NLP models and compositionality

Aleksandr Drozd

blackbird.pw

bkbrd

RIKEN Center For Computational Science,
High Performance Artificial Intelligence Systems Research Team

Tokyo Institute of Technology,
Dept. of Mathematical and Computing Science

March 20, 2020

Research areas/ projects

- Deep learning at scale in various domains
- Creating deep learning software ecosystem for Fugaku
  - Japan's flagship supercomputer being deployed as we speak
  - 150000 nodes with general-purpose ARM-based CPUs which look more like GPUs by performance characteristics.
- **Natural Language Processing / Computational Linguistics**
- Artificial Life and other attempts to pursue AGI :)

past: GPGPU, bioinformatics, sorting, geo-spatial, performance modeling

Natural Language Processing / Computational Linguistics

- interaction of computers and human language
- rapidly developing field
- one of the keys to solve AGI?
- one with huge appetite for compute resources: thousands of GPU-days to train one model.
- self-supervised pre-training / transfer learning is commonly used

So what is this talk about?

- compositionality as an angle to look at NLP models:
- a bit on NLP models / representations of textual data
- a bit on compositionality (and also contextuality)
- and how these are related

Defining “representations”

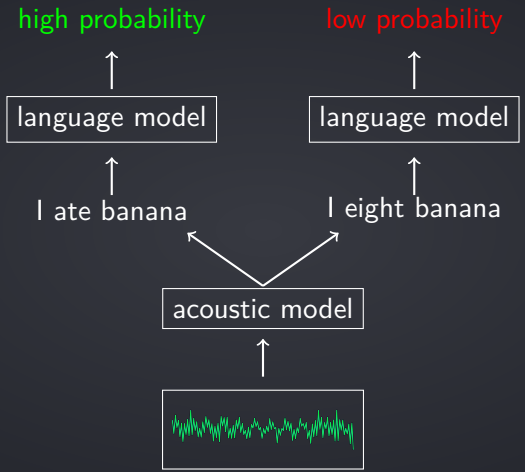
Representations in DL are sets of features that reflect the properties of the target phenomenon (at least partially). In this sense, input, output and intermediate weights can all be considered as representations.

- “representations” are often associated with pre-trained weights such as word2vec or BERT, but both end-to-end and transfer learning can be viewed as representation learning.
- end-to-end learning: representations are learned “on the job” directly;
- transfer learning: representations learned from one task are used for another;
- techniques largely overlap, but different focus;

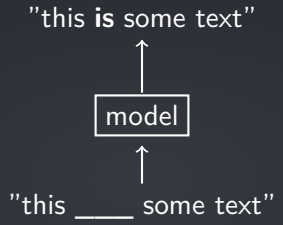
Confusing terms

- "model" in NLP: neural net used for a task or as pre-trained representation
- "language model"
  - explicit or implicit part of the model which relates to language knowledge
  - model (representation) trained on task of language modeling: BERT, ELMO, etc

Language model in sense I



Language model in sense II



language model as pre-training objective

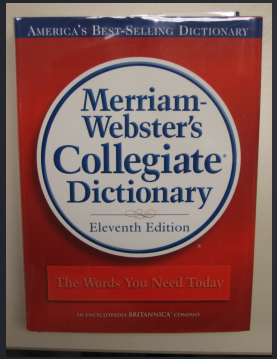
The principle of compositionality

> The meaning of a compound expression is a function of the meanings of its parts and of the way they are syntactically combined.
>
> [Partee, 1984]

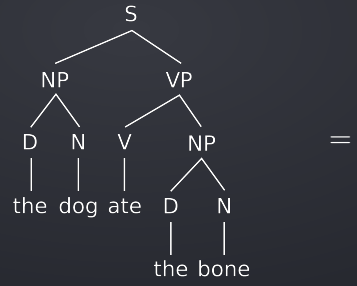
The principle of compositionality

The meaning of a compound expression is a function of the meanings of its parts and of the way they are syntactically combined.

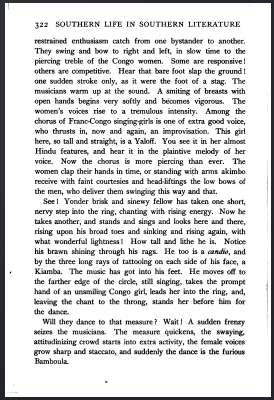
[Partee, 1984]



\+



=


322 SOUTHERN LIFE IN SOUTHERN LITERATURE

restrained enthusiasm catch from one bystander to another. They swing and bow to right and left, in slow time to the piercing treble of the Congo women. Some are responsive! others are competitive. Hear that bare foot slap the ground! one sudden stroke only, as it were the foot of a stag. The musicians warm up at the sound. A smiting of breasts with open hands begins very softly and becomes vigorous. The women's voices rise to a tremulous intensity. Among the chorus of Franc-Congo singing-girls is one of extra good voice, who thrusts in, now and again, an improvisation. This girl here, so tall and straight, is a Yaloff. You see it in her almost Hindu features, and hear it in the plaintive melody of her voice. Now the chorus is more piercing than ever. The women clap their hands in time, or standing with arms akimbo receive with faint courtesies and head-liftings the low bows of the men, who deliver them swinging this way and that.

See! Yonder brisk and sinewy fellow has taken one short, nervy step into the ring, chanting with rising energy. Now he takes another, and stands and sings and looks here and there, rising upon his broad toes and sinking and rising again, with what wonderful lightness! How tall and lithe he is. Notice his brawn shining through his rags. He too is a *canão*, and by the three long rays of tattooing on each side of his face, a Kiamba. The music has got into his feet. He moves off to the farther edge of the circle, still singing, takes the prompt hand of an unsmiling Congo girl, leads her into the ring, and, leaving the chant to the throng, stands her before him for the dance.

Will they dance to that measure? Wait! A sudden frenzy seizes the musicians. The measure quickens, the swaying, attitudinizing crowd starts into extra activity, the female voices grow sharp and staccato, and suddenly the dance is the furious Bamboula.

Productivity argument

Argument: humans can understand previously unseen complex expressions, so they must possess knowledge out of which the meanings of new expressions can be deduced.

Productivity argument

Argument: humans can understand previously unseen complex expressions, so they must possess knowledge out of which the meanings of new expressions can be deduced.

> *"I couldn't be more petrified if a wild Rhinoceros had just come home from a hard day at the swamp and found me wearing his pyjamas, smoking his cigars, and in bed with his wife." (Blackadder III)*

Compositional view of meaning representations in NLP

- Treat words as building blocks for sentences and phrases;
- Treat subword units as building blocks for words;
- Treat dimensions of meaning representations as semantic components;
- The objective is to find the right way to combine the building blocks. The best representation should yield best performance on downstream NLP tasks.

Problems with compositionality

- meaning of the whole does not always depend on the meaning of the constituents (“raining cats and dogs”);
- composition operation may involve a lot of world knowledge (“good car” vs “good party);
- the result of the composition operation has to make sense in the wider context (e.g. sentence in a text), and can be influenced by it.

Did we get it all backwards?

- Words that occur in similar contexts tend to have similar meanings. [Harris, 1954];
- You shall know a word by the company it keeps [Firth, 1957];
- For a large class of cases... the meaning of a word is its use in the language. [Wittgenstein, 1953].
- Never ask for the meaning of a word in isolation, but only in the context of a sentence. [Frege, 1884]

Did we get it all backwards?

- Words that occur in similar contexts tend to have similar meanings. [Harris, 1954];
- You shall know a word by the company it keeps [Firth, 1957];
- For a large class of cases... the meaning of a word is its use in the language. [Wittgenstein, 1953].
- Never ask for the meaning of a word in isolation, but only in the context of a sentence. [Frege, 1884]

If meanings of words come from sentences/texts, how can we use them to represent sentences/texts?

Contextuality vs compositionality: we **need** it both ways

- we do draw on prior knowledge to recognize and process most of what we hear;
- the context does also influence our interpretation, and can even be the **only** source of information.

Contextuality vs compositionality: we **need** it both ways

- we do draw on prior knowledge to recognize and process most of what we hear;
- the context does also influence our interpretation, and can even be the **only** source of information.

> We found a cute little wampimuk sleeping in a tree.
> [Lazaridou et al., 2014]

Storage vs computation trade-off

- Compositionality saves storage: given a vocabulary, we can account represent phrases, sentences, posts, tweets, etc.

Storage vs computation trade-off

- Compositionality saves storage: given a vocabulary, we can account represent phrases, sentences, posts, tweets, etc.

> *"I couldn't be more petrified if a wild Rhinoceros had just come home from a hard day at the swamp and found me wearing his pyjamas, smoking his cigars, and in bed with his wife."*

Storage vs computation trade-off

- Compositionality saves storage: given a vocabulary, we can account represent phrases, sentences, posts, tweets, etc.

> *"I couldn't be more petrified if a wild Rhinoceros had just come home from a hard day at the swamp and found me wearing his pyjamas, smoking his cigars, and in bed with his wife."*

- Full compositionality could result in too much computation, "unless the balance between storage and computation is upset in favor of storage" [Baggio et al., 2012]

NLP models though a prism of compositionality

- how to select the representation components for composition?
- how to compose them?
- bonus question: how meaningful are individual dimensions?

