

Text Representation Learning and Compositional Semantics

ACML 2019 Tutorial

Aleksandr Drozd



RIKEN Center For Computational Science

blackbird.pw [@bkbrd](https://twitter.com/bkbrd)

Anna Rogers



University of Massachusetts Lowell

hackingsemantics.xyz [@annargrs](https://twitter.com/annargrs)

November 17, 2019

Outline

Defining compositionality

Compositional representations in NLP

Explicit composition techniques

Neural Net as a “Black Box” Composition Device

Opening The Black Box

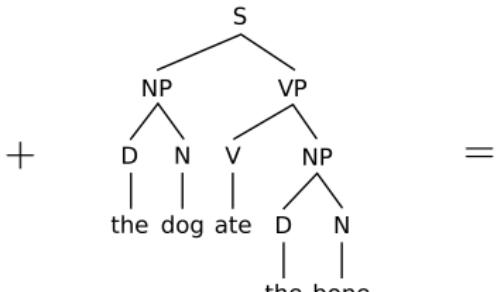
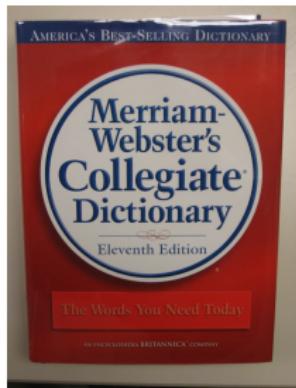
Modular architectures

Bonus: Subcharacter Composition

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]



372 SOUTHERN LIFE IN SOUTHERN LITERATURE

remained estuaries catch from one bystander to another. They swing and how to right and left, in slow time to the piping treble of the singing women. Some are responsive! others are not. Here have not been no ground! one sudden strike only, as it were the foot of a man. The reeakers warm up at the sound. A string of bassoons with open hands begin very softly and becomes vigorous. The women's voices rise to a tremulous imitation. Among the shrill voices of the women is one of coarse gaitor notes, who thumps in, now and again, an improvisation. This girl here, so tall and straight, is a Yulet. You see it in her stout African features, and hear it in the plaintive melody of her voice. When the music begins to grow louder and louder, the women clap their hands in time, or stamping with feet also receive with first crotchetts and headlings the low booms of the men, who deliver them swinging this way and that.

See! Yonder bright and slavish fellow has taken one step, swaying into the ring, and begins to raise his hands. Now he takes another, and sways and sings and looks here and there; riding upon his broad toes and striking and rising again, with what wonderful lightness! How tall and lithe he is. Notice his brown skin through his rags. He too is a slave, and by the look of him, says the woman, he is one of those from Africa. The music has got into his feet. He moves off to the farther edge of the circle, still singing, takes the proudest hand of an unsmiling Congo girl, leads her into the ring, and, leaving the chart 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, antiphonally crowd starts into extra activity, the female voices grow sharp and staccato, and suddenly the dance is the furious Barnards.

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]

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.

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.

“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]

Defining compositionality

Compositional representations in NLP

Explicit composition techniques

Neural Net as a “Black Box” Composition Device

Opening The Black Box

Modular architectures

Bonus: Subcharacter Composition

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.

One could argue that the main power of deep-learning is the ability to learn good representations.

[Goldberg, 2017]

DL perspectives

- ▶ 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;
- ▶ “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.

Types of “representations”

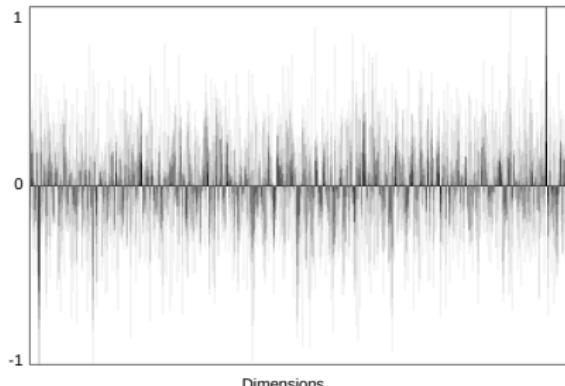
- ▶ representations at different levels: words, subword units, phrases, sentences, texts;
- ▶ compositionality within representations: interpretable semantic components.

Components of meaning

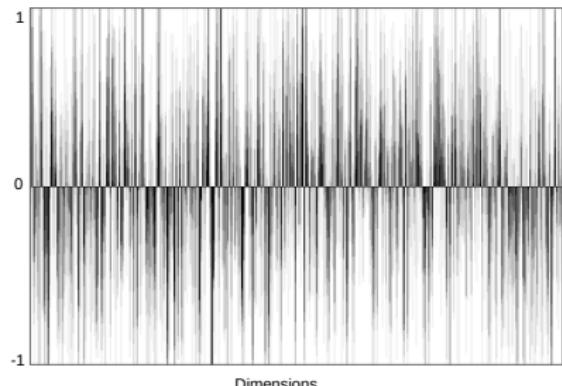
Direction 1:

vector dimensions as semes

We already have meaning components!



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]

“Interpretable” dimensions

Model	Top 5 Words
CBOW	beat, finish, wedding, prize, read
	rainfall, footballer, breakfast, weekdays, angeles
	landfall, interview, asked, apology, dinner
	becomes, died, feels, resigned, strained
Sparse CBOW	best, safest, iucn, capita, tallest
	poisson, parametric, markov, bayesian, stochastic
	ntfs, gzip, myfile, filenames, subdirectories
	hugely, enormously, immensely, wildly, tremendously
CBOW	earthquake, quake, uprooted, levees, spectacularly
	bosons, accretion, higgs, neutrinos, quarks

Interpreting dimensions with top 5 words for a given dimension: dense vs sparse CBOW [Sun et al., 2016]

See also: [Murphy et al., 2012], [Luo et al., 2015]

Compositional Non-negative Sparse Embedding (CNNSE) [Fyshe et al., 2015]

- ▶ sparsity constraint
- ▶ non-negativity constraint
- ▶ composition constraint: given words i and j , and a phrase p consisting of i and j , the rows of matrix A are constrained so that $A_{(p,:)} = f(A_{(i,:)}, A_{(j,:)})$. f is some composition function (weighted addition in this paper).

CNSE: interpretable phrase similarity

Dissimilar phrases: no shared semantic component

behav sim score **1.00**

national government		cold air	
0.1666	biennial future national	0.3317	balmy chilly damp
0.0806	bureaucratic congressional devolved	0.1011	climate climates continental_climate
0.0656	affiliated collaborating for-profit	0.0878	allergens contaminants dust

CNNSE: interpretable phrase similarity

Similar phrases: shared semantic component

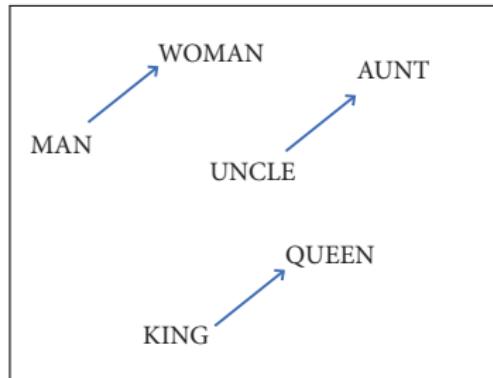
behav sim score **4.39**

vast amount		high price	
0.2162	azure cascading glacial	0.1562	big_bucks cent dime
0.1390	big_bucks cent dime	0.0598	buy buy fedex
0.1338	appreciable demonstrable enormous	0.0372	airfare cheap_flights cheap_tickets

