Deep Contextualized Word Representations

M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer

Slides by D. Kalev and M. Spaconi

Static vs Contextualized Embeddings

- Static word embeddings (e.g. Glove, Word2Vec) do not consider context. Issues:

  Polysemy: a word can have multiple meanings

  Part of speech: a token can belong to different parts of speech (e.g. **play** can be a verb)

Chico Ruiz made a spectacular **play** on Alusik 's grounder	Olivia De Havilland signed to do a Broadway **play** for Garson

- Idea: allow embeddings to capture context.

Embeddings from Language Model (ELMo)

- Contextual: representation depends on the entire context in which it is used
- Deep: employs deep pre-trained model for representations
- Character based: allows out-of-vocabulary words and can use morphological rules

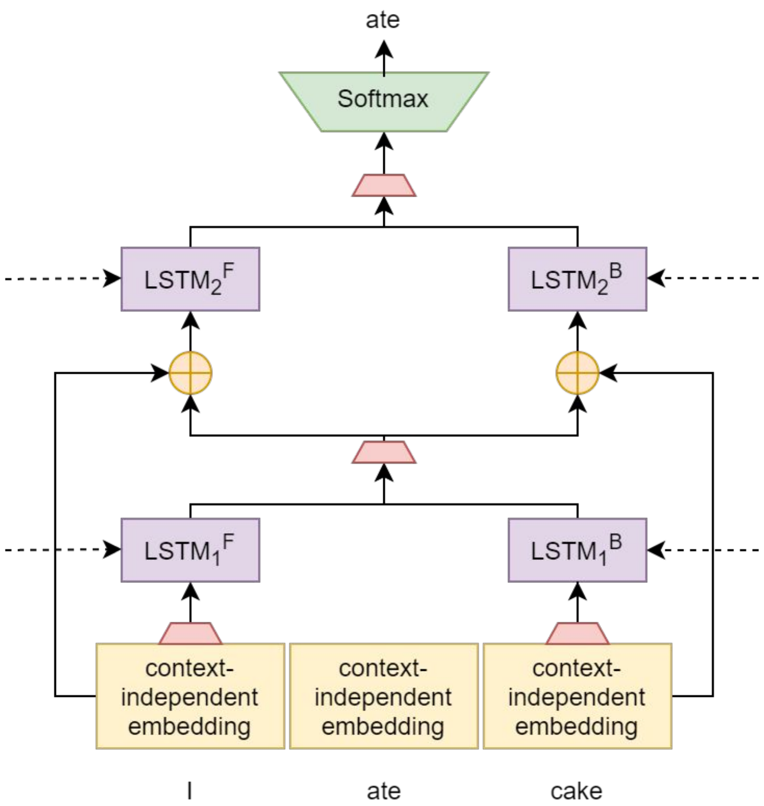
ELMo’s Bidirectional Language Model (biLM)

- Unsupervised task:
  Predict next (previous) word for forward (backward) LSTM
- Shared weights for context-independent embeddings and softmax layer, but different directional LSTM weights.

$$\sum_{k=1}^{N} \big(\log p(t_k \mid t_1, \ldots, t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) + \log p(t_k \mid t_{k+1}, \ldots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) \big) .$$

Training ELMo’s biLM

- Trained on a large dataset (1B words benchmark (Chelba et al., 2014)).
- Importance sampling for softmax
- Residual connection from 1st to 2nd layer
- Based on the work of Jozefowicz, Rafal, et al. "Exploring the limits of language modeling." (2016).



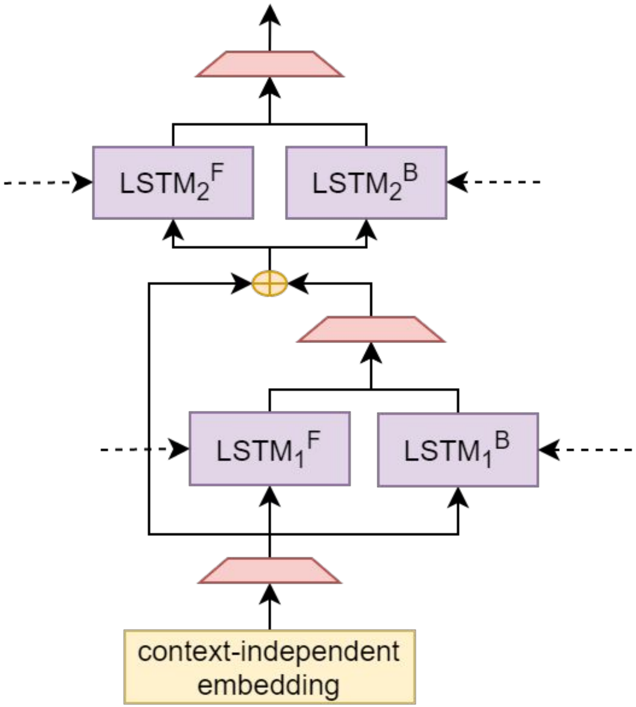
Contextualized Embeddings

- Differently from training, to get the embeddings we also feed the target-word embedding

