

# NMLM

## NLP

Jeff Abrahamson

6 septembre 2020

# Disclaimer

**We have 13 minutes for 6 years.**

This will simplify quite a lot. **Maybe too much.** The goal is to trace the evolution of techniques and to give hints about why.

I'm trying to avoid the non-essential stuff, like optimisation improvements that aren't core to why the state of the art advanced.

All the papers are on github.

# Overview

What's changed:

- (Much) better performance
- Polysemy
- Pre-trained models

# Overview

- $\leq 2014$
- 2014: word2vec and improvements
- 2018: ELMo
- 2019: BERT
- 2018–2020: GPT

# Overview

## Pre-trained models:

- Context-free (word2vec, FastText, GloVe)
- Contextual
  - Unidirectional (ELMo (I think))
  - Bidirectional (BERT, GPT-2,3 (I think))

## TF-IDF (classic example)

- 1-hot encoding to get high-dimensional vectors
- TF-IDF to scale (weight)
- PCA or t-SNE etc. (if we want lower dimensionality)

## TF-IDF (classic example)

- 1-hot encoding to get high-dimensional vectors
- TF-IDF to scale (weight)
- PCA or t-SNE etc. (if we want lower dimensionality)

## What do we have?

- A space of likely significant words.
- Likely no good semantic knowledge.

# Word2vec

- NN, but not deep: think auto-encoder or RBL
- Train on context: similar context  $\Rightarrow$  similar embedding
- Want: distortion  $\approx$  meaning



# Word2vec

- NN, but not deep: think auto-encoder or RBL
- Train on context: similar context  $\Rightarrow$  similar embedding
- Want: distortion  $\approx$  meaning

Mikolov: trade complexity for efficiency  $\Rightarrow$  learn bigger datasets.

# Word2vec

- NN, but not deep: think auto-encoder or RBL
- Train on context: similar context  $\Rightarrow$  similar embedding
- Want: distortion  $\approx$  meaning

Mikolov: trade complexity for efficiency  $\Rightarrow$  learn bigger datasets.

Uses: search, translation, ...

*Tomas Mikolov et al., Efficient Estimation of Word Representations in Vector Space, ICLR Workshop, 2013.*

# Word2vec

Encode words at learning time.

$$(\mathbb{R}^N \rightarrow \mathbb{R}^n \quad \text{for } N \approx 5 \times 10^5 \text{ and } n \approx 300)$$

Problem: no polysemy

# Word2vec

How does it work?

It's learning, so we're optimising something. What?

# Word2vec

How does it work?

It's learning, so we're optimising something. What?

similarity of context

# Word2vec

How does it work?

It's learning, so we're optimising something. What?

cosine distance for similar contexts

How does it work?

It's learning, so we're optimising something. What?

cosine distance for similar contexts

Look at words around the target word. Think of as an n-gram. We want words with similar context to project close together, words with different context (ideally) not to.

Interesting properties:

- Magnitude is related to importance
  - Too common: rarely learns large magnitude
  - Too rare: rarely grows large
  - So Goldilocks property (mid-range is just right)



Interesting properties:

- Magnitude is related to importance
  - Too common: rarely learns large magnitude
  - Too rare: rarely grows large
  - So Goldilocks property (mid-range is just right)
- Learns stop words. Because their contexts are uncorrelated, they optimise near zero.

# Word2vec

How does it work?

It's not a single algorithm, but presented in multiple variants ("efficiency improvements"). Two notable elements are

- CBOW (continuous bag-of-words) – context predicts word
- SG (skipgram) – word predicts context (SGNS = skipgram negative sampling)

*Tomas Mikolov et al., Distributed Representations of Words and Phrases and their Compositionality, NIPS 2013.*

How does it work?

- SGNS factorises a word-context PMI matrix (This is not the most important take-away, but to point out that the mathematical foundations are slowly being understood.)

PMI = pointwise mutual information

$$pmi(x; y) = \log \left( \frac{\mathbf{Pr}(x, y)}{\mathbf{Pr}(x) \mathbf{Pr}(y)} \right) = \log \left( \frac{\mathbf{Pr}(x | y)}{\mathbf{Pr}(x)} \right) = \log \left( \frac{\mathbf{Pr}(y | x)}{\mathbf{Pr}(y)} \right)$$

*Sanjeev Arora et al., A latent variable model approach to word embeddings, Trans. Assoc. Comput. Linguistics 4: 385-399 (2016).*

# Word2vec

How does it work?

Output is a softmax: cost is proportional to number of classes (50K words).

# Word2vec

How does it work?

