Neural Vector Representations beyond Words: Sentence and Document Embeddings

Gerard de Melo
http://gerard.demelo.org

Rutgers University

Outline

- Word Representations
- Phrase Representations
- Sentence Representations
- Document Representations
- Applications and Outlook

Structured (Non-Vector) Representations

Formal Semantics

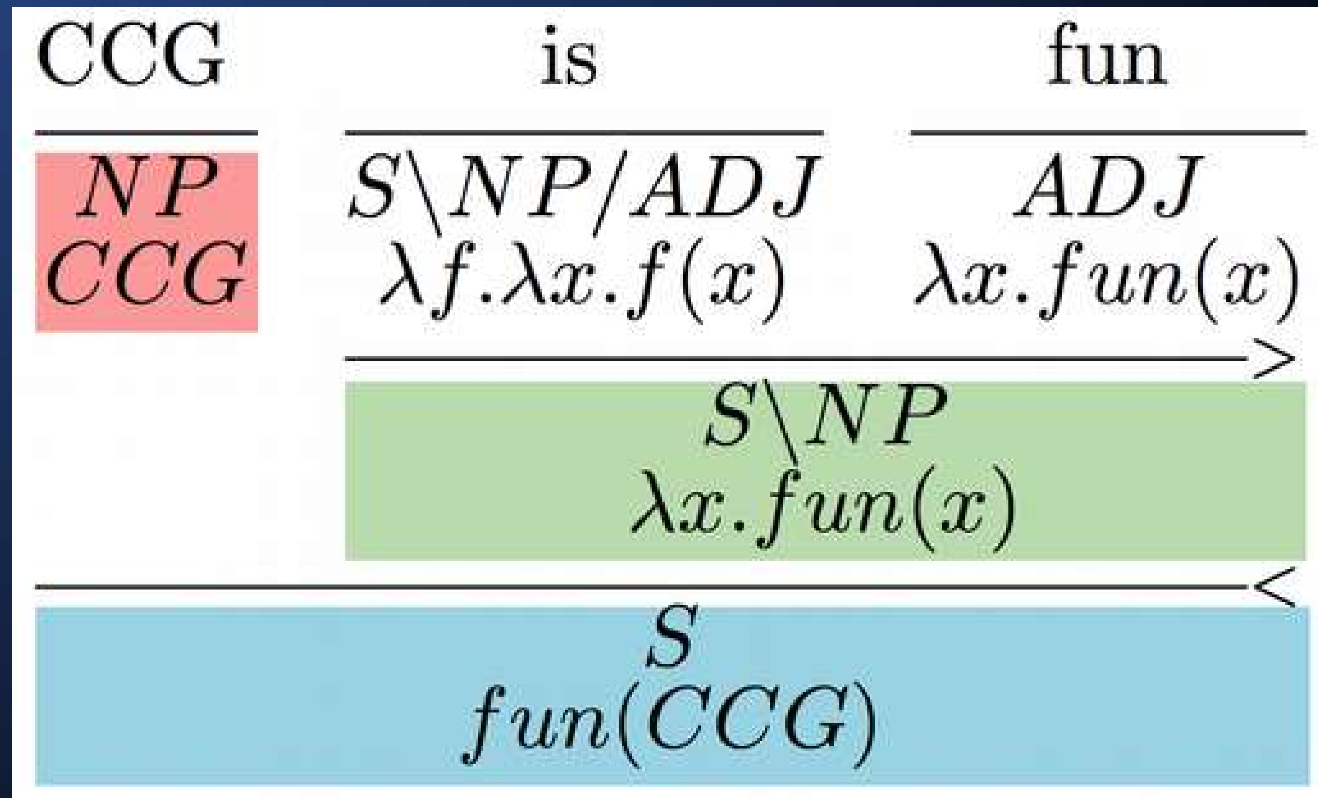
John likes everything that is interesting

$$\forall x\ (\mathrm{likes}(\mathrm{John}, x) \rightarrow \mathrm{interesting}(x))$$

Traditional Grammar Frameworks, e.g. CCG

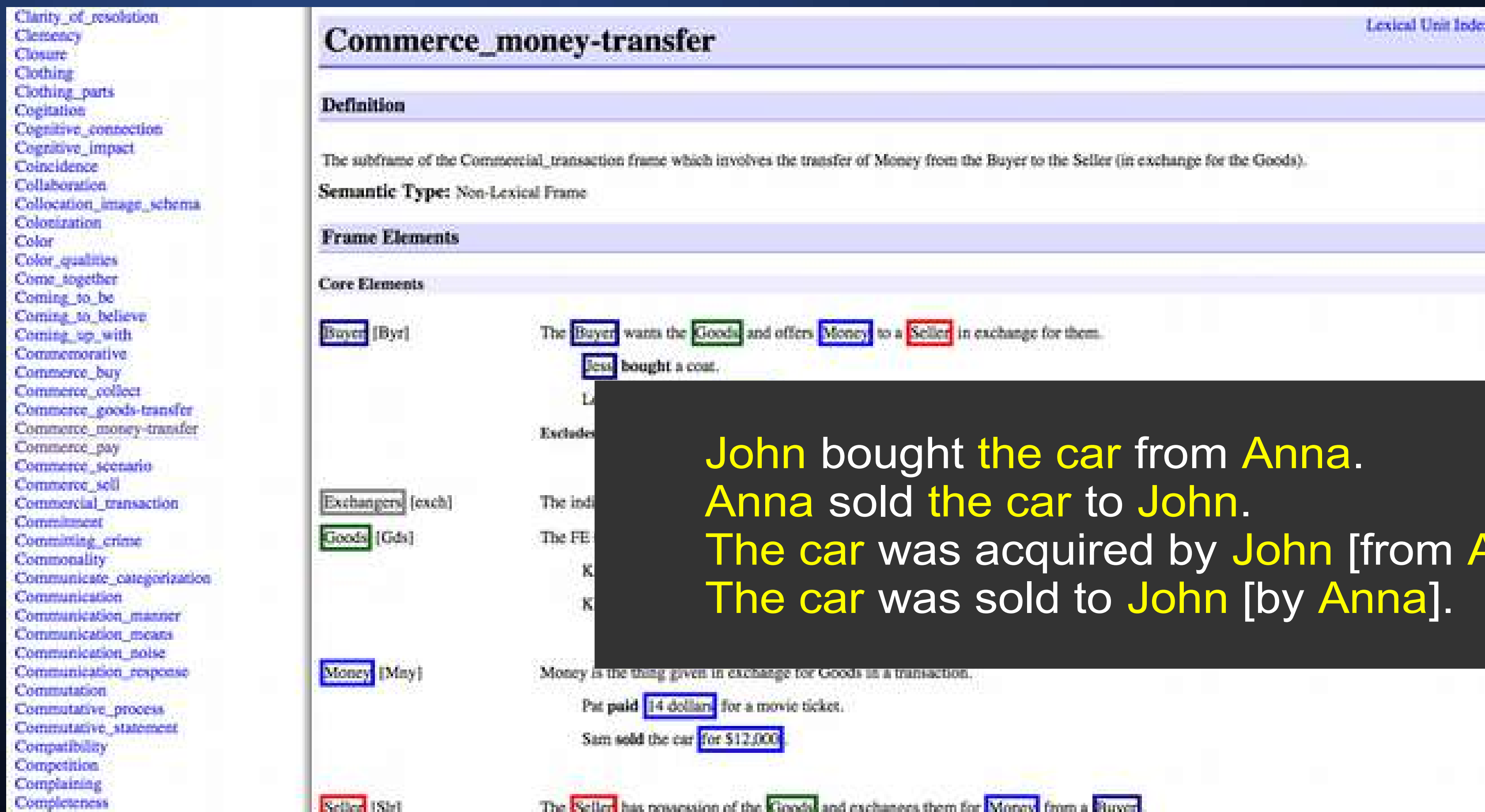
Supervised Learning from Sentence–logic Pairs.

E.g. using CCG (Combinatory Category Grammar)



Source: Yoav Artzi, Nicholas FitzGerald and Luke Zettlemoyer!. Semantic Parsing with Combinatory Categorial Grammars

Frame Semantics and Semantic Role Labeling



John bought the car from Anna.
Anna sold the car to John.
The car was acquired by John [from Anna].
The car was sold to John [by Anna].

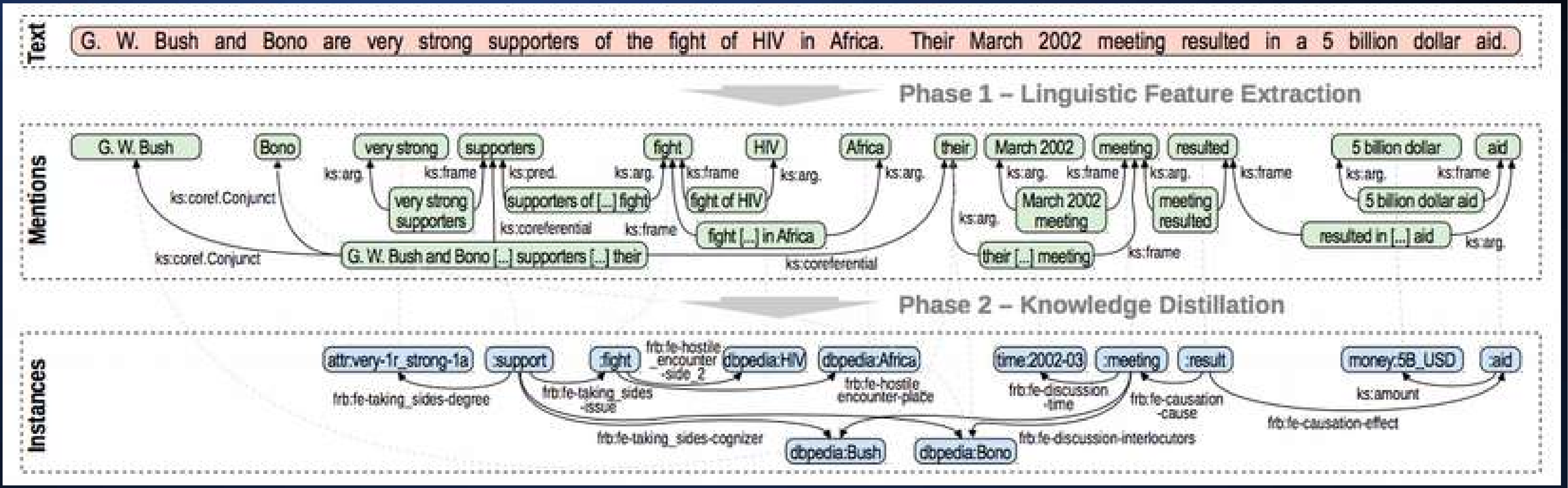
Frame Semantics and Semantic Role Labeling

Microsoft bought the patent from Nokia.
Nokia sold the patent to Microsoft.
The patent was acquired by Microsoft [from Nokia].
The patent was sold [by Nokia] to Microsoft.

Underlying frame:
Commercial transfer

Buyer:	Microsoft
Seller:	Nokia
Product:	The patent

FrameBase.org: Text to FrameBase



PIKES: Corcoglioniti et al. 2016

Video:
https://www.youtube.com/watch?v=D0mcnUKc3sg

FrameBase.org: Text to FrameBase

KnEWS

Knowledge Extraction With Semantics

KNEWS is a composite tool that bridges semantic parsing (using C&C tools and Boxer), word sense disambiguation (using UKB or Babelfy) and entity linking (using Babelfy or DBpedia Spotlight) to produce a unified, LOD-compliant abstract representation of meaning.

KNEWS can produce several kinds of output:

1. Frame instances, based on the FrameBase scheme
2. Word-aligned semantics, based on lexicalized Discourse Representation Graphs)
3. First-order logic formulae with WordNet synsets and DBpedia ids as symbols

The source code of KNEWS is freely available at https://github.com/ valeriobasile/learningbyreading.

KnEWS: Basile et al. 2016 (INRIA/CNRS)

https://github.com/valeriobasile/learningbyreading
http://gingerbeard.alwaysdata.net/knews/

Neural Frame Semantic Parsing

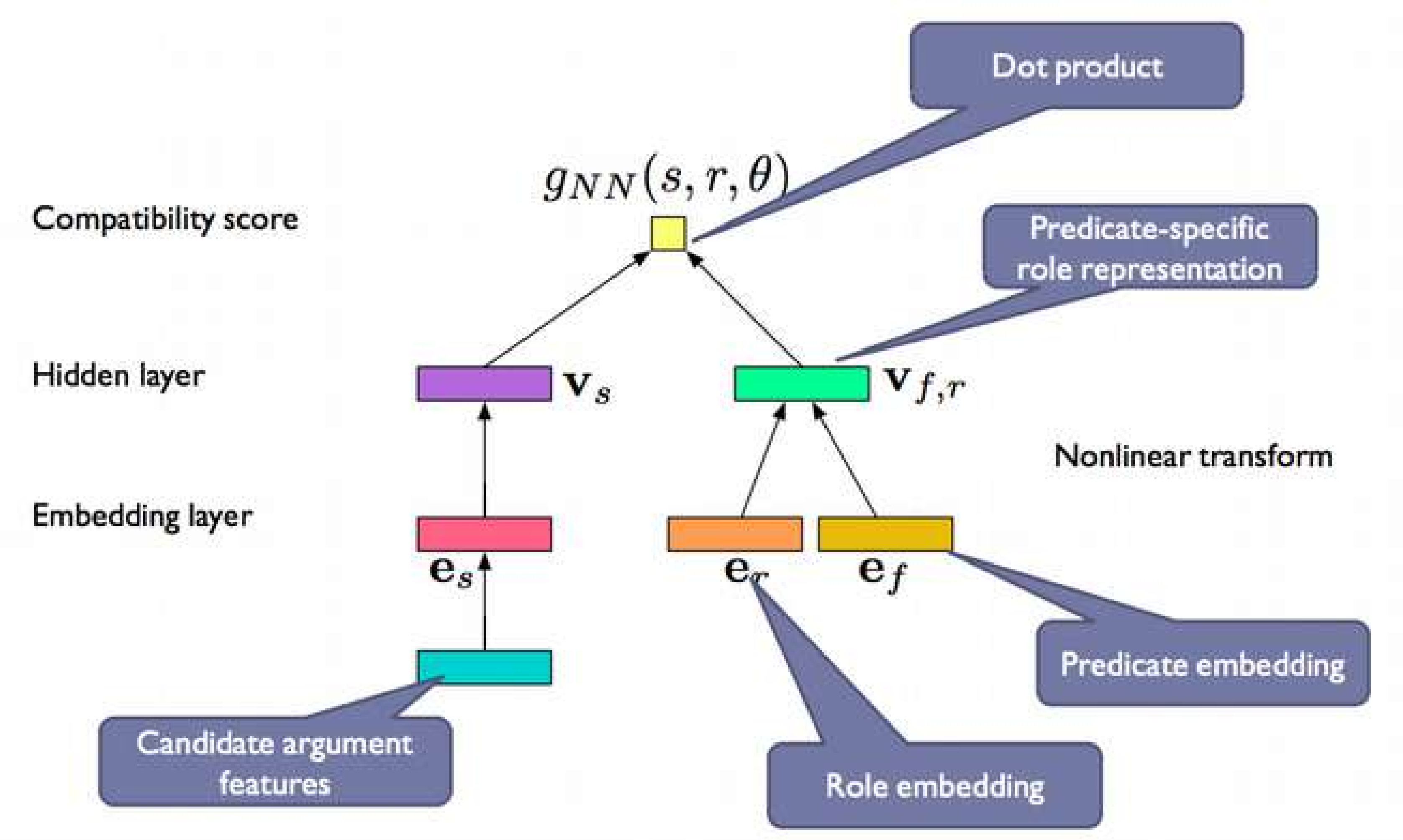
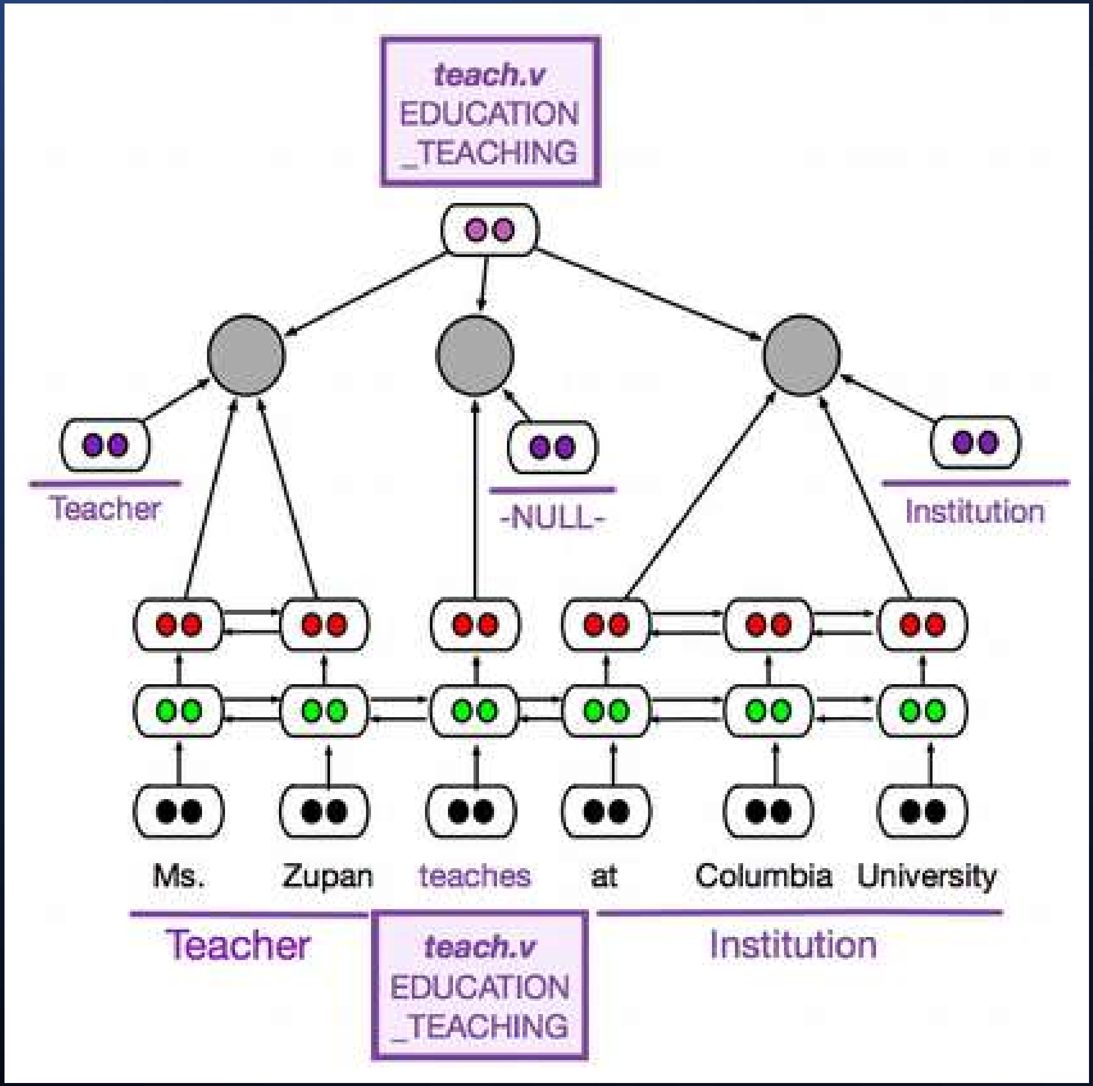


Image: Diego Marcheggiani, EMNLP 2017 Tutorial

Nicholas FitzGerald, Oscar Tackström, Kuzman Ganchev & Dipanjan Das. Semantic role labeling with neural network factors. EMNLP 2015

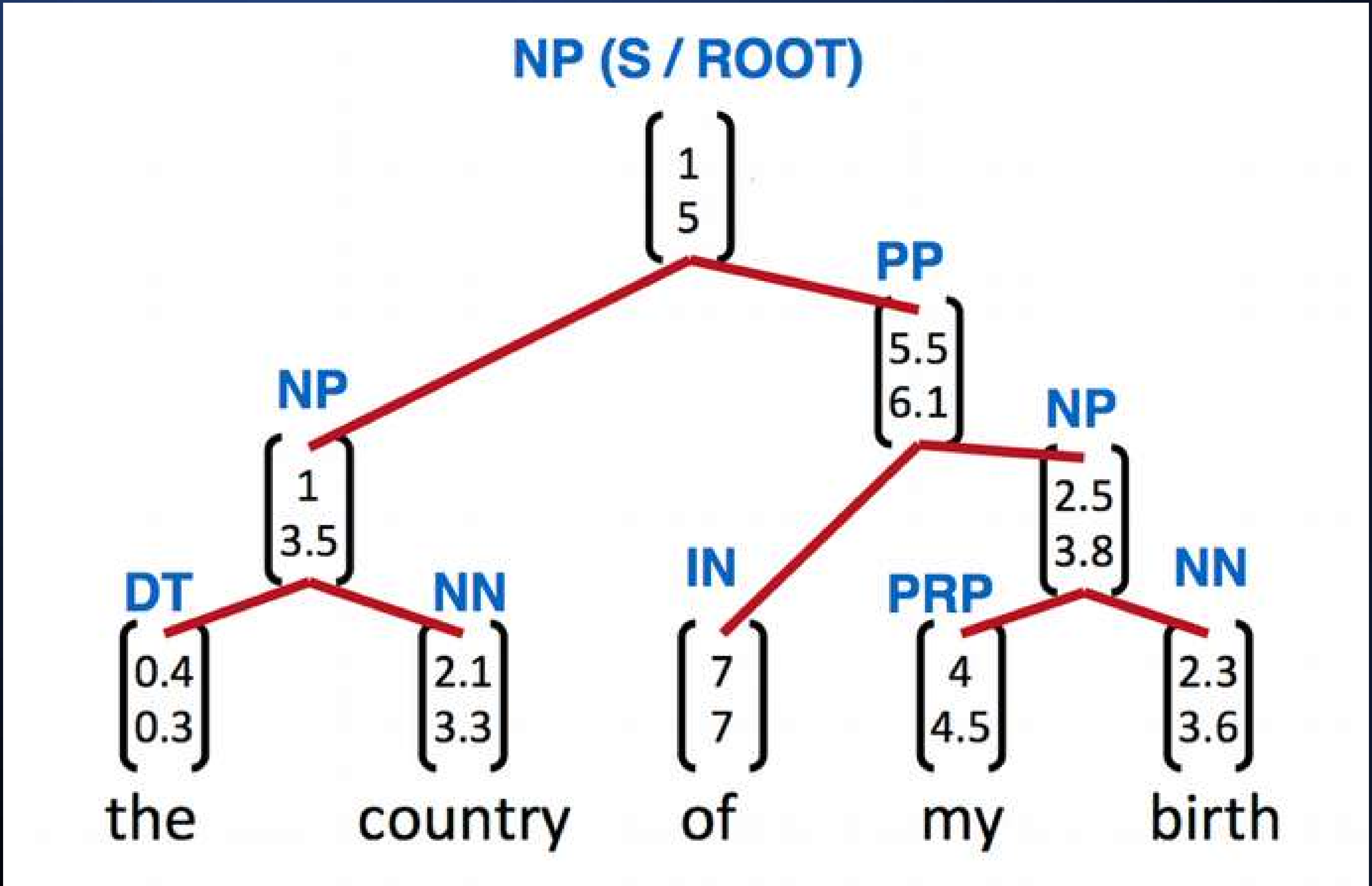
Neural Frame Semantic Parsing



Swabha Swayamdipta, Sam Thomson, Chris Dyer, Noah A. Smith. Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold. https://arxiv.org/pdf/1706.09528.pdf

Vector Compositionality

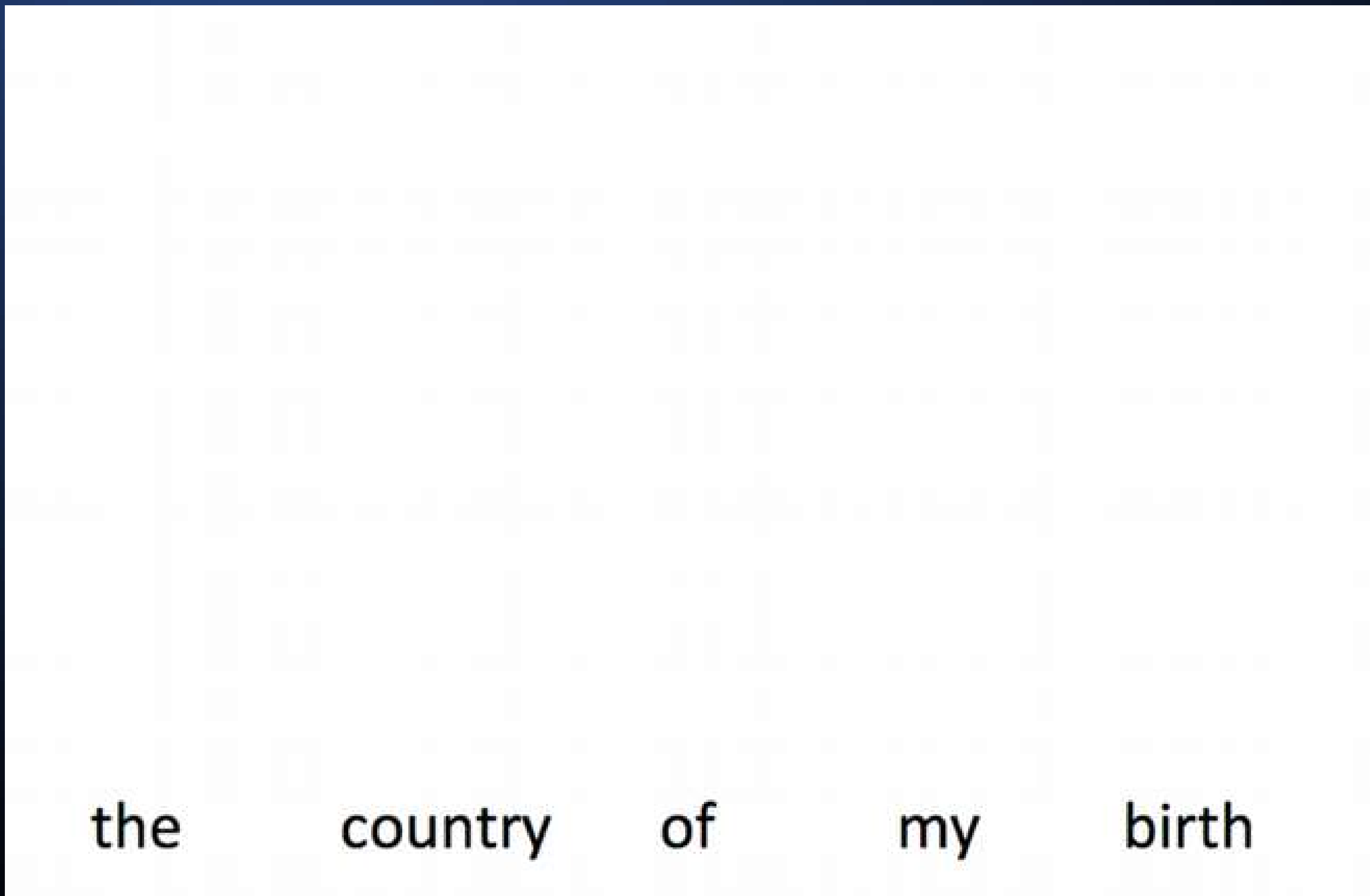
Compositional Representations



Recursive Neural Network approach by Socher et al.