- 3 different level of word embeddings (after each linear projection): independent, syntactic and semantic

- They can capture different information

- We can collapse them to provide a single embedding



Usage for downstream tasks

- Plug-in replacement for static embeddings
- Embeddings can be frozen or let train;
  training typically improves performance on downstream task

Linear combination of ELMo's outputs.

γ and s_j are learnt when training the model for the downstream task. s_j values are softmaxed.

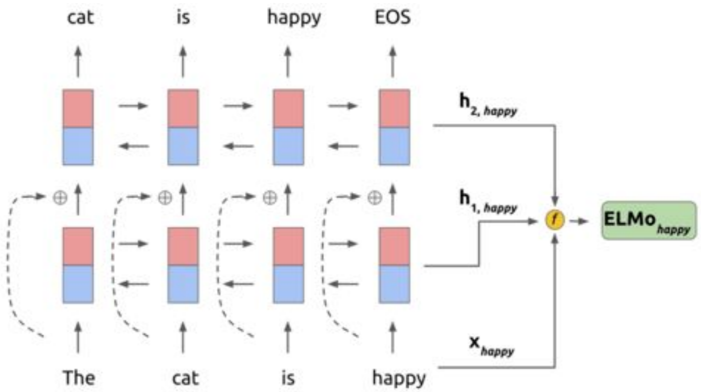
$$\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM}. \quad (1)$$

https://appliedmachinelearning.blog/2019/11/30/training-elmo-from-scratch-on-custom-data-set-for-generating-embeddings-tensorflow/

Model

biLM

- 4096 units in each biLSTM
- 512 dimension projections
- residual connection from 1st to 2nd layer



Character level word embeddings

- character level embeddings: size 16
- n-gram CNN: `[1, 32], [2, 32], [3, 64], [4, 128], [5, 256], [6, 512], [7, 1024]`
- max-pooling: 2048
- 2 highway layers
- projection to 512

https://s3-us-west-2.amazonaws.com/allennlp/models/elmo/2x4096_512_2048cnn_2xhighway_5.5B/elmo_2x4096_512_2048cnn_2xhighway_5.5B_options.json

Experiments

- Question answering
  - SQuAD dataset - 100k question-answer pairs
  - answer is a span of a Wikipedia article
- Textual entailment
  - SNLI dataset - 550k hypothesis-premise pairs
- Semantic Role Labeling
  - OntoNotes dataset - 2.9 mln words; predicate - argument structure
  - various genres of text (news, talk shows, phone conversations, etc)
  - 3 languages (English, Mandarin, Arabic)
- Coreference resolution
  - coreference annotations in CoNLL 2012 dataset
- Named Entity Recognition (NER)
  - CoNLL 2003 - news from the Reuters RCV1 corpus
  - tagged with 4 different entity types (PER,LOC,ORG, MISC)
- Sentiment Analysis
  - SST-5 - describe a sentence from a movie review with a label (from very negative to very positive)

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.

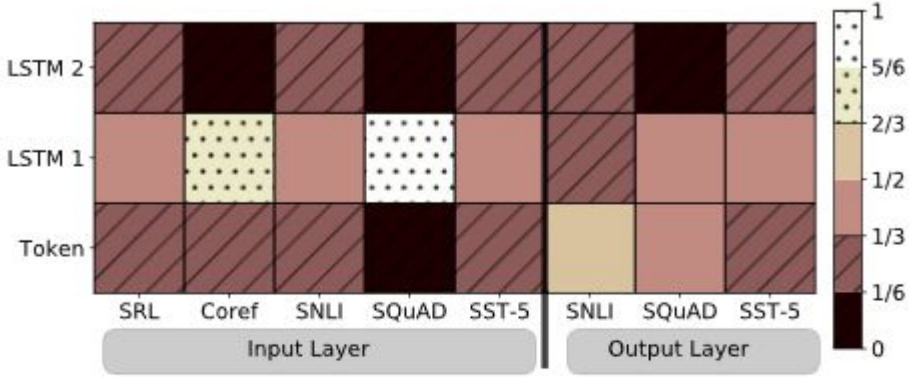
Modeling polysemy

	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.

