Deep Contextualized Word Representation

Matthew E. Peters†, Mark Neumann†, Mohit Iyyer†, Matt Gardner†
Christopher Clark∗, Kenton Lee∗, Luke Zettlemoyer†∗

†Allen Institute for Artificial Intelligence
∗Paul G. Allen School of Computer Science & Engineering, University of Washington

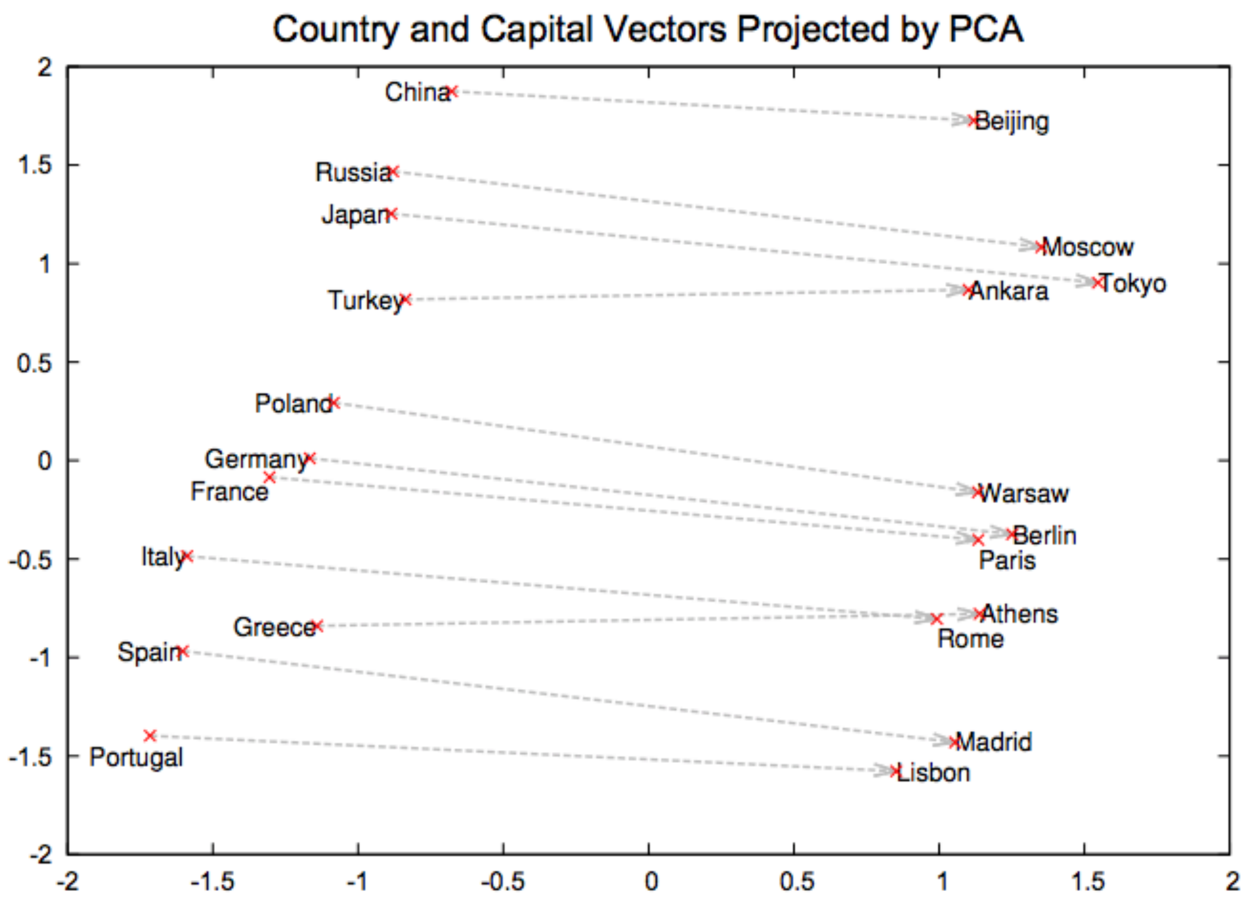
Presenter: Liyuan Liu (Lucas)