Image: Roelof Pieters

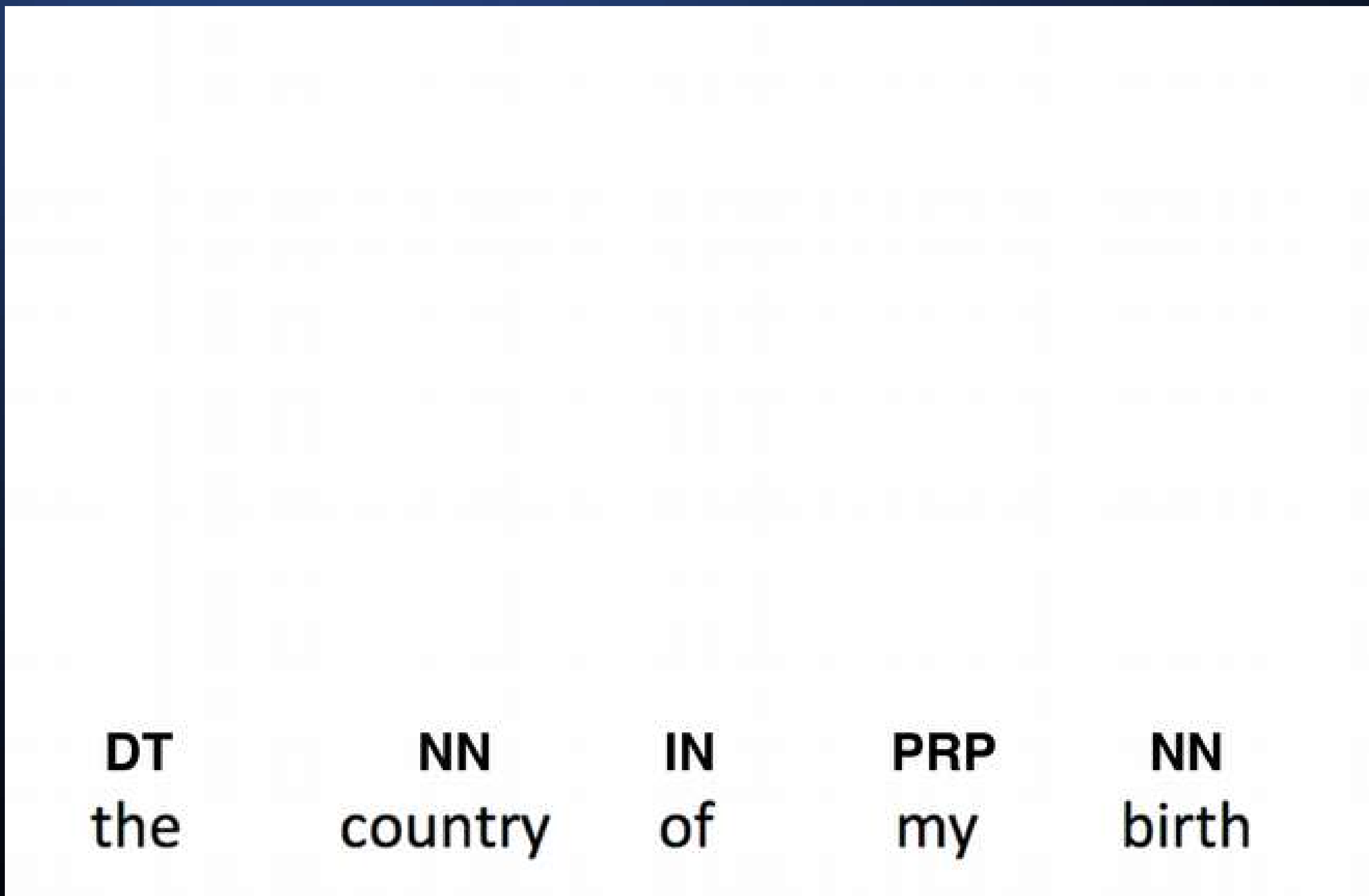
Compositional Representations



Recursive Neural Network approach by Socher et al.

Image: Roelof Pieters

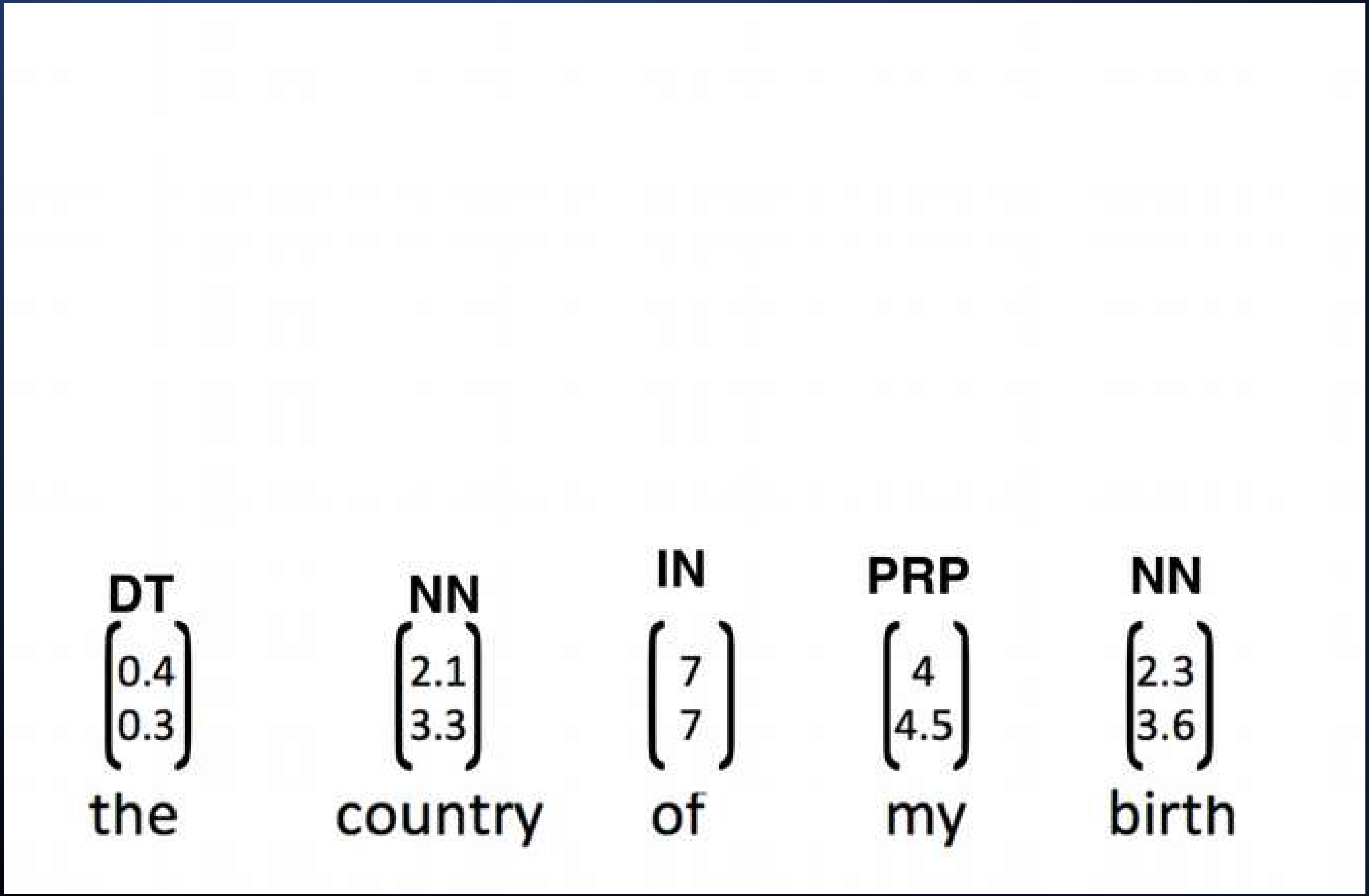
Compositional Representations



Recursive Neural Network approach by Socher et al.

Image: Roelof Pieters

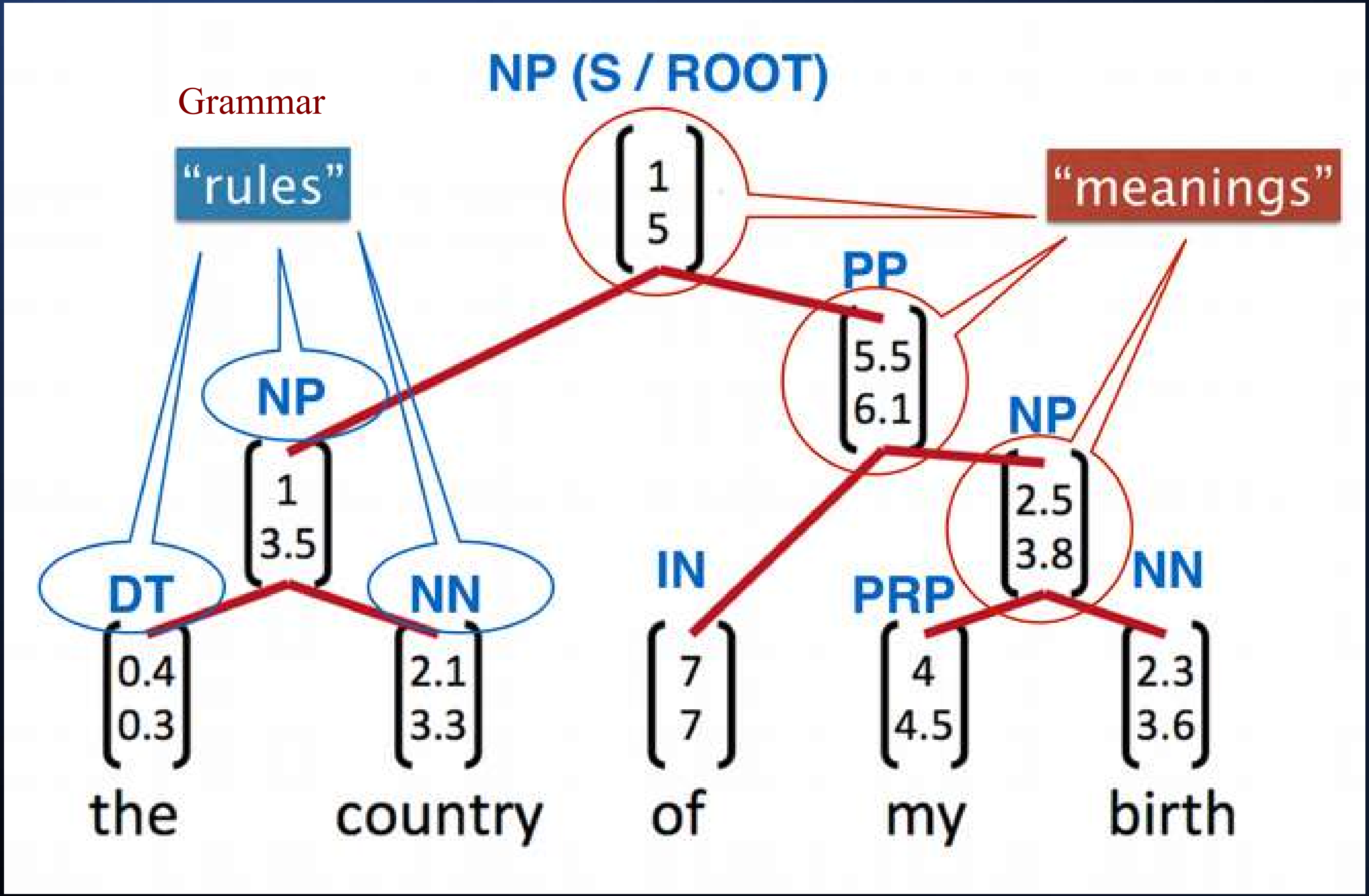
Compositional Representations



Recursive Neural Network approach by Socher et al.

Image: Roelof Pieters

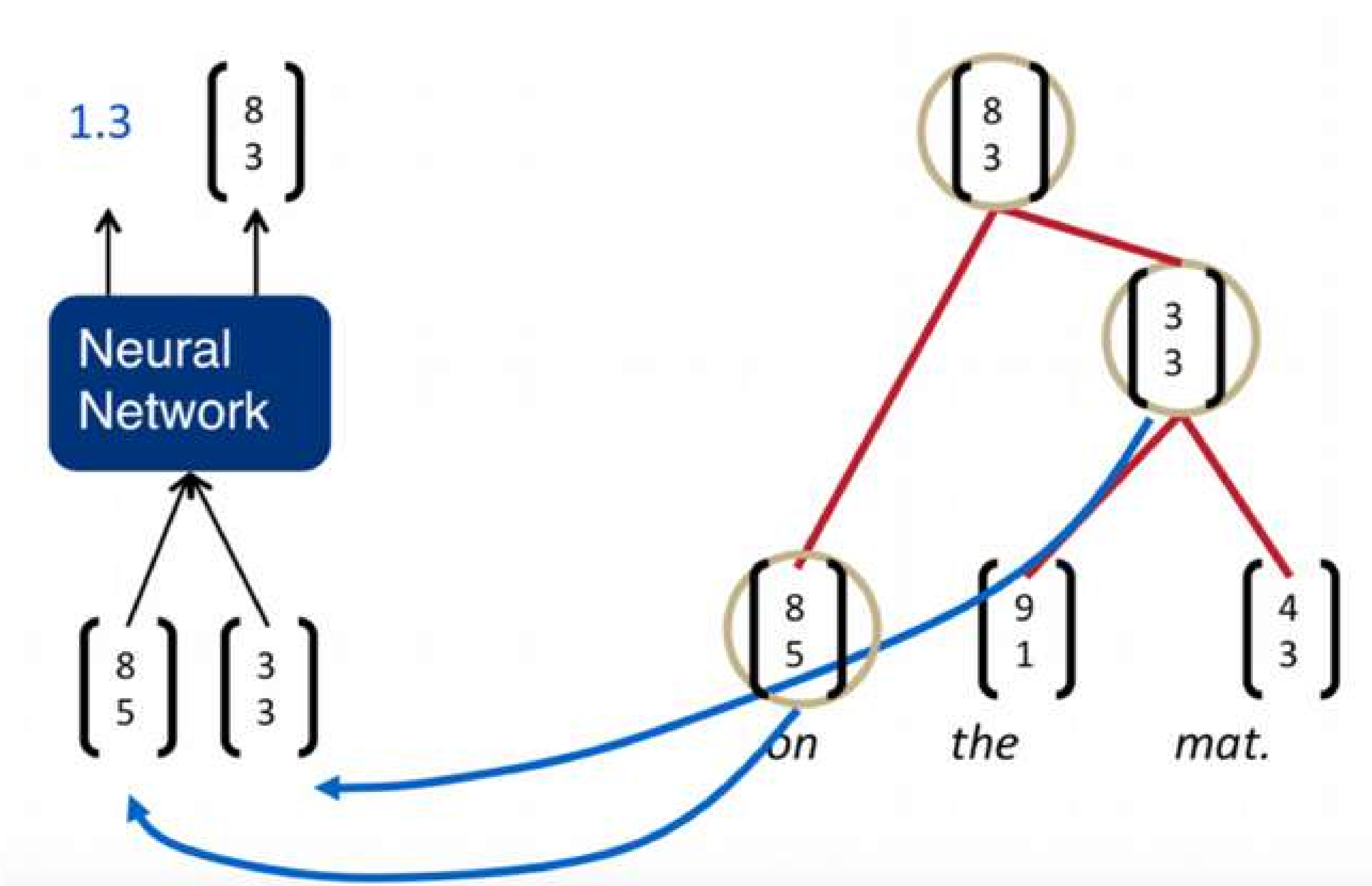
Compositional Representations



Recursive Neural Network approach by Socher et al.

Image: Roelof Pieters

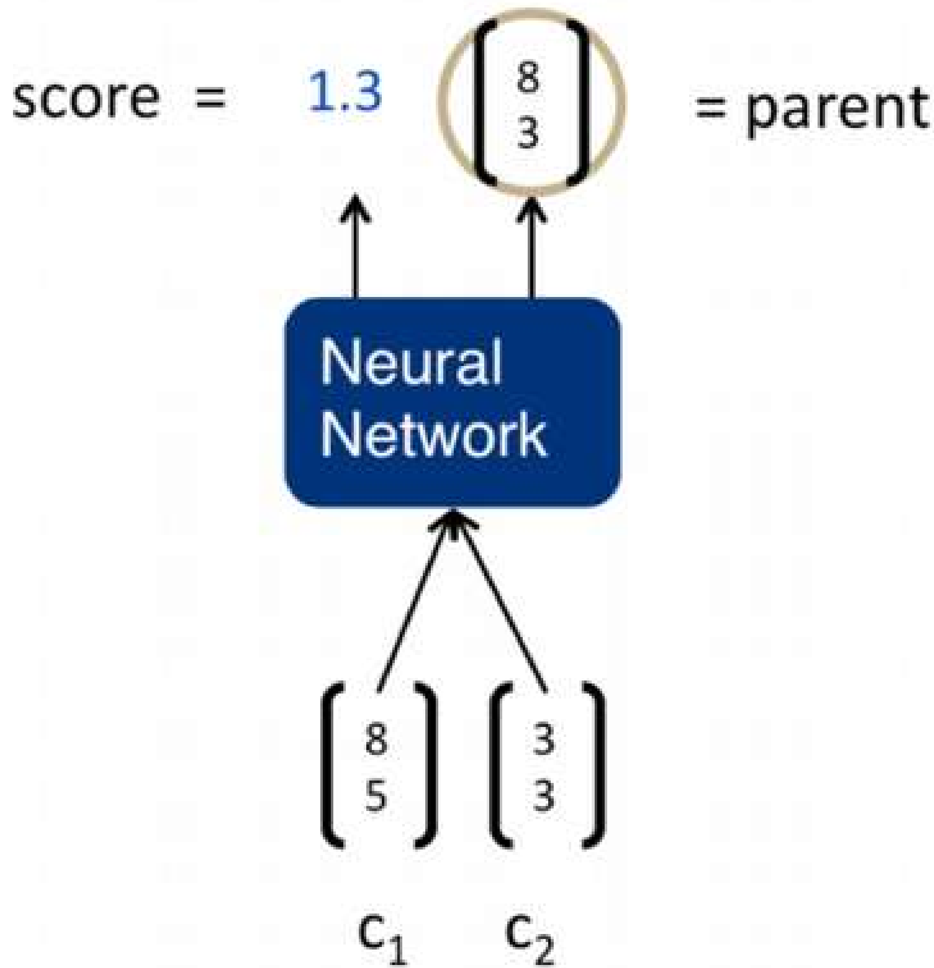
Compositional Representations



Recursive Neural Network approach by Socher et al.

Image: Roelof Pieters

Compositional Representations



$$\text{score} = U^T p$$

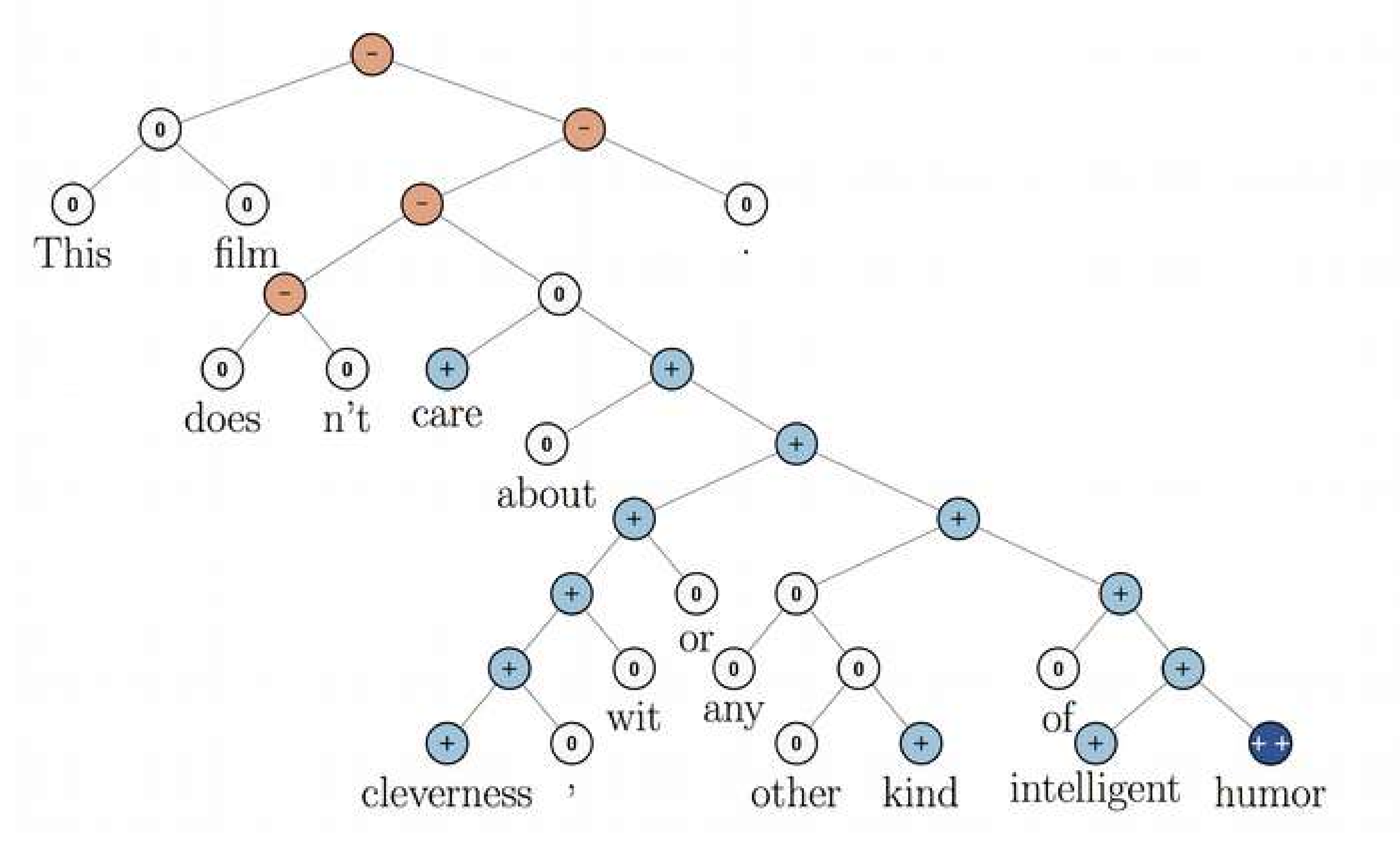
$$\boxed{p = \tanh\left(W \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + b\right),}$$

Same W parameters at all nodes of the tree

Recursive Neural Network approach by Socher et al.

