  - Sentiment classification
  - Semantic matching
  - Text Summarization
  - Machine Translation

How can we learn sentence representations?

NBOW

paraVector

CNN

vRNN

Skip-Thought

RecNN

LSTM

Transformer

GRU

GNN

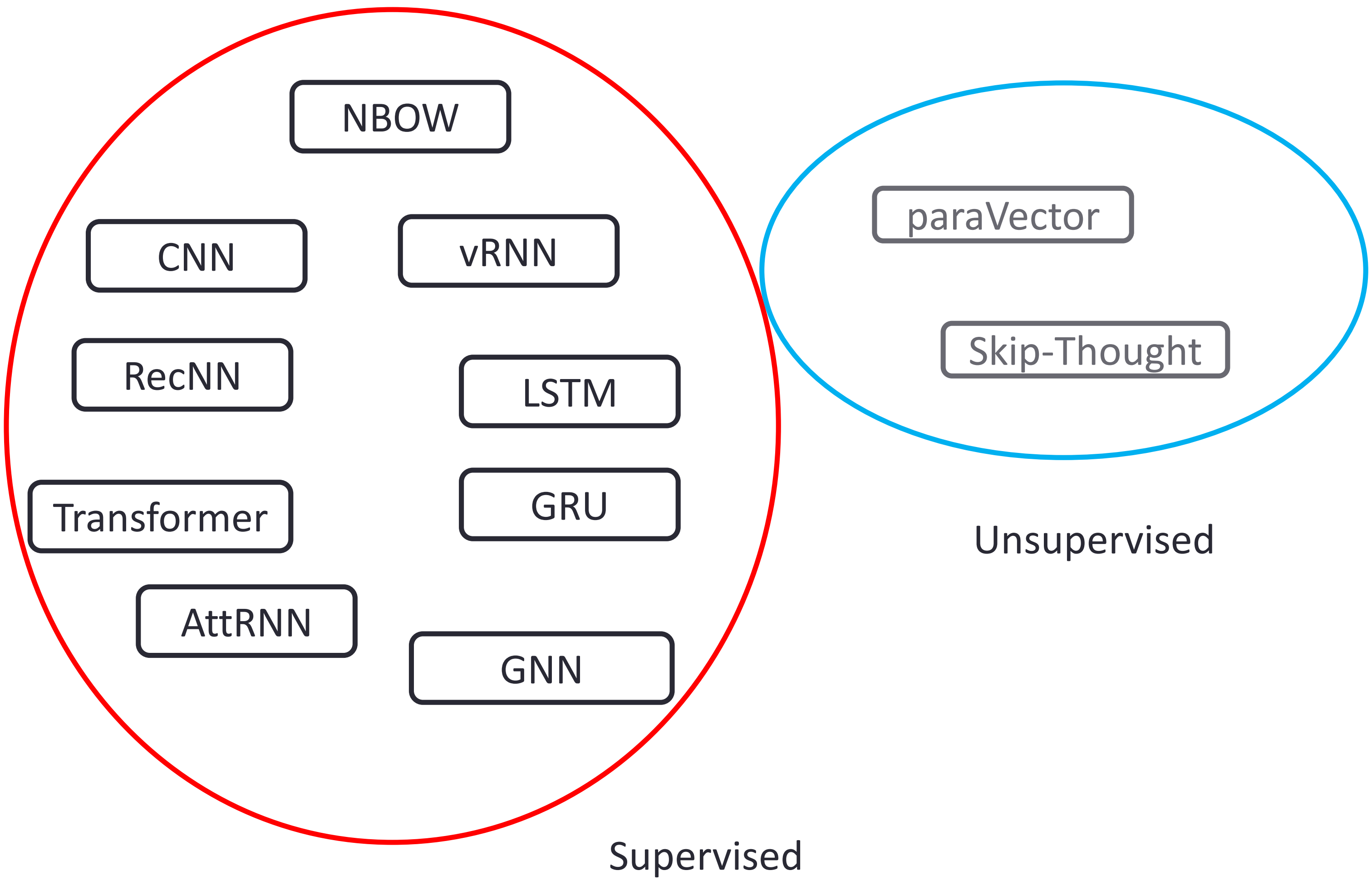
A lot of approaches!