Word Representation

Represent word with distributed vectors while retaining their semantic meaning:

1. Resulting vectors are usually treated as the input layer of NNs for NLP tasks.

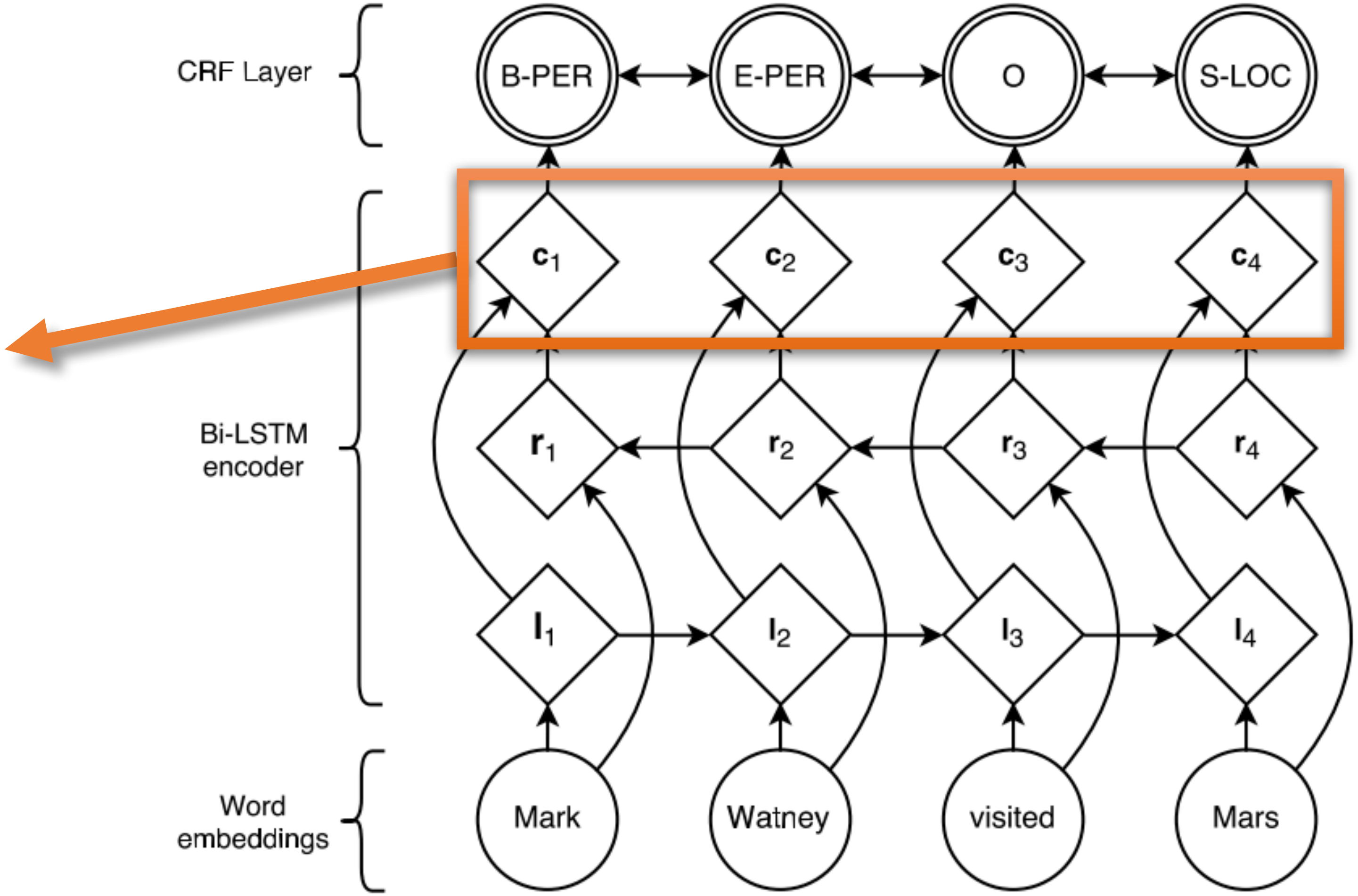
2. It's context agnostic and usually requires additional parameters for the end task.