Image: Roelof Pieters

Compositional Representations



Socher et al. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank

Compositional Representations

Model	Error rate (Positive/ Negative)	Error rate (Fine-grained)
Naïve Bayes (Socher et al., 2013b)	18.2 %	59.0%
SVMs (Socher et al., 2013b)	20.6%	59.3%
Bigram Naïve Bayes (Socher et al., 2013b)	16.9%	58.1%
Word Vector Averaging (Socher et al., 2013b)	19.9%	67.3%
Recursive Neural Network (Socher et al., 2013b)	17.6%	56.8%
Matrix Vector-RNN (Socher et al., 2013b)	17.1%	55.6%
Recursive Neural Tensor Network (Socher et al., 2013b)	14.6%	54.3%
Paragraph Vector	**12.2%**	**51.3%**

Results on Stanford Sentiment Treebank

Modern methods easily outperform Recursive Neural Networks

Note: A few recent works again use trees quite successfully

Quoc Le & Mikolov (2014). Distributed Representations of Sentences and Documents

Modifying word2vec

Paragraph Vector Approach

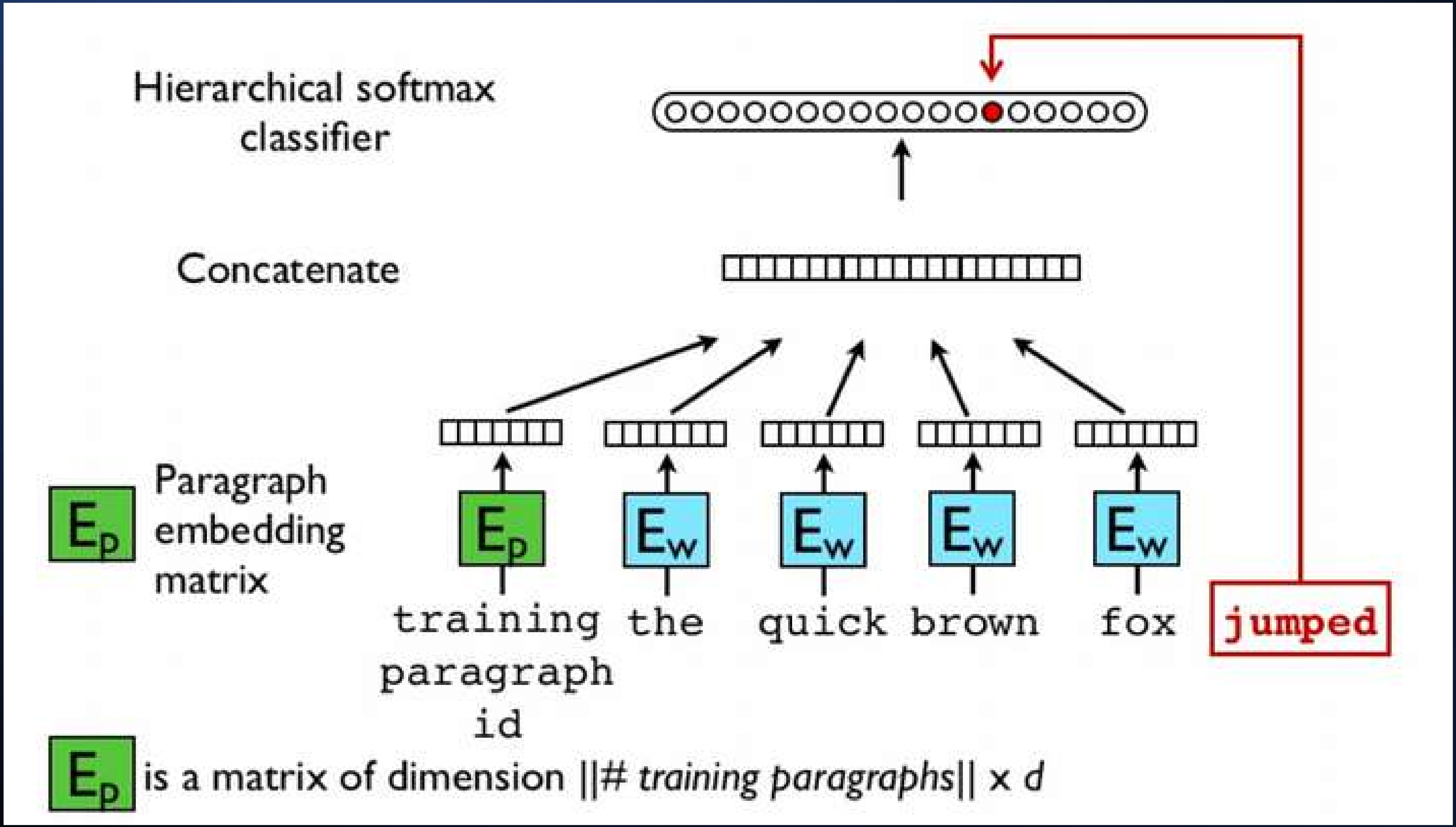
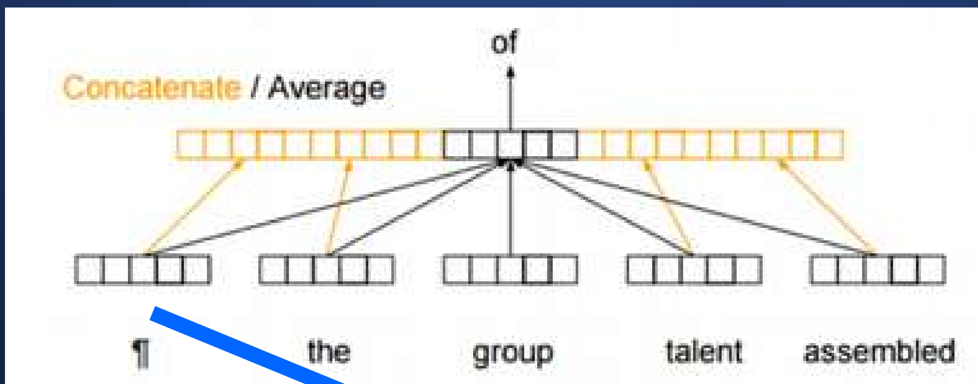
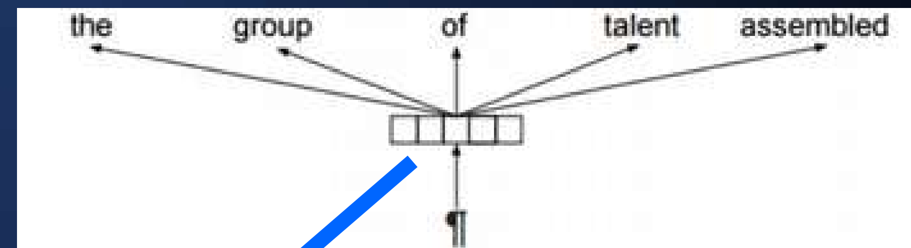


Image: Jeff Dean, Google

Quoc Le & Mikolov (2014). Distributed Representations of Sentences and Documents

Paragraph Vector Approach





PV-DM
(Distributed Memory):
CBOW-like

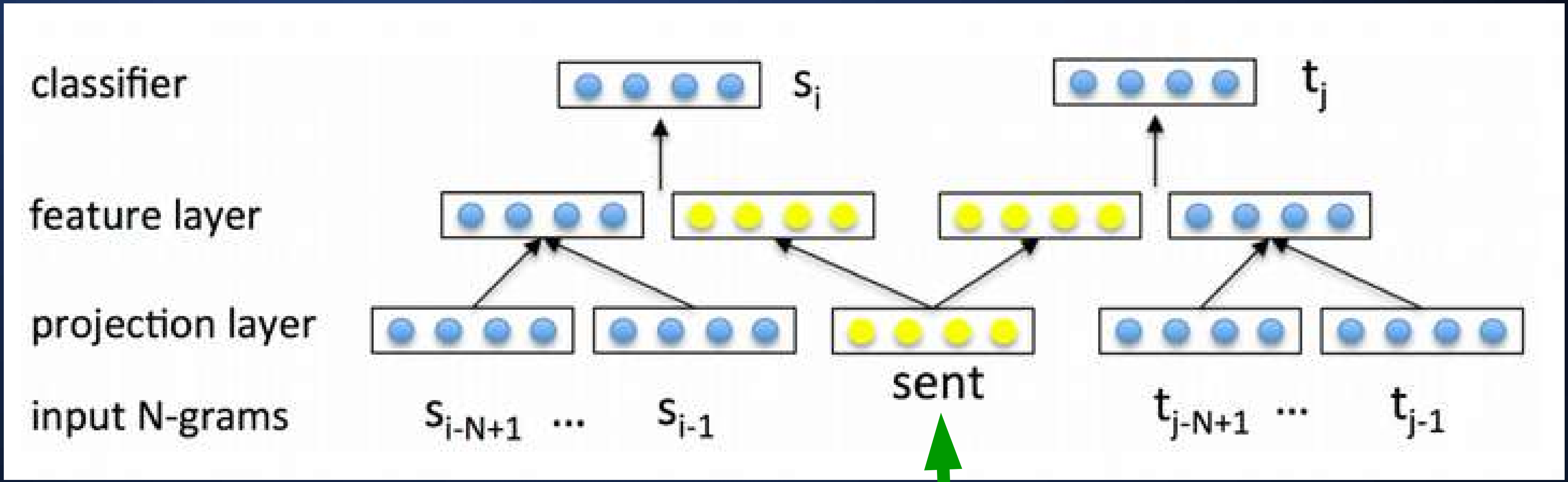
PV-DBOW
(Distributed Bag of Words):
SGNS-like

Concatenation

Quoc Le & Mikolov (2014). Distributed Representations of Sentences and Documents

Images: Sergey I. Nikolenko

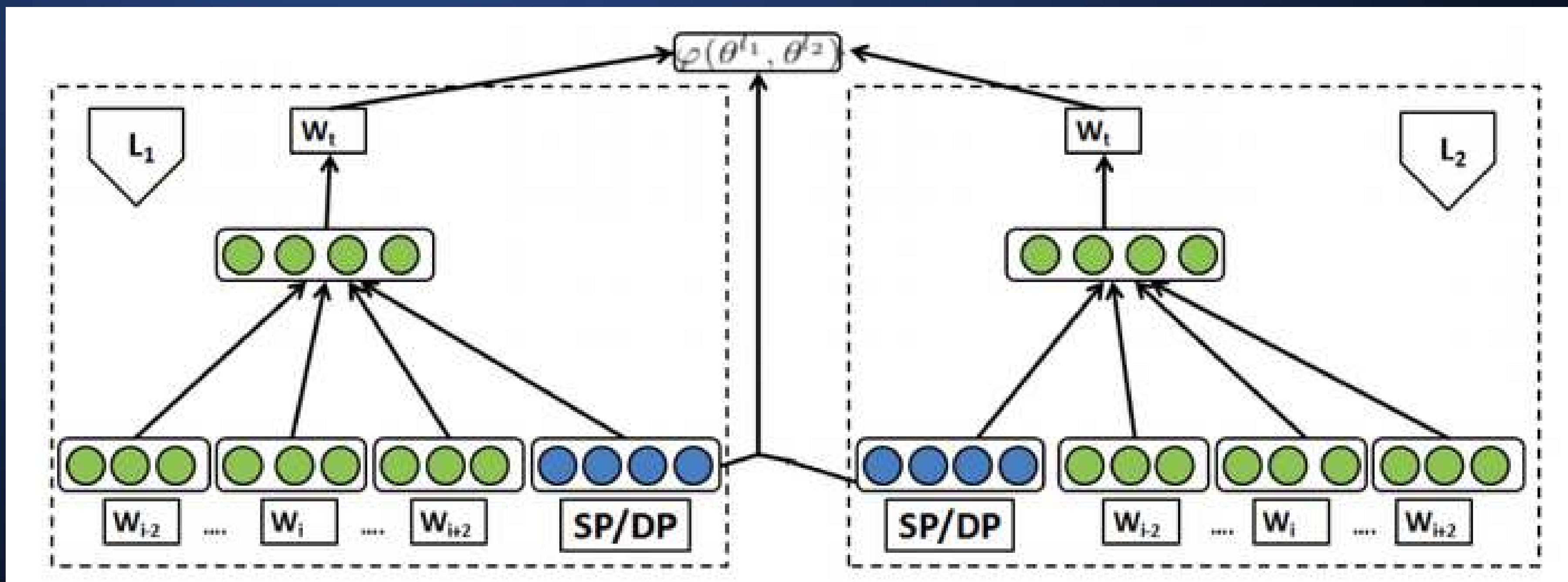
Bilingual Paragraph Vectors



Shared representation for aligned sentences.

Pham, Luong, Manning. Learning Distributed Representations for Multilingual Text Sequences. NAACL-HLT 2015

Bilingual Paragraph Vectors



$$\mathcal{L} = \min_{\theta^{l_1}, \theta^{l_2}} \sum_{l \in \{l_1, l_2\}} \sum_{C^l} \mathcal{M}^l(w_t, h; \theta^l) + \frac{\lambda \varphi(\theta^{l_1}, \theta^{l_2})}{2}$$

Bilingual correlation constraint
for paragraph vector and
mean vector of aligned sentences.

BRAVE Approach

Also present a heuristic when
only aligned documents available

Aditya Mogadala & Achim Rettinger. Bilingual Word Embeddings from Parallel and Non-parallel Corpora for Cross-Language Text Classification

Recurrent Models for Compositionality?

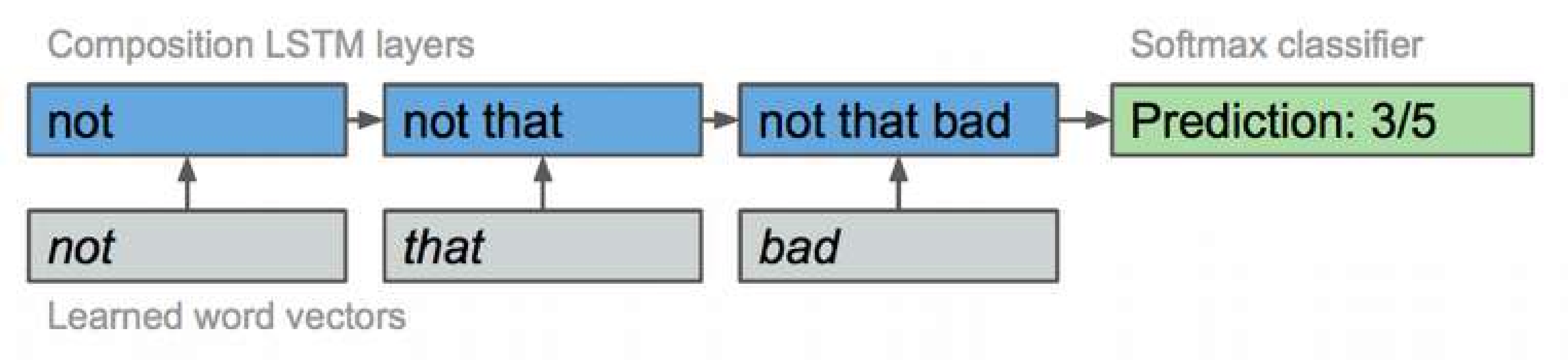
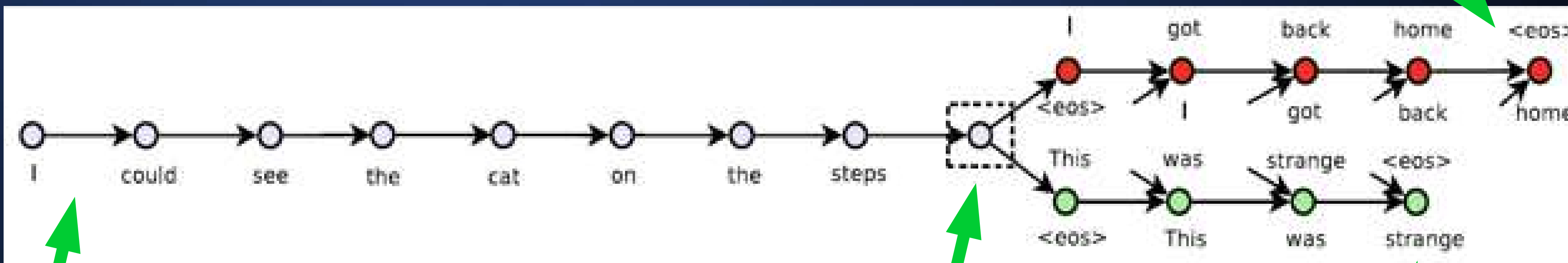


Image: Sam Bowman, NYU

Skip-Thought Vectors

GRU to decode
prev. sentence



Encode input
sentence using
recurrent model

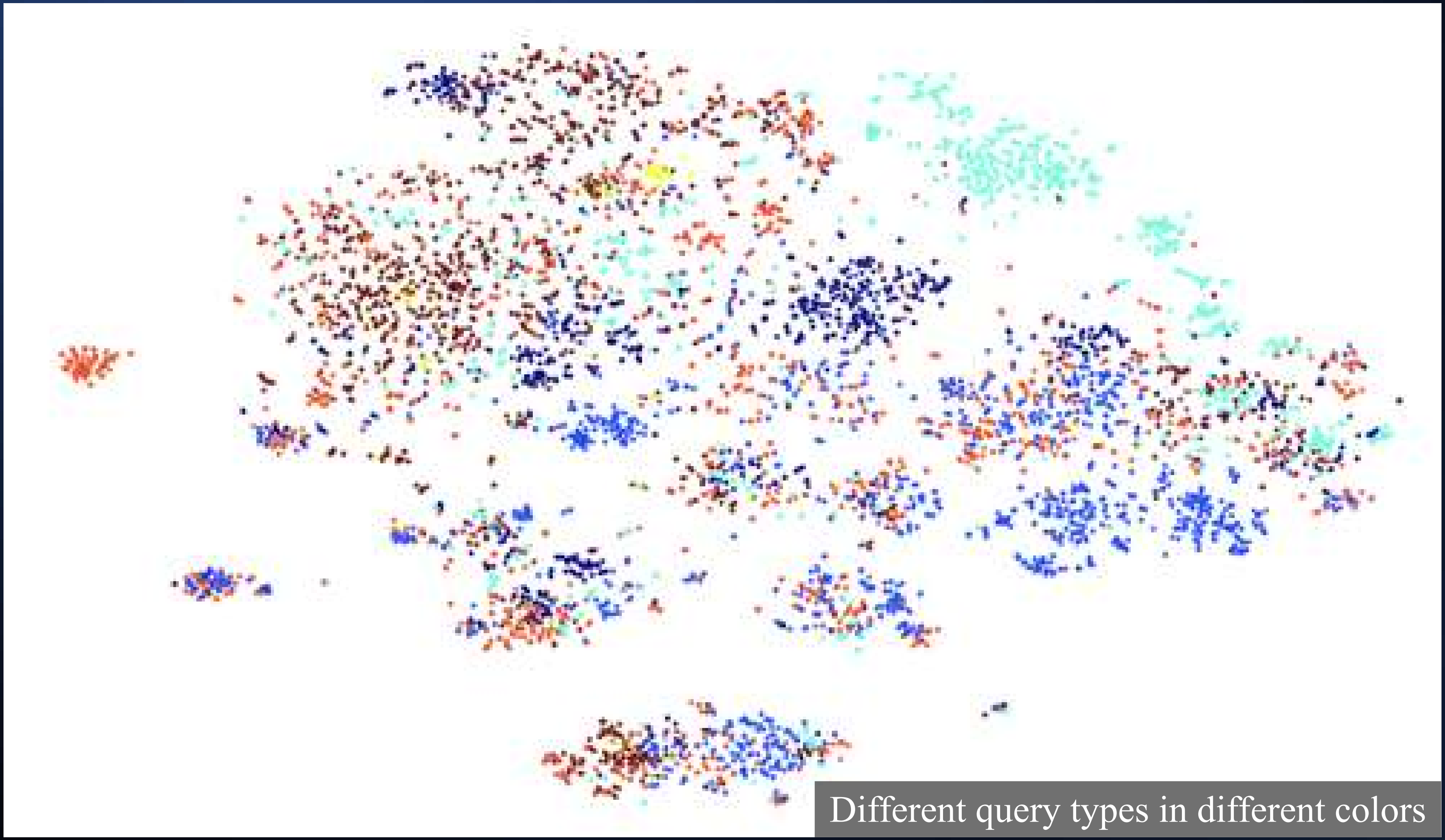
Representation:
final state

GRU to decode
next sentence

**Like word2vec Skip-Gram model
but at the level of sentences
(representation of current sentence should
enable predicting neighbour sentences)**

Kiros et al. (2015). Skip-Thought Vectors. arXiv

Skip-Thought Vectors



Kiros et al. (2015). Skip-Thought Vectors. arXiv

Skip-Thought Vectors

Method	r	ρ	MSE
Illinois-LH [18]	0.7993	0.7538	0.3692
UNAL-NLP [19]	0.8070	0.7489	0.3550
Meaning Factory [20]	0.8268	0.7721	0.3224
ECNU [21]	0.8414	–	–
Mean vectors [22]	0.7577	0.6738	0.4557
DT-RNN [23]	0.7923	0.7319	0.3822
SDT-RNN [23]	0.7900	0.7304	0.3848
LSTM [22]	0.8528	0.7911	0.2831
Bidirectional LSTM [22]	0.8567	0.7966	0.2736
Dependency Tree-LSTM [22]	**0.8676**	**0.8083**	**0.2532**
uni-skip	0.8477	0.7780	0.2872
bi-skip	0.8405	0.7696	0.2995
combine-skip	0.8584	0.7916	0.2687
combine-skip+COCO	0.8655	0.7995	0.2561

Results on SICK

Kiros et al. (2015). Skip-Thought Vectors. arXiv

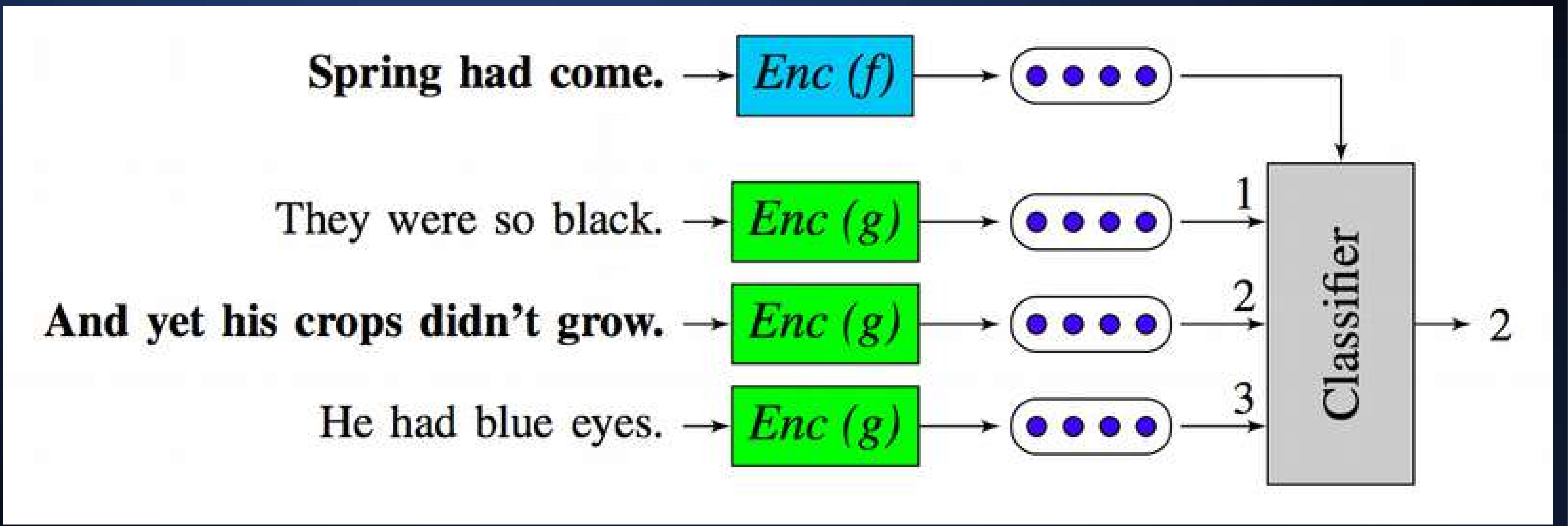
Skip-Thought Vectors

Query and nearest sentence
he ran his hand inside his coat , double-checking that the unopened letter was still there . he slipped his hand between his coat and his shirt , where the folded copies lay in a brown envelope .
im sure youll have a glamorous evening , she said , giving an exaggerated wink . im really glad you came to the party tonight , he said , turning to her .
although she could tell he had n't been too invested in any of their other chitchat , he seemed genuinely curious about this . although he had n't been following her career with a microscope , he 'd definitely taken notice of her appearances .
an annoying buzz started to ring in my ears , becoming louder and louder as my vision began to swim . a weighty pressure landed on my lungs and my vision blurred at the edges , threatening my consciousness altogether .
if he had a weapon , he could maybe take out their last imp , and then beat up errol and vanessa . if he could ram them from behind , send them sailing over the far side of the levee , he had a chance of stopping them .
then , with a stroke of luck , they saw the pair head together towards the portaloos . then , from out back of the house , they heard a horse scream probably in answer to a pair of sharp spurs digging deep into its flanks .
" i 'll take care of it , " goodman said , taking the phonebook . " i 'll do that , " julia said , coming in .
he finished rolling up scrolls and , placing them to one side , began the more urgent task of finding ale and tankards . he righted the table , set the candle on a piece of broken plate , and reached for his flint , steel , and tinder .