Direction 2:

searching for words with a given seme

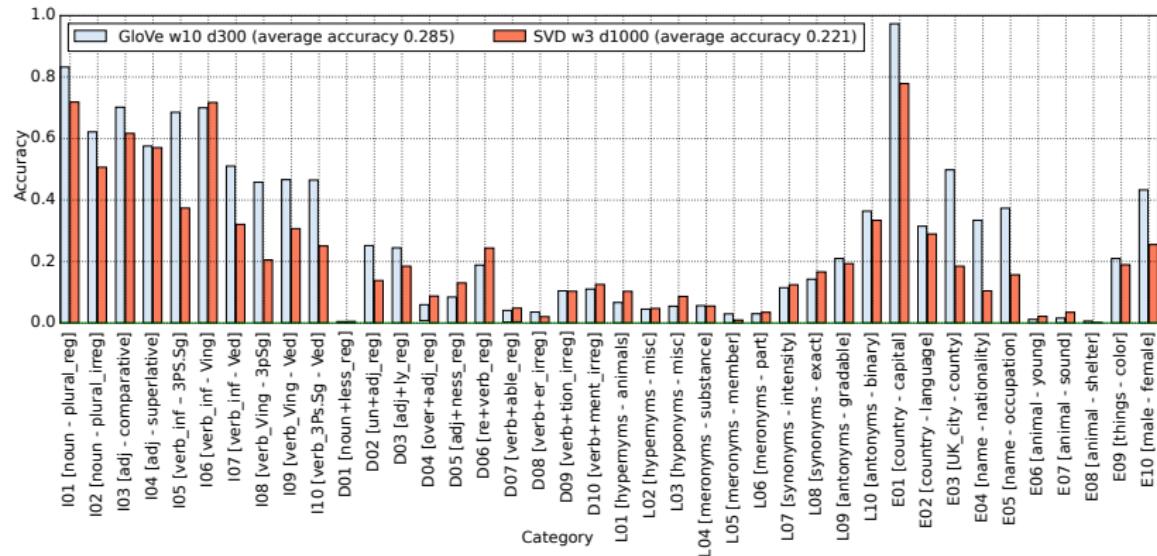
Inducing meaning shifts by analogy



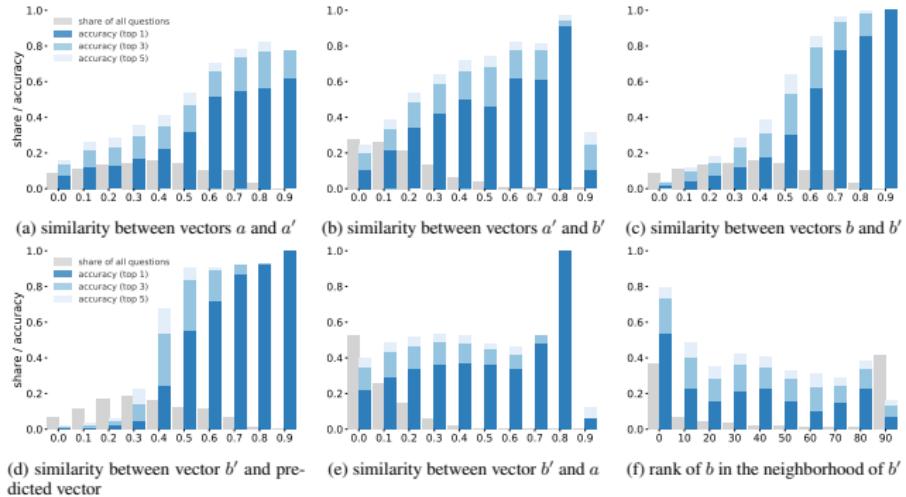
$$\text{king} - \text{man} + \text{woman} = \text{queen}$$

Moving between word pairs with a shared semantic component with linear vector offset [Mikolov et al., 2013b]

Only in some relations [Gladkova et al., 2016]



Only for close targets [Rogers et al., 2017]



Accuracy on analogy task depends on the vector distance between the source words [Rogers et al., 2017]

The information might be detectable with a smarter method [Drozd et al., 2016]

Analogy solving methods: average accuracy per relation type for GloVe and Skip-Gram in BATS [Drozd et al., 2016]

Method	Encyclopedia		Lexicography		Inflections		Derivation	
	GloVe	SG	GloVe	SG	GloVe	SG	GloVe	SG
3CosAdd	31.5%	26.5%	10.9%	9.1%	59.9%	61.0%	10.2%	11.2%
3CosAvg	44.8	34.6%	13.0%	9.6%	68.8%	69.8	11.2%	15.2%
LRCos	40.6%	43.6	16.8%	15.4	74.6%	87.2%	17.0%	45.6
LRCosF	40.8%	42.6%	17.6	14.4%	75.0	87.4	17.1	44.2%

Defining compositionality

Compositional representations in NLP

Explicit composition techniques

Neural Net as a “Black Box” Composition Device

Opening The Black Box

Modular architectures

Bonus: Subcharacter Composition

Two problems to solve

- ▶ how to select the representations for composition?
- ▶ how to compose them?

Problem 1: selecting representations for composition

- ▶ manual pre-selection;
- ▶ using syntactic/morphological parsers;
- ▶ constrained by network design (CNN, LSTM);
- ▶ learned attention or controller for tree traversal.

Problem 2: composition mechanisms

- ▶ explicit compositional operations;
- ▶ encoding with neural networks.

Defining compositionality

Compositional representations in NLP

Explicit composition techniques

Explicit compositionality

Algebraic approaches

Syntax-aware proposals

Neural Net as a “Black Box” Composition Device

Opening The Black Box

Modular architectures

Bonus: Subcharacter Composition

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

Model	Function
Additive	$p_i = u_i + v_i$
Kintsch	$p_i = u_i + v_i + n_i$
Multiplicative	$p_i = u_i \cdot v_i$
Tensor product	$p_{i,j} = u_i \cdot v_j$
Circular convolution	$p_i = \sum_j u_j \cdot v_{i-j}$
Weighted additive	$p_i = \alpha v_i + \beta u_i$
Dilation	$p_i = v_i \sum_j u_j u_j + (\lambda - 1) u_i \sum_j u_j v_j$
Head only	$p_i = v_i$
Target unit	$p_i = v_i(t_1 t_2)$

Example: 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 Check crown Polish zolty CTK	Hanoi Ho Chi Minh City Viet Nam Vietnamese	airline Lufthansa carrier Lufthansa flag carrier Lufthansa Lufthansa	Moscow Volga River upriver Russia	Juliette Binoche Vanessa Paradis Charlotte Gainsbourg Cecile De

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

Averaging

- ▶ simple mean [Hill et al., 2016, Adi et al., 2017]
- ▶ weighted mean, e.g. 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

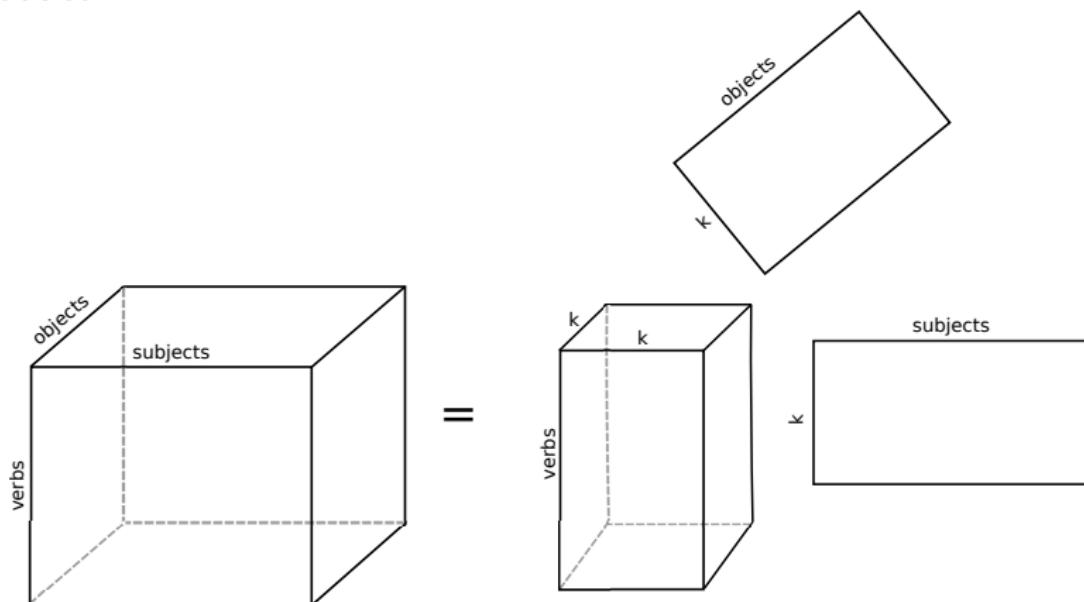
Discussion: [Clark and Pulman, 2007, Mitchell and Lapata, 2010, Clarke, 2012, Van de Cruys et al., 2013]