Problem 1: selecting representations for composition

- manually
- using syntactic/morphological parsers
- constrained by network design (CNN, LSTM)
- learned mechanism: e.g. attention or controller for tree traversal

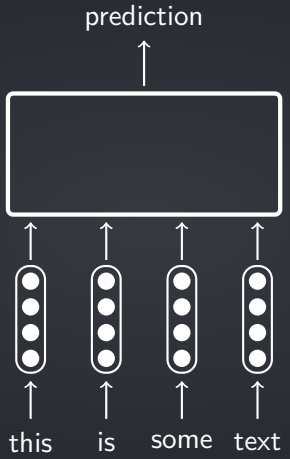
Problem 2: composition mechanisms

- explicit compositional operations
- learned with neural networks

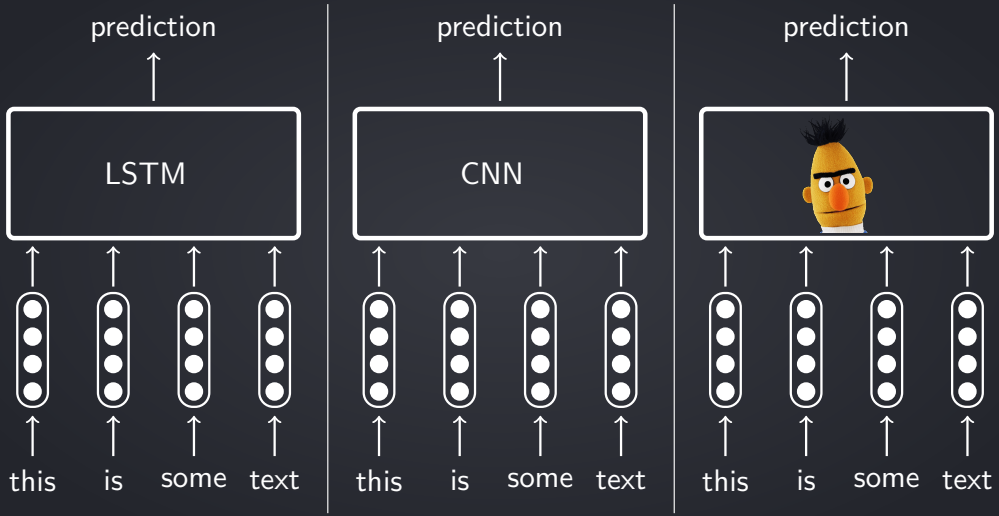
Embeddings


```
from torch.nn import Embedding
```

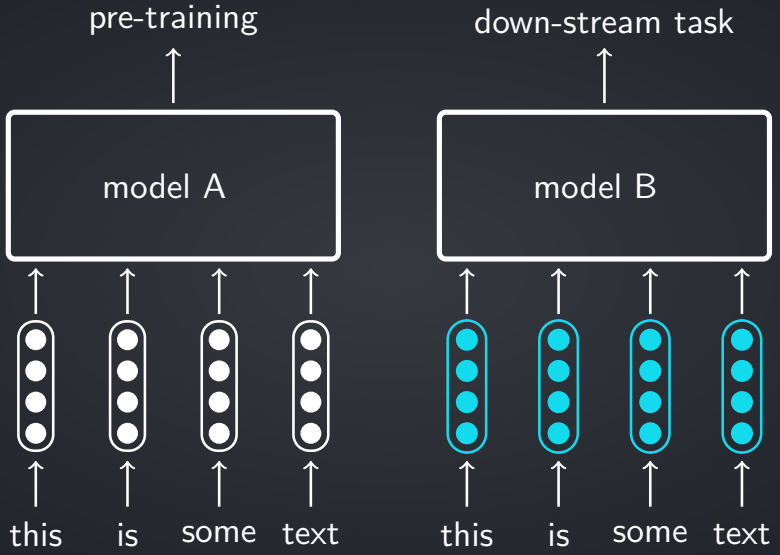
Embeddings



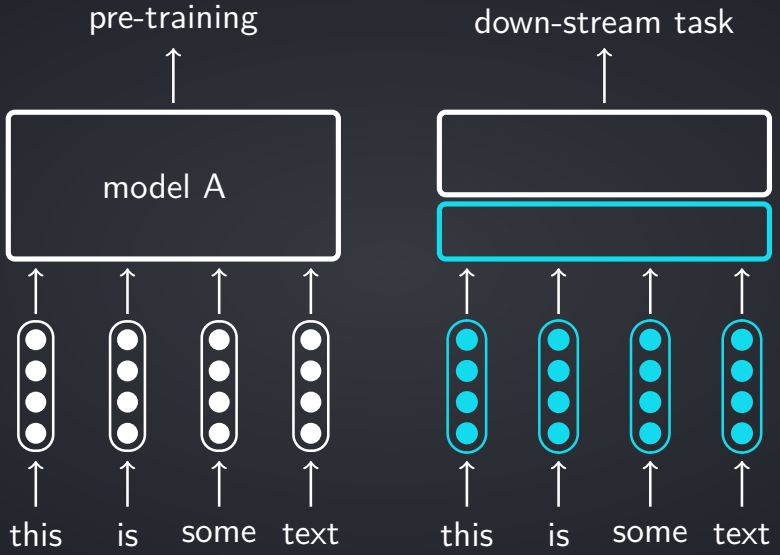
Embeddings


prediction
LSTM
this
is
some
text
prediction
CNN
this
is
some
text
prediction
this
is
some
text

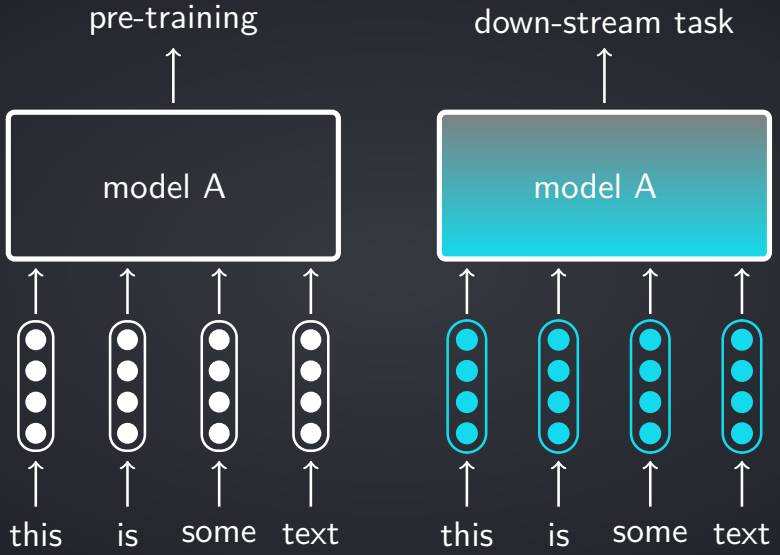
Transfer learning



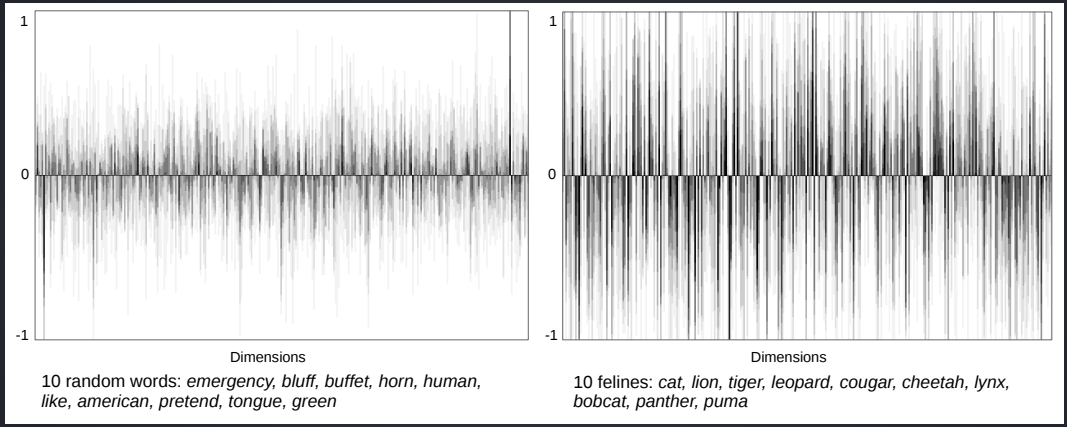
Transfer learning



Transfer learning



Components of meaning

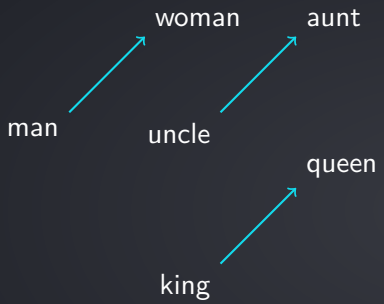


10 random words: *emergency, bluff, buffet, horn, human, like, american, pretend, tongue, green*

10 felines: *cat, lion, tiger, leopard, cougar, cheetah, lynx, bobcat, panther, puma*

[Gladkova and Drozd, 2016]

Inducing meaning shifts by analogy



$king - man + woman = queen$
(BTW this is not working as advertised!!! [Rogers et al., 2017])

[Mikolov et al., 2013b]

Algebraic methods of combining distributional representations [Mitchell and Lapata, 2010]

model	function
Additive	$p = u + v$
Weighted additive	$p = \alpha u + \beta v$
Multiplicative	$p = uv$

Addition

- classic [Widdows, 2004] and intuitive method, where all components of source vectors contribute towards the resulting representation;
- [Mikolov et al., 2013a]: trained Skip-gram representations of phrases are similar to the results of the addition of the corresponding token vectors.

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Averaging

- simple mean [Hill et al., 2016, Adi et al., 2017]
- weighted mean, e.g. with with TF-IDF scores [Boom et al., 2015, Corrêa Júnior et al., 2017]
- The overall pattern of results is the same for sums and average [White et al., 2015]

> Sentence representations based on averaged word vectors are **surprisingly effective**, and encode a non-trivial amount of information regarding sentence length. [Adi et al., 2017]

Multiplication

- **Pointwise multiplication**: keeps only the components which had corresponding non-zero values
- not clear how semantically motivated that is, to just get rid of non-shared components.

> car has many semantic components not present in blue, but arguably should keep them in the representation of blue car

- less popular than sums and means, but was the best in phrase similarity and paraphrase tasks [Blacoe and Lapata, 2012].
- More discussion: [Mitchell and Lapata, 2010, Clarke, 2012, among others]

Attempts to explicitly introduce algebraic compositionality in training objectives

> in a good model of distributive semantics representation and composition must go hand in hand, i.e., **they must be mutually learned.** [Blacoe and Lapata, 2012]

- linear combination [Peng and Gildea, 2016]
- weighted addition [Fyshe et al., 2015]
- [Rudolph and Giesbrecht, 2010] propose a semantic space consisting of quadratic matrices rather than vectors, with matrix multiplication as the only compositional operation

Pros and cons of simple algebraic methods

- **Con:** not expressive;
- **Con:** all words contribute equally;
- **Con:** no syntactic finesse;

- **Pro:** trivial to implement, fast and often works well enough!

Problem for simple algebraic methods: language is not commutative!