Results after c. 2 weeks of training on books corpus

Kiros et al. (2015). Skip-Thought Vectors. arXiv

Quick-Thought Vectors



https://github.com/lajanugen/S2V

Lajanugen Logeswaran, Honglak Lee, An efficient framework for learning sentence representations. ICLR 2018.

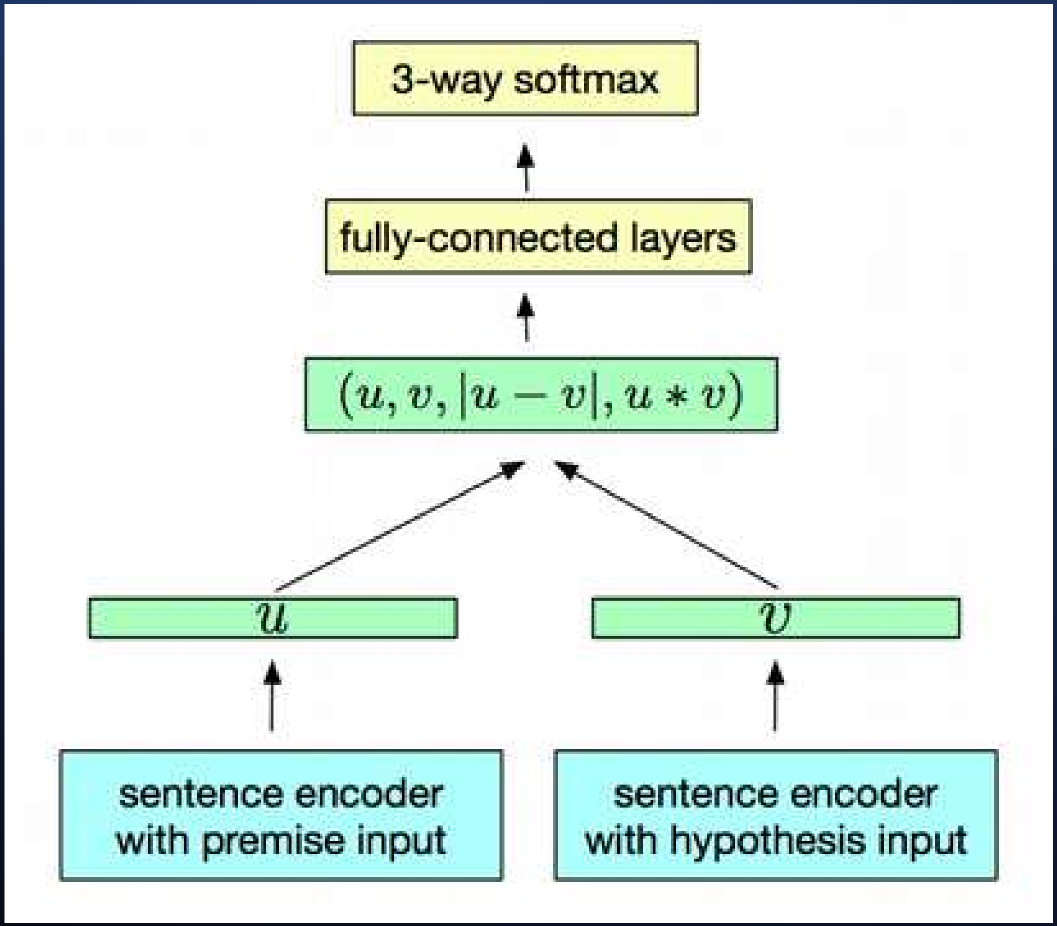
Supervised Approaches

Supervision from Textual Entailment

- **Entailment:** if A is true, then B is true (c.f. paraphrase, where opposite is also true)
  - The woman bought a sandwich for lunch
    → The woman bought lunch
- **Contradiction:** if A is true, then B is not true
  - The woman bought a sandwich for lunch
    → The woman did not buy a sandwich
- **Neutral:** cannot say either of the above
  - The woman bought a sandwich for lunch
    → The woman bought a sandwich for dinner

Source: Graham Neubig

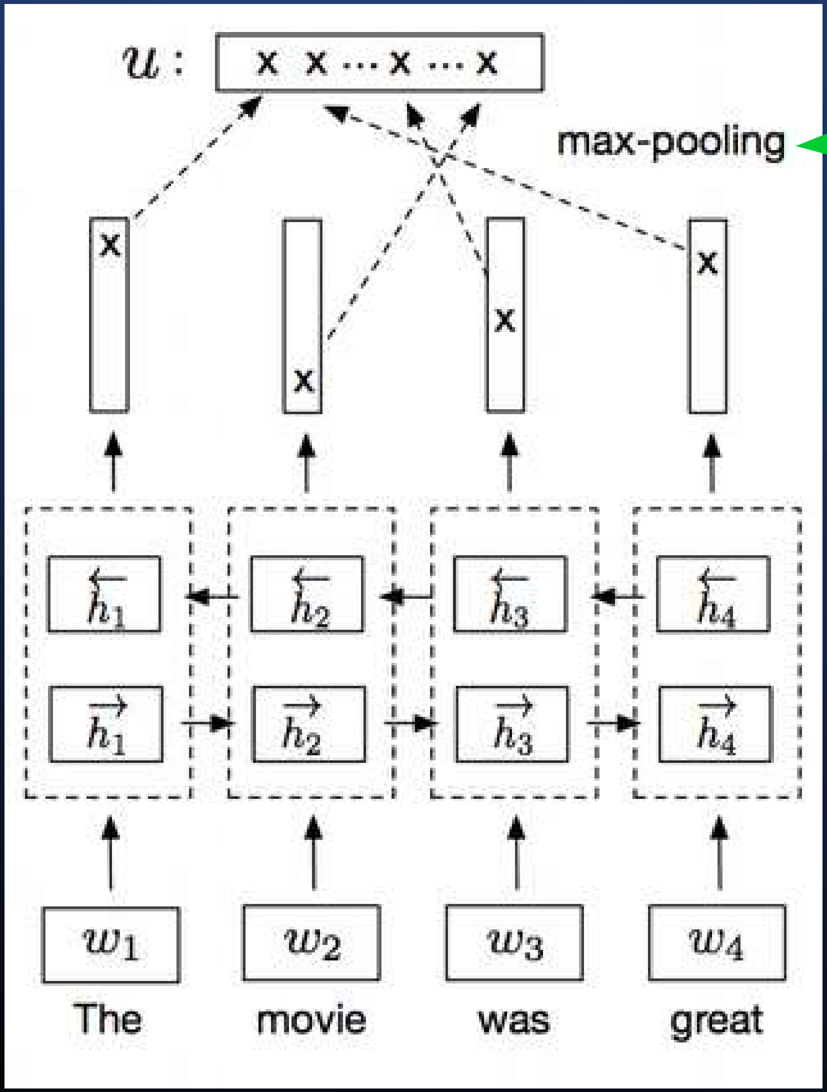
Supervision from Textual Entailment: InferSent



Supervision via SNLI

A. Conneau, D. Kiela, H. Schwenk, L. Barrault, A. Bordes, Supervised Learning of Universal Sentence Representations from Natural Language Inference Data

Supervision from Textual Entailment: InferSent



BiLSTM with dimension-wise Max-Pooling

Downside for non-English: NLI-style training data not readily available

A. Conneau, D. Kiela, H. Schwenk, L. Barrault, A. Bordes, Supervised Learning of Universal Sentence Representations from Natural Language Inference Data

Supervision from Semantic Similarity

- Do two sentences mean something similar?

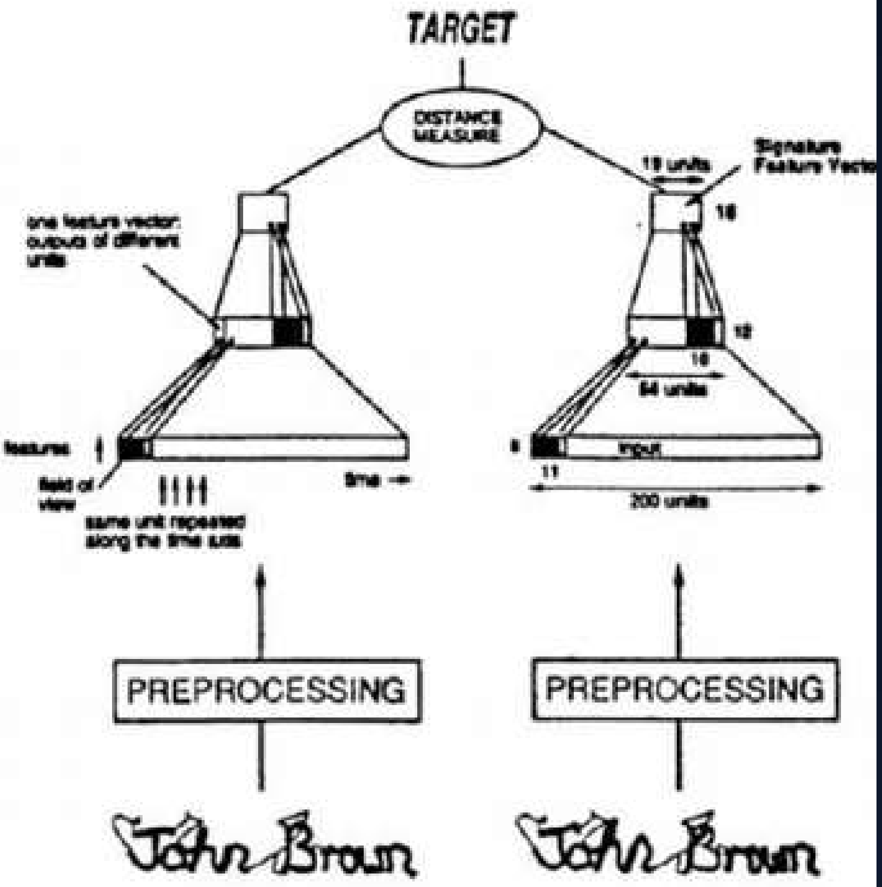
Relatedness score	Example
1.6	A: "*A man is jumping into an empty pool*" B: "*There is no biker jumping in the air*"
2.9	A: "*Two children are lying in the snow and are making snow angels*" B: "*Two angels are making snow on the lying children*"
3.6	A: "*The young boys are playing outdoors and the man is smiling nearby*" B: "*There is no boy playing outdoors and there is no man smiling*"
4.9	A: "*A person in a black jacket is doing tricks on a motorbike*" B: "*A man in a black jacket is doing tricks on a motorbike*"

- Like paraphrase identification, but with shades of gray.

Source: Graham Neubig

Supervision from Semantic Similarity: Siamese Networks

- Use the same network, compare the extracted representations
- (e.g. Time-delay networks for signature recognition)



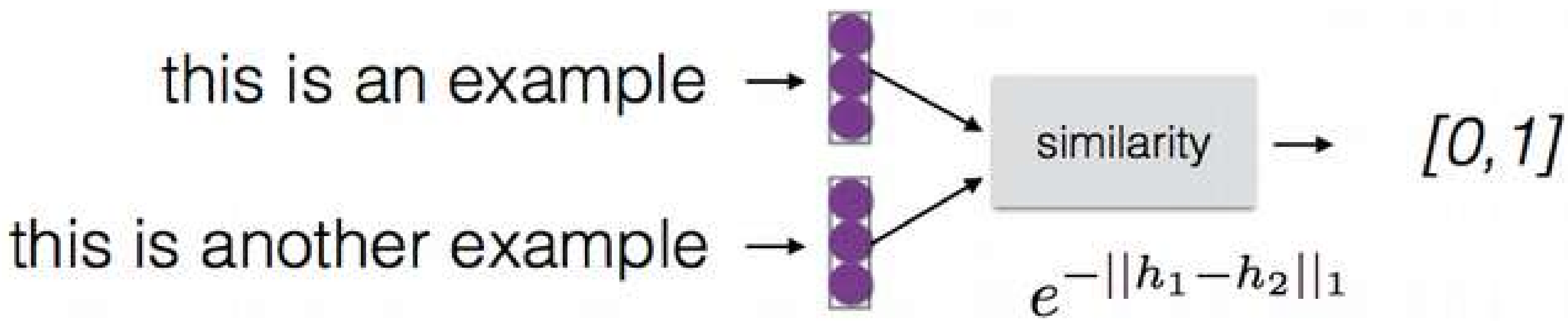
Bromley et al. 1993

Source: Graham Neubig

Supervision from Semantic Similarity: Siamese Networks

Supervision via Semantic Relatedness

- Use **siamese LSTM architecture** with e^-L1 as a similarity metric

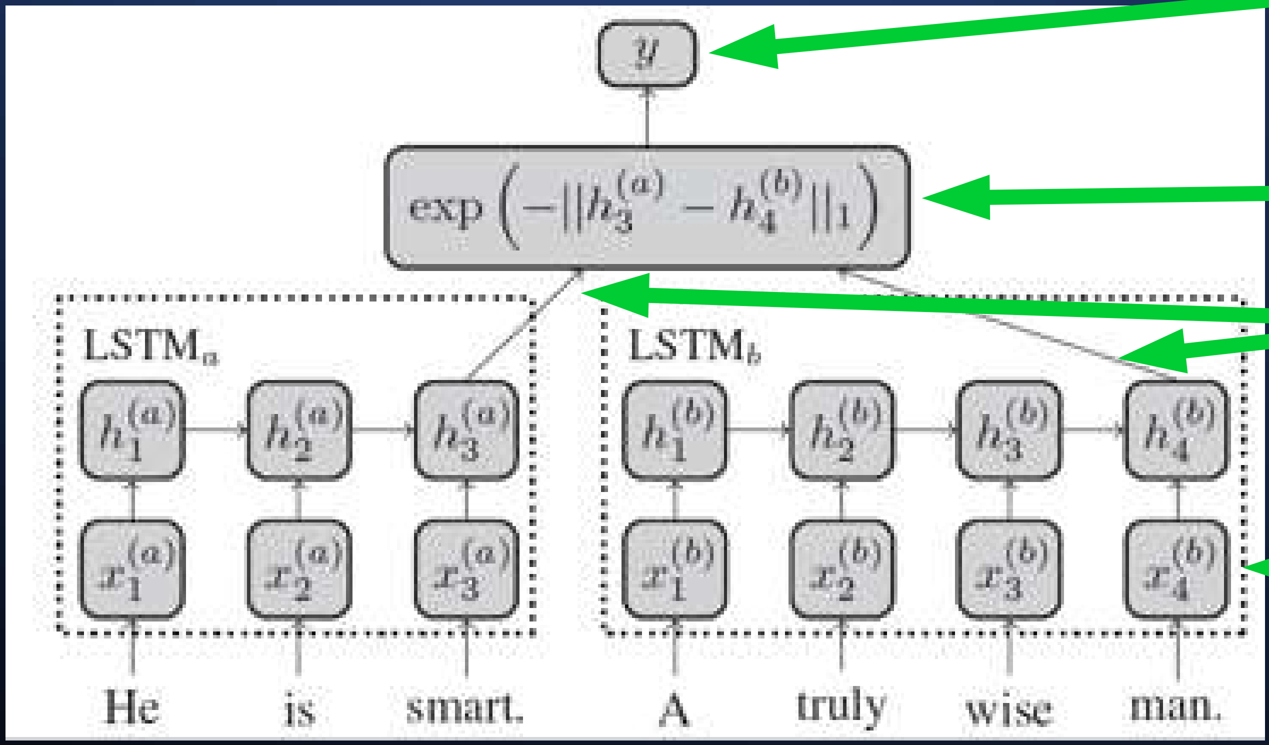


- **Simple model!** Good results due to engineering? Including pre-training, using pre-trained word embeddings, etc.

Jonas Mueller, Aditya Thyagarajan. Siamese Recurrent Architectures for Learning Sentence Similarity. AAAI 2016

Source: Graham Neubig

Supervision from Semantic Similarity: Siamese Networks



Train on SemEval data.
Augment by replacing random words with WordNet synonyms

Comparison via Manhattan distance (L1)

50-dim. final hidden state vectors

300-dim. word2vec embeddings as input

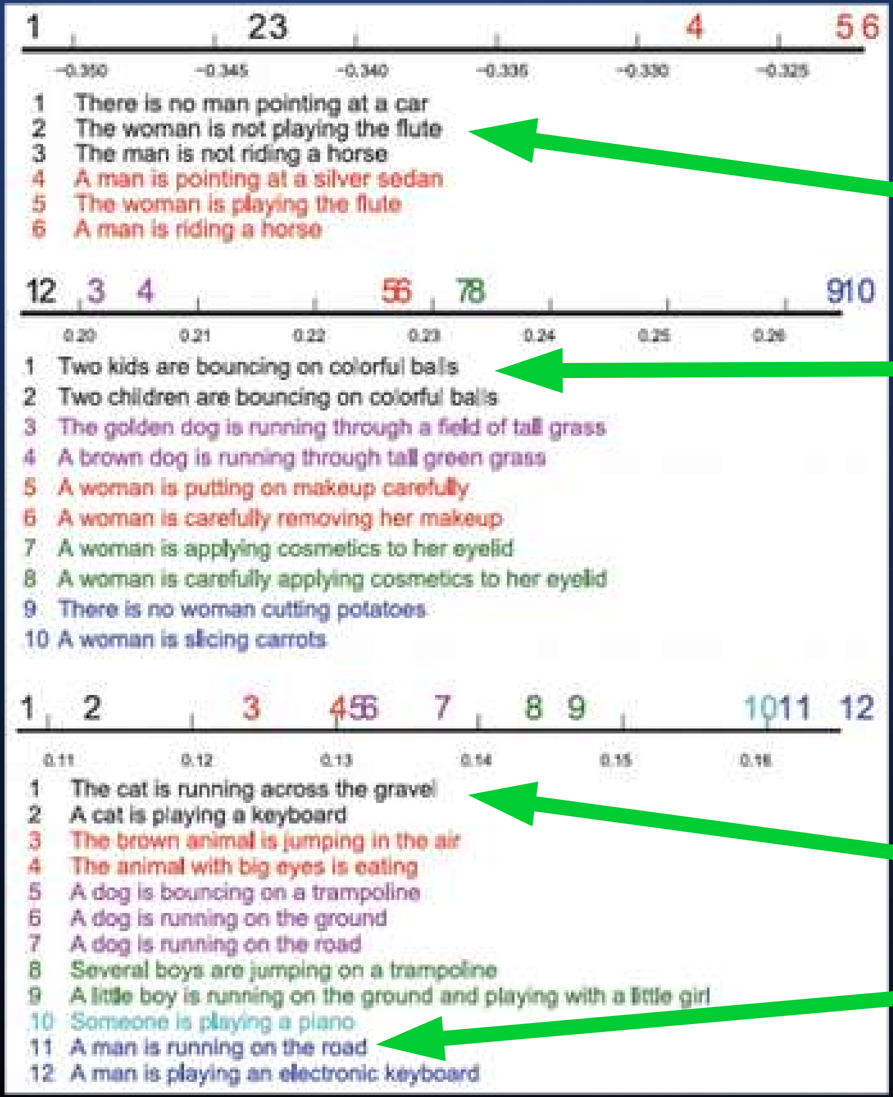
Jonas Mueller, Aditya Thyagarajan. Siamese Recurrent Architectures for Learning Sentence Similarity. AAAI 2016

Supervision from Semantic Similarity: Siamese Networks

Method	r	ρ	MSE
Illinois-LH (Lai and Hockenmaier 2014)	0.7993	0.7538	0.3692
UNAL-NLP (Jimenez et al. 2014)	0.8070	0.7489	0.3550
Meaning Factory (Bjerva et al. 2014)	0.8268	0.7721	0.3224
ECNU (Zhao, Zhu, and Lan 2014)	0.8414	–	–
Skip-thought+COCO (Kiros et al. 2015)	0.8655	0.7995	0.2561
Dependency Tree-LSTM (Tai, Socher, and Manning 2015)	0.8676	0.8083	0.2532
ConvNet (He, Gimpel, and Lin 2015)	0.8686	0.8047	0.2606
MaLSTM	**0.8822**	**0.8345**	**0.2286**

Jonas Mueller, Aditya Thyagarajan. Siamese Recurrent Architectures for Learning Sentence Similarity. AAAI 2016

Supervision from Semantic Similarity: Siamese Networks



3 specific hidden units

Negation vs. no negation

Kind of activity, irrespective of subject

Kind of subject, irrespective of activity

Jonas Mueller, Aditya Thyagarajan. Siamese Recurrent Architectures for Learning Sentence Similarity. AAAI 2016

Supervision from Parallel Corpora: Inducing Monolingual Paraphrases

	Sentence	P(R)
R:	Room was comfortable and the staff at the front desk were very helpful.	1.0
T:	The staff were very nice and the room was very nice and the staff were very nice.	<0.01
R:	The enchantment of your wedding day, captured in images by Flore-Ael Surun.	0.98
T:	The wedding of the wedding, put into images by Flore-Ael A.	<0.01
R:	Mexico and Sweden are longstanding supporters of the CTBT.	1.0
T:	Mexico and Sweden have been supporters of CTBT for a long time now.	0.06
R:	We thought Mr Haider ’ s Austria was endangering our freedom.	1.0
T:	We thought that our freedom was put at risk by Austria by Mr Haider.	0.09

R: Reference, T: Backtranslation

Use MT to translate aligned sentences back to English, as rephrasing of original English sentence

Wieting et al. Learning Paraphrastic Sentence Embeddings from Back-Translated Bitext. EMNLP 2017