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

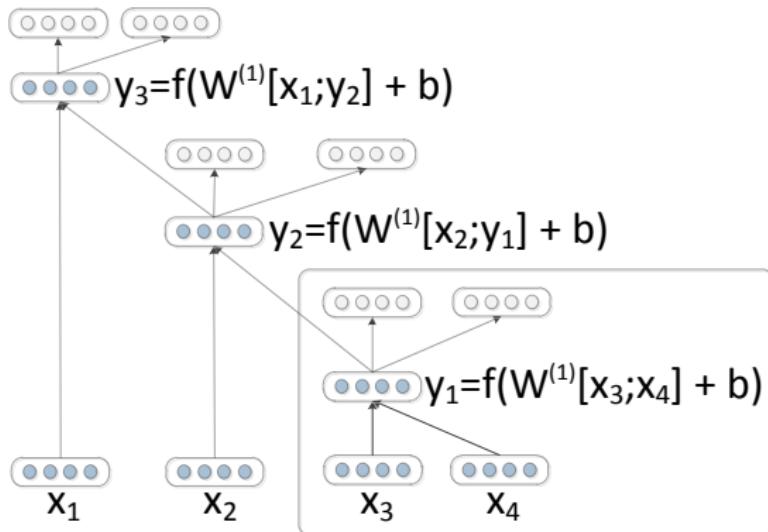
- ▶ **adjective-specific linear map** (alm) method: an adjective-noun phrase is represented by multiplying an adjective weight matrix with a noun (column) vector
- ▶ experiments with phrase generation with SVD embeddings: how close is the generated phrase to a pre-computed one?
- ▶ composed phrases shown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

Representing subject verb object triples

[Van de Cruys et al., 2013] compute the core tensor for interactions between verbs, subjects and objects with n-mode product



Recursive autoencoders for sentence-level prediction of sentiment label distributions [Socher et al., 2011]



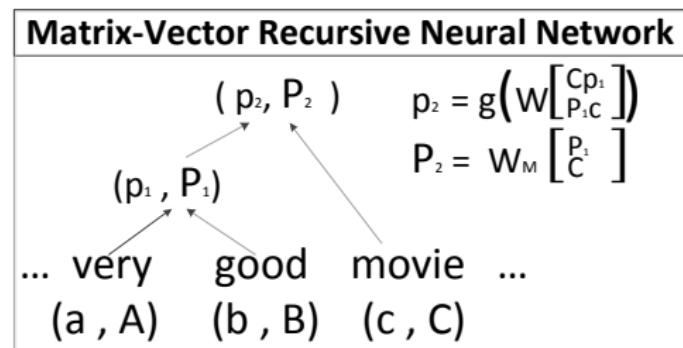
W : matrix for weighting constituent word vectors x, y
 b : bias term
 f : non-linear function (*tanh* is used), applied in a binarized parse tree

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 :

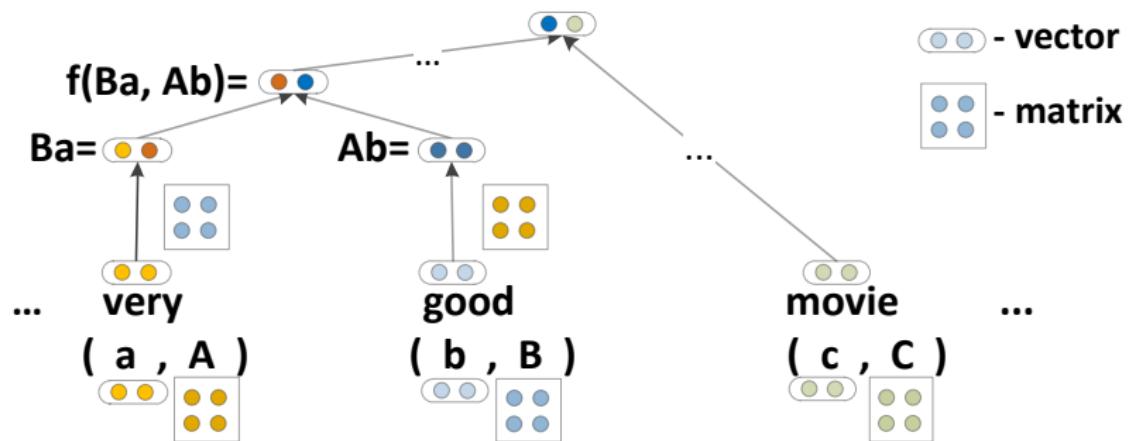
$$p = g \left(W \begin{bmatrix} Ba \\ Ab \end{bmatrix} \right)$$

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



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

Recursive Matrix-Vector Model



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].

Defining compositionality

Compositional representations in NLP

Explicit composition techniques

Neural Net as a “Black Box” Composition Device

Opening The Black Box

Modular architectures

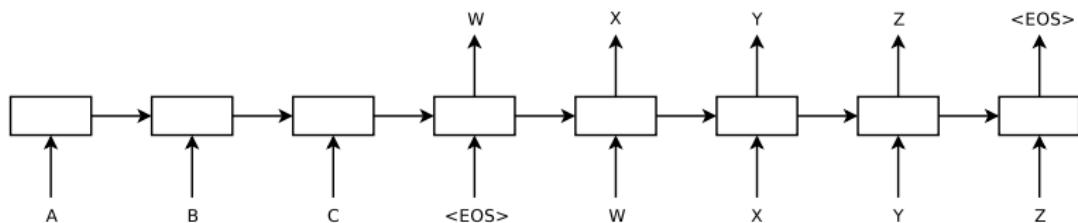
Bonus: Subcharacter Composition

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.

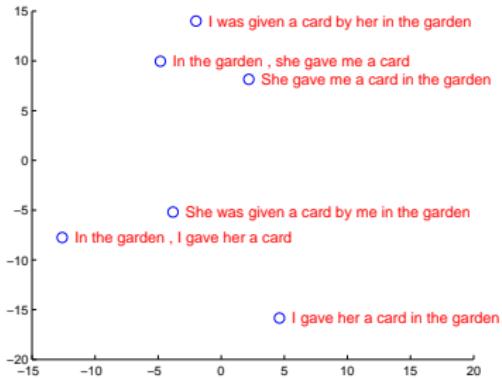
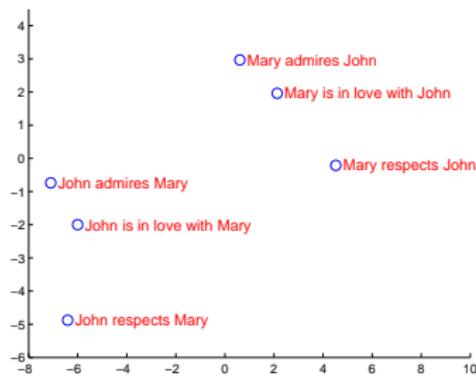
End-to-end sequence encoding [Sutskever et al., 2014]

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.



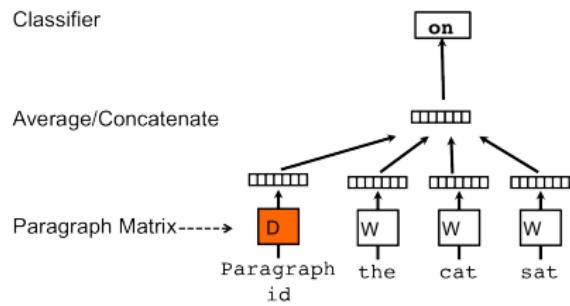
End-to-end sequence encoding [Sutskever et al., 2014]

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

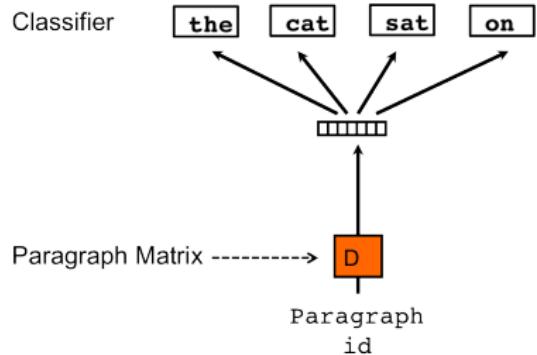


Paragraph vectors [Le and Mikolov, 2014]

Distributed Memory version



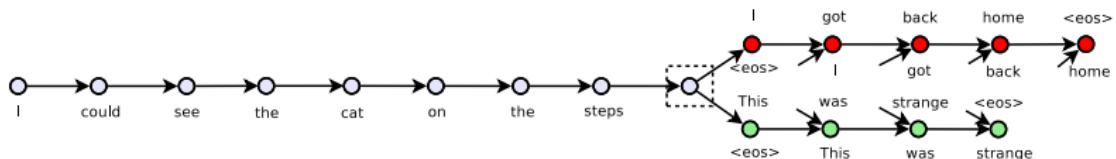
Distributed Bag-of-words version



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} .

