Neural Vector Representations beyond Words:
Sentence and Document Embeddings

Gerard de Melo
http://gerard.demelo.org

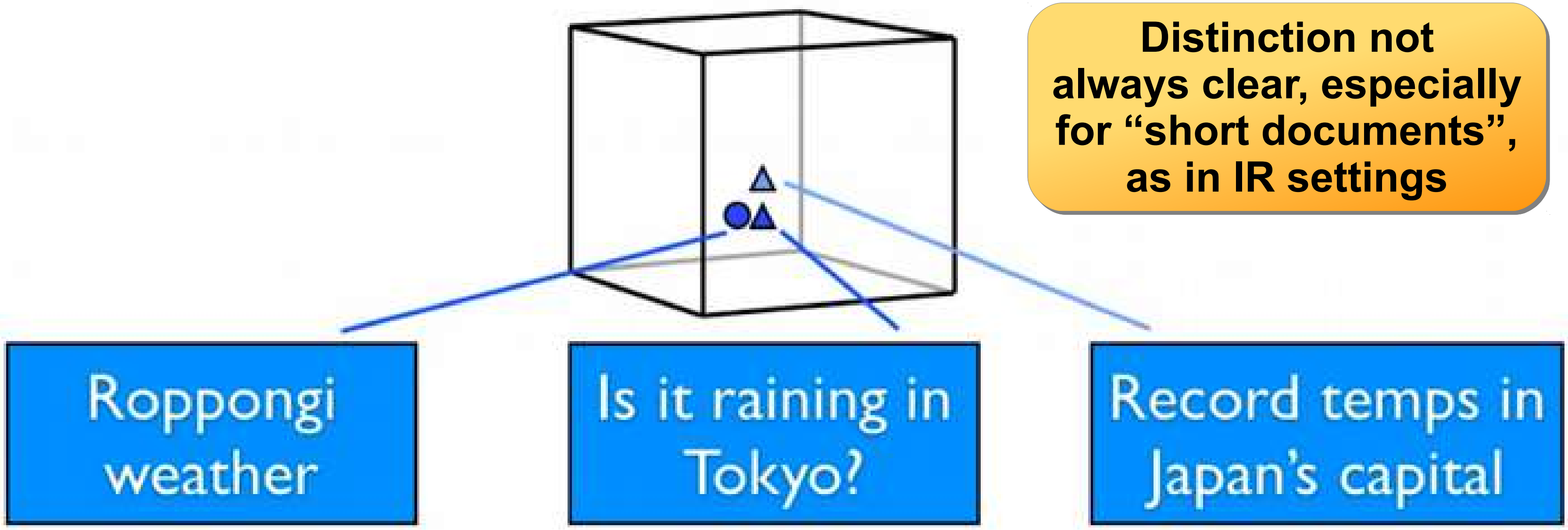
Rutgers University

Outline

- Word Representations
- Phrase Representations
- Sentence Representations
- Document Representations
- Applications and Outlook

Public Domain Image from https://pixabay.com/en/book-text-read-paper-education-451067/