- It was not the sales manager who hit the bottle that day, but the office worker with the serious drinking problem.
- That day the office manager, who was drinking, hit the problem sales worker with the bottle, but it was not serious.

[Landauer et al., 1997]

Particularly poor results for bag-of-word version of paragraph vectors
[Le and Mikolov, 2014]

Tensor products

- the core idea: different words should contribute differently to the resulting representation
- matrices capture only 2-way cooccurrences
- tensors are multidimensional arrays that enable accounting for multi-way cooccurrence

Discussion: [Clark and Pulman, 2007, Mitchell and Lapata, 2010, Clarke, 2012, Van de Cruys et al., 2013]

Success varies by relation type and model

Correlation with human phrase similarity judgements: the basic co-occurrence sparse model [Mitchell and Lapata, 2010]

Model	Adjective-Noun	Noun-Noun	Verb-Object
Additive	.36	.39	.30
Kintsch	.32	.22	.29
Multiplicative	.46	.49	.37
Tensor product	.41	.36	.33
Convolution	.09	.05	.10
Weighted Additive	.44	.41	.34
Dilation	.44	.41	.38
Target Unit	.43	.34	.29
Head only	.43	.17	.24
Humans	.52	.49	.55

Adjectives as weight matrices on noun meanings [Baroni and Zamparelli, 2010]

- **adjective-specific linear map** (alm) method: an adjective-noun phrase is represented by multiplying an adjective weight matrix with a noun (column) vector
- experiments with phrase generation with SVD embeddings: how close is the generated phrase to a pre-computed one?
- composed phrases shown to be similar to observed phrases, e.g. the real neighbor of the phrase *common understanding* was *common approach*, while composition yielded *common vision*

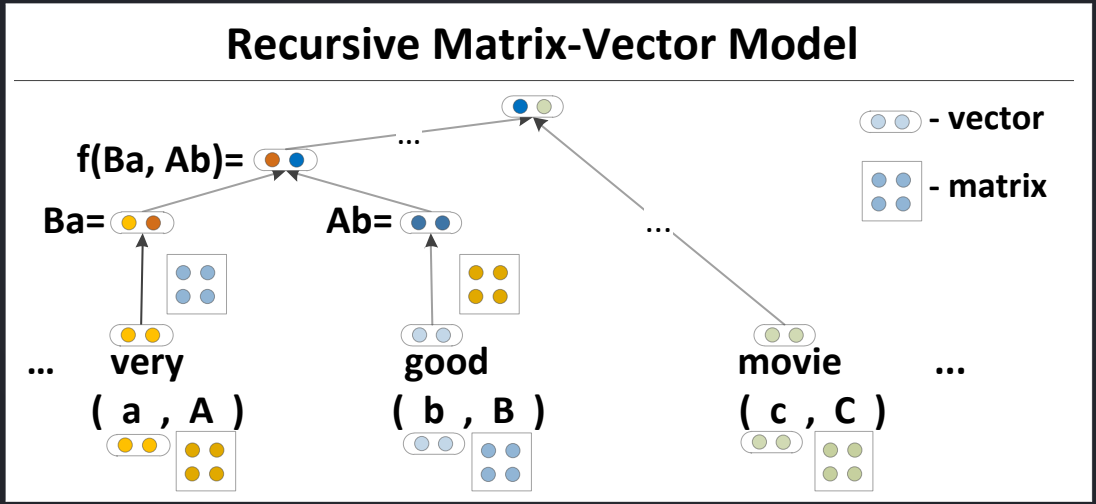
Matrix-vector RNN [Socher et al., 2012]

- builds on the above, but with vector-matrix representations for all words: a, b are vectors, A, B are matrices
- input-specific composition function p:

$$p = g\ (\ W\ (\begin{bmatrix} Ba \\ Ab \end{bmatrix})\ )$$

- for g a non-linear function is chosen, but could be identity function.

Example: “a very good movie” [Socher et al., 2012]



Pros and cons

- **Pro:** expressive;
- **Pro:** linguistically motivated;

Pros and cons

- **Pro:** expressive;
- **Pro:** linguistically motivated;
- **Con:** much explicit linguistic knowledge (and math) required;
- **Con:** linguistic motivation not entirely clear
- **Con:** computationally expensive;
- **Con:** not necessarily outperforming simple algebraic baselines [Mitchell and Lapata, 2010]

Other proposals

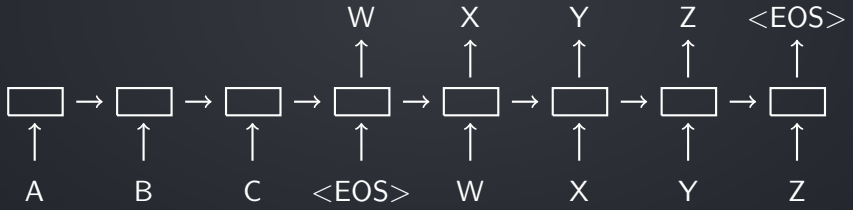
- sentence vectors as Kroenecker products, which enables applying relational words as filters on nouns
  [Grefenstette and Sadrzadeh, 2011, Grefenstette and Sadrzadeh, 2015];
- sentence vectors based on orthogonal basis of the subspace spanned by a word and its surrounding context, efficient and competitive with recent neural baselines [Yang et al., 2019].

Compositionality as a by-product

- DL models for NLP typically receive as input representations for individual words / subword elements;
- At some point all the input information passes through an element serving as an information bottleneck;
- The resulting representation must be compositional in some way.

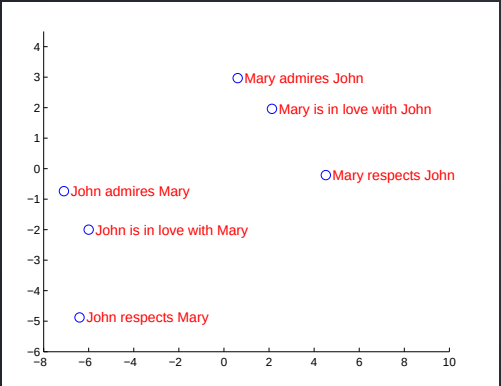
End-to-end sequence encoding [Sutskever et al., 2014]

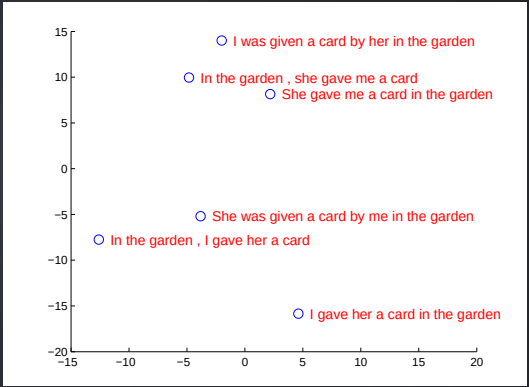
One LSTM reads the input sequence "ABC", one timestep at a time, to obtain large fixed-dimensional vector representation, and then to use another LSTM to extract the output sequence "WXYZ" from that vector.



End-to-end sequence encoding [Sutskever et al., 2014]

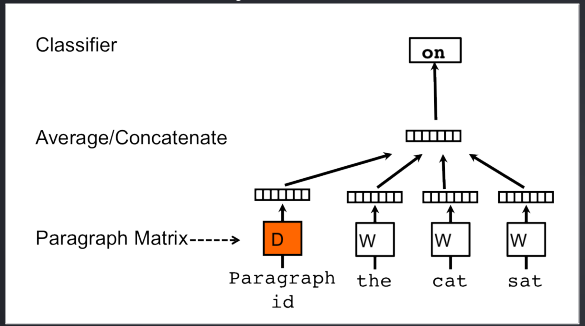
2-dimensional PCA projection of the LSTM hidden states for the phrases in the figures.



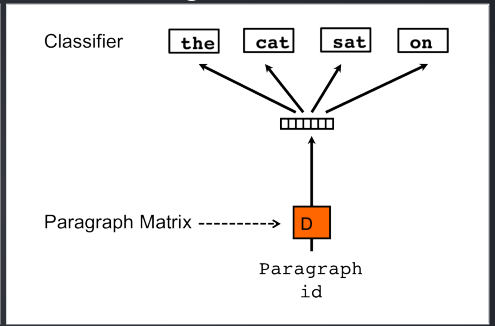


Paragraph vectors [Le and Mikolov, 2014]

Distributed Memory version



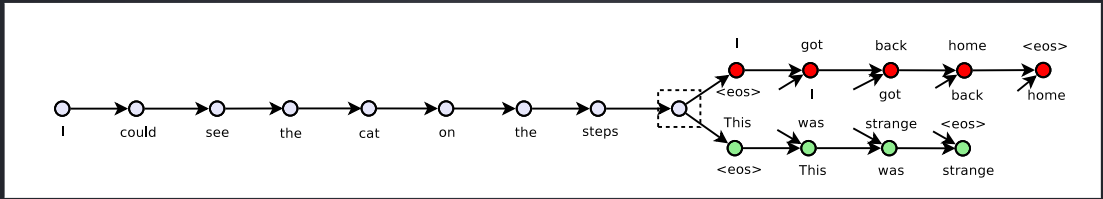
Distributed Bag-of-words version



At prediction time, the paragraph vectors are inferred by fixing the word vectors and training the new paragraph vector until convergence.