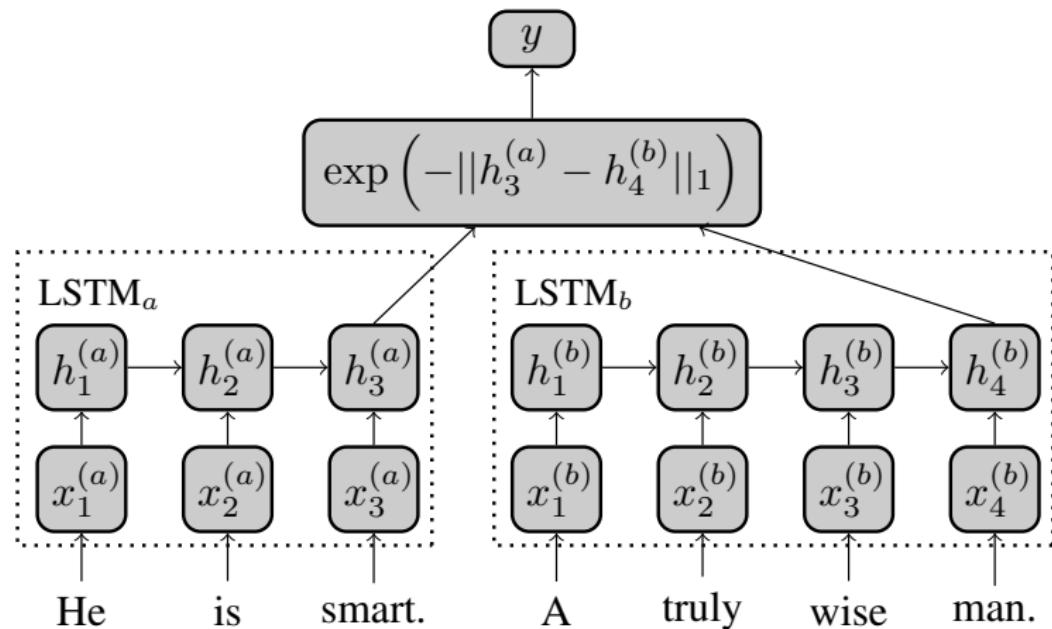
Example tuple: (*I got back home. I could see the cat on the steps. This was strange.*)



Siamese LSTM for sentence similarity

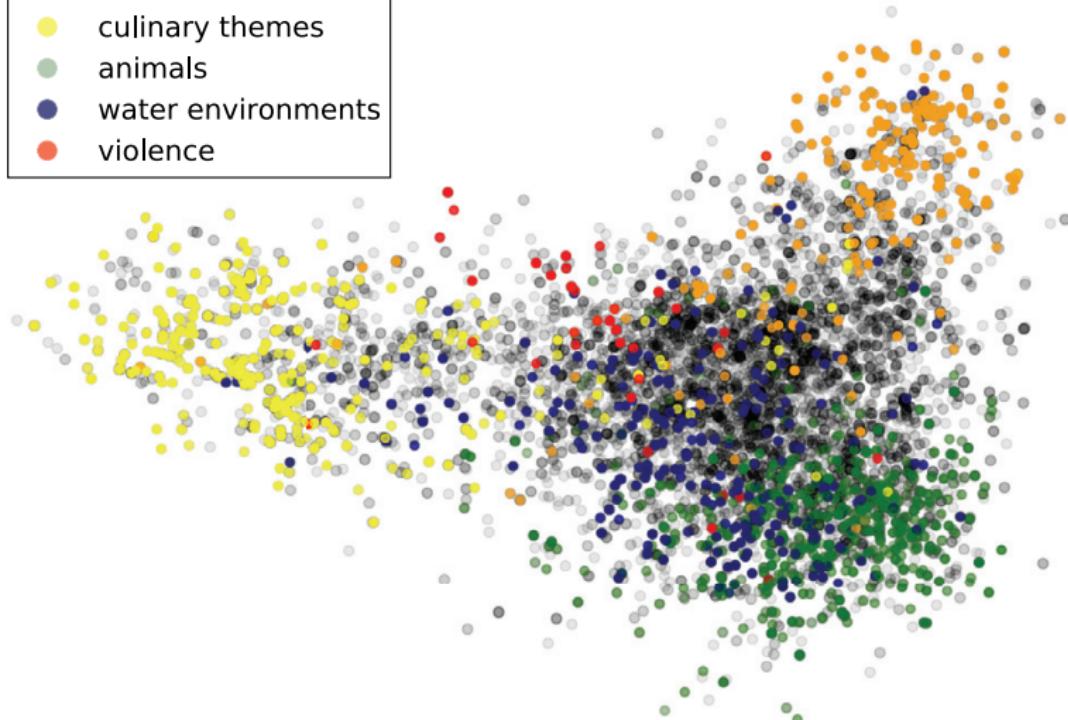
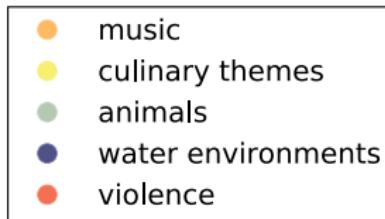
[Mueller and Thyagarajan, 2016]

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

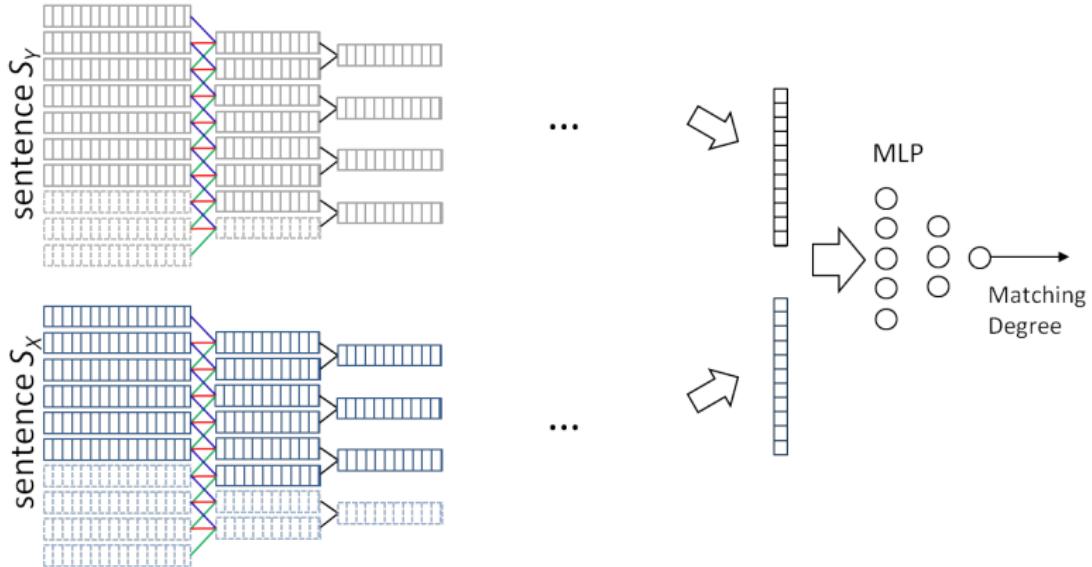


Siamese LSTM [Mueller and Thyagarajan, 2016]