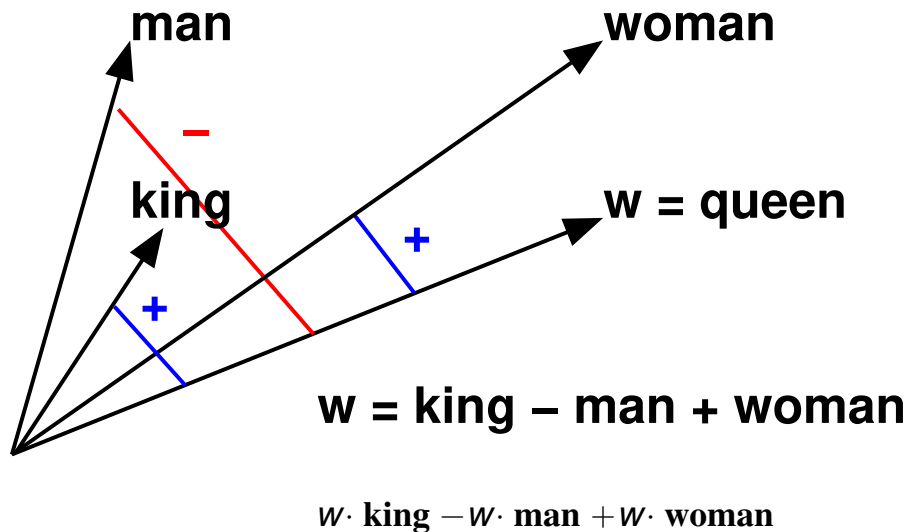
Output is a softmax: cost is proportional to number of classes (50K words).

Approximate the softmax to reduce cost: hierarchical softmax:  $O(N) \Rightarrow O(\log n)$ .

*Yoav Goldberger and Omber Levy, Word2vec Explained, arXiv 2014.*

# Word2vec

The famous “man : woman as king : queen” in maths: we’re maximising two similarities and minimising a dissimilarity.



## More references (hardly exhaustive):

- *Yann LeCun, Bengio, Hinton, Deep Learning, Nature 521, 436–444(2015).*
- *Omer Levy and Yoav Goldberg, Neural Word Embedding as Implicit Matrix Factorization, 2014.*
- *Omer Levy and Yoav Goldberg, Linguistic Regularities in Sparse and Explicit Word Representations, 2014.*
- *Omer Levy et al., A Strong Baseline for Learning Cross-Lingual Word Embeddings from Sentence Alignments, 2017.*



## Improvements:

### ① FastText (Facebook)

*Piotr Bojanowski et al., Enriching Word Vectors with Subword Information, 2017.*

*Armand Joulin, Bag of Tricks for Efficient Text Classification, 2016.*

## Improvements:

- 1 FastText (Facebook)
- 2 GloVe (“Global Vectors”)

*Jeffrey Pennington et al., GloVe: Global Vectors for Word Representation, 2014.*

*<https://nlp.stanford.edu/projects/glove/>*

## Improvements:

- 1 FastText (Facebook)
- 2 GloVe (“Global Vectors”)
- 3 ULMFit (“Universal Language Model Fine Tuning”)

*Jeremy Howard and Sebastian Ruder, Universal Language Model Fine-tuning for Text Classification, 2018.*

*Andrew Dai and Quoc Le, Semi-supervised Sequence Learning, 2015. (← not ULMFiT, actually)*

Improvements:

- ① FastText (Facebook)
- ② GloVe (“Global Vectors”)
- ③ ULMFit (“Universal Language Model Fine Tuning”)

Better performance, similar properties.

What NLP last quarter century, just a lot better.

Improvements:

- ① FastText (Facebook)
- ② GloVe (“Global Vectors”)
- ③ ULMFit (“Universal Language Model Fine Tuning”)

Better performance, similar properties.

What NLP last quarter century, just a lot better. **Don't use any of these, they're obsolete.**



- Encode words before and after (not just at learning time)
  - No: dictionary: map word to vector
  - Yes: on the fly, run context through deep network to produce new context
- Captures linguistic context: polysemy
- Models word use: captures both syntactic and semantic information
- Most tasks better than word2vec and relatives (question answering, named entity exceptions, sentiment analysis, . . .)