Contextualized Word Representation

As most NLP tasks are context related, most of existing methods would contextualize the word embedding before make the final prediction

Contextualized Word Representation



Contextualized Word Representation

- However, complicated neural models requires extensive training data:
  - Models pre-trained on the ImageNet are widely used for Computer Vision tasks
  - What’s the proper way to conduct pre-training for NLP?

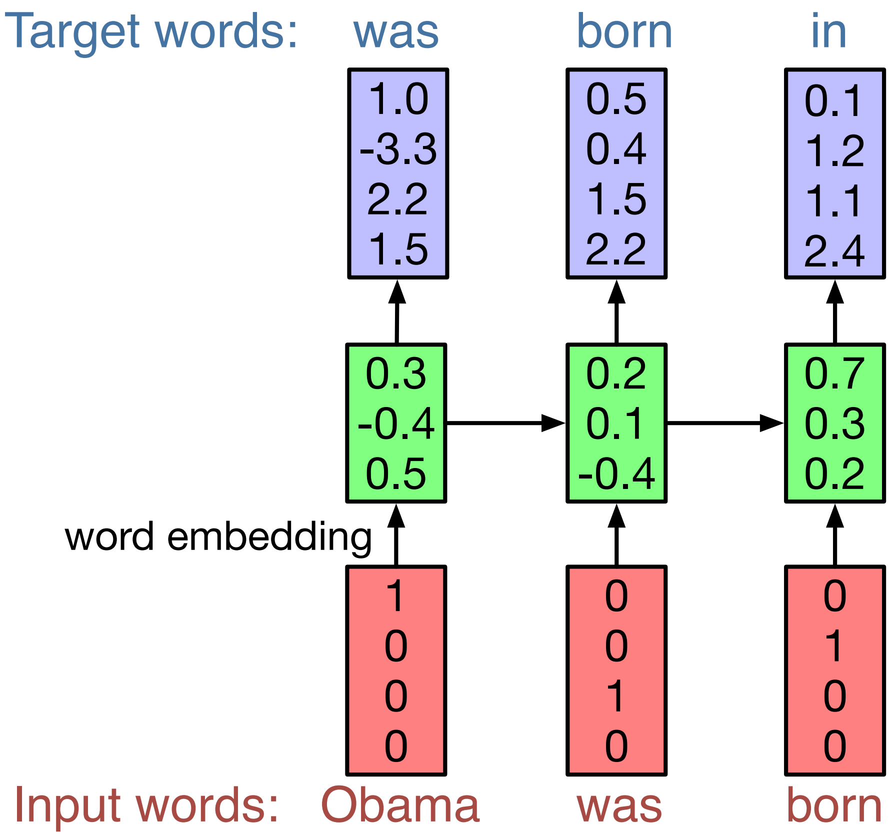
Basic Intuition

- Leveraging **Language Modeling** to get **pre-trained contextualized representation models**.

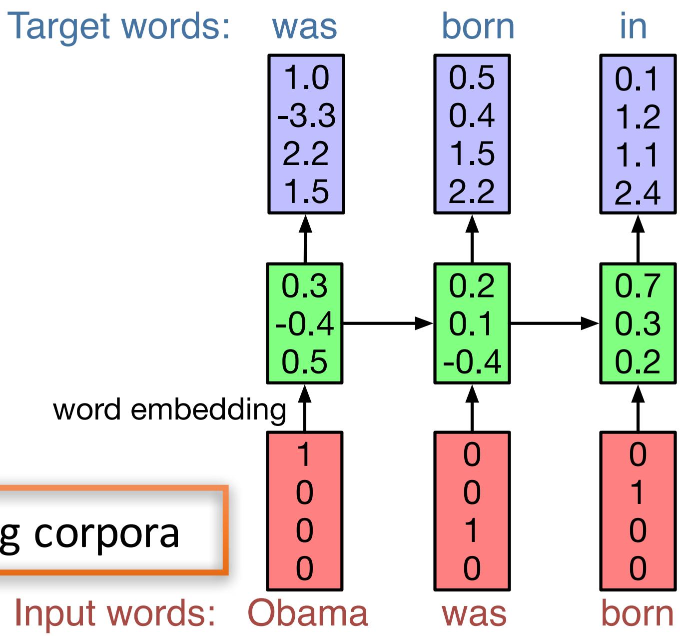
- Highlight:
  - 1. rely on large corpora, instead of human annotations
  - 2. works very well ---- improve the performance of existing SOA methods a lot