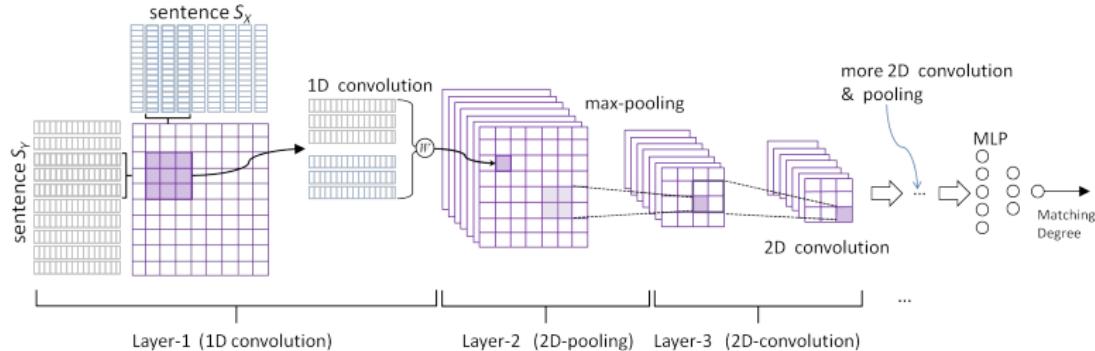
t-SNE representation of the SICK dataset



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



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



Pros and cons

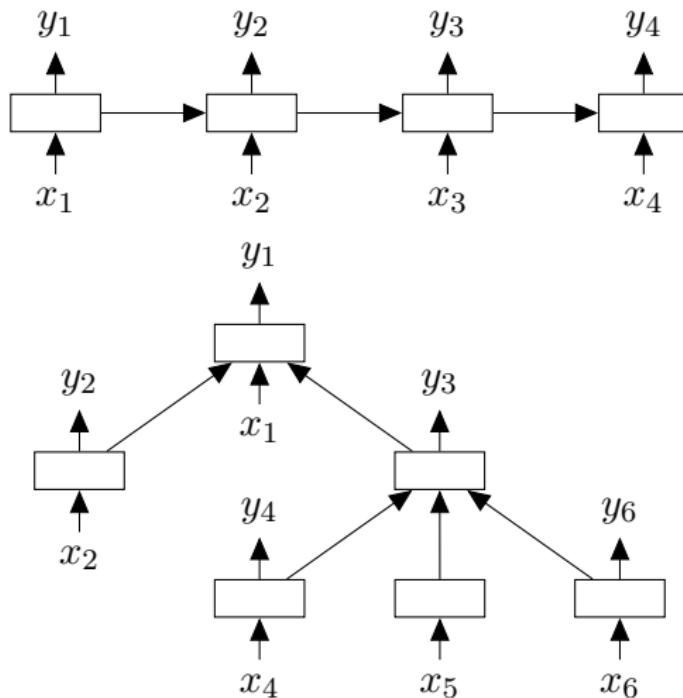
Pros:

- ▶ expressive
- ▶ seems to work on its own

Cons:

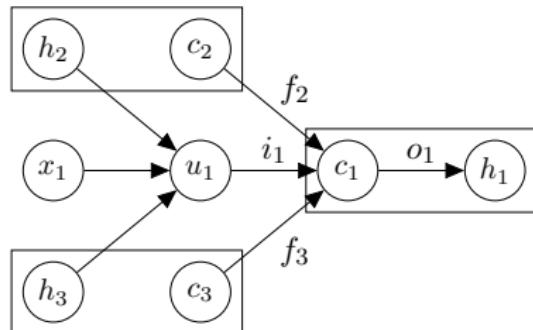
- ▶ sensitive to dozens of parameters;
- ▶ computationally expensive;
- ▶ lack of syntactic finesse, unless cheating;
- ▶ not clear that it's really not compositional, as long as the tokenized words are fed one by one, and especially if word embeddings are used;
- ▶ since it is a black box, hard to understand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

Defining compositionality

Compositional representations in NLP

Explicit composition techniques

Neural Net as a “Black Box” Composition Device

Opening The Black Box

Modular architectures

Bonus: Subcharacter Composition

Rough summary:

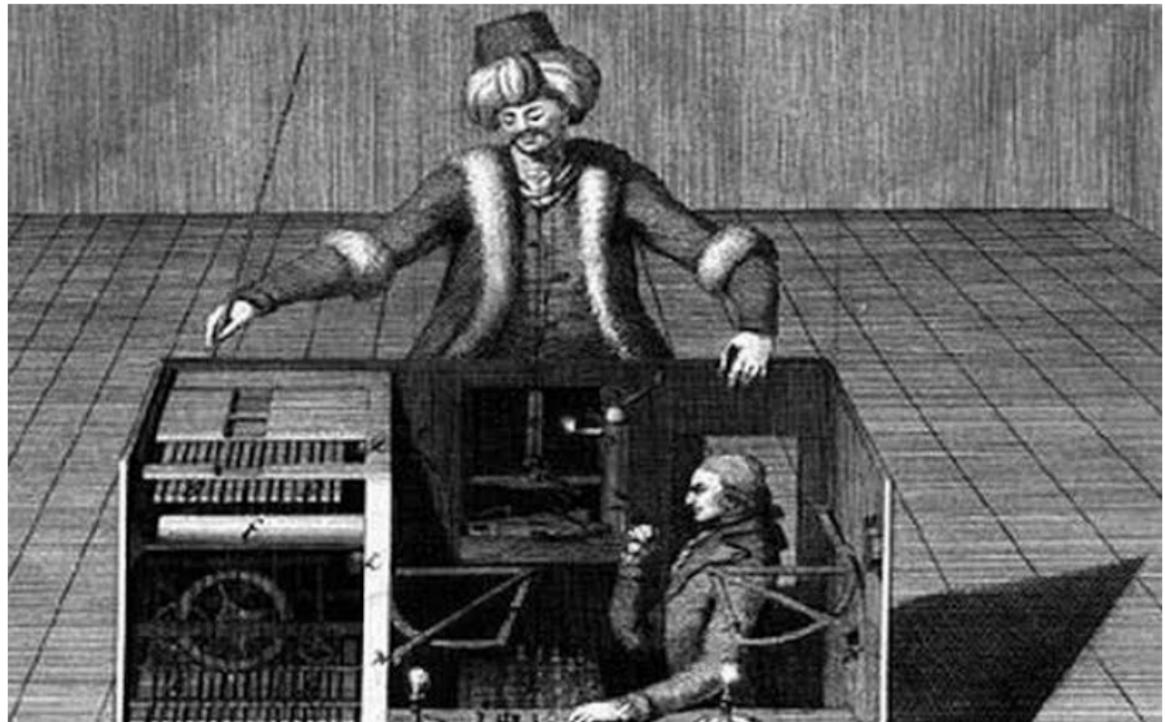
Overall architecture / objective

- ▶ classification
- ▶ seq-2-seq
- ▶ auto-regressive (language model)
- ▶ siamese networks
- ▶ AEs, GANs, etc

Building blocks

- ▶ dense blocks
- ▶ convolutional layers
- ▶ recurrent cells
- ▶ attention
- ▶ etc.

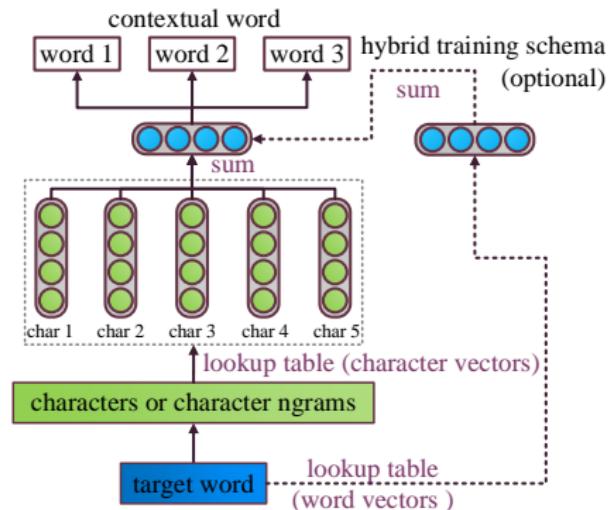
"arithmetical" methods lurking in neural nets