Sentences vs. Documents



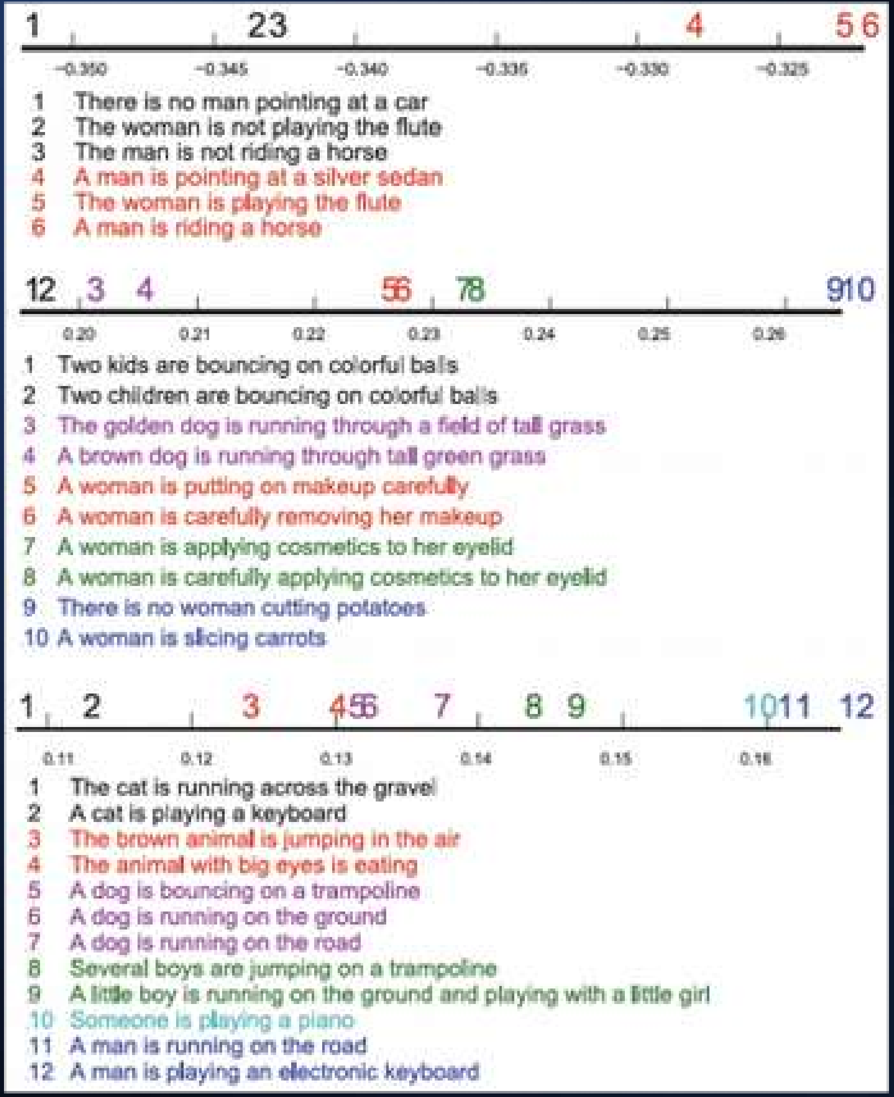
Distinction not always clear, especially for "short documents", as in IR settings

- Query similarity / Query-Document scoring
- Machine translation
- Question answering
- Natural language *understanding*?

Source: Jeff Dean, Google

Sentences vs. Documents

For sentences, we care about **detailed semantics**



Sentences vs. Documents

For sentences, we care about **detailed semantics**

For documents, we typically need to capture **aboutness**

Bag-of-Words Vectors

D1
dog food
and
cat food

D2
good food
for dogs
and cats

	D1	D2
dog	1	1
food	2	1
cat	1	1
good	0	1
...	0	0

TF-IDF Bag-of-Words Vectors

D

good
dog food
and good
cat food

Assume
N=10
documents

	f(t)				
dog	1				
food	2				
cat	1				
good	2				
...	0				

$$tfidf(t) = (1 + \log f(t)) \times \log \frac{N}{n(t)}$$

Conceptual Vector Spaces

"new"	1.0
"york"	1.0
"jaguar"	1.0
"automobile"	0.0
"car"	0.0
"10th"	1.0
"street"	1.0
"show"	1.0
...	...