Again, let’s try to cluster them!

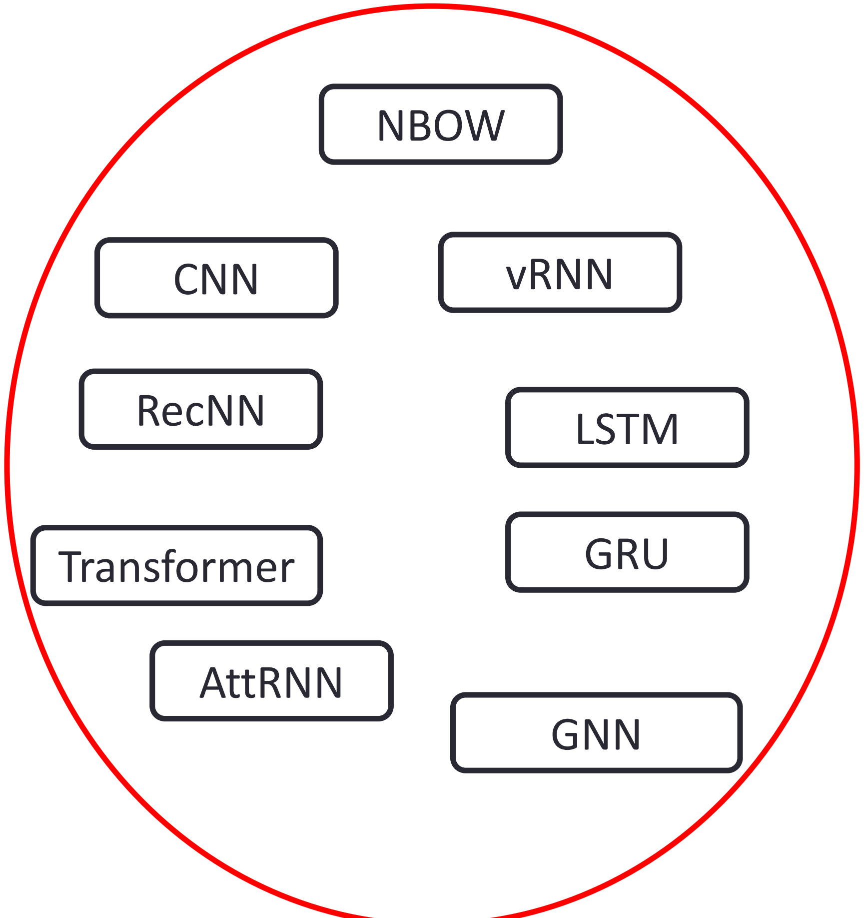
Supervised or Unsupervised?

- Supervised
  - labeled data
- Unsupervised
  - unlabeled data

Clusters of Approaches



Clusters of Approaches



Supervised

Different Structural Biases

- **Structural Bias** : a set of prior knowledge incorporated into your model design
  - Connection ways
  - Topological structures

Two perspectives

Different Structural Biases

Connection ways

Interact with nearest neighbors

Local $w_1 \quad w_2 \quad w_3 \quad \boxed{w_4} \quad w_5 \quad w_6 \quad w_7$

Non-local $w_1 \quad w_2 \quad w_3 \quad \boxed{w_4} \quad w_5 \quad w_6 \quad w_7$

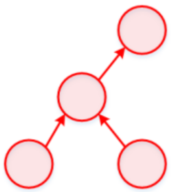
Interact with distant neighbors

Different Structural Biases

Topological structure

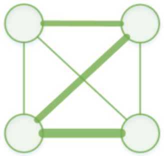
Sequential

Tree



(syntactic parsing tree)