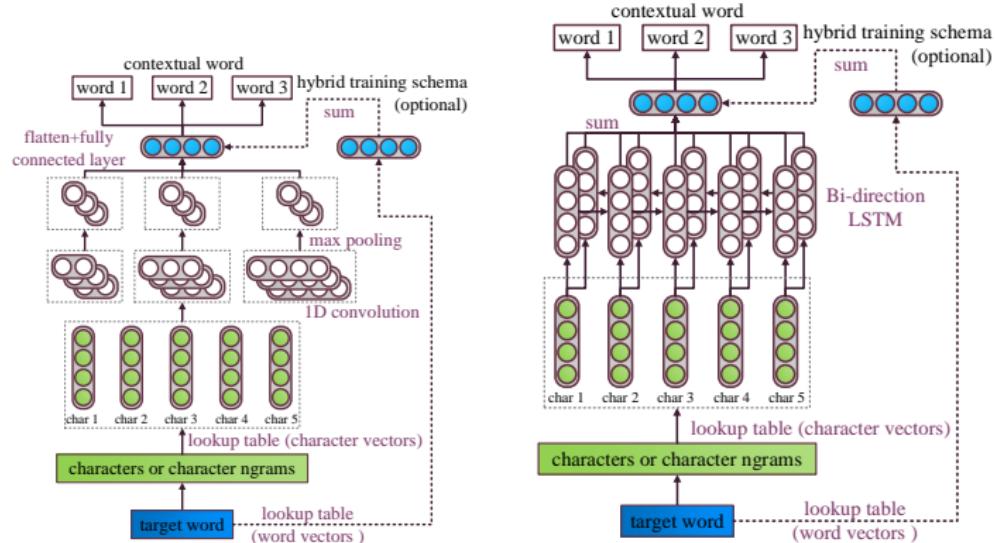
Case of FastText [Bojanowski et al., 2017]

- ▶ **vocabulary is not fixed**: each word is a bag-of-character n-grams, thus allowing to build embeddings for out-of-vocabulary words by summing the trained embeddings of character n-grams.
- ▶ 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]



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



Neural net as compositional function

in- re- un- -able -al -er -ful -ic -ion

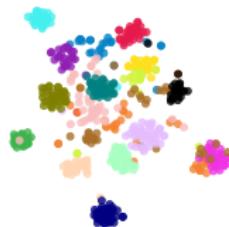
-ist -ity -ize -less -ly -ment -ness -ous -y



(a) Skip-Gram



(b) FastText_{subword}



(c) CNN_{subword}



(d) RNN_{subword}

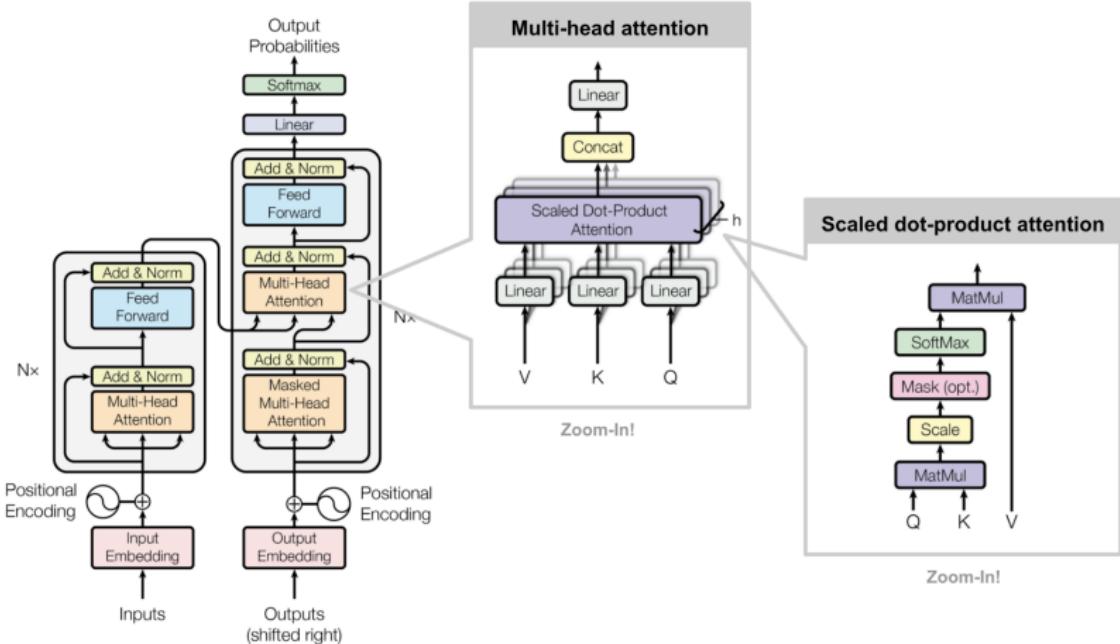
t-SNE projection of representations

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

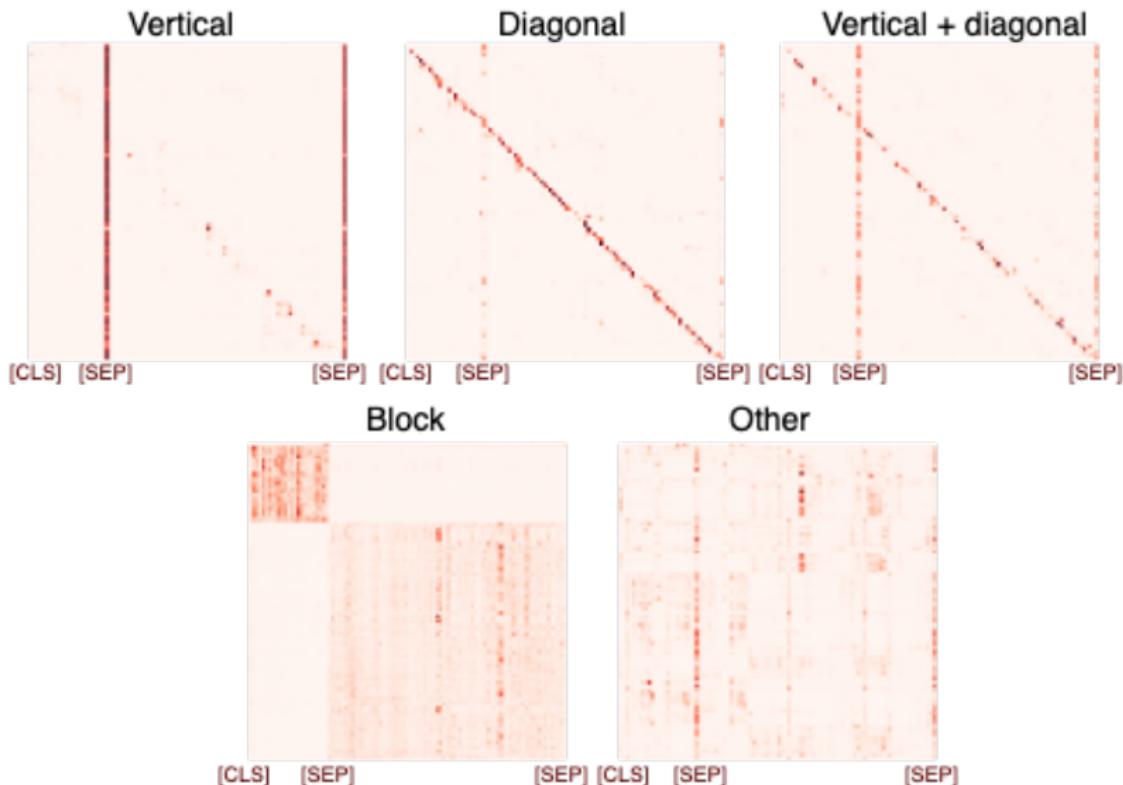
Transformers [Vaswani et al., 2017]

- ▶ How to select: self-attention
- ▶ How to compose: addition (in one head)

Transformers [Vaswani et al., 2017]