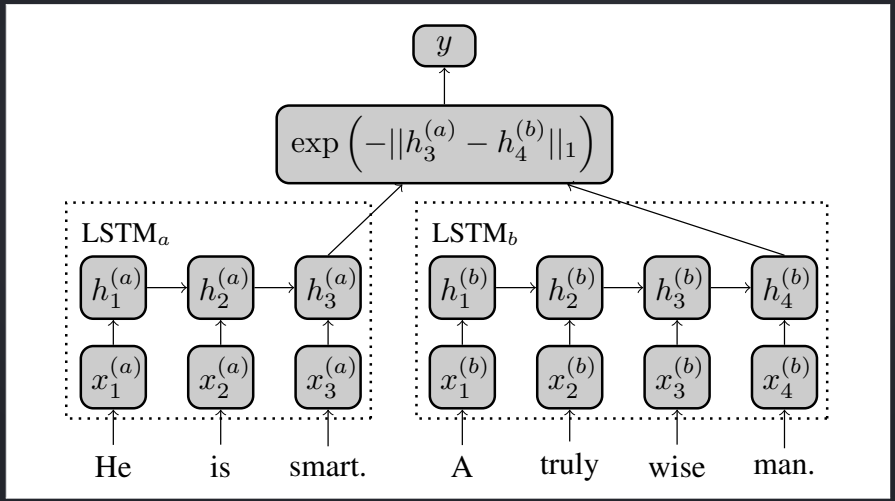
Encoder/decoder: Skip-Thought [Kiros et al., 2015]

Given a tuple of sentences (s_i-1, s_i , s_i+1), the sentence s_i is encoded and used to reconstruct s_i-1 and next sentence s_i+1.
Example tuple: (*I got back home. I could see the cat on the steps. This was strange.*)



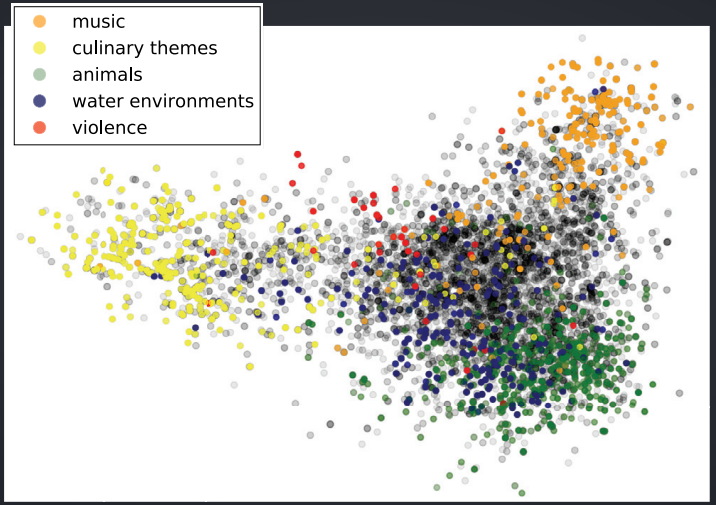
Siamese LSTM for sentence similarity [Mueller and Thyagarajan, 2016]

The final hidden state of LSTM acts as representation of sentences and is used to compare them.

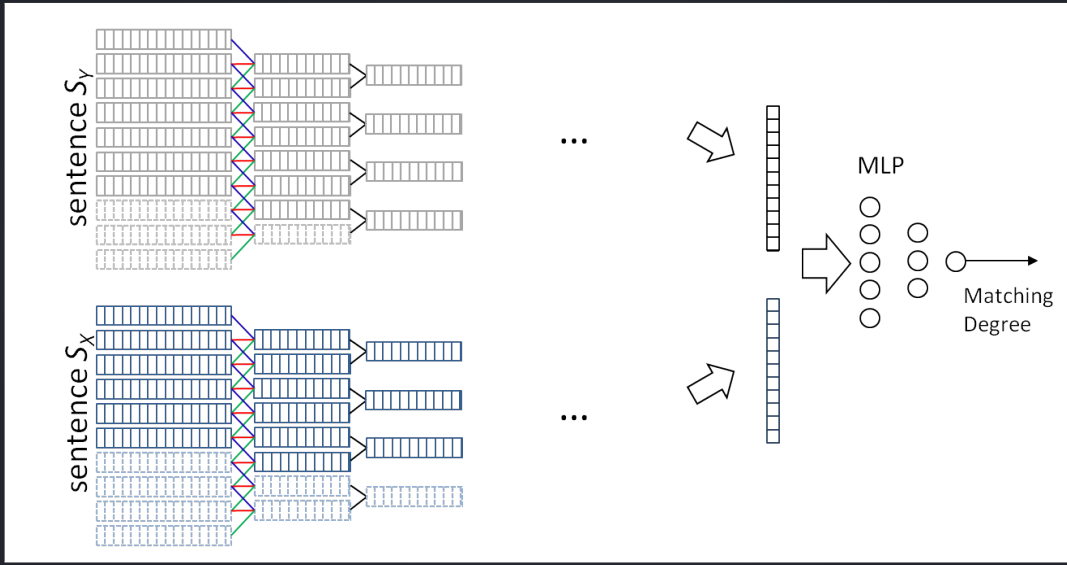


Siamese LSTM [Mueller and Thyagarajan, 2016]

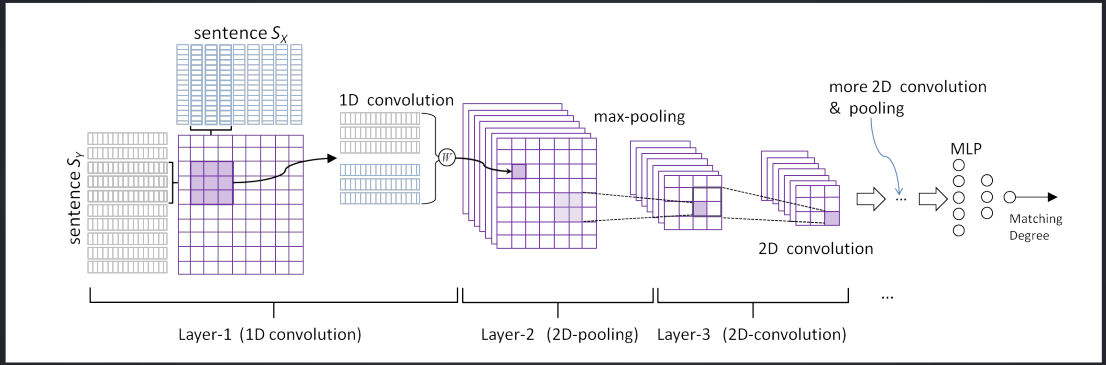
t-SNE representation of the SICK dataset



Sentence matching: Architecture 1 [Hu et al., 2014]



Sentence matching: Architecture 2 [Hu et al., 2014]



Pros and cons

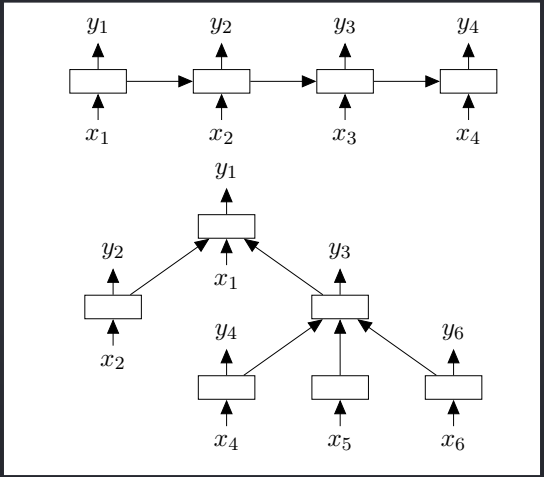
Pros:

- expressive
- seems to work on its own

Cons:

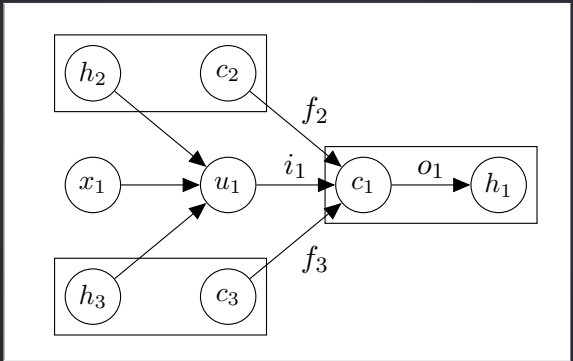
- sensitive to dozens of parameters;
- computationally expensive;
- lack of syntactic finesse, unless cheating;
- not clear that it's really not compositional, as long as the tokenized words are fed one by one, and especially if word embeddings are used;
- since it is a black box, hard to understand how to improve them.

Tree-LSTM: LSTM generalized to tree-structured network topologies [Tai et al., 2015]



Tree-LSTM [Tai et al., 2015]

- gating vectors and memory cell updates can depend on several child units;
- there is a forget gate for each child.



Honourable Mentions

InferSent [Conneau et al., 2017] - BiLSTM with max-pooling trained on NLI task

Sent2Vec [Pagliardini et al., 2018] Averaging embeddings of words and n-grams of words in a sentence.

Quick Thoughts [Logeswaran and Lee, 2018] - Simplification of Skip Though: classifying whether candidate sentences belongs to adjacent sentences.

a la carte [Khodak et al., 2018] - learn a matrix to predict target words from context words and use it to predict representations for rare targets or n-grams

Universal Sentence Encoder [Cer et al., 2018] - two models (transformer and FF layers on top of averaged bi-grams) + multi-task training

Birds-eye view on NLP models

- High-level structure: classifier, seq-2-seq, Siamese, etc
- Building blocks: LSTM, CNN, self-attention etc
- Training objective: language modeling, labeling, generation etc
- Transfer learning: removing parts, adding parts, gradual freezing while fine-tuning etc.
- Down-stream task.