Graph



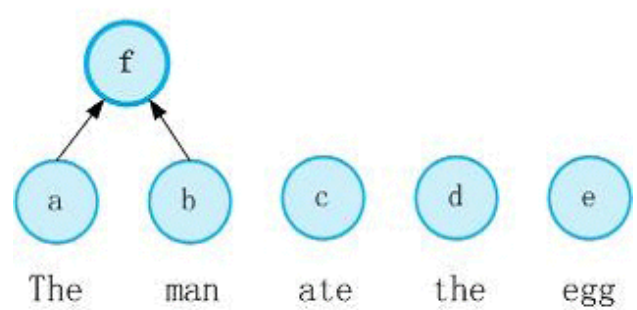
(entity relation)

Along what structure are sentences modeled

Clusters of Approaches

Connection ways | Local | Non-local

Topological Structure | Seq. | Tree | Graph



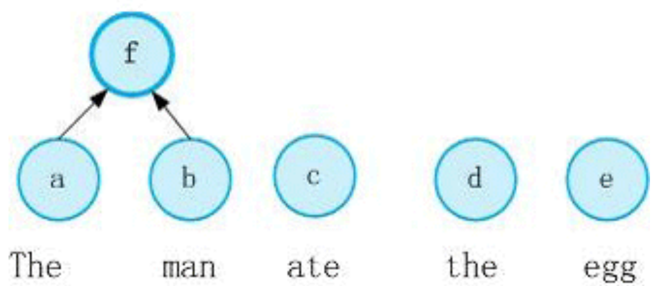
RNN | LSTM | GRU

CNN

Clusters of Approaches

Connection ways | Local | Non-local

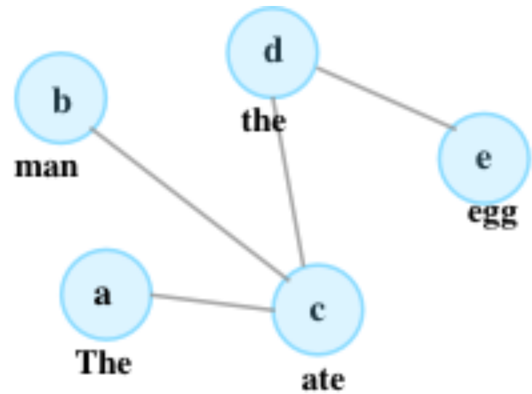
Topological Structure | Seq. | Tree | Graph



Recursive NN | TreeLSTM | TreeCNN

Clusters of Approaches

Connection ways	Local		Non-local
Topological Structure	Seq.	Tree	Graph

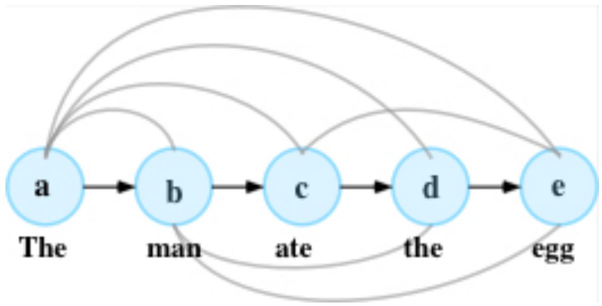


Graph Neural Nets

Clusters of Approaches

Connection ways | Local | Non-local

Topological Structure | Seq. | Tree | Graph

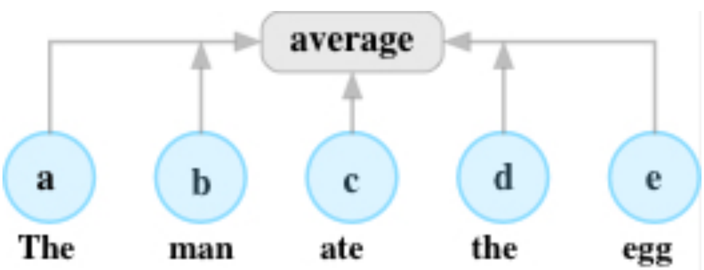


Attention LSTM

Clusters of Approaches

Connection ways | Local | Non-local

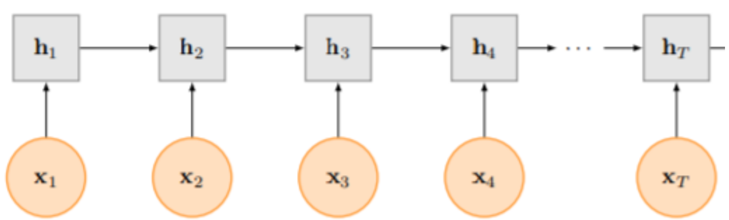
Topological Structure | Seq. | Tree | Graph