What is language modeling?

- Describing the generation of text:
  - predicting the next word based on previous contexts



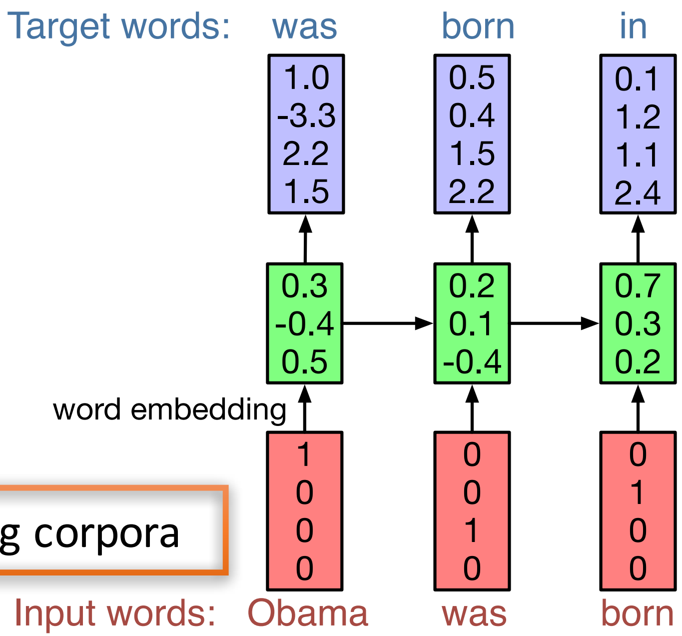
What is language modeling?



- Describing the generation of text:
  - predicting the next word based on previous contexts
- Pros:
  - does not require any human annotations
  - resulting models can generate sentences of an unexpectedly high quality:

Nearly unlimited training corpora

What is language modeling?



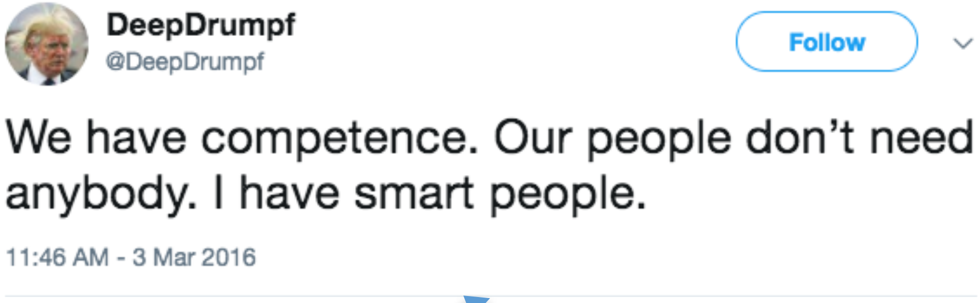
- Describing the generation of text:
  - predicting the next word based on previous contexts
- Pros:
  - does not require any human annotations
  - resulting models can generate sentences of an unexpectedly high quality:

Nearly unlimited training corpora

'''See also''': [[List of ethical consent processing]]

== See also ==
*[[lender dome of the ED]]
*[[Anti-autism]]

===[[Religion|Religion]]===
*[[French Writings]]
*[[Maria]]
*[[Revelation]]
*[[Mount Agamul]]