Supervision from Parallel Corpora: Neural Machine Translation

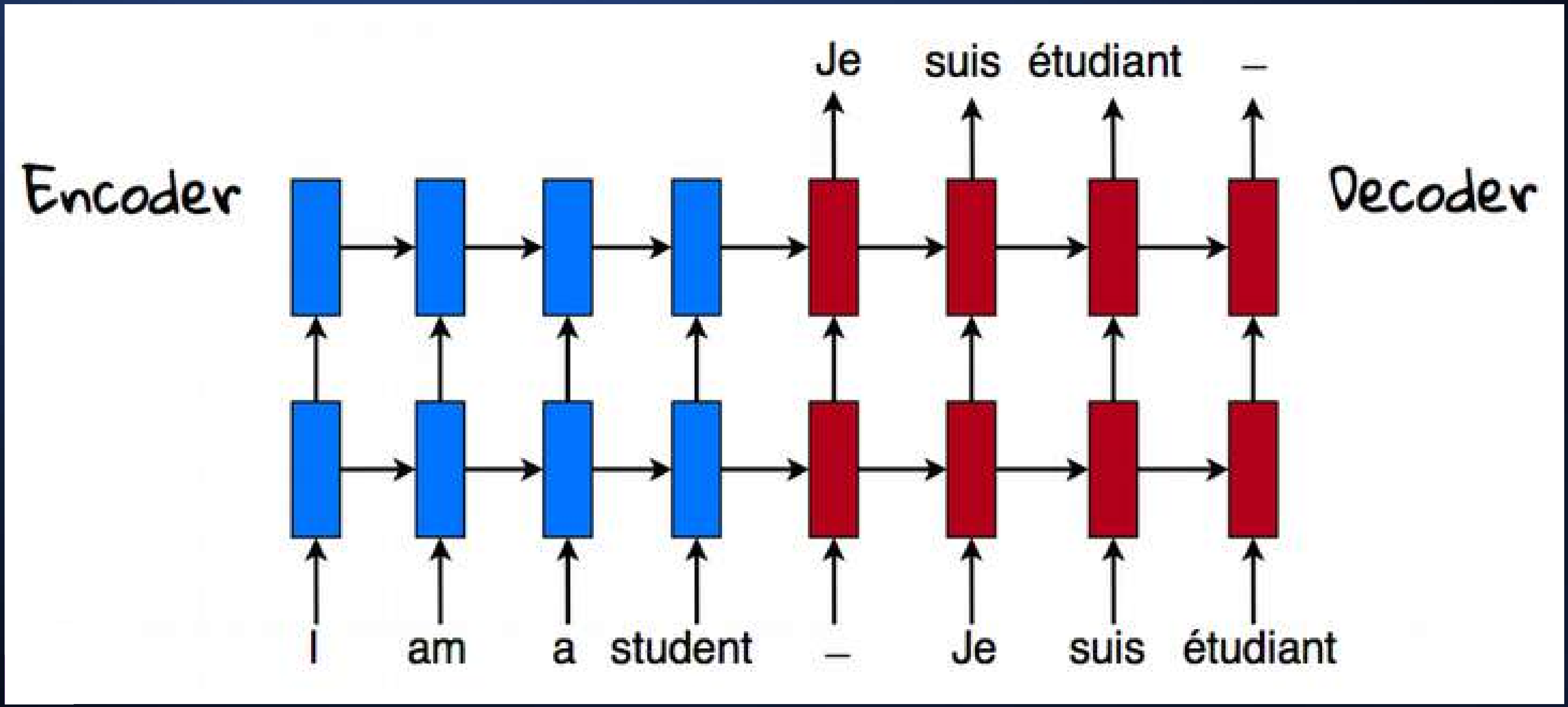
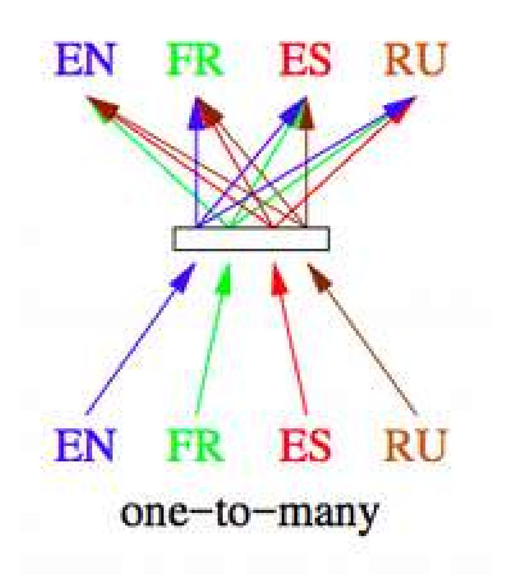


Image: Christopher Manning

Supervision from Parallel Corpora: NMT for Cross-Lingual Embeddings



One-to-many strategy

- Translate from one to all other language, source excluded

⇒ Always at least one common target language

- Sentence embeddings for all languages

– Needs N-way parallel training corpora

- Extension to *"many-to-many strategy"* straightforward

Holger Schwenk et al.

Supervision from Parallel Corpora: NMT vs. Sentence Representations

NMT

- BLSTM, the deeper the better
- Quite complicated architectures (short-cut connections)
- Convolutional networks

Sentence representations

- Deep networks doesn't seem to be useful
- Sentence representation:
  - last LSTM layer (original seq2seq)
  - BLSTM + element-wise max-pooling
- the proposed framework is generic:
  any type of encoder and decoder can be used

Holger Schwenk et al.

Supervision from Parallel Corpora: NMT for Cross-Lingual Embeddings

Training Strategies: One-to-One

System	Average Similarity Error			
	efs	efsr	efsra	efsraz
#pairs:	6	10	15	21
LSTM nhid=512 + last state:				
efs-a	2.14	–	–	–
efs-r	1.97	–	–	–
efsr-a	1.90	2.40	–	–
efsra-z	1.91	2.26	2.51	–
efsraz-all	1.70	1.97	2.38	2.59
LSTM nhid=1024 + last state:				
efsraz-all	1.36	1.64	1.89	1.95
BLSTM nhid=512 + max pooling:				
efsra-z	1.03	1.20	1.26	–
efsraz-all	0.92	1.07	1.15	1.20

- Error decreases with the number of languages covered
- Training strategy one-to-many is slightly better
- BLSTM + max pooling is considerably better

e=English, f=French, s=Spanish, r=Russian, a=Arabic, z=Chinese

Holger Schwenk et al.

Supervision from Parallel Corpora: NMT for Cross-Lingual Embeddings

Query:	**All kinds of obstacles must be eliminated.**
D_2=0.905	All kinds of barriers have to be removed.
D_3=0.682	All forms of violence must be prohibited.
D_4=0.673	All forms of provocation must be avoided.
D_5=0.636	All forms of social dumping must be stopped.
Query:	**I did not find out why.**
D_2=0.836	I do not understand why.
D_3=0.821	I fail to understand why.
D_4=0.786	I cannot understand why.
D_5=0.780	I have no idea why.

- Five closest sentences found by monolingual similarity search in English (D_1 = query, not shown)
- All are some of form para-phrasing $\rightarrow$ **linguistic** similarity

Holger Schwenk et al.

Supervision from Parallel Corpora: NMT for Cross-Lingual Embeddings

Monolingual Similarity Search: Examples

Query	**All citizens who commit sexual crimes against children must be punished, regardless of whether the crime is committed within or outside the EU.**
D_2=0.662	The second proposal is to protect children against child sex tourism by all member states criminalising sexual crimes both within and outside the EU.
D_3=0.655	We need standard national legislation throughout Europe which punishes union citizens who engage in child sex tourism, irrespective of where the offence was committed.
D_4=0.655	The impunity of those who commit terrible crimes against their own citizens and against other people regardless of their citizenship must be ended.
D_5=0.609	Any person who commits a criminal act should be punished, including those who employ the third-country nationals, illegally and under poor conditions.

- A more complicated English sentence (25 words)
- All closest sentences cover the punishment of (sexual) crimes.
- The similarity is at the overal sentence level not simple paraphrasing or synonymes

Holger Schwenk et al.

Supervision from Parallel Corpora: NMT for Cross-Lingual Embeddings

EN_{59177}	**Query**	**Allow me, however, to comment on certain issues raised by the honourable Members.**
FR_{59177}	$D_1=0.739$	Permettez-moi toutefois de commenter certaines questions soulevées par les députés.
FR_{394434}	$D_2=0.643$	Je voudrais commenter quelques-unes des questions soulevées par les députés.
FR_{791798}	$D_3=0.618$	Je voudrais faire les commentaires suivants sur plusieurs aspects spécifiques soulevés par certains orateurs.
FR_{666349}	$D_4=0.615$	Permettez-moi de dire quelques mots sur certaines questions qui ont été soulevées.
FR_{444790}	$D_5=0.609$	Je voudrais juste faire quelques commentaires sur certaines des questions qui ont été soulevées.
ES_{59177}	$D_1=0.719$	No obstante, permítanme comentar ciertas cuestiones planteadas por sus señorías.
ES_{394434}	$D_2=0.628$	Me gustaría comentar algunas de las cuestiones planteadas por algunos diputados.
ES_{271614}	$D_3=0.615$	No obstante, quisiera hacer algunos comentarios sobre el debate que nos ocupa.
ES_{661451}	$D_4=0.605$	Por ultimo, permítanme que añada algunos comentarios sobre las enmiendas presentadas.
ES_{666285}	$D_5=0.605$	No obstante, permítanme que conteste a algunos comentarios que se han realizado.

- All the cosine distances are close and the sentences are indeed semantically related.

Holger Schwenk et al.

Supervision from Parallel Corpora: NMT for Cross-Lingual Embeddings

EN_{77622}	**Query**	**And yet the report on the fight against racism does not demonstrate that the necessary conclusions have been drawn.**
FR_{77622}	D_1=0.767	Pourtant, le rapport sur la lutte contre le racisme n'indique pas que l'on en ait tiré les conclusions qui s'imposent.
$FR_{1094939}$	D_2=0.746	Ainsi, le rapport sur la lutte contre le racisme n'indique pas que l'on en a tiré les conclusions qui s'imposent.
FR_{73928}	D_3=0.491	Et, comme le démontrent les faits, ce n'est pas en interdisant que l'on va obtenir des résultats.
$FR_{1249269}$	D_4=0.476	Ce rapport, qui se propose de lutter contre la corruption, ne fait qu'illustrer votre incapacité à le faire.
ES_{77622}	D_1=0.820	Sin embargo, el informe sobre la lucha contra el racismo no muestra que se hayan extraído las conclusiones necesarias.
$ES_{1094939}$	D_2=0.797	Así, el informe sobre la lucha contra el racismo no muestra que se hayan extraído las conclusiones necesarias.
ES_{287052}	D_3=0.517	No obstante, el informe deja mucho que desear en lo que se refiere a las medidas necesarias para combatir el cambio climático y, por tanto, pone de relieve que el parlamento europeo no se encuentra a la vanguardia de esta batalla.
ES_{74892}	D_4=0.515	Y el informe de los expertos demuestra que no había el control y el seguimiento necesarios.

- Correct French and Spanish translation were retrieved
- Second closest sentences are also semantically well related to the query
- Other have smaller distance and only cover some aspect of the query

Holger Schwenk et al.

Supervision from Multiple Tasks

Kitchen Sink Approach:

Learn from all kinds of tasks

Multi-Task Learning Approach by MILA/MSR Montreal

1. Skip Thoughts
2. NLI
3. Neural Machine Translation
4. Syntactic Constituency Parsing

Including weakly labeled data output by existing parser

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, Christopher Pal. Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning. ICLR 2018

Supervision from Multiple Tasks

Kitchen Sink Approach:

Learn from all kinds of tasks

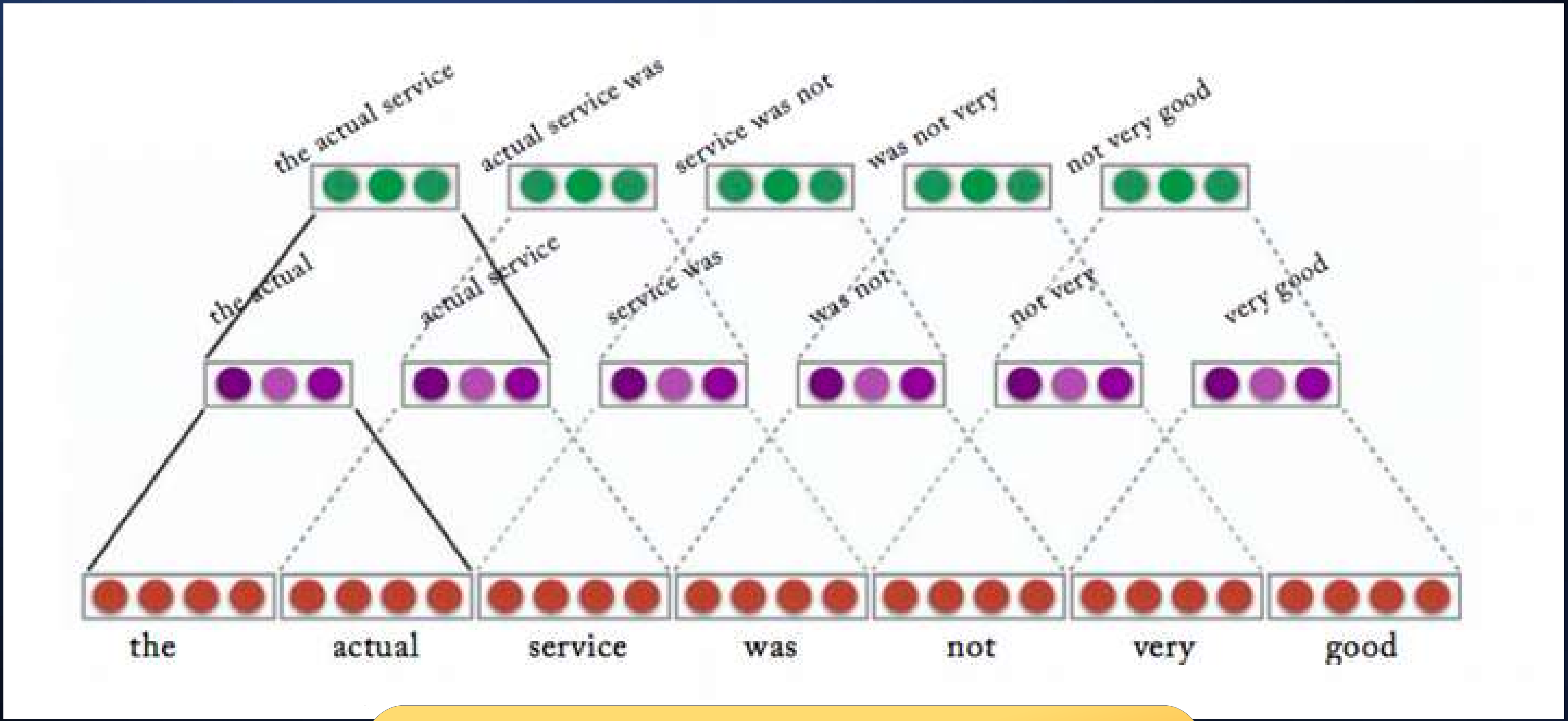
Multi-Task Learning Approach
by MILA/MSR Montreal

Task	Sentence Pairs
En-Fr (WMT14)	40M
En-De (WMT15)	5M
Skipthought (BookCorpus)	74M
AllNLI (SNLI + MultiNLI)	1M
Parsing (PTB + 1-billion word)	4M
Total	124M

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, Christopher Pal. Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning. ICLR 2018

Word Vector-Based Approaches

Sentence Representations

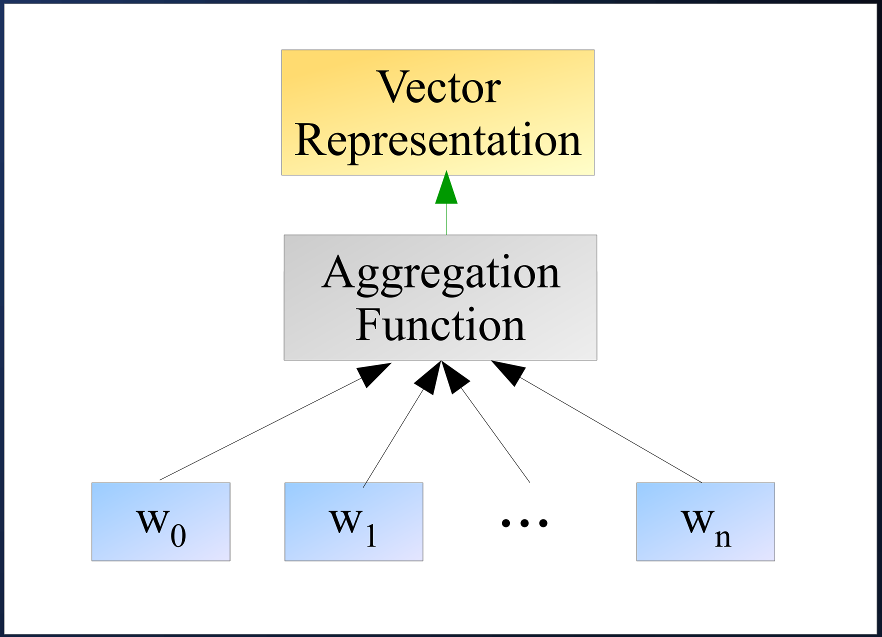


Compose in multiple levels?

Image: Yoav Goldberg

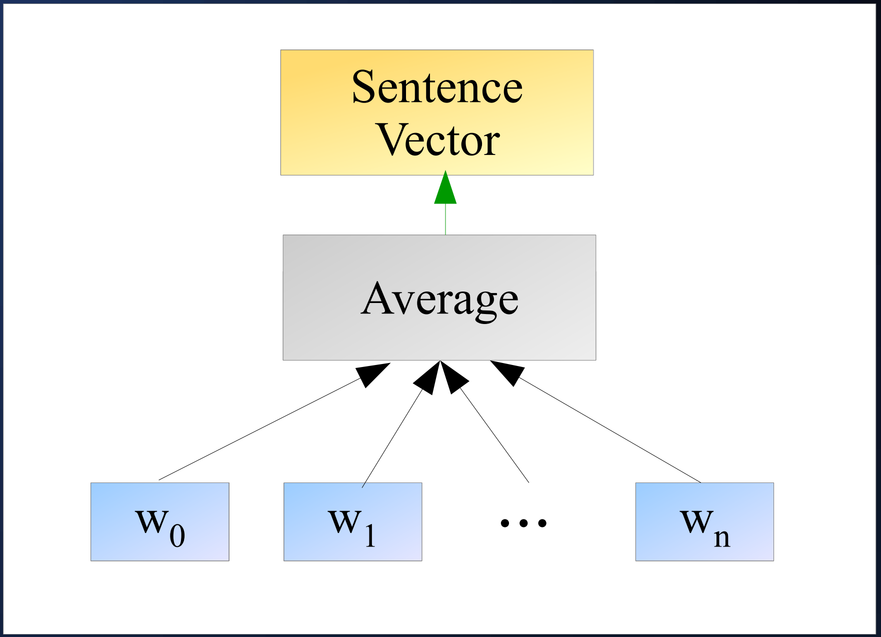
Word Vector Aggregation

Directly aggregate vector for entire sentence in one step.



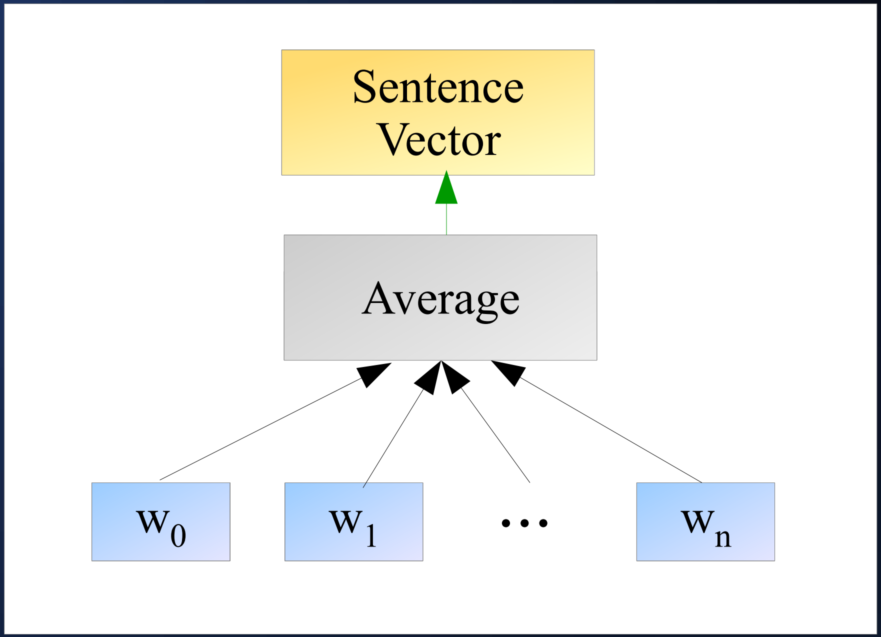
Word Vector Averaging

$$\vec{v}_S = \frac{1}{|S|} \sum_{w \in S} \vec{v}_w$$



Word Vector Averaging

$$\vec{v}_S = \frac{1}{|S|} \sum_{w \in S} \vec{v}_w$$

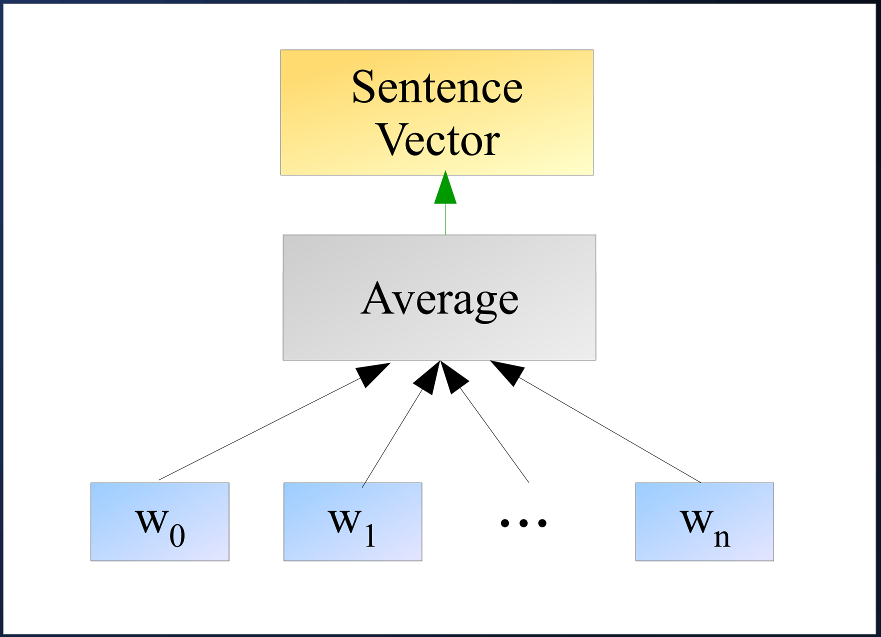


If vectors are first preprocessed via supervision (PPDB paraphrases), then averaging outperforms LSTM's final hidden state

John Wieting, Mohit Bansal, Kevin Gimpel & Karen Livescu. Towards Universal Paraphrastic Sentence Embeddings. ICLR 2016

Word Vector Averaging

$$\vec{v}_S = \frac{1}{|S|} \sum_{w \in S} \vec{v}_w$$

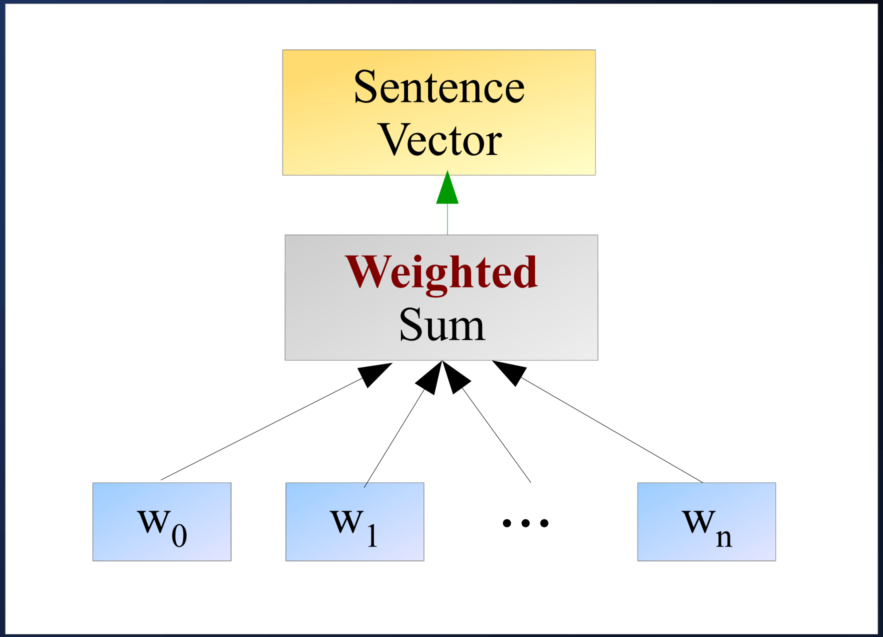


They later found that LSTMs do better when averaging hidden states, adding better supervised data (Simple English Wikipedia), and applying various other small tricks (regularization / preinitialization)

John Wieting, Kevin Gimpel. Revisiting Recurrent Networks for Paraphrastic Sentence Embeddings. ACL 2017

Creating Sentence and Document Vectors

$$\vec{v}_S = \frac{1}{|S|} \sum_{w \in S} \alpha_{S,w} \, \vec{v}_w$$



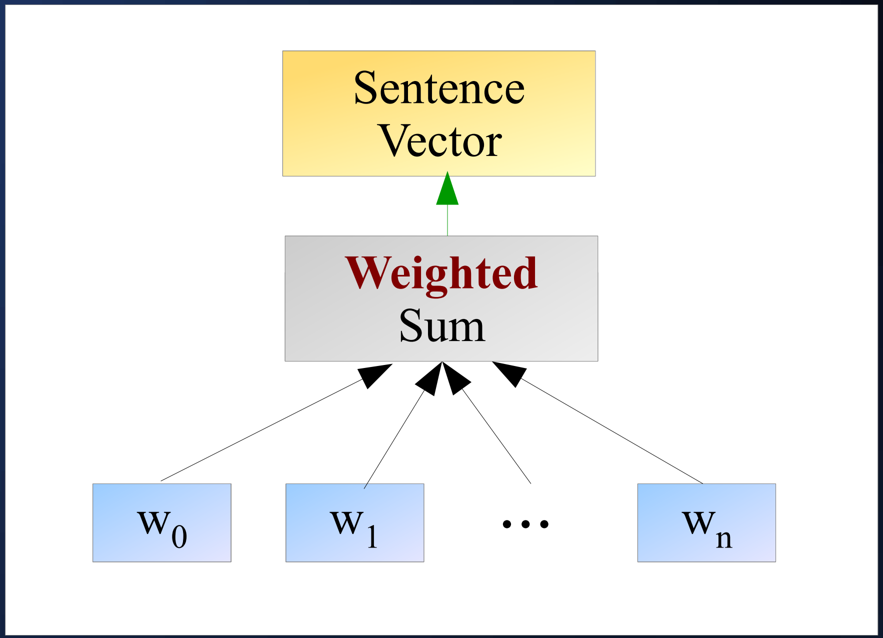
Additional weights

E.g. 0 for stop words
IDF

Word Vector Averaging: Arora et al.

$$\vec{v}_S = \frac{1}{|S|} \sum_{w \in S} \alpha_{S,w} \, \vec{v}_w$$

Smoothed inverse frequency, similar to IDF but with some extra smoothing



for all sentence s **in** S **do**

$$v_s \leftarrow v_s - uu^\top v_s$$

Remove "common component": u is 1st singular value of a matrix that contains all sentence vectors in its columns

Sanjeev Arora, Yingyu Liang, Tengyu Ma. A Simple but tough-to-beat Baseline for Sentence Embeddings. ICLR 2017

Word Vector Averaging: Arora et al.