NBOW

Case Study: Must-know Points about RNN

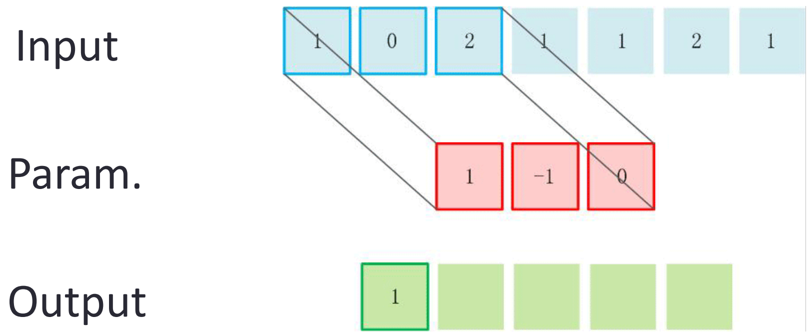
- You can get **word-level** and **sentence-level** representations
- Vanilla RNNs are not good at dealing with long sentences
- There are at least 100 RNN variants … (LSTM, GRU)
- Gating mechanism

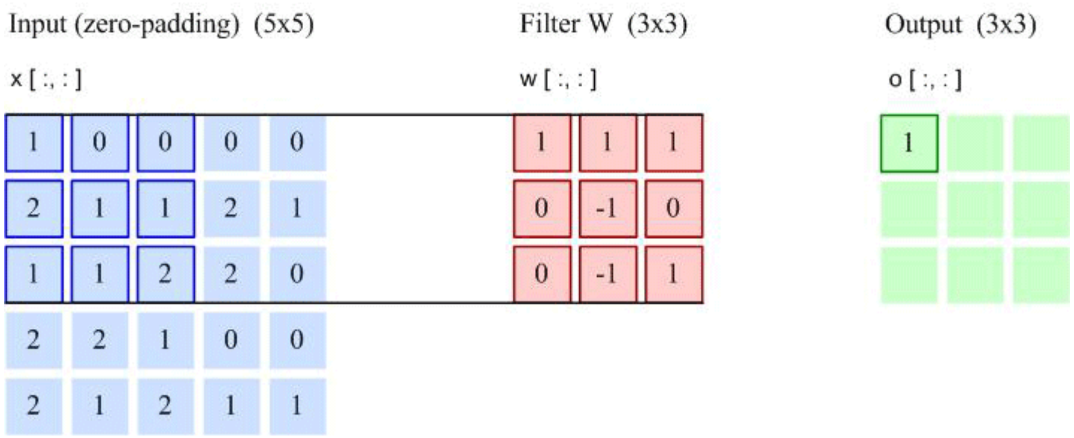


$$\mathbf{h}_t = \begin{cases} 0 & t = 0 \\ f(\mathbf{h}_{t-1}, \mathbf{x}_t) & \text{otherwise} \end{cases}$$