High-level structure

- Classifier (in lack of better word)
  - classify whole text
  - label every token (POS, NER etc.)
- Siamese networks
- Sequence to sequence models
- Generative adversarial network aka GAN

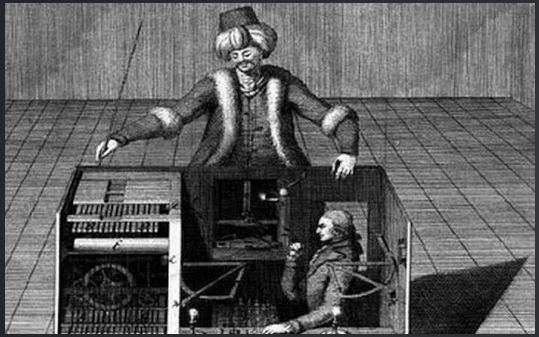
Building blocks

- embedding layer (more on it later)
- dense
- LSTM cells
- convolutional filters
- [self] attention
- other exotic beasts

Objectives

- text classification (e.g. sentiment or effect stock prices)
- sequence labeling
- machine translation
- other forms of text generation
- language modeling: predicting missing words
  - sequential
  - masked
  - bag of word / skipgram-like

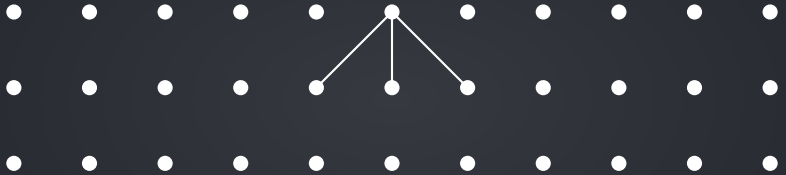
NLP models through a prism of compositionality, again



- how to select the representation components for composition?
- how to compose them?
- this is not to say that a given model as a whole can't do something which is not decomposable to composition in each building blocks

Let's look at convnets

what to compose:



"field of view" is limited

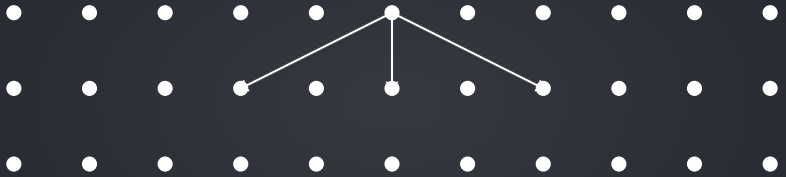
Let's look at convnets

what to compose:

"field of view" is limited

Addressing this with dilated convolutions

[Kalchbrenner et al., 2016, Strubell et al., 2017]



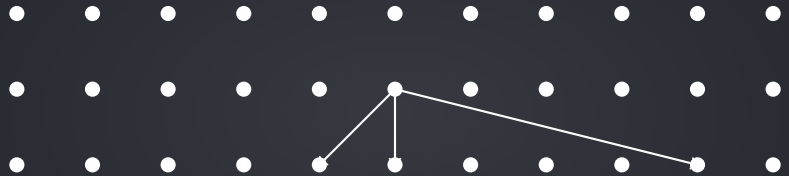
but now we might miss some important pieces

Addressing this with dilated convolutions

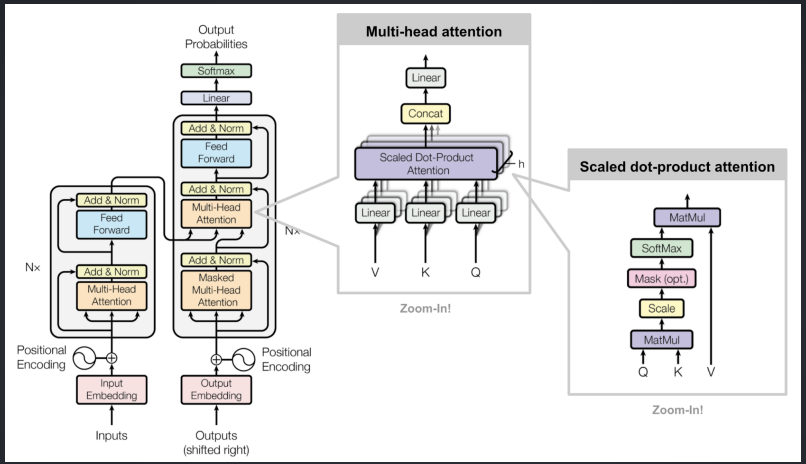
[Kalchbrenner et al., 2016, Strubell et al., 2017]

but now we might miss some important pieces

Would be nice to pick up just relevant bits

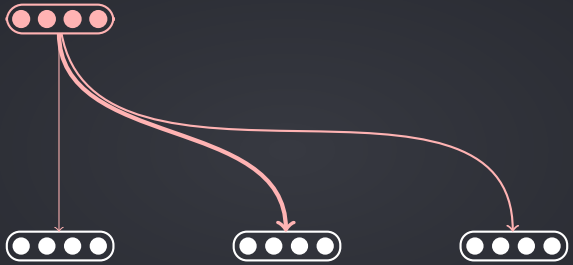


Alone Come Transformers [Vaswani et al., 2017]

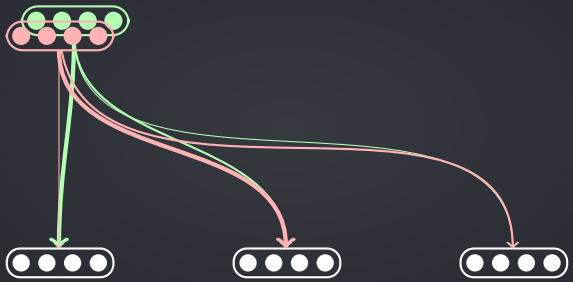


- How to select: self-attention
- How to compose: addition in one head, MLP across heads

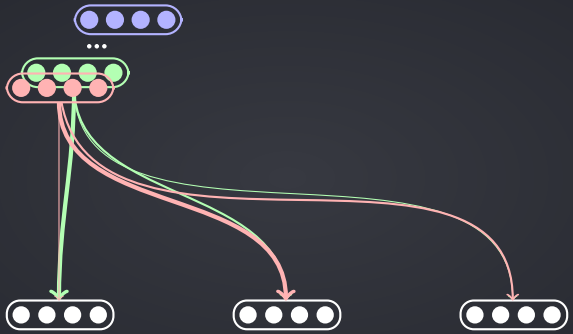
Attention to input



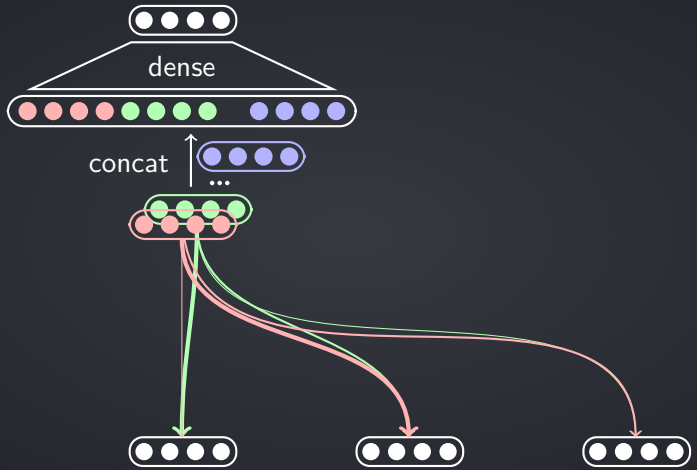
Multi-head attention



Multi-head attention



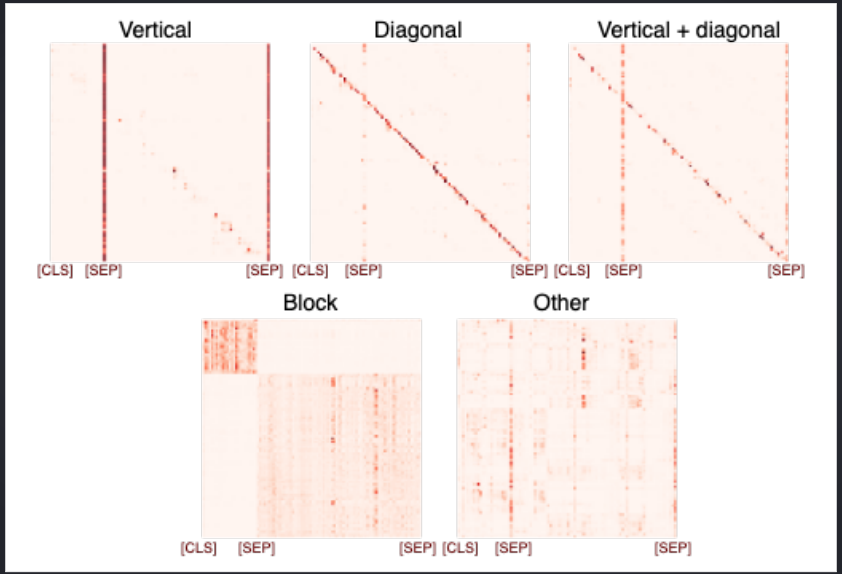
Concatenate



But does it actually "look" at something sensible?

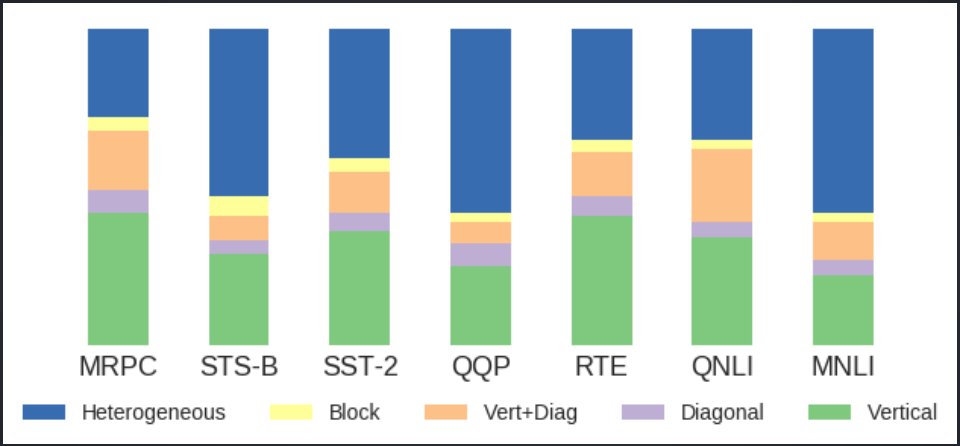
- There's a study showing that syntax trees can be recovered [Hewitt and Manning, 2019]
- Also studies showing that this information is not used in GLUE tasks [Kovaleva et al., 2019]

Attention pattern types in BERT [Kovaleva et al., 2019]



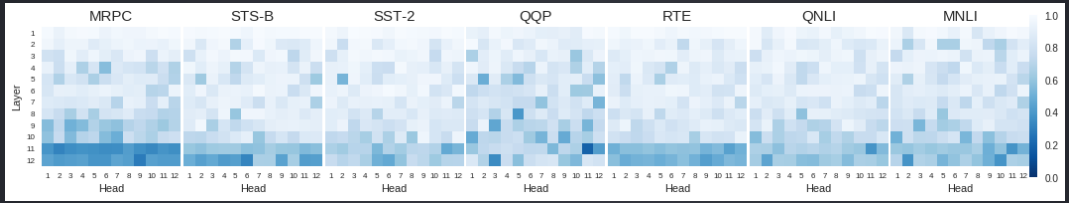
Typical self-attention classes used for training a neural network.