Results on Semantic Textual Similarity

	Results collected from (Wieting et al., 2016) except tfidf-GloVe											Our approach	
Supervised or not	Su.							Un.			Se.	Un.	Se.
Tasks	PP	PP -proj.	DAN	RNN	iRNN	LSTM (no)	LSTM (o.g.)	ST	avg-GloVe	tfidf-GloVe	avg-PSL	GloVe +WR	PSL +WR
STS'12	58.7	**60.0**	56.0	48.1	58.4	51.0	46.4	30.8	52.5	58.7	52.8	56.2	59.5
STS'13	55.8	56.8	54.2	44.7	56.7	45.2	41.5	24.8	42.3	52.1	46.4	56.6	**61.8**
STS'14	70.9	71.3	69.5	57.7	70.9	59.8	51.5	31.4	54.2	63.8	59.5	68.5	**73.5**
STS'15	75.8	74.8	72.7	57.2	75.6	63.9	56.0	31.0	52.7	60.6	60.0	71.7	**76.3**
SICK'14	71.6	71.6	70.7	61.2	71.2	63.9	59.0	49.8	65.9	69.4	66.4	72.2	**72.9**
Twitter'15	52.9	52.8	**53.7**	45.1	52.9	47.6	36.1	24.7	30.3	33.8	36.3	48.0	49.0

Sanjeev Arora, Yingyu Liang, Tengyu Ma. A Simple but tough-to-beat Baseline for Sentence Embeddings. ICLR 2017

Word Vector Averaging: Arora et al.

Results on Sentence Classification

	PP	DAN	RNN	LSTM (no)	LSTM (o.g.)	skip-thought	Ours
similarity (SICK)	84.9	85.96	73.13	85.45	83.41	85.8	**86.03**
entailment (SICK)	83.1	84.5	76.4	83.2	82.0	-	**84.6**
sentiment (SST)	79.4	83.4	86.5	86.6	**89.2**	-	82.2

Word Vector Averaging with PPDB weighting (Wieting et al. 2016)

Deep Averaging Networks (Iyyer et al. 2015)

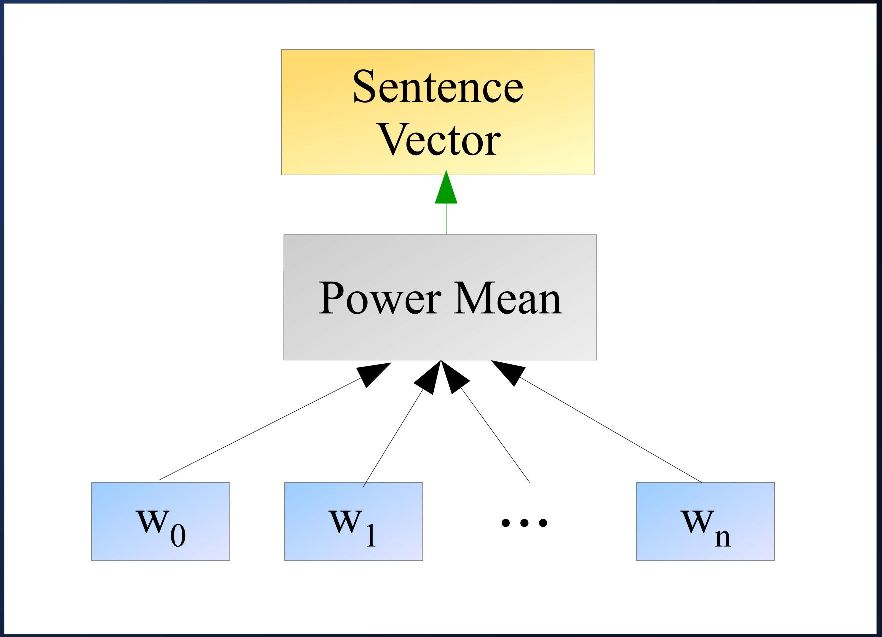
LSTM with output gates

Word Vector Averaging with weights and postprocessing

Sanjeev Arora, Yingyu Liang, Tengyu Ma. A Simple but tough-to-beat Baseline for Sentence Embeddings. ICLR 2017

k Power Means

$$v_i = \left(\frac{1}{|S|} \sum_{w \in S} x(w)_i^p \right)^{\frac{1}{p}}$$



for each dimension

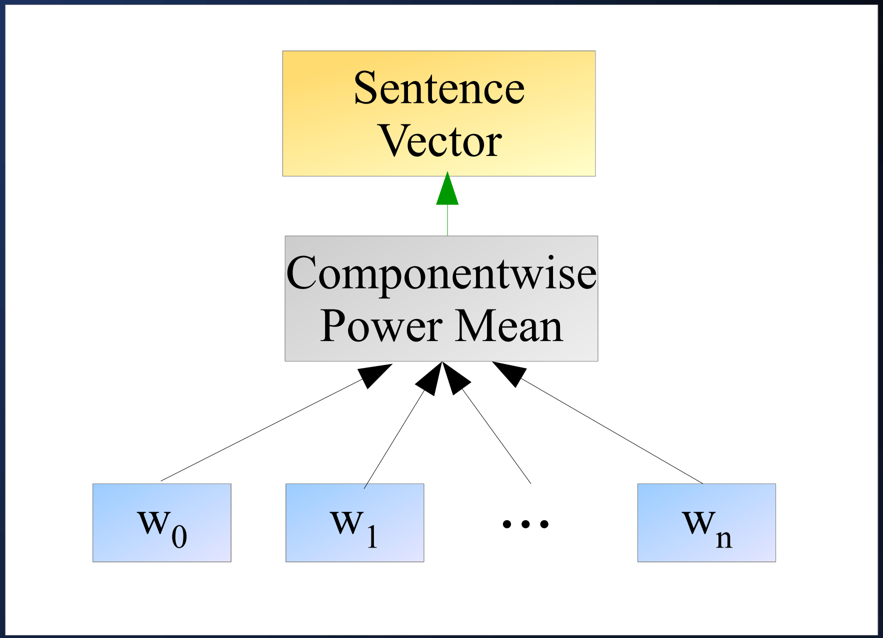
Component-wise power mean for different p

p = 1: Arithmetic Mean
p = –∞: Min.
p = ∞: Max.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, Iryna Gurevych. Concatenated Power Mean Word Embeddings as Universal Cross-Lingual Sentence Representations. https://arxiv.org/abs/1803.01400

k Power Means

$$v_i = \left(\frac{1}{|S|}\sum_{w\in S} x(w)_i^p\right)^{\frac{1}{p}}$$



for each dimension

Component-wise power mean for different p

Finally concatenate different versions

$p = 1$:	Arithmetic Mean
$p = -\infty$:	Min.
$p = \infty$:	Max.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, Iryna Gurevych. Concatenated Power Mean Word Embeddings as Universal Cross-Lingual Sentence Representations. https://arxiv.org/abs/1803.01400

Word Vector Averaging

```
val sentences = spark.createDataFrame(Seq(
  ("en", tokenize("en", "There are at least ten sparrows in the backyard.")),
  ("de", tokenize("de", "Im Garten sind mindestens zehn Sperlinge.")),
  ("fr", tokenize("fr", "Il y a au moins dix moineaux dans le jardin.")),
  ("en", tokenize("en", "It is an arid region, almost a desert.")),
  ("he", tokenize("he", "זה איזור צחיח, כמעט מדברי.")),
  ("ru", tokenize("ru", "Колодец высох.")), // The well ran dry
  ("zh", tokenize("zh", "這口井乾涸了。")),
  ("es", tokenize("es", "El Desierto de Atacama es el más árido del planeta.")),
  ("nl", tokenize("nl", "De Atacama is de droogste woestijn ter wereld."))
)).toDF("language", "text")

val sentenceVectors = sentences.select(
                          $"*",
                          wordVectorUDF($"language", $"text").alias("vector"))
```

G. de Melo. Inducing Conceptual Embedding Spaces from Wikipedia. Proc. WWW

Contextual Word Vectors

Key Goal:
Instead of using the original sequence of word embeddings, perform quick on-the-fly adaptation considering the local context

Contextual Word Vectors

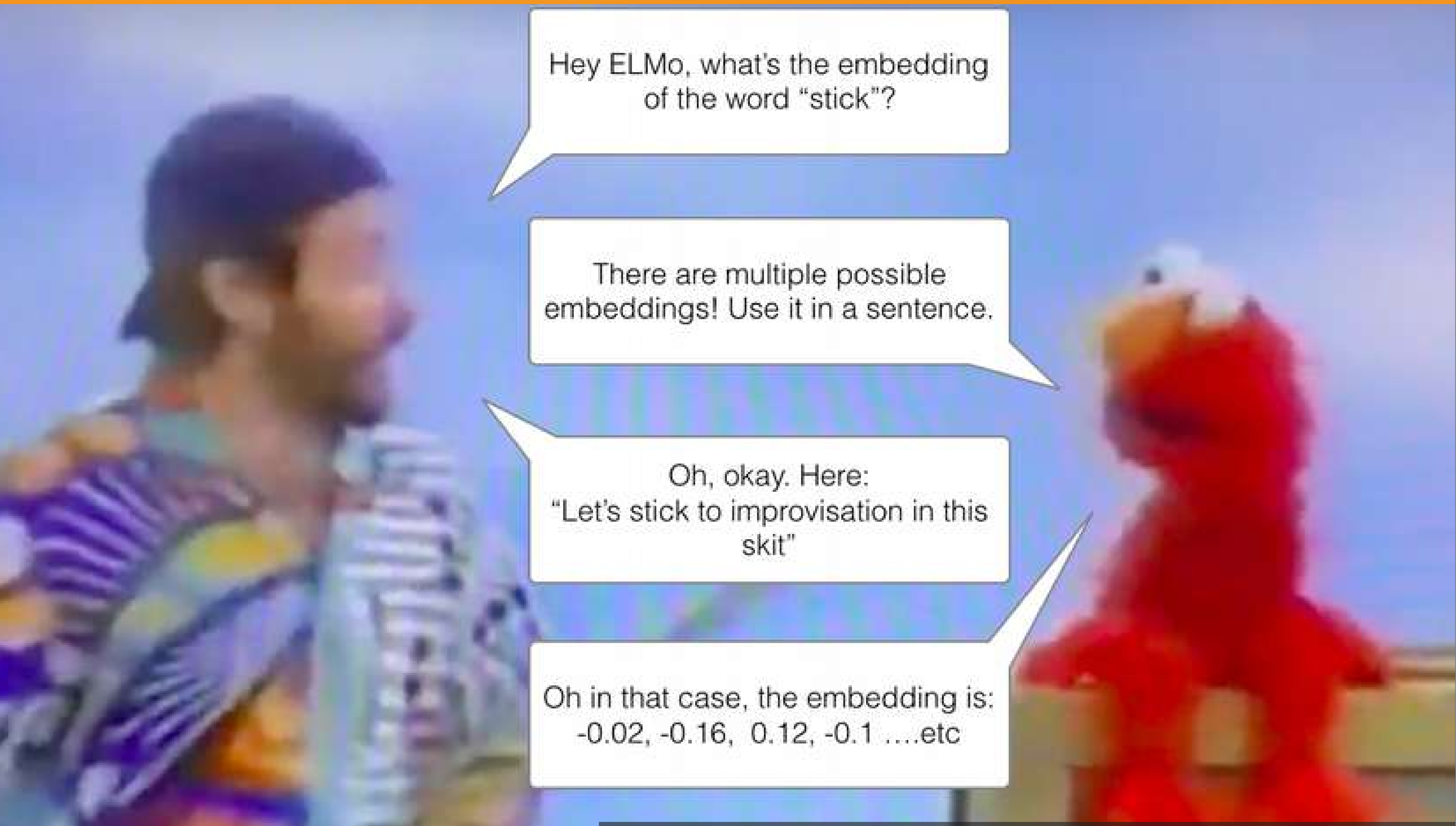


Image: Jay Alammar. http://jalammar.github.io/illustrated-bert/

Contextual Word Vectors

Key Goal:
Instead of using the original sequence of word embeddings, perform quick on-the-fly adaptation considering the local context

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer

Peters, Neumann, Iyyer, Gardner, Clark, Lee, Zettlemoyer. Deep contextualized word representations. ACL 2018

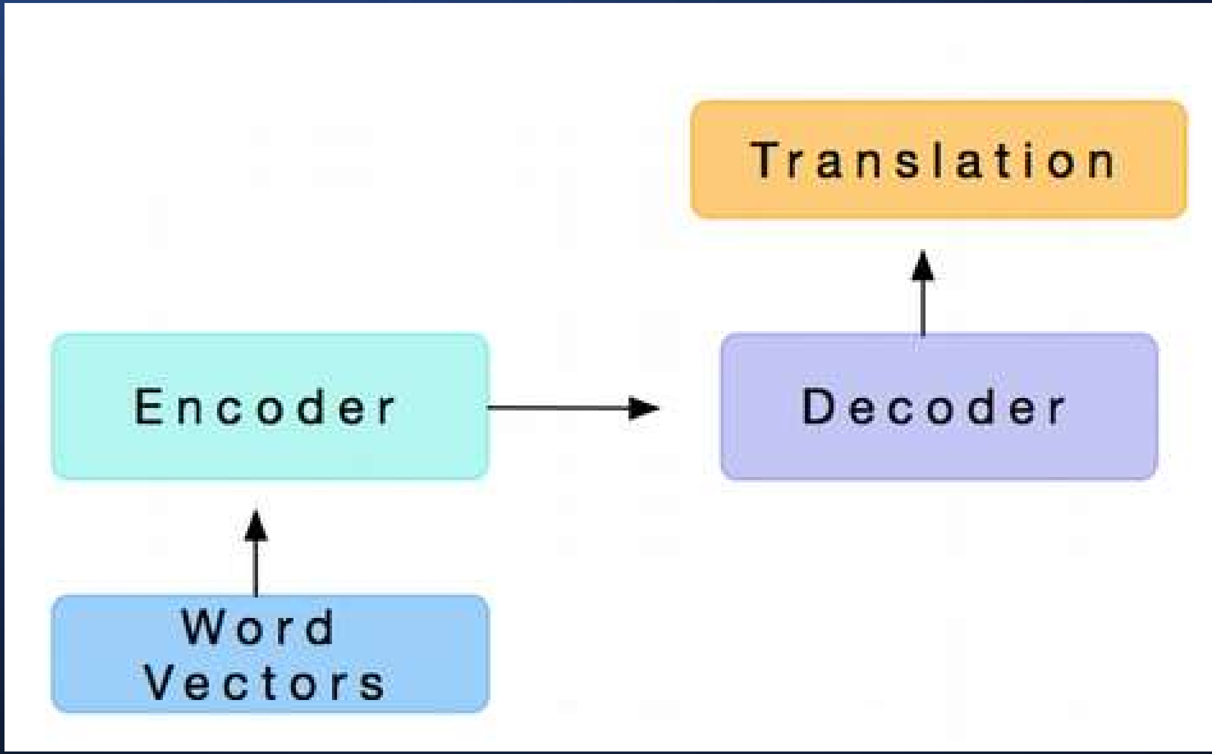
Contextual Word Vectors

Key Goal:
Instead of using the original sequence of word embeddings, perform quick on-the-fly adaptation considering the local context

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
ELMo	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

Peters, Neumann, Iyyer, Gardner, Clark, Lee, Zettlemoyer. Deep contextualized word representations. NAACL 2018

Contextual Word Vectors: Using NMT (COVE)



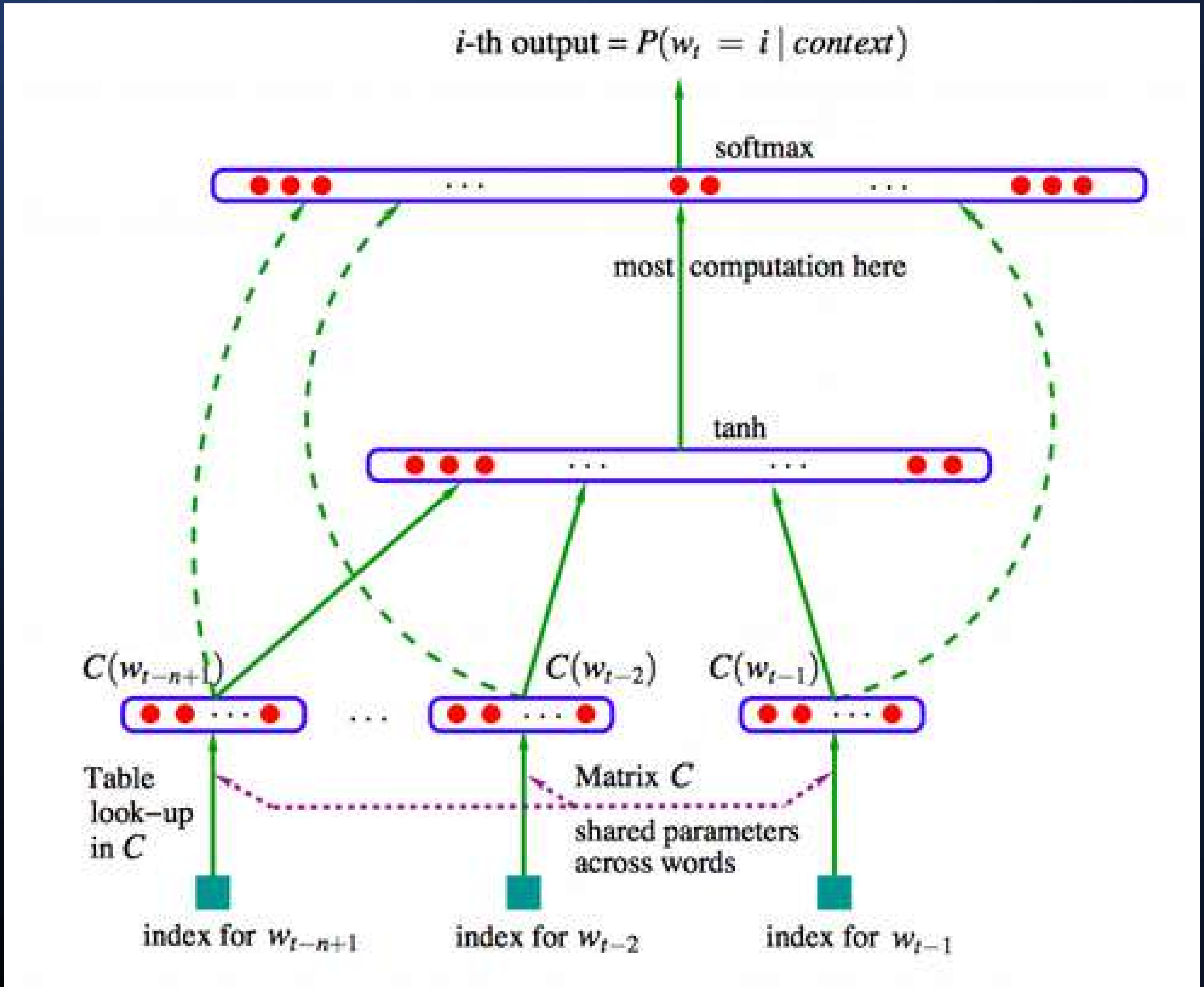
Train NMT model (2-layer Bi-LSTM). Then re-use encoder to obtain encoding of sentence for other downstream tasks (concatenate with regular GloVe vectors)

https://github.com/salesforce/cove

B. McCann, J. Bradbury, C. Xiong, R. Socher, Learned in Translation: Contextualized Word Vectors. NIPS 2017

Reminder: word2vec as Simplified Neural Language Model

Bengio et al. (2003). A Neural Probabilistic Language Model



Contextual Word Vectors: Using Language Modeling (ELMo)

ELMo: Embeddings from Language Models

Peters, Neumann, Iyyer, Gardner, Clark, Lee, Zettlemoyer. Deep contextualized word representations. NAACL 2018

Image: Tiffany Terry

Contextual Word Vectors: Using Language Modeling (ELMo)