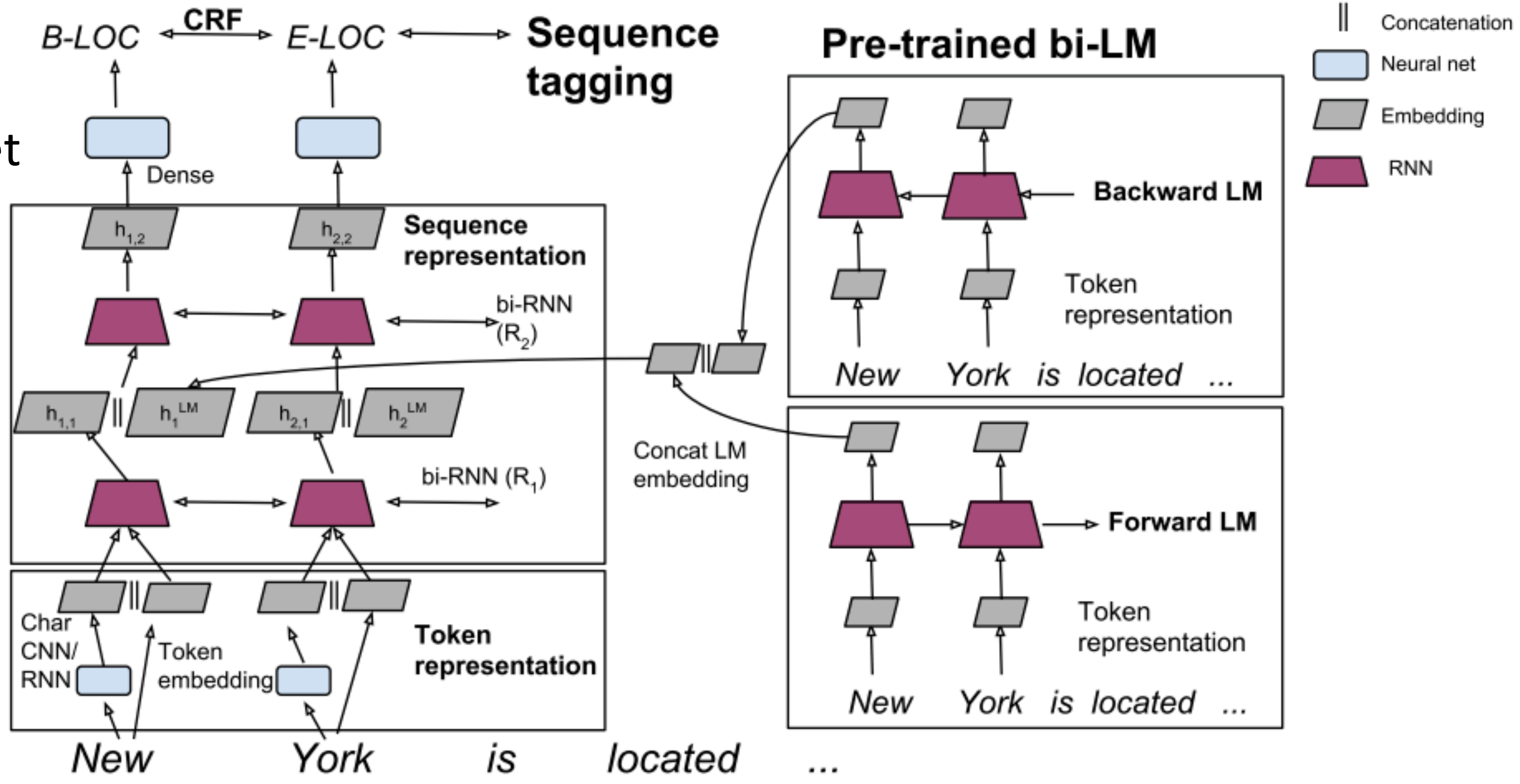


generated character-by-character

How to leverage Language Models?

TagLM:

- Pre-train language models on large dataset
- used the output of the final layer as the LM embedding



Peters, Matthew E., et al. "Semi-supervised sequence tagging with bidirectional language models." *ACL* (2017).