Ratio of different self-attention patterns in BERT for GLUE tasks [Kovaleva et al., 2019]



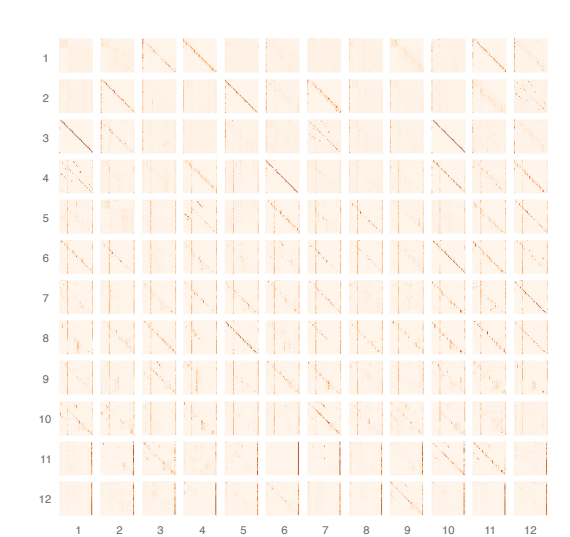
the ratio of potentially useful attention maps varies by task, but in most cases it is less than half of BERT.

Differences in fine-tuned and pre-trained BERT self-attention heads [Kovaleva et al., 2019]



The last two layers change the most in fine-tuning. Ideally, this should reflect the kinds of linguistic relations important for a given task.

Fine-tuning can reinforce non-informative attention patterns, as in this QNLI example
[Kovaleva et al., 2019].



Pre-trained vs fine-tuning [Kovaleva et al., 2019]

Dataset	Pre-trained	Fine-tuned, initialized with normal distr.	Fine-tuned, initialized with pre-trained	Metric	Size
MRPC	0/31.6	81.2/68.3	87.9/82.3	F1/Acc	5.8K
STS-B	33.1	2.9	82.7	Acc	8.6K
SST-2	49.1	80.5	92	Acc	70K
QQP	0/60.9	0/63.2	65.2/78.6	F1/Acc	400K
RTE	52.7	52.7	64.6	Acc	2.7K
QNLI	52.8	49.5	84.4	Acc	130K
MNLI-m	31.7	61.0	78.6	Acc	440K

BERT with completely random self-attention weights still works pretty well on all tasks, except textual similarity!

Still trying to figure out what's going on

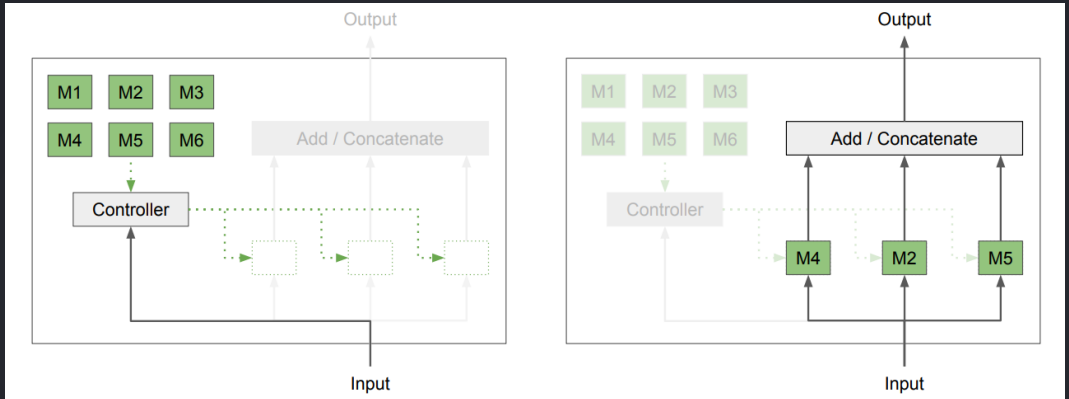
Explosion of papers doing different ways of analysis/ probing tasks etc on BERT-like models. Good survey by [Rogers et al., 2020]

Modular architectures

architecture components as task components

- motivated by analogies with brain modules;
- intuitively, should help with catastrophic forgetting and domain adaptation.

Modular architecture

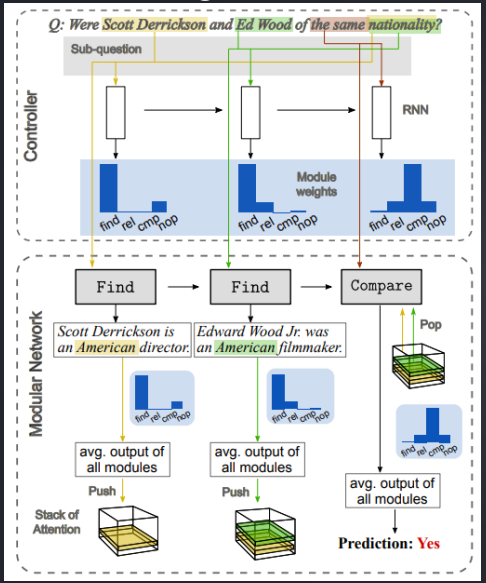


(a) Based on the input, the controller selects K modules from a set of M available modules. In this example, $K = 3$ and $M = 6$.

(b) The selected modules then each process the input, with the results being summed up or concatenated to form the final output of the modular layer.

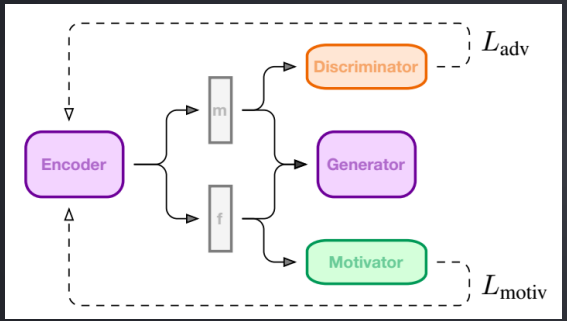
Continuous arrows represent data flow, while dotted arrows represent flow of modules [Kirsch et al., 2018]

Self-assembling modular network for multi-hop QA



Modular network with a controller (top) and the dynamically-assembled modular network (bottom). At every step, the controller produces a sub-question vector and predicts a distribution to weigh the averages of the modules' outputs. [Jiang and Bansal, 2019]

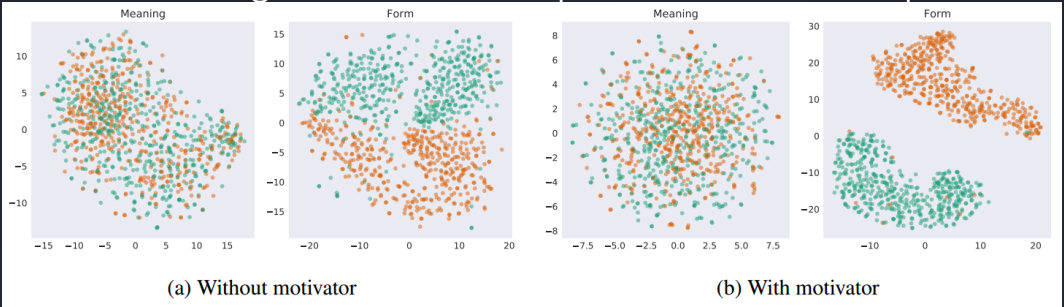
Direction 2: learning separate aspects of the input



- Encoder decomposes the input into "meaning" and "form" vectors;
- Generator takes those vectors to produce an output sequence
- Discriminator tries to classify form based on meaning ($>$ encoder is penalized to make this hard)
- Motivator tries to classify form based on form (encoder is penalized to make this easy)

[Romanov et al., 2019]

Dissociated meaning and form vectors [Romanov et al., 2019]



(a) Without motivator (b) With motivator

Form and meaning embeddings for news (green) and scientific (orange) article headlines

Aye, sir. (EME)	→	Yes, sir. (CE)
Fare thee well, my lord (EME)	→	Fare you well, my lord (CE)
This guy will tell us everything. (CE)	→	This man will tell us everything. (EME)
I've done no more to caesar than you will do to me. (CE)	→	I have done no more to caesar than, you shall do to me. (EME)

Decoding Shakespeare from Early Modern English (EME) into contemporary English (CE)

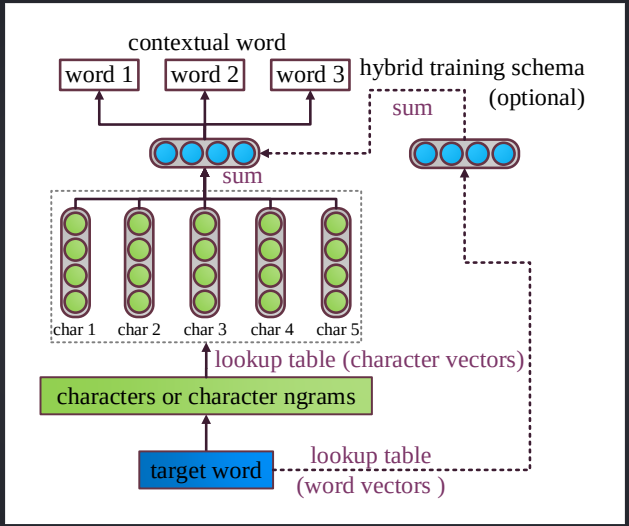
Sub-word level compositionality

- *pencils* = *pencil* + *multiple of*
- essentially the sub-word compositionality allows **vocabulary to not be fixed**
- good fore for rare words
- or languages without spaces

FastText [Bojanowski et al., 2017]