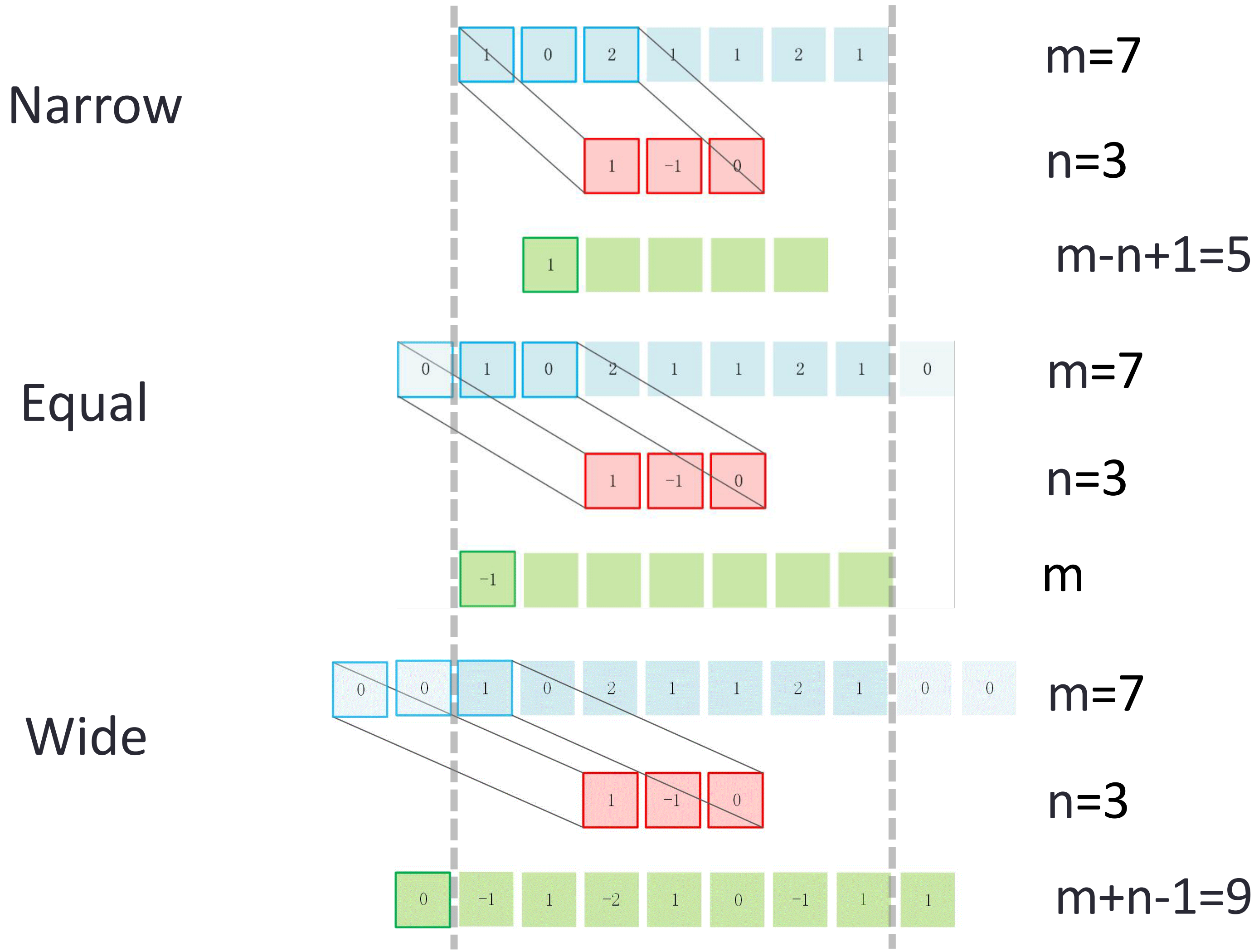
Case Study: Must-know Points about CNN

CNN: 1d and 2d Convolution

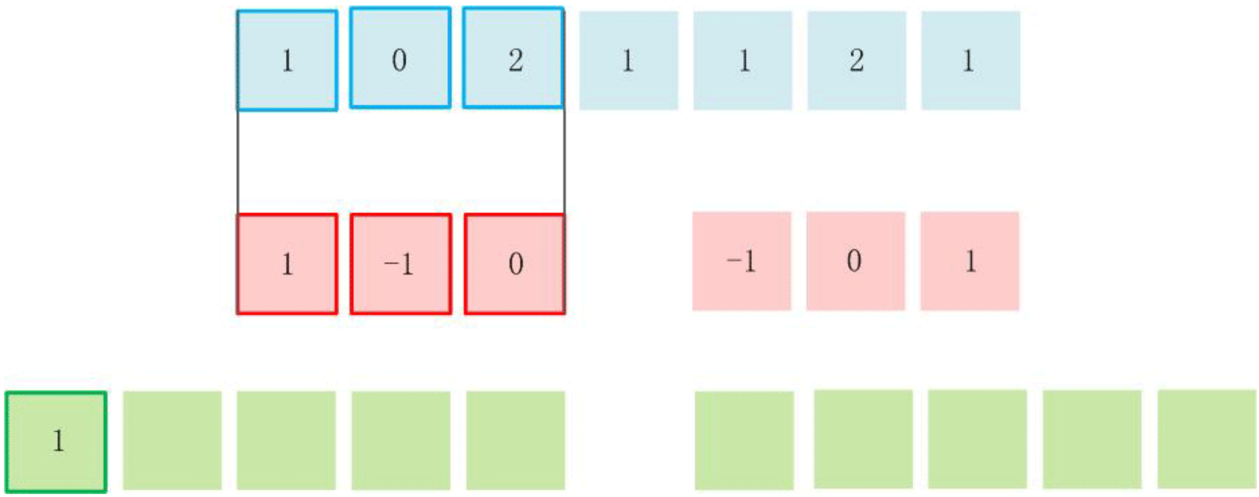




CNN: Narrow/Equal/Wide Convolution

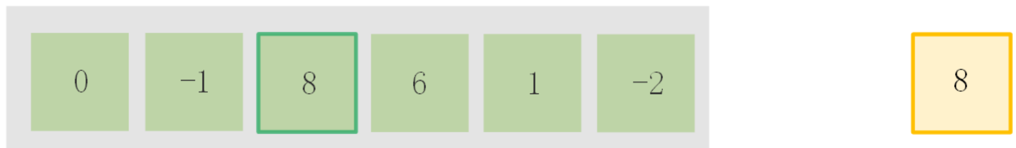