Intrinsic evaluation

- Different layers encode different information
  - Layer 1 - Syntactic
  - Layer 2 - Semantic



Word Sense Disambiguation (WSD)

- compute representations of all words (SemCor 3) using biLM
- take average representation for each sense
- 1-nearest neighbours sense

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.

POS tagging

- The Wall Street Journal part of the Penn Treebank (PTB) dataset
- ELMo embeddings as input to a linear classifier that predicts the POS tags

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.

Sample efficiency

- number of parameter updates
  - from 486 to 10 epochs (98% relative decrease) for SRL
- training set size

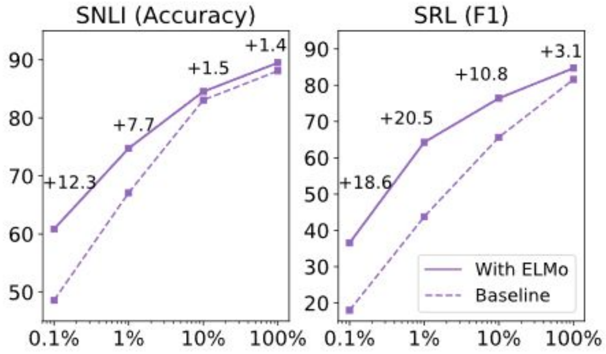


Figure 1: Comparison of baseline vs. ELMo performance for SNLI and SRL as the training set size is varied from 0.1% to 100%.

Layer weighting

- λ=1 reduces to simple average over layers

$$\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM}. \quad (1)$$

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.

Perks

- Capturing context helps with polysemy and POS ambiguity
- Plug-in solution applicable to different models and tasks
- Different layers capture different information that can be used as needed by downstream models
- Higher sample efficiency

Problems

- The paper does not explain some implementation details:

    The softmax on large scale vocabulary: uses importance sampling but it's not clearly stated

    *The context insensitive type representation uses **2048 character n-gram convolutional filters** followed by two highway layers (Srivastava et al., 2015) and a linear projection down to a 512 representation.*
    The full specification is provided on github, not on the paper.

Future Work

- Deeper Language Models
- Transformer instead of biLM (GPT)
- Discriminative fine-tuning (ULMfit)

Neural Metaphor Detection in Context

Christiaan van der Vlist

ATCS 2020

Table of Contents

What are metaphors?

A metaphor is a figure of speech that, for rhetorical effect, directly refers to one thing by mentioning another
- *Wikipedia*

The problem

- Semantics of a word change without the word itself changing
- Traditional word representations cannot deal with this
- Important for NLP tasks such as machine translation or sentiment analysis

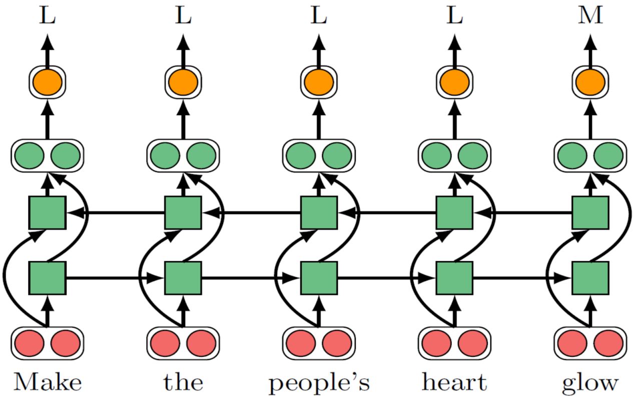
Idea

- Previous approaches used limited linguistic context or contextual expressivity
- The idea of this paper: *What if we use* better *and* more *linguistic context?*
- Specifically, they use ELMo embeddings and train a model to predict the metaphoricity of *all* words in a sentence
- ELMo embeddings are context-*dependent*

Methods

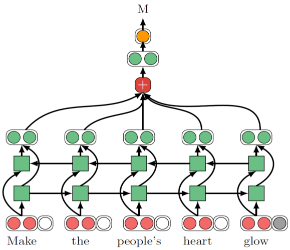
- Two architectures:
  1. Sequence labeling (SEQ)
  2. Classification (CLS)
- Both architectures:
  1. Take a pre-trained word embedding (300d GloVe) plus an 1024d ELMo vector as input per word
  2. Use a BiLSTM to encode sentences
  3. Use a feedforward network to classify

Sequence labeling



Model architecture for sequence labeling model (SEQ)

Classification



Architecture for classification model (CLS)

- CLS also:
  1. Takes an additional index embedding with information about whether the word is the target verb
  2. Uses an attention layer after the BiLSTM, before the feedforward network