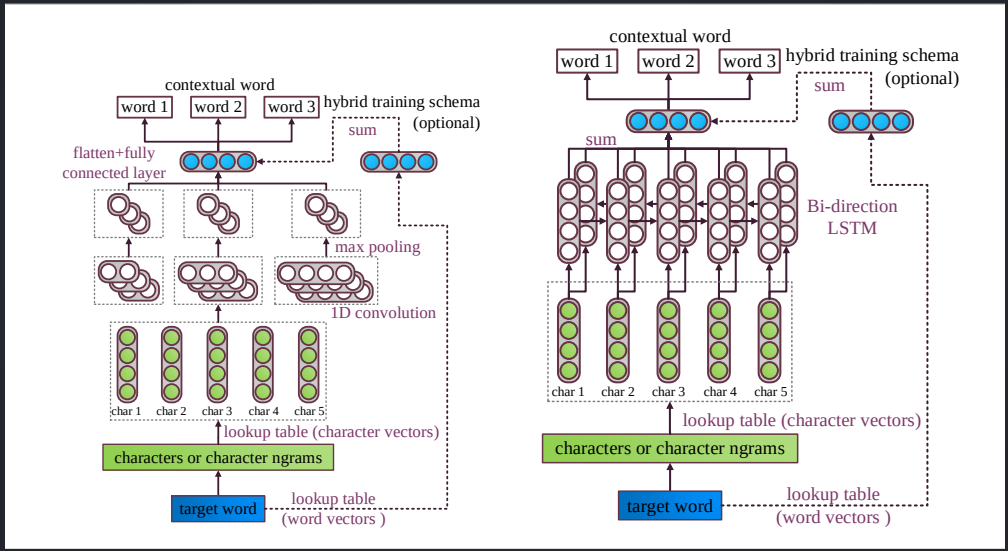
- one of the first popular subword models
- composition function is defined as $f(w) = \sum_{g \in \mathcal{G}_w} \vec{g}$, where g is the character n-gram, and $\vec{g}$ is its corresponding n-gram vector with length N.
- $\mathcal{G}_w$ is the set of character n-grams for word w. For example, when $n = 3$, $\mathcal{G}_w$ for word "bigger" is defined as `<bi, big, igg, gge, ger, er>`.

Case of FastText [Bojanowski et al., 2017]

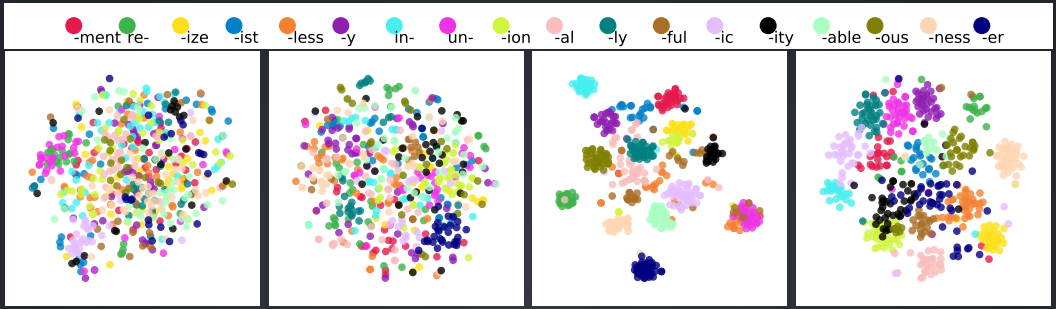


characters and character ngrams belonging to a word

Neural net as compositional function [Li et al., 2018]



Neural net as compositional function



(a) Skip-Gram (b) FastText$_{subword}$ (c) CNN$_{subword}$ (d) RNN$_{subword}$

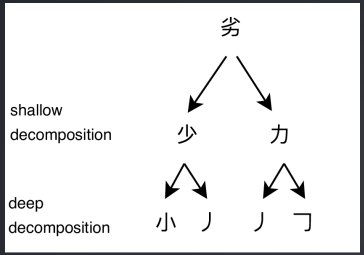
t-SNE projection of representations

- performing on par/better on a range of morphological tasks even when train on char-unigrams
- while taking order of magnitude smaller memory footprint

Subcharacter composition

- Subword embeddings provide building blocks for rare vocabulary;
- What about logographic languages?

Kanji decomposition



Original sentence:
彼は数学の天才だが、パソコンには劣る。

Shallow decomposition:
(彳 皮) 彼は (娄 攵) 数 (丷 冖 子) 学の
(一 大) 天 (才) オだが、パソコンには (少 力) 劣る。

Translation:
He is a genius at math, but will lose to a computer.

for language modelling: [Nguyen et al., 2017]

Subcharacter embeddings in Japanese [Karpinska et al., 2018]

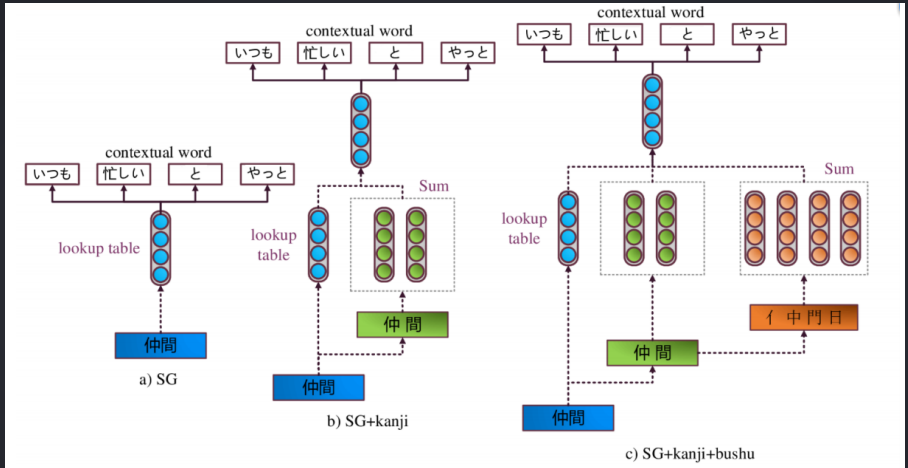


Figure 2: Model architecture of SG, SG+kanji, and SG+kanji+bushu. Example sentence: いつも 忙しい 仲間 と やっと 会え た (“I have finally met with my busy colleague.”), window size 2.

Are the bushus worth it? [Karpinska et al., 2018]

- Some improvement over vanilla skip-gram for kanji-rich newspaper domains on tasks involving single-character words;
- A new Japanese analogy dataset and improved Japanese similarity dataset (`http://vecto.space/projects/`).

Shared semantic space for kanji, kana and bushu

疒 *yamaidare* (the roof from illness)
患(sickness) 症(disease) 妊 (pregnancy)
臓 (internal organs, bowels) 腫 (tumor)
インフルエザ (influenza)
関節リウマチ (articular rheumatism)
リューマチ (rheumatism) リウマチ(rheumatism)
メタボリックシンドローム (metabolic syndrome)

Example bushu: closest single kanji (upper row) and multiple kanji/katakana (lower row) for SG+kanji+bushu model [Karpinska et al., 2018].

Recent sub-word level approaches

Byte Pair encoding [Sennrich et al., 2016]: start from a vocabulary of individual bytes/chars; merge together most frequent pairs; repeat until vocabulary size limit is reached. Rooted in

Wordpiece encoding [Kudo, 2018]: almost identical to BPE. merges the bigram that, when merged, would increase the likelihood of a unigram language model trained on the training data.

- The overall idea is to have common words in once peace, and decompose rare words into whatever pieces available.

Recent sub-word level approaches

original:

```
hi, how do you do nonetkn?
```

BertWP:

```
'hi', ',', 'how', 'do', 'you', 'do', 'none', '##t', '##k', '##n', '?'
```

BERT-like models

Recap: background

- Compositionality - intuitive idea, but lots of theoretical issues
- tension between compositionality and contextuality

  - And don't go bothering him – he's never really re-covered from Fanny's death. [COCA]
  - The re-covered books were three times more likely to be checked out than those with plain cloth covers. [COCA]

- Directly relates to problems which are hard to current SOTA models

Recap: compositionality methods

- macro level
  - black box model with some sort of bottlneck
- micro level
  - simple algebraic methods
  - syntax-aware methods
  - building blocks od DNN

Recap: compositionality at different levels

- distributional representations of phrases, sentences, paragraphs and texts
- subword-level embeddings
- character and subcharacter embeddings

- identifying meaning components: interpretable dimensions and controllable meaning shifts

Challenges for compositionality

- It's still not clear what a representation of a complex unit is even supposed to be
- Distributional representations and composition functions for function words.
- **Informative** evaluation metrics
- More error analysis
- Understanding based on both compositionality and contextuality

Special thanks to collaborators:

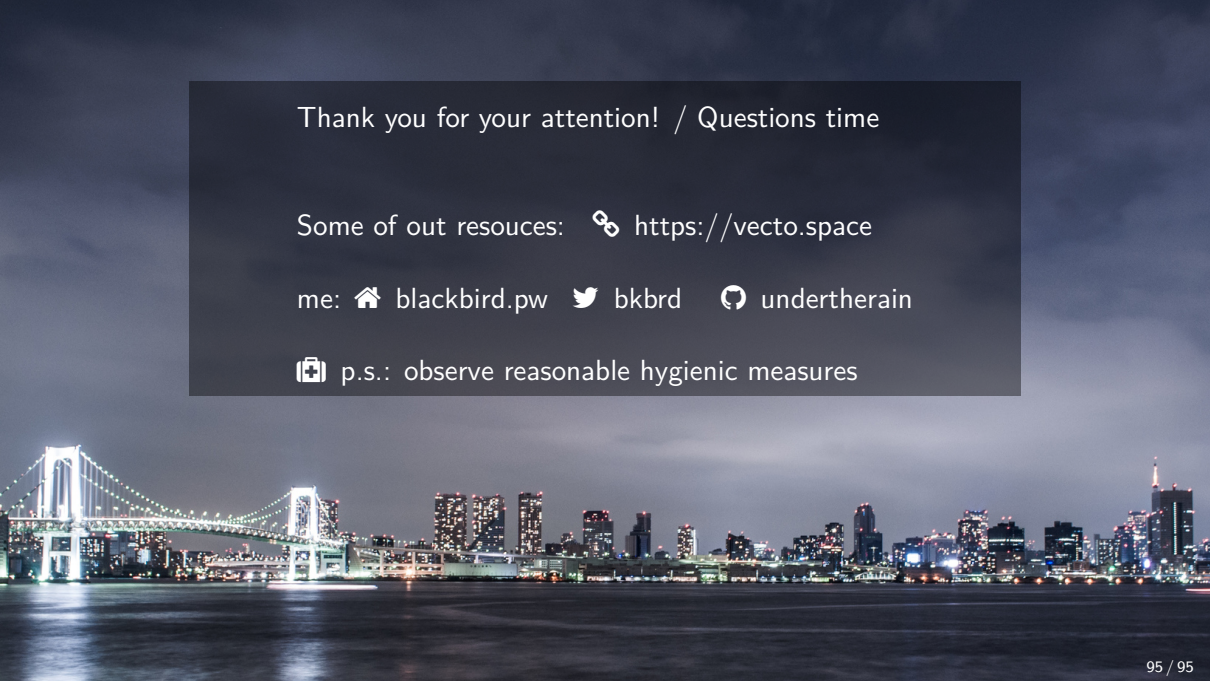
Anna Rogers, Bofang Li, Marzena Karpinska, Anna Rumshisky, Amir Bakarov, Roman Suvorov, Shashwath Hosur Ananthakrishna

Thank you for your attention! / Questions time

Some of out resouces: https://vecto.space

me: blackbird.pw bkbrd undertherain

p.s.: observe reasonable hygienic measures



Bibliography I

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., and Goldberg, Y. (2017).
Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks.
In *ICLR*, pages 1–13, Toulon, France.