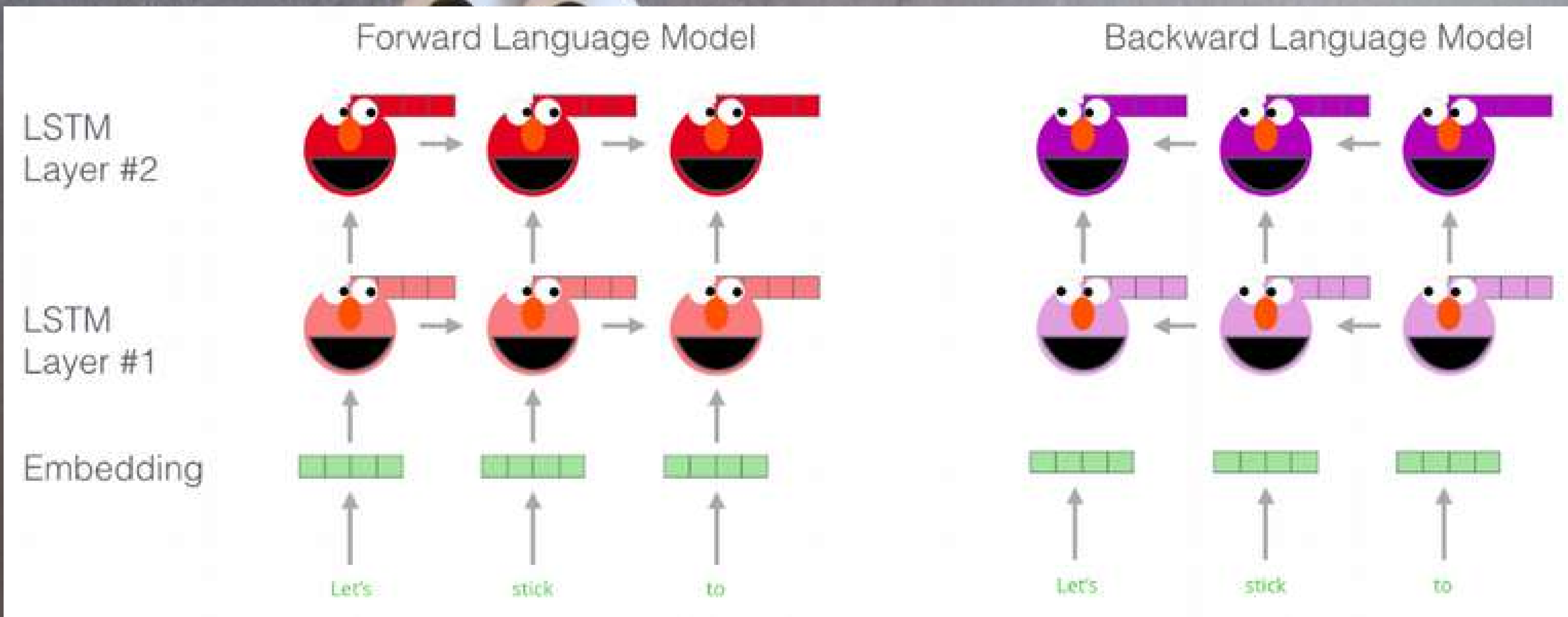
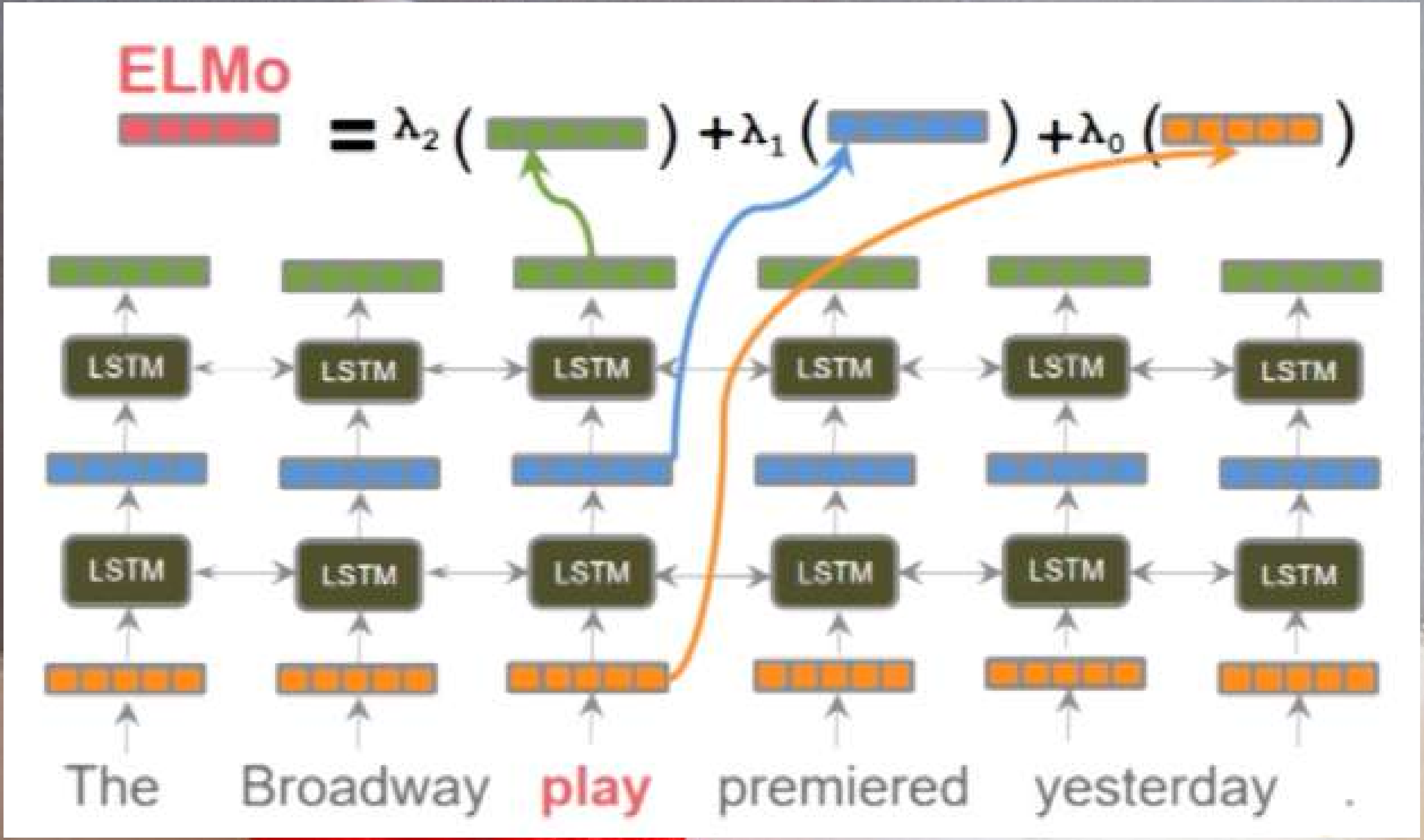


Image: Jay Alammar. http://jalammar.github.io/illustrated-bert/

Peters, Neumann, Iyyer, Gardner, Clark, Lee, Zettlemoyer. Deep contextualized word representations. NAACL 2018

Image: Tiffany Terry

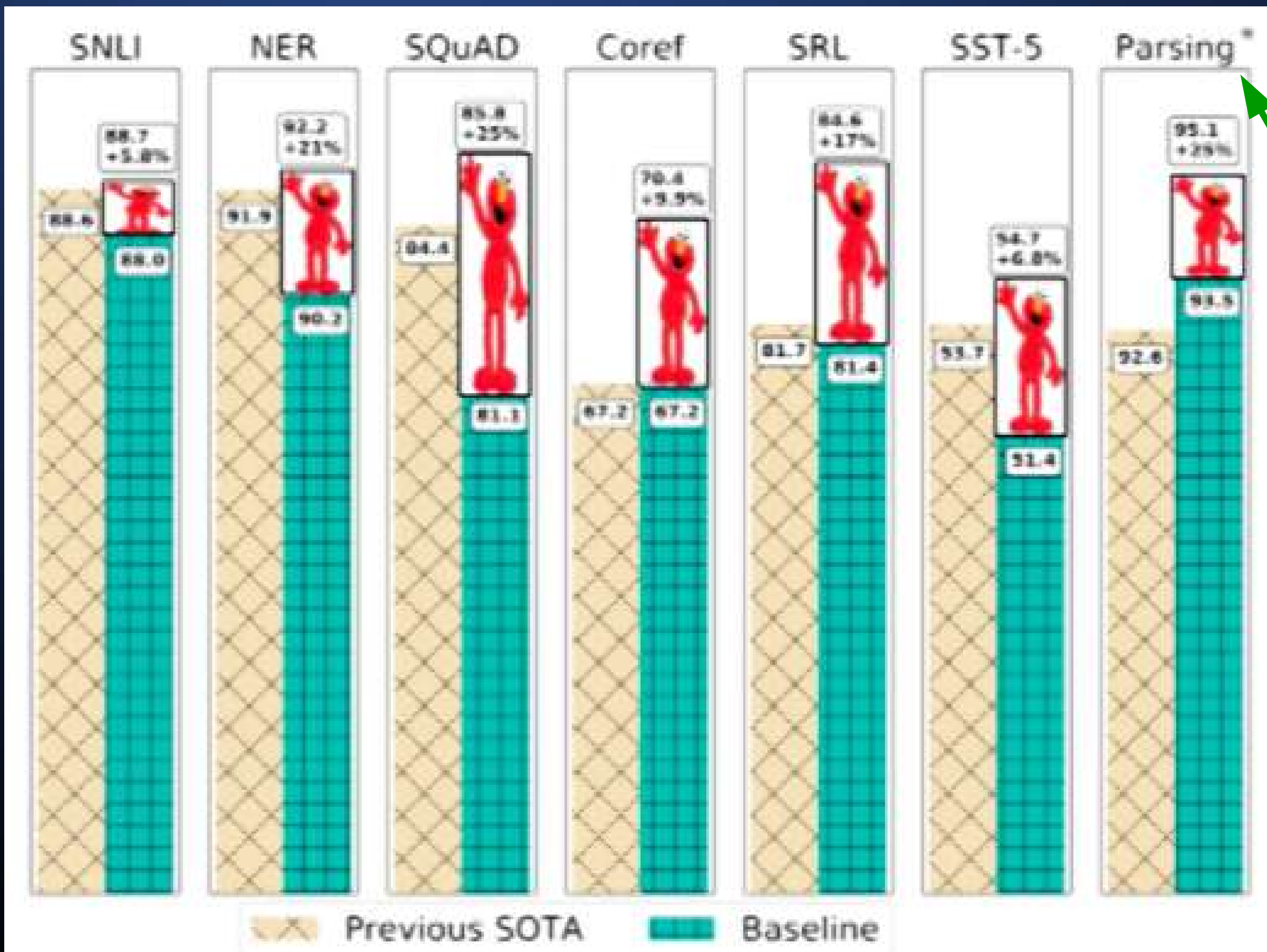
Contextual Word Vectors: Using Language Modeling (ELMo)



Peters, Neumann, Iyyer, Gardner, Clark, Lee, Zettlemoyer. Deep contextualized word representations. NAACL 2018

Image: Tiffany Terry

Contextual Word Vectors: Using Language Modeling (ELMo)



Parsing results from Kitaev & Klein. ACL 2018

Model/Code: http://allennlp.org/elmo

Peters, Neumann, Iyyer, Gardner, Clark, Lee, Zettlemoyer. Deep contextualized word representations. NAACL 2018

Contextual Word Vectors: Using Cloze Task (BERT)

Image: Sesame Street

Contextual Word Vectors: Using Cloze Task (BERT)

Image: Jay Alammar. http://jalammar.github.io/illustrated-bert/

Image: Sesame Street

Contextual Word Vectors: Using Cloze Task (BERT)

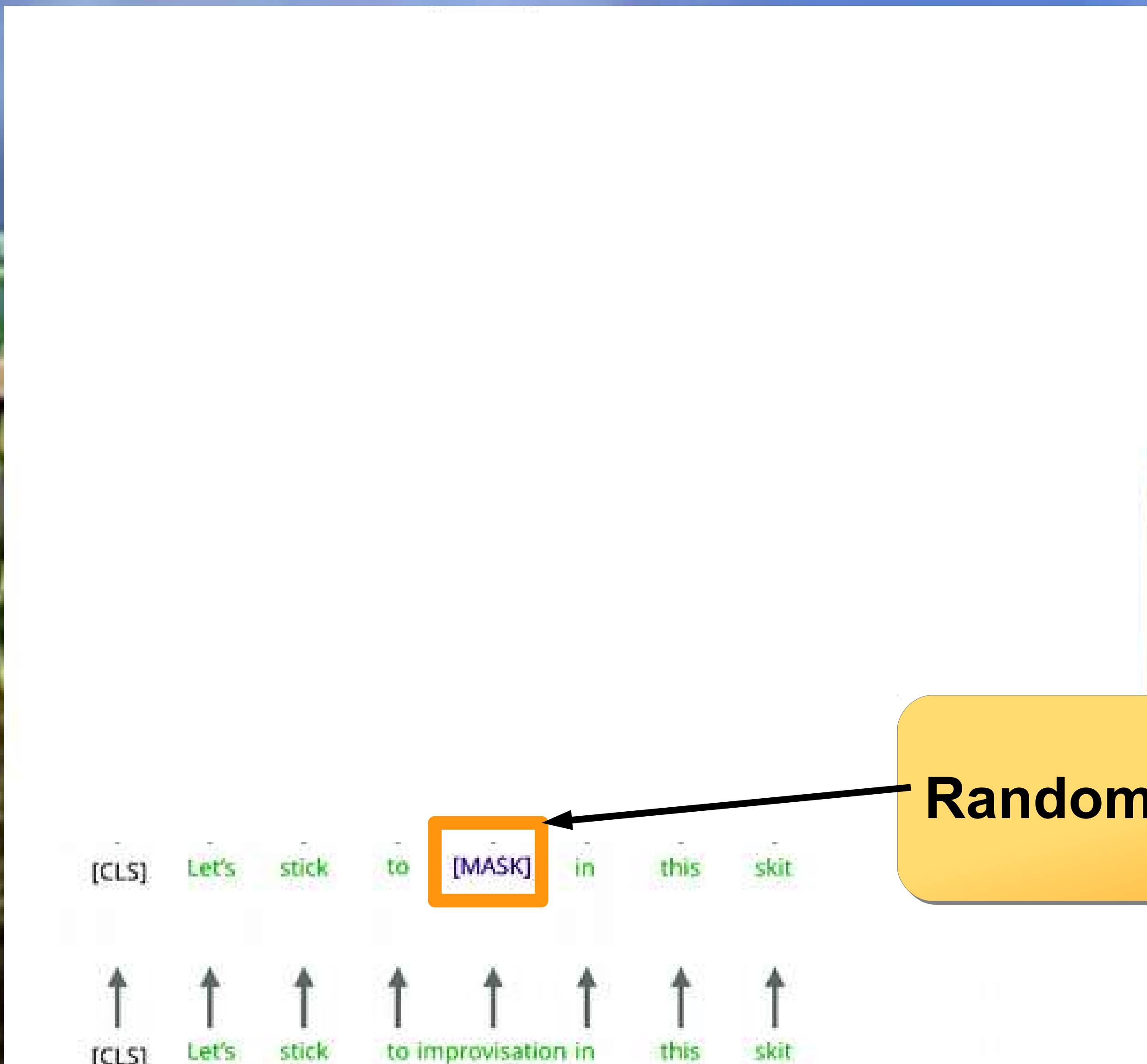


Image: Jay Alammar. http://jalammar.github.io/illustrated-bert/

Randomly hide 15% of tokens

Image: Sesame Street

Contextual Word Vectors: Using Cloze Task (BERT)

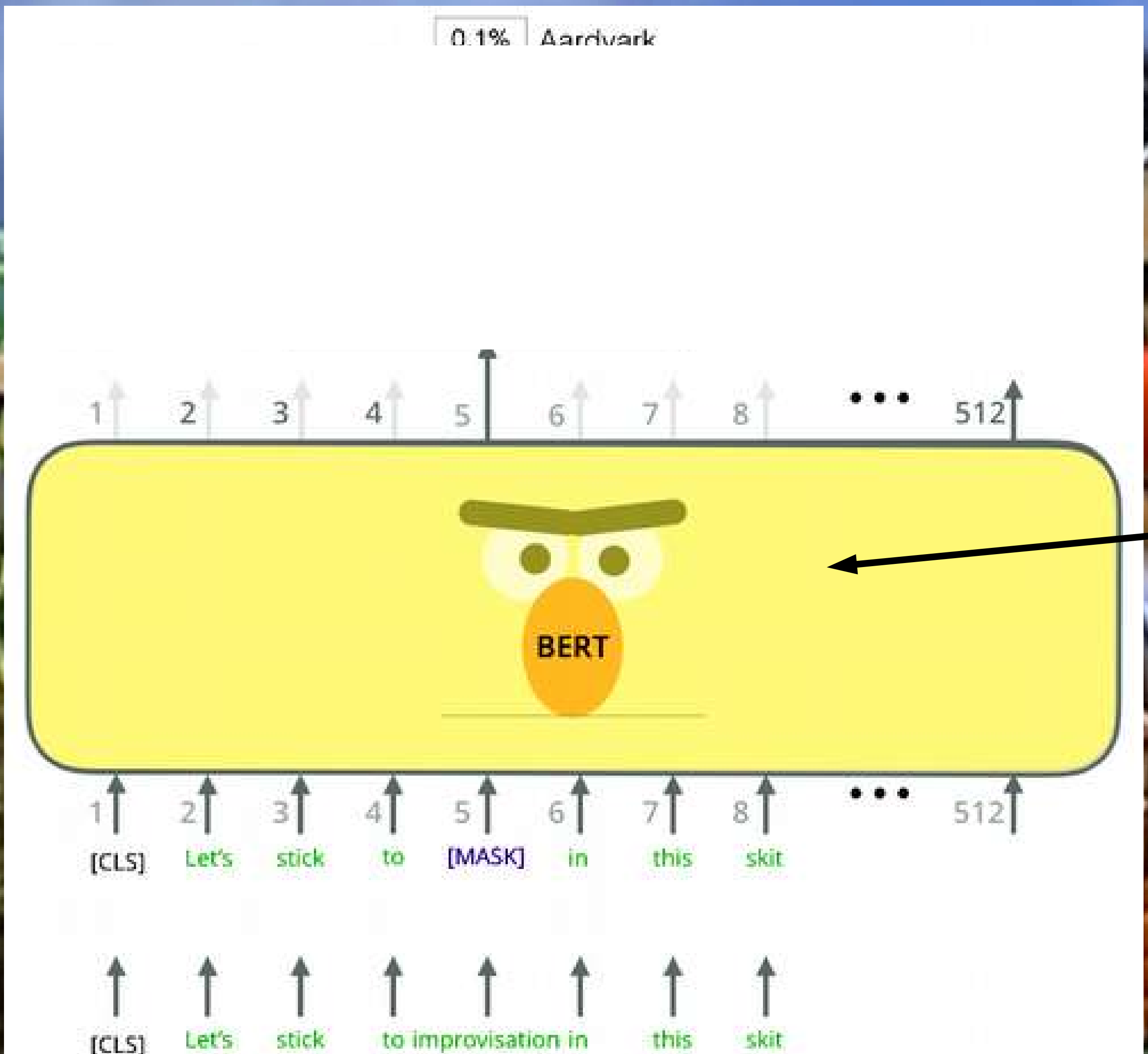


Image: Jay Alammar. http://jalammar.github.io/illustrated-bert/

Transformer model: Self-attention applied repeatedly

Image: Sesame Street

Contextual Word Vectors: Using Cloze Task (BERT)

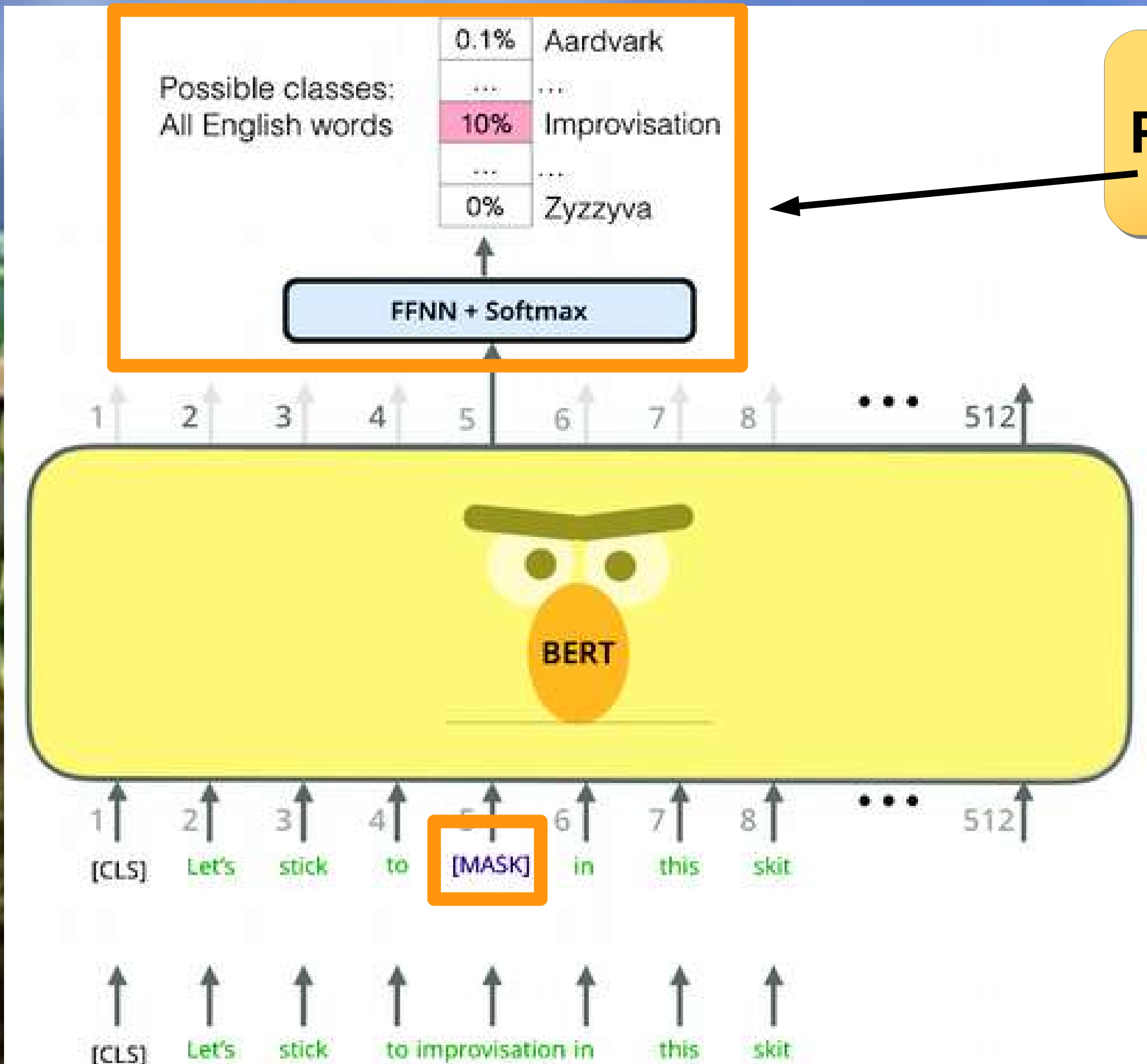


Image: Jay Alammar. http://jalammar.github.io/illustrated-bert/

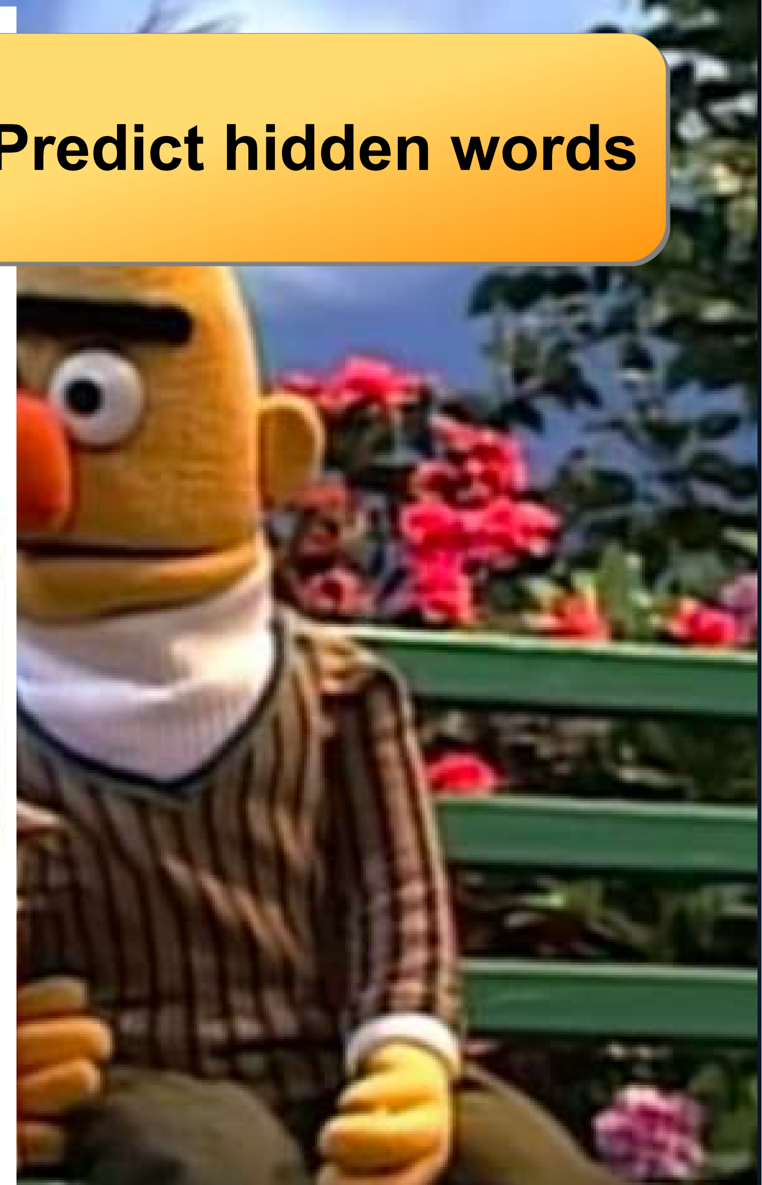


Image: Sesame Street

Contextual Word Vectors: Using Cloze Task (BERT)