Experiment

- Two tasks:
  1. Sequence labeling
  2. Classification
- Sequence labeling is only performed by SEQ
- Classification task is performed by both models
- Note that sequence labeling is a generalized classification task
- Baseline labels a word as metaphorical if it is metaphorical more often than not

Data

- Sequence labeling task:
  1. VUA
- Classification task:
  1. VUA
  2. MOH-X
  3. TroFi
- Unlabeled words are considered literal

	N	% metaphor	# uniq verb	avg sent len
MOH-X	647	49%	214	8.0
TroFi	3737	43%	50	28.3
VUA	23113	28%	2047	24.5

Properties of datasets used in the paper

Results - Sequence labeling

	P	R	F1	Acc.
Baseline	68.6	45.2	54.5	90.6
Theirs (SEQ)	**71.6**	**73.6**	**72.6**	**93.1**

Results obtained on the VUA test set

- Baseline has high precision because some words are exclusively literal
- SEQ mostly improves on recall

Results - Sequence labeling

POS	#	% metaphor	P	R	F1
VERB	**20K**	18.1	68.1	71.9	69.9
NOUN	20K	13.6	59.9	60.8	60.4
ADP	13K	**28.0**	**86.8**	**89.0**	**87.9**
ADJ	9K	11.5	56.1	60.6	58.3
PART	3K	10.1	57.1	59.1	58.1

The breakdown of performance on the VUA sequence labeling test set by POS tags.

- Adposition is easiest to identify, it also has the highest percentage of metaphors
- Particles are difficult to identify because they often appear in expressions

Results - Classification on MOH-X

	P	R	F1	Acc.
Baseline	39.1	26.7	31.3	43.6
Theirs - CLS	75.3	**84.3**	**79.1**	**78.5**
Theirs - SEQ	**79.1**	73.5	75.6	77.2

Results obtained on MOH-X with 10-fold cross validation

- CLS outperforms SEQ
- Only verbs are annotated

Results - Classification on TroFi

	P	R	F1	Acc.
Baseline	**72.4**	55.7	62.9	71.4
Regression with abstractness[1]	-	-	**75.0**	-
Theirs - CLS	68.7	**74.6**	72.0	73.7
Theirs - SEQ	70.7	71.6	71.1	**74.6**

Results obtained on TroFi with 10-fold cross validation

- Köper et al. outperform CLS and SEQ
- Concreteness labels are correlated to metaphor labels
- TroFi has only 50 verbs, Köper et al. look at verb lemmas

[1] Köper and Walde, "Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses".

Results - Classification on VUA

	P	R	F1	Acc.	MaF1
Baseline	67.9	40.7	50.9	76.4	48.9
CNN-LSTM[2]	60.0	**76.3**	67.2	-	-
Theirs - CLS	53.4	65.6	58.9	69.1	53.4
Theirs - SEQ	**68.2**	71.3	**69.7**	**81.4**	**66.4**

Results obtained on VUA test set

- SEQ outperforms CLS
- Metaphorical labels of context are important

[2] Wu et al., "Neural metaphor detecting with cnn-lstm model".

Error analysis - Metaphor types

- Indirect metaphor:
  - Contrast between basic and contextual meaning
  - The results could prove **valuable** to researchers.
- Direct metaphor:
  - No contrast between basic and contextual meaning
  - John is like a **ferret**.
- Personification:
  - Based on a comparison between human and non-human
  - He thought of thick <u>motorways</u> **carving** up that land.

Error Analysis - SEQ

- Error analysis performed on a sample of errors on the VUA validation set for classification
- Many indirect metaphors (most common type) and personifications were mistaken for literal verbs
- Many literal verbs with implicit arguments were mistaken for metaphors
    1. To *throw* up an impenetrable Berlin Wall between you and them could be tactless.

Results - Error Analysis

- SEQ outperforms CLS on:
  1. Personifications
  2. Indirect metahpors
  3. Direct metaphors with uncommon verbs

Their Conclusion

- Using contextualized word embeddings improves metaphor detection
- Predicting the metaphoricity of all words in a sentence also improves metaphor detection

My Opinion

- Labeling all words gives better insight into the metaphoricity
- Model architectures were relatively simple
- A large part of the improvement comes from ELMo
- Error analysis lacks an interpretation

Future Research

- How well does SEQ aid in other NLP tasks, such as machine translation or sentiment analysis?
- How to identify metaphors types that SEQ has trouble with?
- Transformer based architectures
  - Faster to train
  - Can be fine-tuned for this task
  - Multi-headed attention may capture more nuance

Thank you for your attention

Any questions?