Baggio, G., Van Lambalgen, M., and Hagoort, P. (2012).
The processing consequences of compositionality.
In *The Oxford Handbook of Compositionality*, pages 655–672. Oxford University Press.

Baroni, M. and Zamparelli, R. (2010).
Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space.
In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, MIT, Massachusetts, USA, 9-11 October 2010.

Blacoe, W. and Lapata, M. (2012).
A Comparison of Vector-based Representations for Semantic Composition.
In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 546–556, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017).
Enriching Word Vectors with Subword Information.
Transactions of the Association for Computational Linguistics, 5(0):135–146.

Boom, C. D., Canneyt, S. V., Bohez, S., Demeester, T., and Dhoedt, B. (2015).
Learning Semantic Similarity for Very Short Texts.
In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1229–1234.

Bibliography II

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018).
Universal Sentence Encoder.
arXiv:1803.11175 [cs].

Clark, S. and Pulman, S. (2007).
Combining Symbolic and Distributional Models of Meaning.
In *AAAI Spring Symposium: Quantum Interaction*, pages 52–55.

Clarke, D. (2012).
A context-theoretic framework for compositionality in distributional semantics.
Computational Linguistics, 38(1):41–71.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017).
Supervised Learning of Universal Sentence Representations from Natural Language Inference Data.
In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 7–11, 2017. Association for Computational Linguistics.

Corrêa Júnior, E. A., Marinho, V. Q., and dos Santos, L. B. (2017).
NILC-USP at SemEval-2017 Task 4: A Multi-view Ensemble for Twitter Sentiment Analysis.
In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 611–615. Association for Computational Linguistics.

Firth, J. R. (1957).
A synopsis of linguistic theory 1930-55.
1952-59:1–32.

Bibliography III

Fyshe, A., Wehbe, L., Talukdar, P. P., Murphy, B., and Mitchell, T. M. (2015).
A Compositional and Interpretable Semantic Space.
Proceedings of the NAACL-HLT, Denver, USA.

Gladkova, A. and Drozd, A. (2016).
Intrinsic evaluations of word embeddings: What can we do better?
In *Proceedings of The 1st Workshop on Evaluating Vector Space Representations for NLP*, pages 36–42, Berlin, Germany. ACL.

Grefenstette, E. and Sadrzadeh, M. (2011).
Experimental support for a categorical compositional distributional model of meaning.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404. Association for Computational Linguistics.

Grefenstette, E. and Sadrzadeh, M. (2015).
Concrete Models and Empirical Evaluations for the Categorical Compositional Distributional Model of Meaning.
Computational Linguistics, 41(1):71–118.

Harris, Z. (1954).
Distributional structure.
Word, 10(23):146–162.

Hewitt, J. and Manning, C. D. (2019).
A structural probe for finding syntax in word representations.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Bibliography IV

Hill, F., Cho, K., and Korhonen, A. (2016).

Learning Distributed Representations of Sentences from Unlabelled Data.

In *Proceedings of NAACL-HLT 2016*, pages 1367–1377, San Diego, California, June 12-17, 2016. Association for Computational Linguistics.

Hu, B., Lu, Z., Li, H., and Chen, Q. (2014).

Convolutional Neural Network Architectures for Matching Natural Language Sentences.

In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2042–2050. Curran Associates, Inc.

Jiang, Y. and Bansal, M. (2019).

Self-Assembling Modular Networks for Interpretable Multi-Hop Reasoning.

In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4464–4474, Hong Kong, China. Association for Computational Linguistics.

Kalchbrenner, N., Espeholt, L., Simonyan, K., van den Oord, A., Graves, A., and Kavukcuoglu, K. (2016).

Neural machine translation in linear time.

Karpinska, M., Li, B., Rogers, A., and Drozd, A. (2018).

Subcharacter Information in Japanese Embeddings: When Is It Worth It?

In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 28–37, Melbourne, Australia. Association for Computational Linguistics.

Khodak, M., Saunshi, N., Liang, Y., Ma, T., Stewart, B., and Arora, S. (2018).

A La Carte Embedding: Cheap but Effective Induction of Semantic Feature Vectors.

In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Melbourne, Australia. Association for Computational Linguistics.

Bibliography V


Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015).
Skip-Thought Vectors.
Advances in Neural Information Processing Systems 28 (NIPS 2015), page 9.


Kirsch, L., Kunze, J., and Barber, D. (2018).
Modular Networks: Learning to Decompose Neural Computation.
In *NIPS 2018*, page 11.


Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. (2019).
Revealing the Dark Secrets of BERT.
In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4356–4365, Hong Kong, China. Association for Computational Linguistics.


Kudo, T. (2018).
Subword regularization: Improving neural network translation models with multiple subword candidates.
In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.


Landauer, T. K., Laham, D., Rehder, B., and Schreiner, M. E. (1997).
How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans.
In *Proceedings of the 19th Annual Meeting of the Cognitive Science Society*, pages 412–417.


Lazaridou, A., Bruni, E., and Baroni, M. (2014).
Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world.
In *ACL (1)*, pages 1403–1414.

Bibliography VI


Le, Q. and Mikolov, T. (2014).
Distributed Representations of Sentences and Documents.
In *International Conference on Machine Learning - ICML 2014*, volume 32, pages 1188–1196.


Li, B., Drozd, A., Liu, T., and Du, X. (2018).
Subword-level composition functions for learning word embeddings.
In *Proceedings of The 2nd Workshop on Subword and Character level models in NLP (SCLeM)*, pages 38–48. ACL.


Logeswaran, L. and Lee, H. (2018).
An efficient framework for learning sentence representations.
In *ICLR 2018*.


Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013a).
Distributed representations of words and phrases and their compositionality.
In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 3111–3119.


Mikolov, T., Yih, W.-t., and Zweig, G. (2013b).
Linguistic Regularities in Continuous Space Word Representations.
In *HLT-NAACL*, pages 746–751.


Mitchell, J. and Lapata, M. (2010).
Composition in distributional models of semantics.
Cognitive science, 34(8):1388–1429.


Mueller, J. and Thyagarajan, A. (2016).
Siamese Recurrent Architectures for Learning Sentence Similarity.
In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, pages 2786–2792, Phoenix, Arizona, February 12-17, 2016. AAAI Press.

Bibliography VII

Nguyen, V., Brooke, J., and Baldwin, T. (2017).
Sub-character Neural Language Modelling in Japanese.
In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 148–153.

Pagliardini, M., Gupta, P., and Jaggi, M. (2018).
Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540, New Orleans, Louisiana. Association for Computational Linguistics.

Partee, B. (1984).
Compositionality.
In Landman, F. and Veltman, F., editors, *Varieties of Formal Semantics: Proceedings of the 4th Amsterdam Colloquium, Sept. 1982*, pages 281–311. Foris Pubs., Dordrecht.

Peng, X. and Gildea, D. (2016).
Exploring phrase-compositionality in skip-gram models.
arXiv:1607.06208 [cs].

Rogers, A., Drozd, A., and Li, B. (2017).
The (Too Many) Problems of Analogical Reasoning with Word Vectors.
In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 135–148.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020).
A primer in bertology: What we know about how bert works.

Bibliography VIII

Romanov, A., Rumshisky, A., Rogers, A., and Donahue, D. (2019).
Adversarial Decomposition of Text Representation.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 815–825.

Rudolph, S. and Giesbrecht, E. (2010).
Compositional Matrix-Space Models of Language.
In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 907–916, Uppsala, Sweden 11-16 June 2010.

Sennrich, R., Haddow, B., and Birch, A. (2016).
Neural machine translation of rare words with subword units.
In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Socher, R., Huval, B., Manning, C. D., and Ng, A. Y. (2012).
Semantic Compositionality through Recursive Matrix-Vector Spaces.
In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211, Stroudsburg, PA, USA.

Strubell, E., Verga, P., Belanger, D., and McCallum, A. (2017).
Fast and accurate entity recognition with iterated dilated convolutions.
In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2670–2680, Copenhagen, Denmark. Association for Computational Linguistics.

Bibliography IX

Sutskever, I., Vinyals, O., and Le, Q. V. (2014).
Sequence to Sequence Learning with Neural Networks.
In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Tai, K. S., Socher, R., and Manning, C. D. (2015).
Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks.
In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1556–1566, Beijing, China, July 26-31, 2015. Association for Computational Linguistics.

Van de Cruys, T., Poibeau, T., and Korhonen, A. (2013).
A tensor-based factorization model of semantic compositionality.
In *Conference of the North American Chapter of the Association of Computational Linguistics (HTL-NAACL)*, pages 1142–1151.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017).
Attention is all you need.
In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

White, L., Togneri, R., Liu, W., and Bennamoun, M. (2015).
How Well Sentence Embeddings Capture Meaning.
In *Proceedings of the 20th Australasian Document Computing Symposium*, ADCS '15, pages 9:1–9:8, New York, NY, USA. ACM.

Widdows, D. (2004).
Word-Vectors and Search Engines.
In *Geometry and Meaning*. Stanford: Center for the Study of Language and Information.

Bibliography X

Yang, Z., Zhu, C., and Chen, W. (2019).
Parameter-free Sentence Embedding via Orthogonal Basis.
In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 638–648, Hong Kong, China. Association for Computational Linguistics.