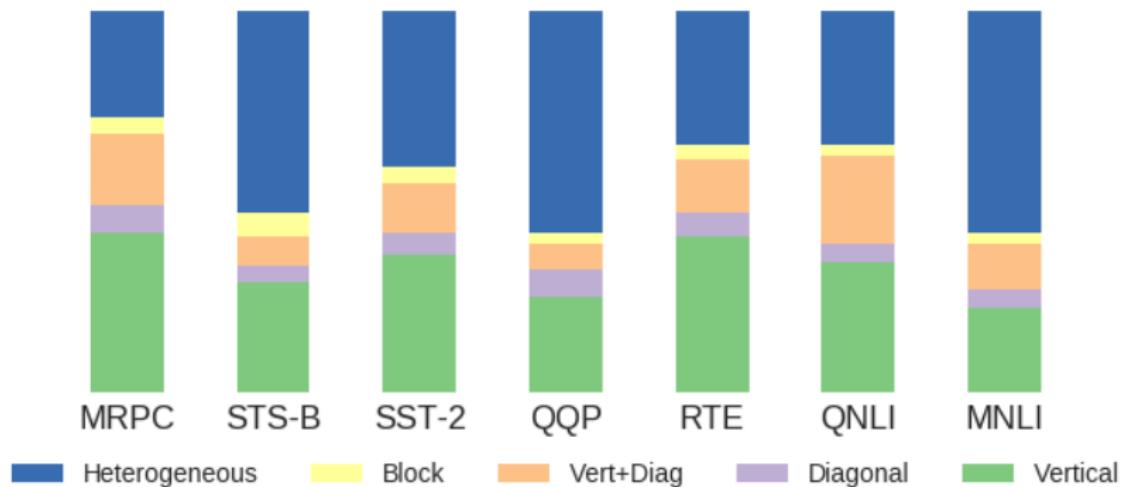


Attention pattern types in BERT [Kovaleva et al., 2019]



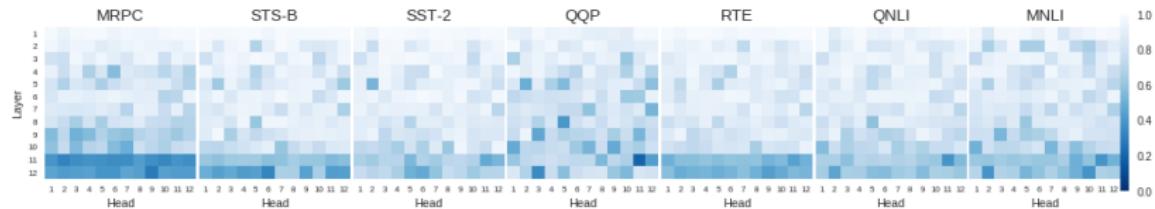
Typical self-attention classes used for training a neural network.

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



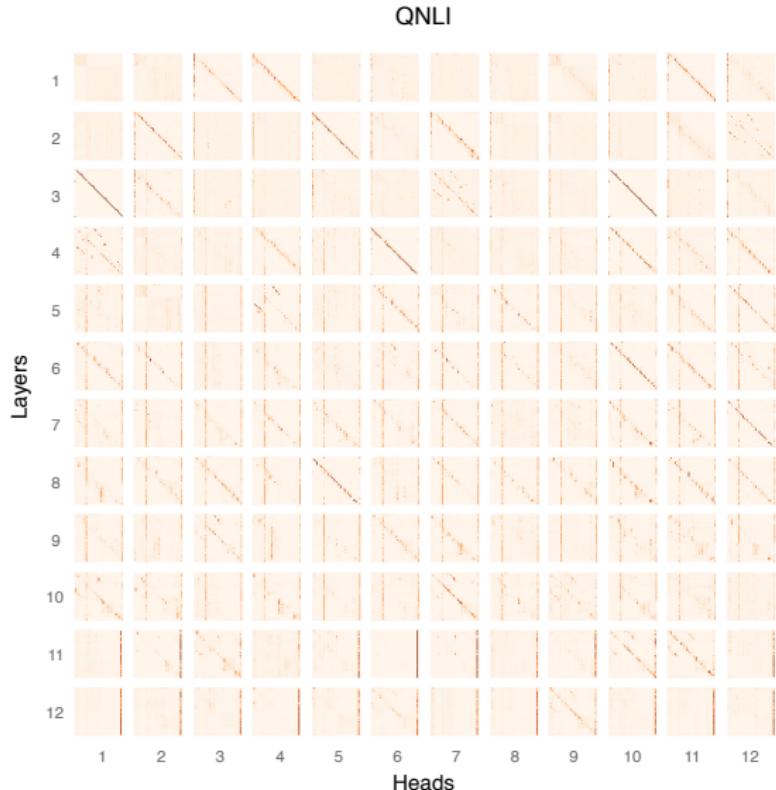
Takeaway: the ratio of potentially useful attention maps varies by task, but in most cases it is less than half of BERT.

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



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

Takeaway:
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. 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

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

Defining compositionality

Compositional representations in NLP

Explicit composition techniques

Neural Net as a “Black Box” Composition Device

Opening The Black Box

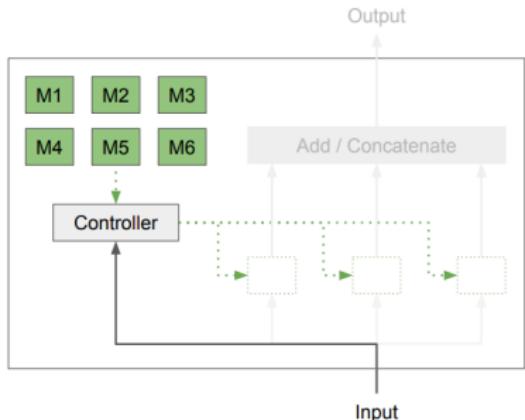
Modular architectures

Bonus: Subcharacter Composition

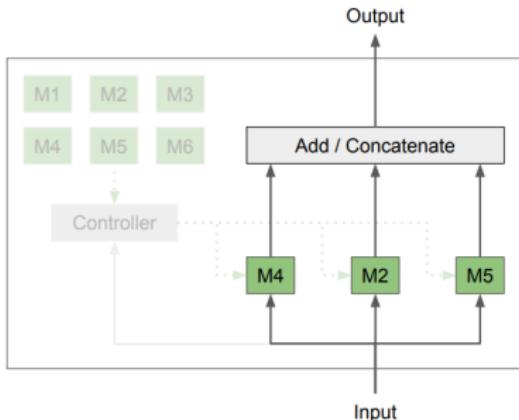
Direction 1: architecture components as task components

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

Modular architecture



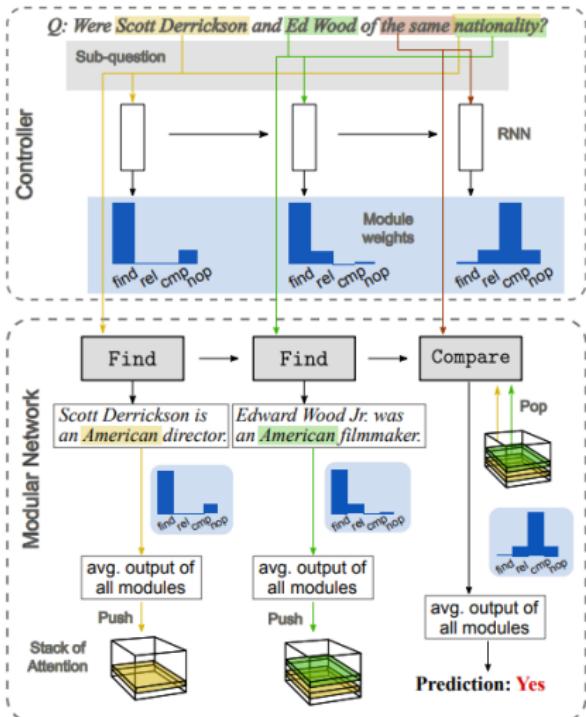
(a) Based on the input, the controller selects K modules from a set of M available modules. In this example, $K = 3$ and $M = 6$.