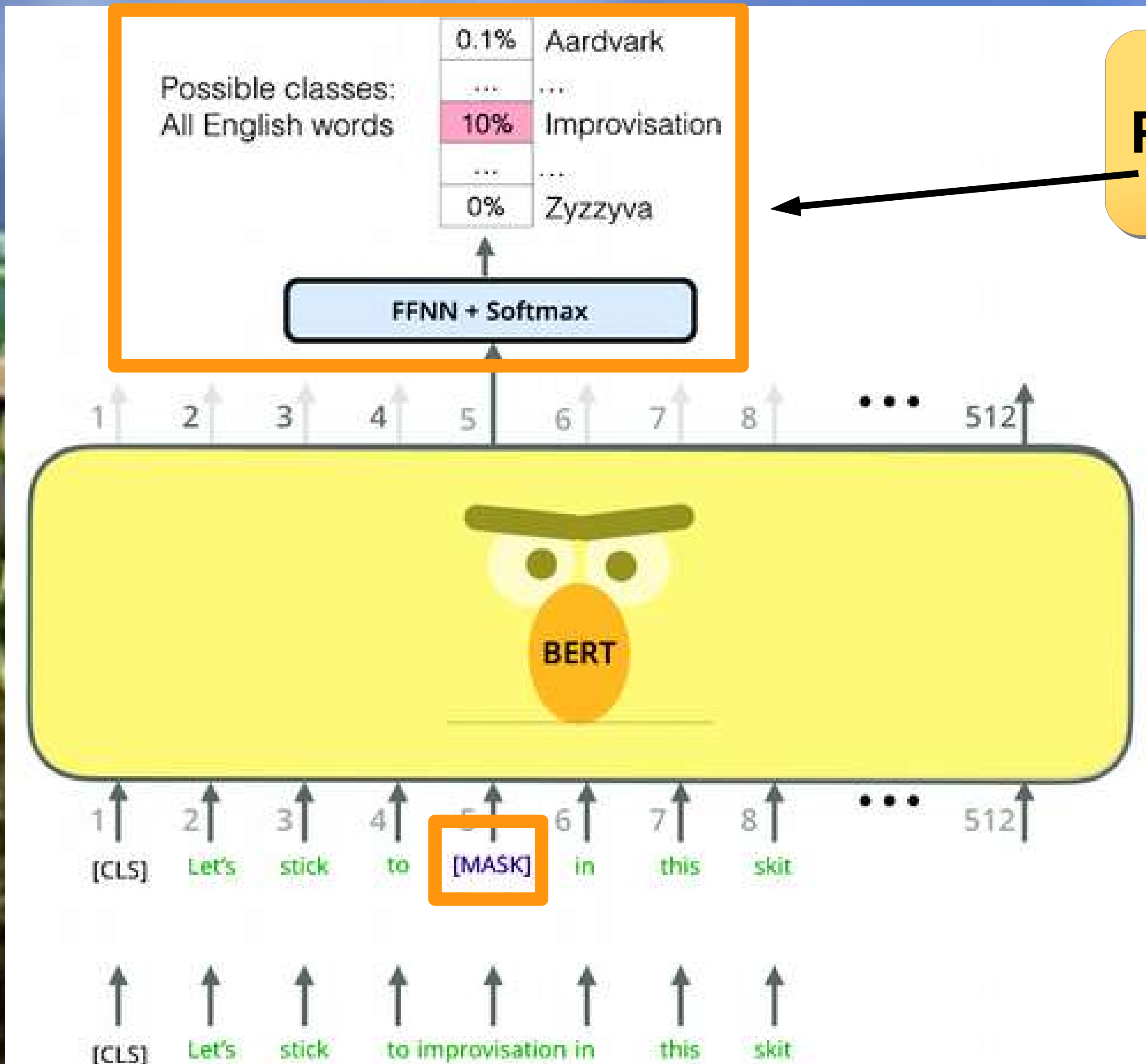


Image: Jay Alammar. http://jalammar.github.io/illustrated-bert/

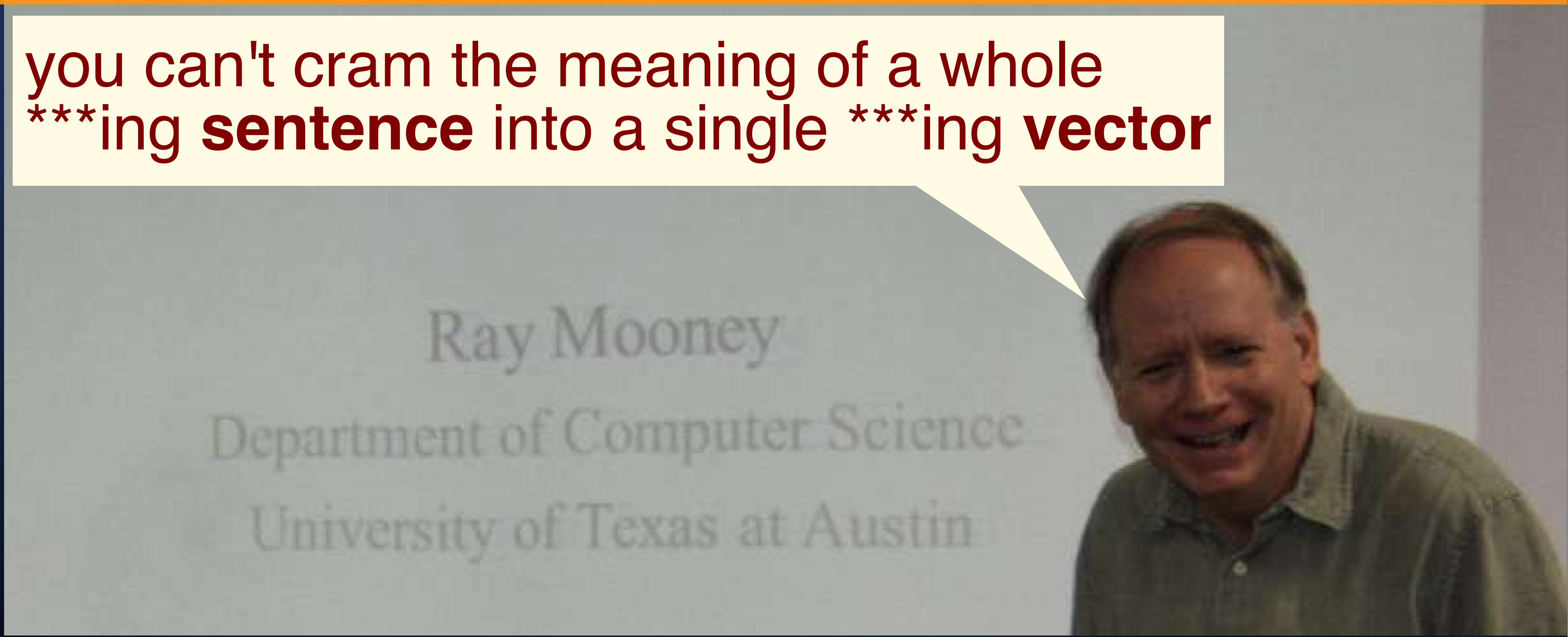
Predict hidden words

Additionally, BERT also predicts next sentence (like SkipThoughts)

Image: Sesame Street

What are they capturing?

What are they capturing?



CVSC Workshop at ACL 2013

Image: Adapted from http://www.cse.unt.edu/people/StudentNewsletters/2013_Apr_StudentEmailNewsletter.html

Evaluation via "Probing"

What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties

Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

German Kruszewski
Facebook AI Research
germank@fb.com

Guillaume Lample
Facebook AI Research
Sorbonne Universités
glample@fb.com

Loïc Barrault
Université Le Mans
loic.barrault@univ-lemans.fr

Marco Baroni
Facebook AI Research
mbaroni@fb.com

Also: Adi et al. ICLR 2016

These test whether enough information is kept to learn something from 100,000 training examples.

Our Approach: Inspect Proximity Structure

S_0 A person is slicing an onion.

$\mathrm{sim}(S_0, S_=)$

$S_=$ A person is cutting an onion.

S_* A person is not slicing an onion.

Zhu, Li, de Melo. Exploring Semantic Properties of Sentence Embeddings. Proc. ACL 2018

Our Approach: Inspect Proximity Structure

S_0 A person is slicing an onion.

$S_=$ A person is cutting an onion.

S_* A person is not slicing an onion.

$sim(S_0,S_=)$

$sim(S_0,S_*)$

All three sentences are closely related, but arguably the second should be even more similar than the third.

Our Approach: Inspect Proximity Structure

S_0 A person is slicing an onion.

$S_=$ A person is cutting an onion.

S_* A person is not slicing an onion.

sim(S_0,$S_=$)

sim(S_0,S_*)

sim(S_0,$S_=$) > sim(S_0,S_*) ?

Negation Detection

S_0 A person is slicing an onion.

$S_=$ A person is cutting an onion.

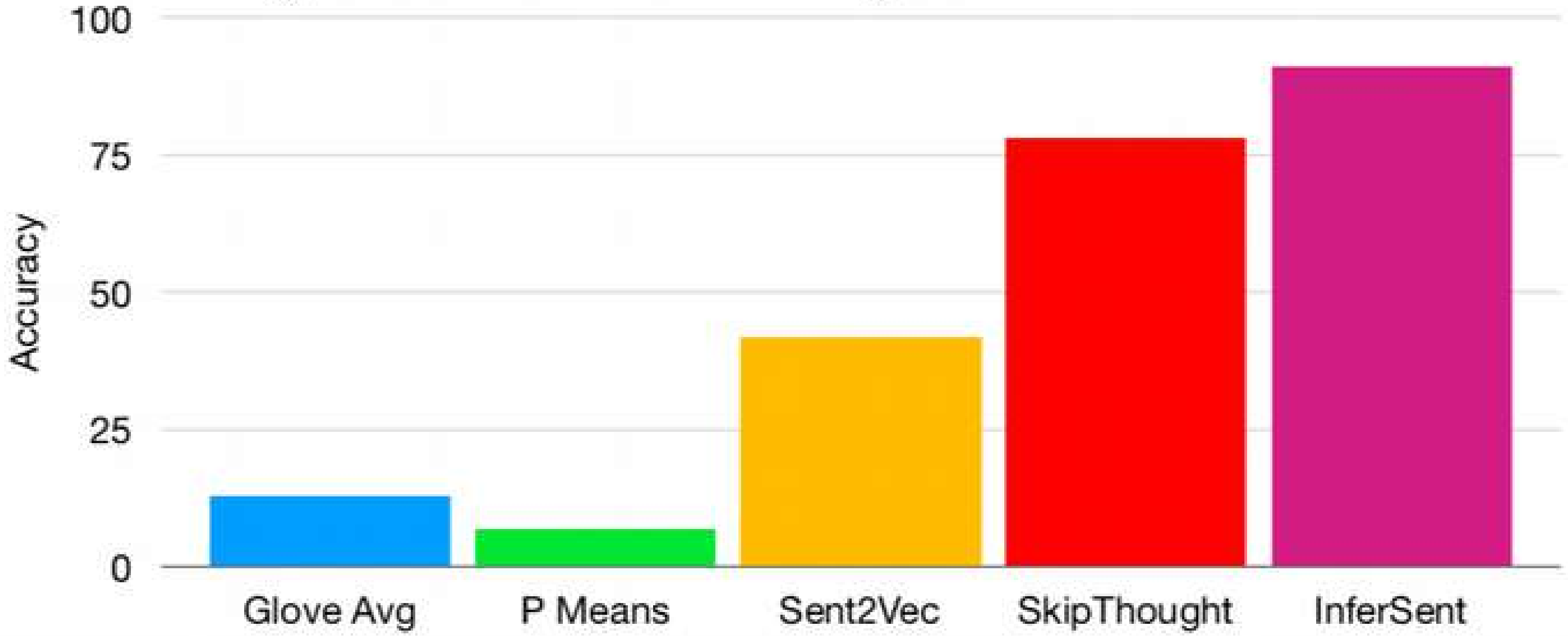
S_* A person is not slicing an onion.

Zhu, Li, de Melo. Exploring Semantic Properties of Sentence Embeddings. Proc. ACL 2018

Negation Detection

- Average of Word Embeddings is more easier misled by negation.

- Both InferSent and SkipThought succeed in distinguishing unnegated sentences from negated ones.



Zhu, Li, de Melo. Exploring Semantic Properties of Sentence Embeddings. Proc. ACL 2018

Negation Variant

S_0
(Negation)

A man is not standing on his head under water.

$S_=$
(Negated Existential)

There is no man standing on his head under water.

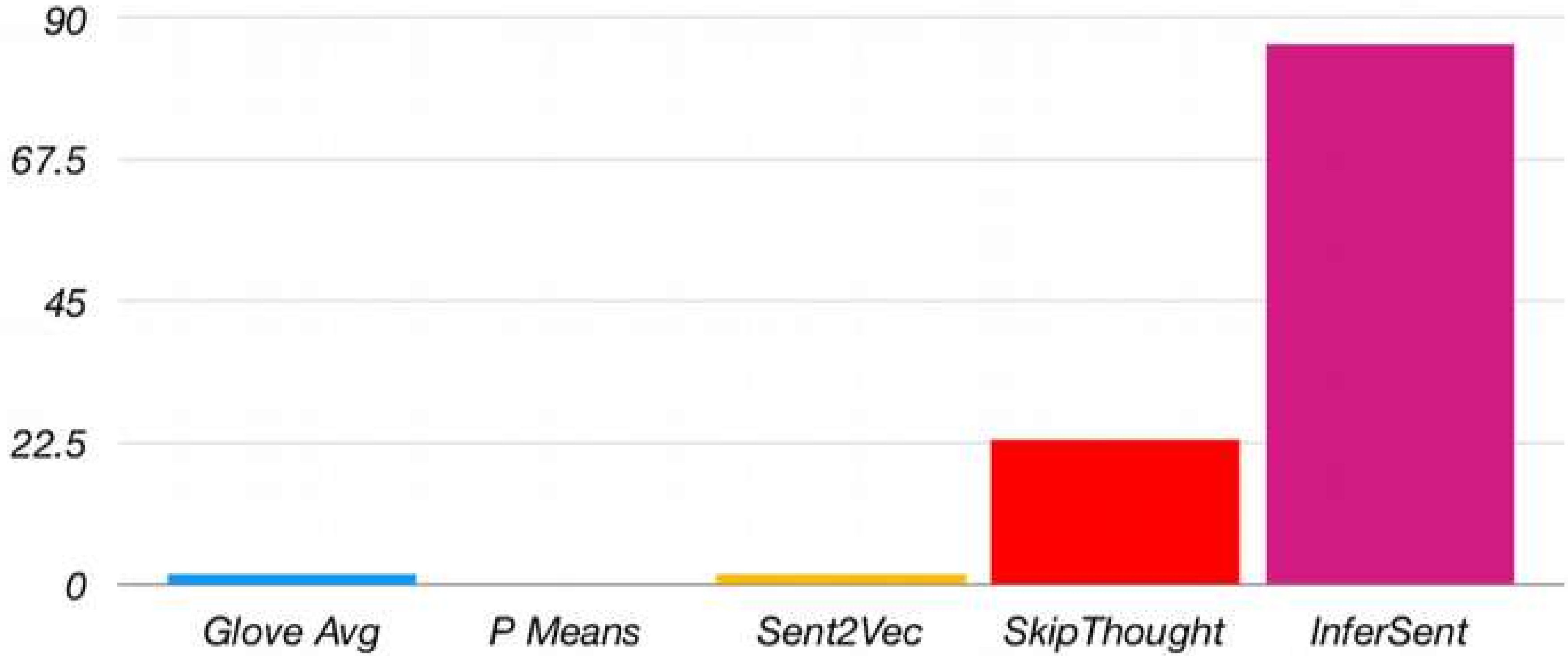
S_*

A man is standing on his head under water.

Zhu, Li, de Melo. Exploring Semantic Properties of Sentence Embeddings. Proc. ACL 2018

Negation Variant

- Both averaging of word embeddings and SkipThought are dismal in terms of the accuracy.
- InferSent appears to have acquired a better understanding of negation quantifiers, as these are commonplace in many NLI datasets.



Zhu, Li, de Melo. Exploring Semantic Properties of Sentence Embeddings. Proc. ACL 2018

Clause Relatedness

S_0 Octel said the purchase was expected.

Clause Extraction (for suitable head verbs only)

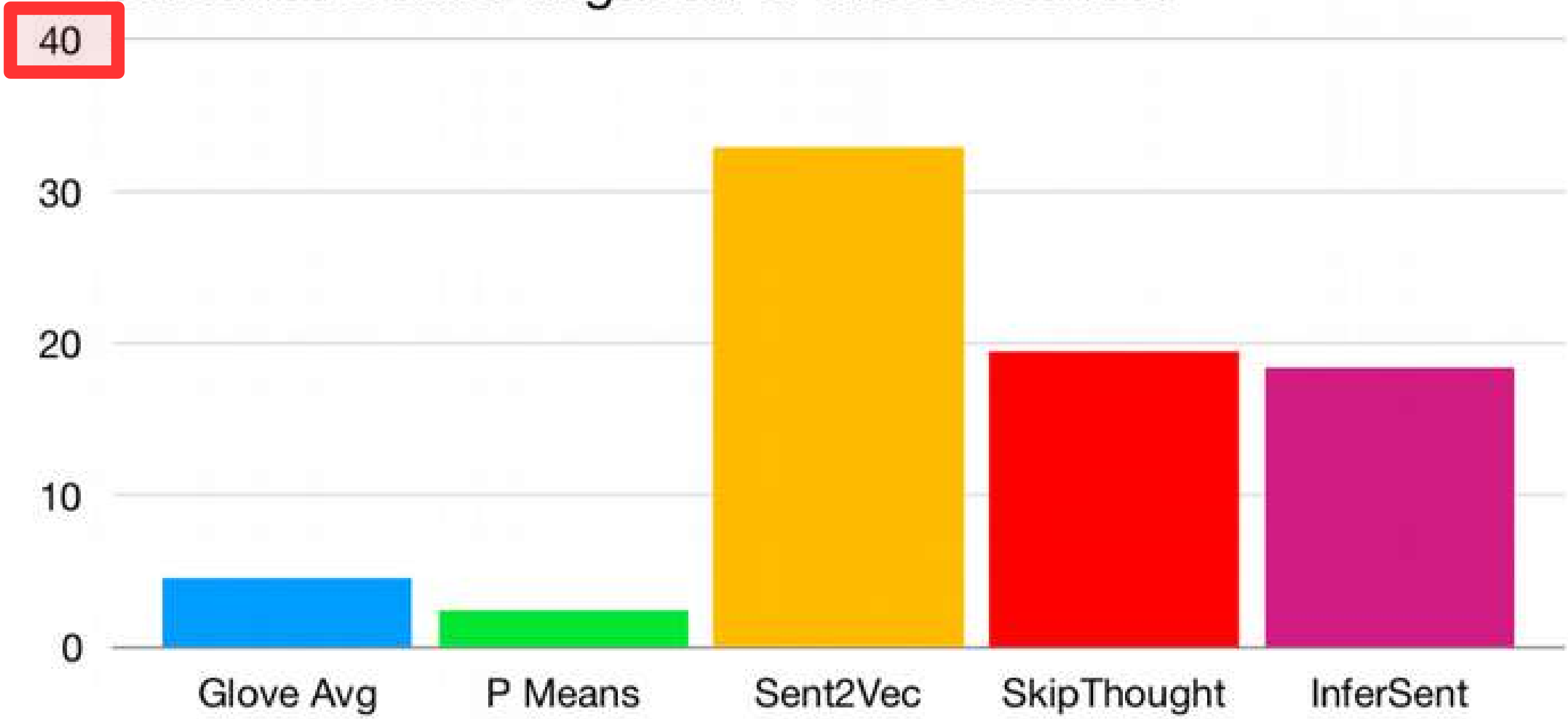
$S_=$ The purchase was expected.

S_* Octel said the purchase was not expected.

Zhu, Li, de Melo. Exploring Semantic Properties of Sentence Embeddings. Proc. ACL 2018

Clause Relatedness

- Both SkipThought vectors and InferSent works poorly when sub clause is much shorter than original one.
- Sent2vec best in distinguishing the embedded clause of a sentence from a negation of that sentence.



Zhu, Li, de Melo. Exploring Semantic Properties of Sentence Embeddings. Proc. ACL 2018

ELMo and BERT

BERT better than ELMo, but worse than InferSent

Image: Sesame Street

Argument Sensitivity

S_0

Francesca teaches Adam to adjust the microphone on his stage.

$S_=$

(Passive)

Adam is taught to adjust the microphone on his stage.

S_*

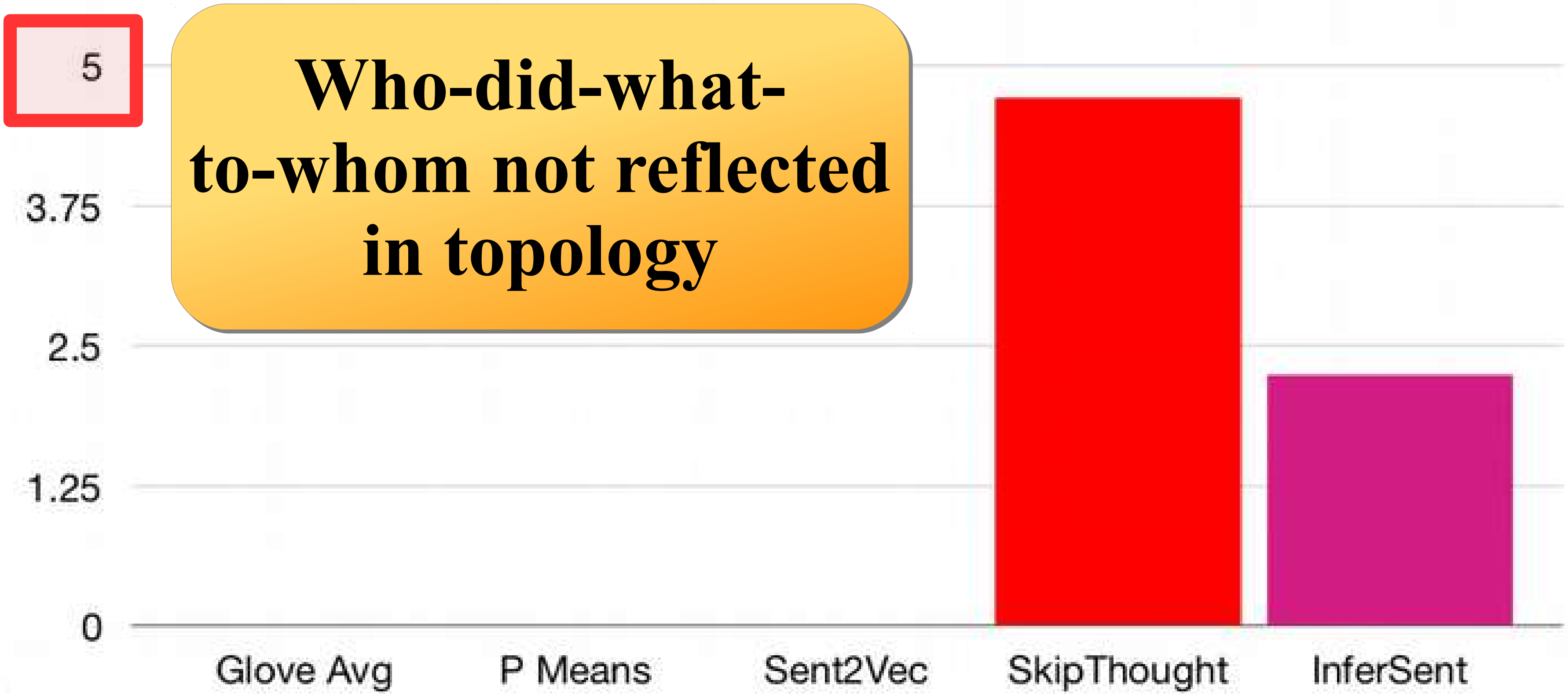
(Argument Inversion)

Adam teaches Francesca to adjust the microphone on his stage.

Zhu, Li, de Melo. Exploring Semantic Properties of Sentence Embeddings. Proc. ACL 2018

Argument Sensitivity

- None of the analyzed approaches prove adept at distinguishing the semantic information from structural information in this case.



Zhu, Li, de Melo. Exploring Semantic Properties of Sentence Embeddings. Proc. ACL 2018

Questions?

Image: https://www.flickr.com/photos/opensourceway/5556249000