CNN: Multiple Filter Convolution

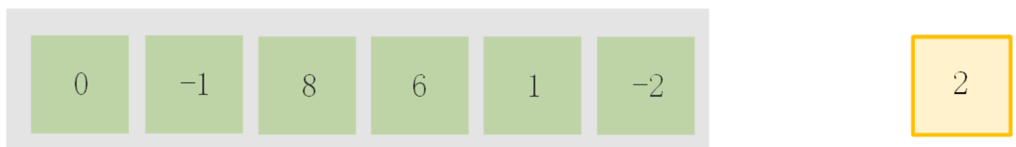


CNN: Pooling

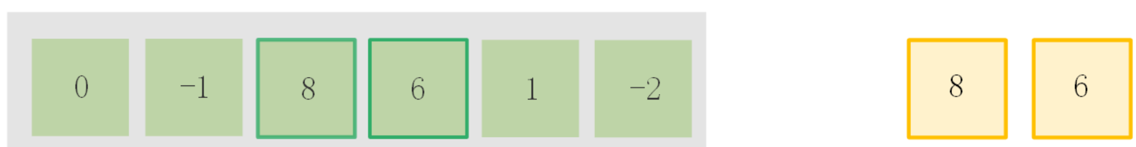
Max pooling:



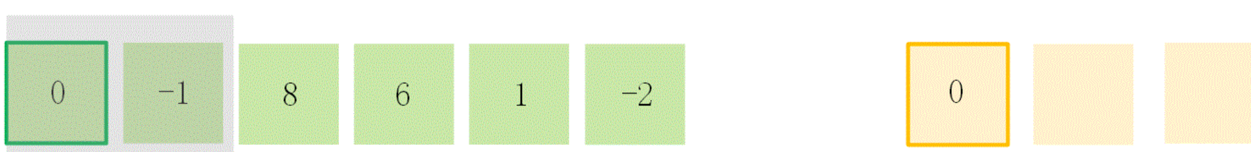
Mean pooling:



K-max pooling

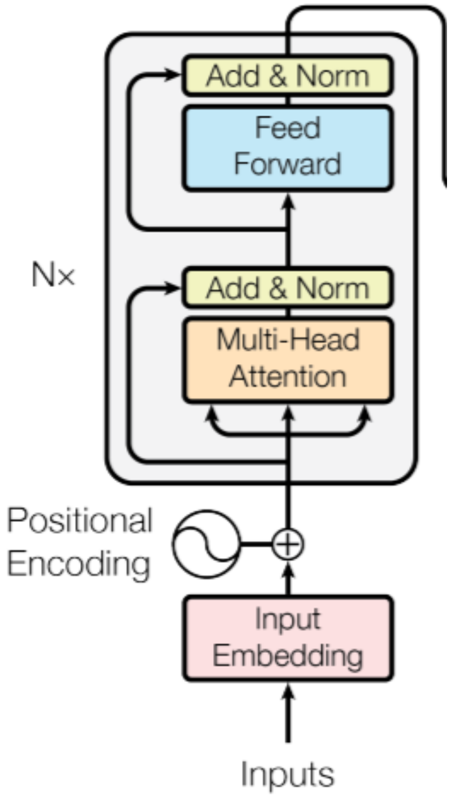


Dynamic pooling:



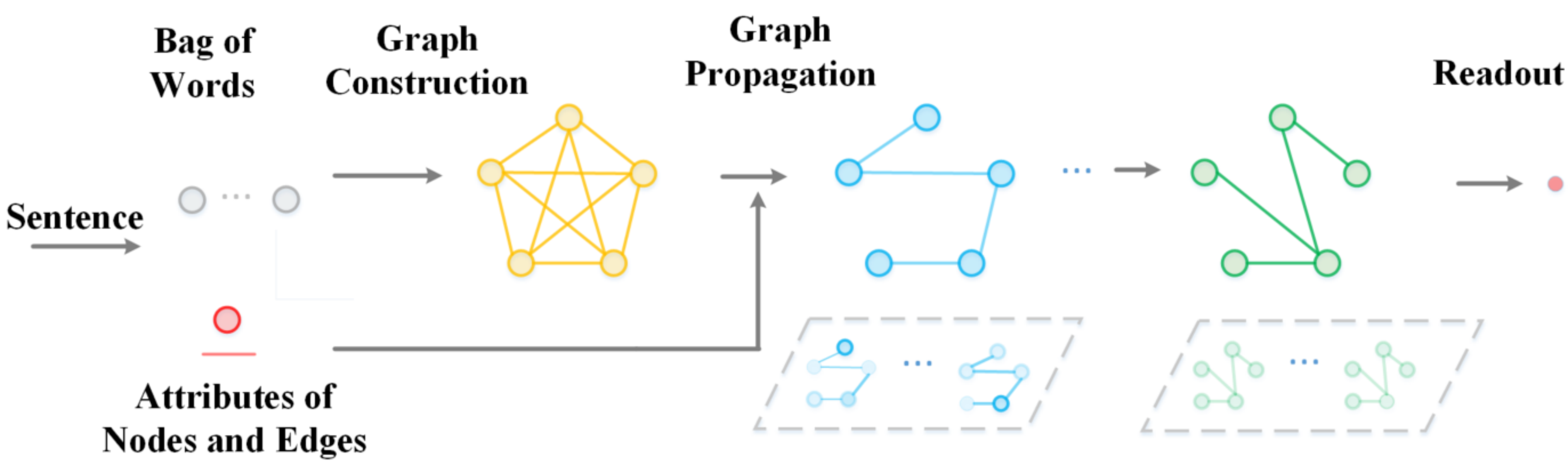
Case Study: Must-know Points about Transformer

- It’s a composite module
- No CNN! No RNN! Only Attention
- Fast and parallel training (BERT)
- Lack of local bias, require more data



Case Study: Must-know Points about GNNs

- Help us introduce relational bias
- Transformer is a fully-connected graph
- Not very efficient to train



Which one should I choose?

Transformer is suggested when…

- ✓ Machine translation
- ✓ If you have more training data
- ✓ Extremely deep neural nets
- ✓ Best of both worlds

LSTM is suggested when…

- ✓ In most cases …
- ✓ Modestly-sized data (tagging)
- ✓ Best of both worlds

Graph Neural Net is suggested when …

- ✓ More complicated relational biases
  - QA: entity
  - Summarization: structure of doc.
- ✓ Modestly-sized data
- ✓ Best of both worlds

CNN is suggested when …

- ✓ Word encoder
- ✓ …

Task-wisely, if you can use BERT…

Tagging, Text Classification

BERT (FLAIR) + MLP

BERT (FLAIR) + LSTM + MLP

BERT (FLAIR) + Transformer + MLP

Note:

1) BERT should be fine-tuned
2) For tagging tasks, FLAIR performs better than BERT

Task-wisely, if you can use BERT…

Text Generation

BART + Seq2Seq

Task-wisely, if you can't use BERT...