(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]

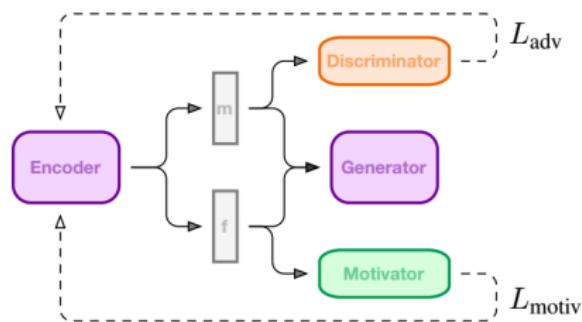
Self-assembling modular network for multi-hop QA



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

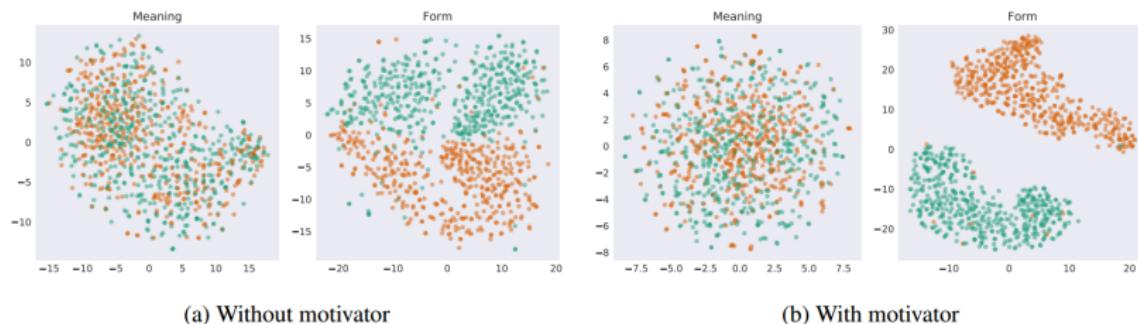
- ▶ Encoder decomposes the input into “meaning” and “form” vectors;
- ▶ Generator takes those vectors to produce an output sequence
- ▶ Discriminator tries to classify form based on meaning (> encoder is penalized to make this hard)
- ▶ Motivator tries to classify form based on form (encoder is penalized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



[Romanov et al., 2019]

Dissociated meaning and form vectors

[Romanov et al., 2019]



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)

Defining compositionality

Compositional representations in NLP

Explicit composition techniques

Neural Net as a “Black Box” Composition Device

Opening The Black Box

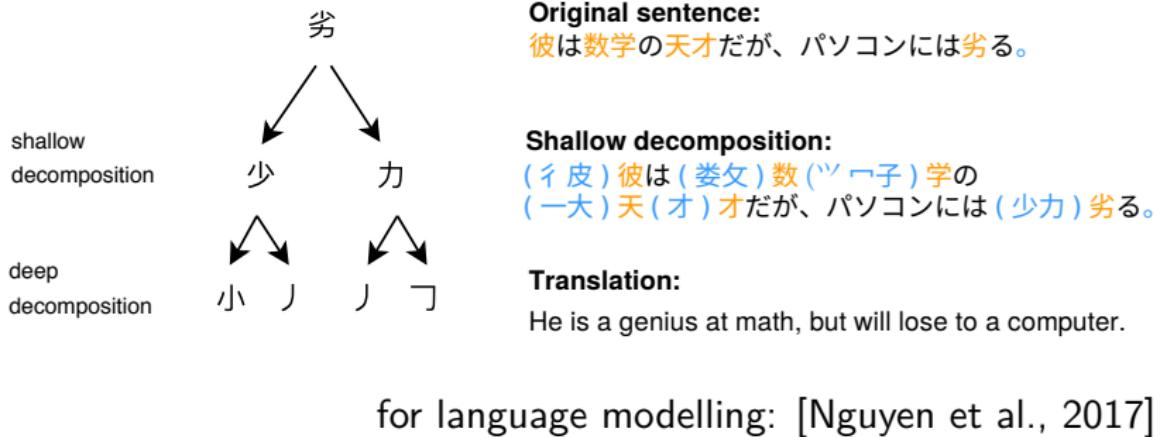
Modular architectures

Bonus: Subcharacter Composition

Subcharacter composition

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

Kanji decomposition



Subcharacter embeddings in Japanese

[Karpinska et al., 2018]

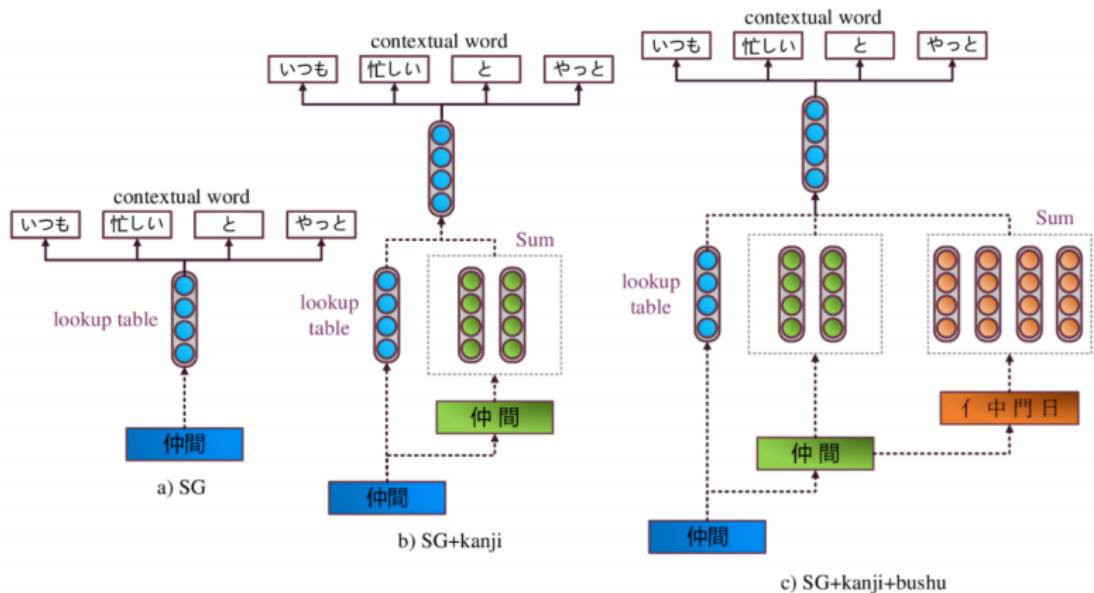


Figure 2: Model architecture of SG, SG+kanji, and SG+kanji+bushu. Example sentence: いつも 忙しい 仲間 と やっと 会え た (“I have finally met with my busy colleague.”), window size 2.

Shared semantic space for kanji, kana and bushu

病 *yamaidare* (the roof from illness)

患(sickness) 症(disease) 妊 (pregnancy)

臓 (internal organs, bowels) 肿 (tumor)

インフルエンザ (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].

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;
- ▶ Not beating single-character FastText baseline, not used in most other work on subcharacter embeddings;
- ▶ A new Japanese analogy dataset and improved Japanese similarity dataset (<http://vecto.space/projects/>).



Hiring Post-Doc/Researcher at RIKEN CCS
High Performance Artificial Intelligence Systems Team
Tokyo, Japan
<http://bit.ly/AITOKYO>

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.
-  Drozd, A., Gladkova, A., and Matsuoka, S. (2016). Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530, Osaka, Japan, December 11–17.

Bibliography III

-  Firth, J. R. (1957).
A synopsis of linguistic theory 1930-55.
1952-59:1-32.
-  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.
-  Gladkova, A., Drozd, A., and Matsuoka, S. (2016).
Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't.
In *Proceedings of the NAACL-HLT SRW*, pages 47–54, San Diego, California, June 12-17, 2016. ACL.
-  Goldberg, Y. (2017).
Neural Network Methods for Natural Language Processing.
Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, San Rafael.
-  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.

Bibliography IV

-  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.
-  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.
-  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.

Bibliography V

-  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.
-  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.
-  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*.
-  Luo, H., Liu, Z., Luan, H., and Sun, M. (2015).
Online Learning of Interpretable Word Embeddings.
In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1687–1692, Lisbon, Portugal, 17–21 September 2015. Association for Computational Linguistics.
-  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.

Bibliography VII

-  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.
-  Murphy, B., Talukdar, P., and Mitchell, T. (2012).
Learning effective and interpretable semantic models using non-negative sparse embedding.
Proceedings of COLING 2012, pages 1933–1950.
-  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].

Bibliography VIII



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.



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.



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.



Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011).

Semi-supervised Recursive Autoencoders for Predicting Sentiment Distributions.

In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 151–161, Stroudsburg, PA, USA. Association for Computational Linguistics.



Sun, F., Guo, J., Lan, Y., Xu, J., and Cheng, X. (2016).

Sparse Word Embeddings Using L1 Regularized Online Learning.

In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2915–2921, New York, New York, USA. AAAI Press.

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.