- Bidirectional LSTM + residual connectors between 1st and second layer
- Character embedding rather than word: helps with out-of-vocabular words
- Convolutional filters instead of n-gram features

Up to here, similar to Jozfowicz

*Rafal Jozfowicz, Exploring the Limits of Language Modeling, 2016.*



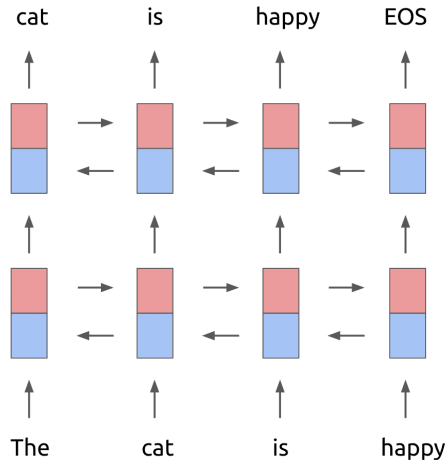
# ELMo

- Bidirectional LSTM + residual connectors between 1st and second layer
- Character embedding rather than word: helps with out-of-vocabular words
- Convolutional filters instead of n-gram features
- Learn different representations for different tasks
- Transfer learning

In CV, it's standard to learn ImageNet and transfer to problem at hand.

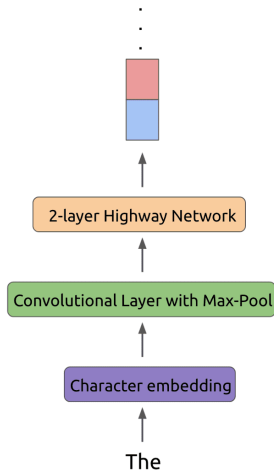
ELMo shows we can do this for NLP problems.

# ELMo



*https:*

*//www.mihaileric.com/posts/deep-contextualized-word-representations-elmo/*



*https:*

*//www.mihaileric.com/posts/deep-contextualized-word-representations-elmo/*

## See also

- *Matt Gardner et al., AllenNLP: A Deep Semantic Natural Language Processing Platform*
- *Yoon Kim, Character-Aware Neural Language Models, 2015.*
- *Matthew Peters et al., Deep contextualized word representations, 2018.*
- *Rupesh Kumar Srivastava et al., Highway Networks, 2015.*

# BERT



# BERT

- Builds on ELMo and first GPT
- Encoder-decoder: use self-attention to encode and attention to decode

*Jacob Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019.*

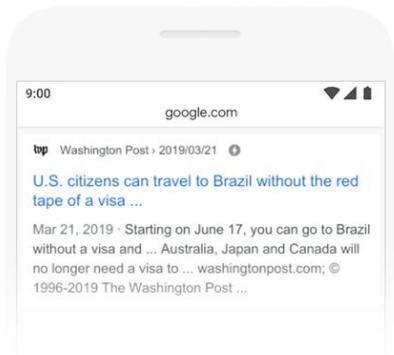
*Ashish Vaswani, Attention is All You Need, 2017.*

# BERT

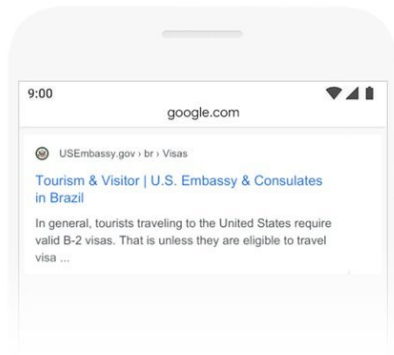


2019 brazil traveler to usa need a visa

BEFORE



AFTER

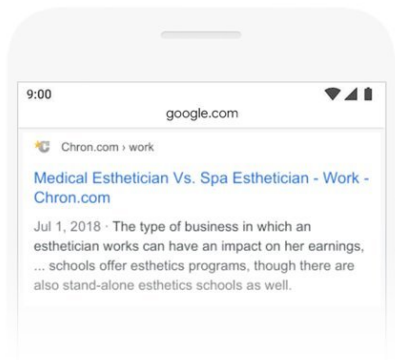


# BERT

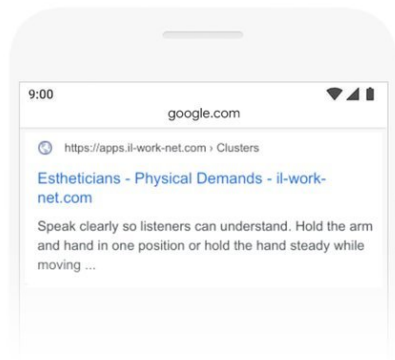


do estheticians stand a lot at work

BEFORE



AFTER





## Generative Pre-trained Transformer

- The first GPT (6/2018) was in the ELMo/BERT world (*very* roughly). It demonstrated real-world knowledge acquisition and long-range dependencies.

## Generative Pre-trained Transformer

- The first GPT (6/2018) was in the ELMo/BERT world (*very* roughly). It demonstrated real-world knowledge acquisition and long-range dependencies.
- GPT-2 and GPT-3 distinguished themselves by scaring people.

**Questions?**