Tagging, Text Classification

GloVe + NBOW + MLP

GloVe + BiLSTM + MLP

GloVe + CNN + BiLSTM + MLP (or CRF)

Note:

1) Glove can be replaced with word2vec, and should be fine-tuned
2) For tagging tasks, replace MLP with CRF layer

Year: Nothing selected ▾ Conference: Nothing selected ▾ Concept: Nothing selected ▾

Year	Conf.	Concept	Cited	Paper
2014	nips	none	19365	Distributed Representations of Words and Phrases and their Compositionality Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean
2013	arxiv	none	15383	Efficient Estimation of Word Representations in Vector Space Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean
2014	emnlp	none	13069	Glove: Global Vectors for Word Representation Jeffrey Pennington, Richard Socher, Christopher Manning
2003	jmlr	none	6074	A Neural probabilistic language model Yoshua Bengio, Rejean Ducharme, Pascal Vincent
2019	naacl	none	5292	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
2018	naacl	none	2913	Deep Contextualized Word Representations Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer
2013	naacl	none	2578	Linguistic Regularities in Continuous Space Word Representations Tomas Mikolov, Wen-tau Yih, Geoffrey Zweig
2012	acl	none	1079	Improving Word Representations via Global Context and Multiple Word Prototypes Eric Huang, Richard Socher, Christopher Manning, Andrew Ng
2014	arxiv	none	971	word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method Yoav Goldberg and Omer Levy
2015	tacl	none	903	Improving Distributional Similarity with Lessons Learned from Word Embeddings Omer Levy, Yoav Goldberg, Ido Dagan
2014	acl	none	862	Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification

http://pfliu.com/paperlist/wordemb.html

Year: Nothing selected   Conference: Nothing selected   Concept: Nothing selected

All | None

setting
- setting-crossLingual
- setting-endangered
- setting-lowResource

Year	Conf.	Concept	Cited	Paper
2009	acl	setting-crossLingual	486	Co-Training for Cross-Lingual Sentiment Xiaojun Wan
2007	acl	setting-crossLingual	436	Learning Multilingual Subjective Language via Cross-Lingual Projections Rada Mihalcea, Carmen Banea, Janyce Wiebe
2006	cl	setting-crossLingual	378	Unsupervised Multilingual Sentence Boundary Detection Tibor Kiss, Jan Strunk
2013	acl	setting-crossLingual	369	Universal Dependency Annotation for Multilingual Parsing Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzma
2008	emnlp	setting-crossLingual	265	Multilingual Subjectivity Analysis Using Machine Translation Carmen Banea, Rada Mihalcea, Janyce Wiebe, Samer Hassan
2014	acl	setting-crossLingual	252	Multilingual Models for Compositional Distributed Semantics Karl Moritz Hermann, Phil Blunsom
2016	emnlp	setting-lowResource	234	Transfer Learning for Low-Resource Neural Machine Translation Barret Zoph, Deniz Yuret, Jonathan May, Kevin Knight
1997	cl	setting-crossLingual	228	Adaptive Multilingual Sentence Boundary Disambiguation David D. Palmer, Marti A. Hearst
2013	acl	setting-crossLingual	221	Linking and Extending an Open Multilingual Wordnet Francis Bond, Ryan Foster
2016	acl	setting-crossLingual	215	Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Barbara Plank, Anders Søgaard, Yoav Goldberg
2004	naacl	setting-crossLingual	208	A Statistical Model for Multilingual Entity Detection and Tracking R. Florian, H. Hassan, A. Ittycheriah, H. Jing, N. Kambhatla, X. Luo, N. Nicolov, S. Roukos
1994	coling	setting-crossLingual	186	MULTEXT: Multilingual Text Tools and Corpora Nancy Ide, Jean Veronis

http://pfliu.com/paperlist/lowsource.html

Thanks !