ELMo: Embeddings from Language Models

- For k-th token, L-layer bi-directional Language Models computes 2L+1 representations

$$\begin{aligned} R_k &= \{\mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} \mid j = 1, \ldots, L\} \\ &= \{\mathbf{h}_{k,j}^{LM} \mid j = 0, \ldots, L\}, \end{aligned}$$

- For a specific down-stream task, ELMo would learn a weight to combine these representations

$$\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM}$$

Use ELMo for supervised NLP tasks

- Add ELMo at the input of RNN. For some tasks (SNLI, SQuAD), including ELMo at the output brings further improvements
- Keypoint**:**
  - **freeze** the weight of the biLM
  - Regularization is necessary: $\lambda \|\mathbf{w}\|_2^2$

$$\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} \boxed{s_j^{task}} \mathbf{h}_{k,j}^{LM}$$

Experiments

TASK	PREVIOUS SOTA		OUR BASELINE	ELMO + BASELINE	INCREASE (ABSOLUTE/ RELATIVE)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

Table 1: Test set comparison of ELMo enhanced neural models with state-of-the-art single model baselines across six benchmark NLP tasks. The performance metric varies across tasks – accuracy for SNLI and SST-5; F_1 for SQuAD, SRL and NER; average F_1 for Coref. Due to the small test sizes for NER and SST-5, we report the mean and standard deviation across five runs with different random seeds. The "increase" column lists both the absolute and relative improvements over our baseline.

Experiments

Where to include the ELMo embedding

Task	Input Only	Input & Output	Output Only
SQuAD	85.1	**85.6**	84.8
SNLI	88.9	**89.5**	88.7
SRL	**84.7**	84.3	80.9

Table 3: Development set performance for SQuAD, SNLI and SRL when including ELMo at different locations in the supervised model.

Alternate Layer Weighting Schemes

Task	Baseline	Last Only	All layers λ=1	All layers λ=0.001
SQuAD	80.8	84.7	85.0	**85.2**
SNLI	88.1	89.1	89.3	**89.5**
SRL	81.6	84.1	84.6	**84.8**

Table 2: Development set performance for SQuAD, SNLI and SRL comparing using all layers of the biLM (with different choices of regularization strength λ) to just the top layer.

Experiments

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

Table 4: Nearest neighbors to "play" using GloVe and the context embeddings from a biLM.

Experiments

- Word sense disambiguation:
  - Calculate the average of representation of each sense in the training data
  - Conduct 1-nearest neighbor search at the test set

Model	F_1
WordNet 1st Sense Baseline	65.9
Raganato et al. (2017a)	69.9
Iacobacci et al. (2016)	**70.1**
CoVe, First Layer	59.4
CoVe, Second Layer	*64.7*
biLM, First layer	67.4
biLM, Second layer	*69.0*

Table 5: All-words fine grained WSD F_1. For CoVe and the biLM, we report scores for both the first and second layer biLSTMs.

Experiments

- POS-Tagging:
  - Directly learn a multi-class classifier for the POS-tagging

Model	Acc.
Collobert et al. (2011)	97.3
Ma and Hovy (2016)	97.6
Ling et al. (2015)	**97.8**
CoVe, First Layer	*93.3*
CoVe, Second Layer	92.8
biLM, First Layer	*97.3*
biLM, Second Layer	96.8

Table 6: Test set POS tagging accuracies for PTB. For CoVe and the biLM, we report scores for both the first and second layer biLSTMs.

Misc

- Podcast by the author (Matthew E. Peters):
  - https://soundcloud.com/nlp-highlights/56-deep-contextualized-word-representations-with-matthew-peters
- A follow-up work on further improving the efficiency:
  - Efficient Contextualized Representation: Language Model Pruning for Sequence Labeling (https://arxiv.org/abs/1804.07827)

Take aways…

- Language Modeling is effective in constructing contextualized representation (could be helpful for a variety of tasks);

- Outputs of all Layers are useful;