"10th street new york jaguar show"

Similar:
"10th New show in York"
"New Jaguar show"
"Show New Street in York"

New_York	1.0
Jaguar (car)	0.0
Jaguar (animal)	1.0
Automobile/Car	0.0
10th Street	1.0
Performance	1.0
...	...
Animal	0.5
Vehicle	0.0

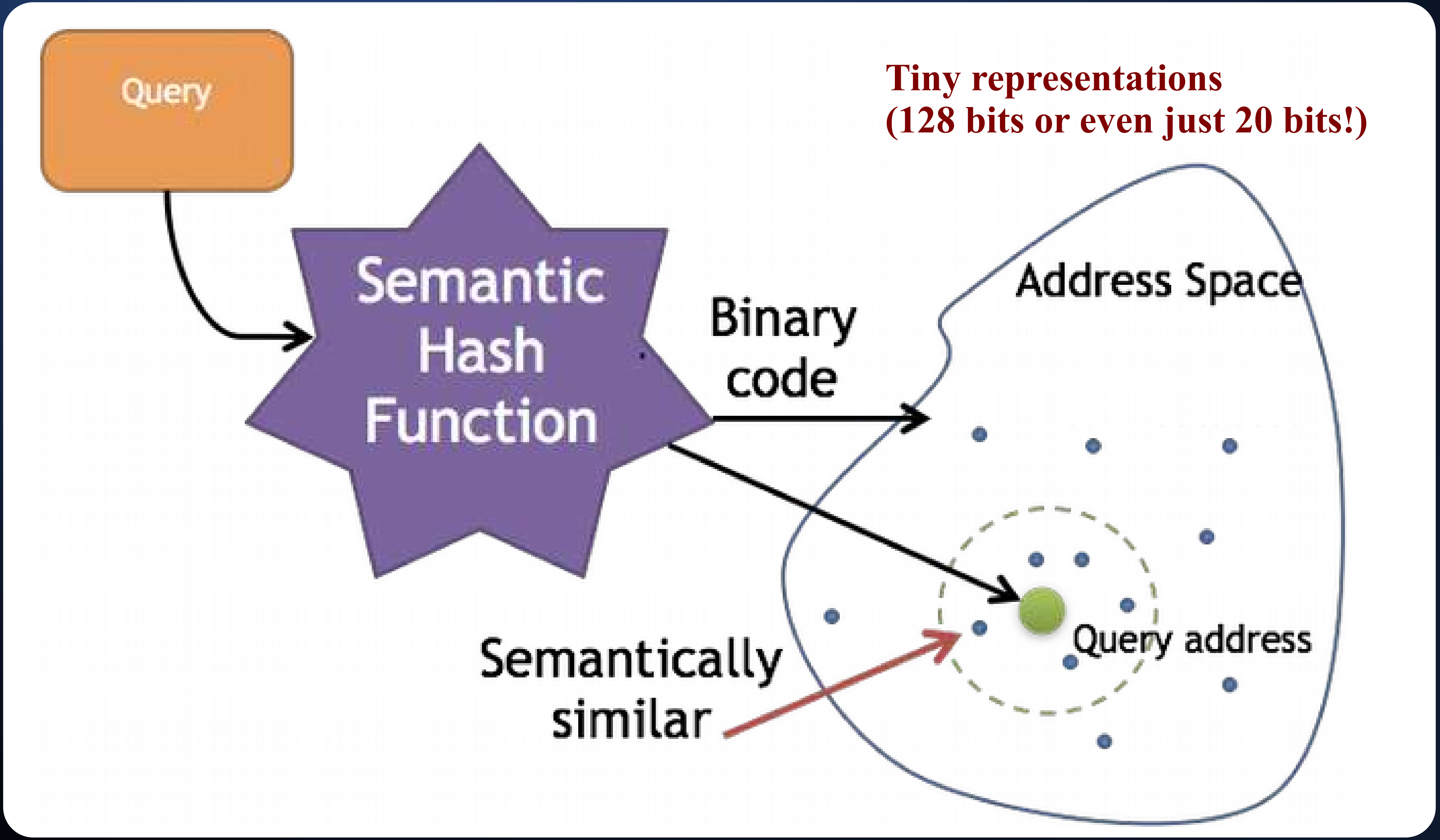
Expansion (de Melo & Siersdorfer)

"10th street new york jaguar show"

Similar:
"10th street nyc jaguar show"
"10th street nyc animal show"
"Exposición de jaguares Nueva York"

Gerard de Melo, Stefan Siersdorfer. Multilingual Text Classification using Ontologies

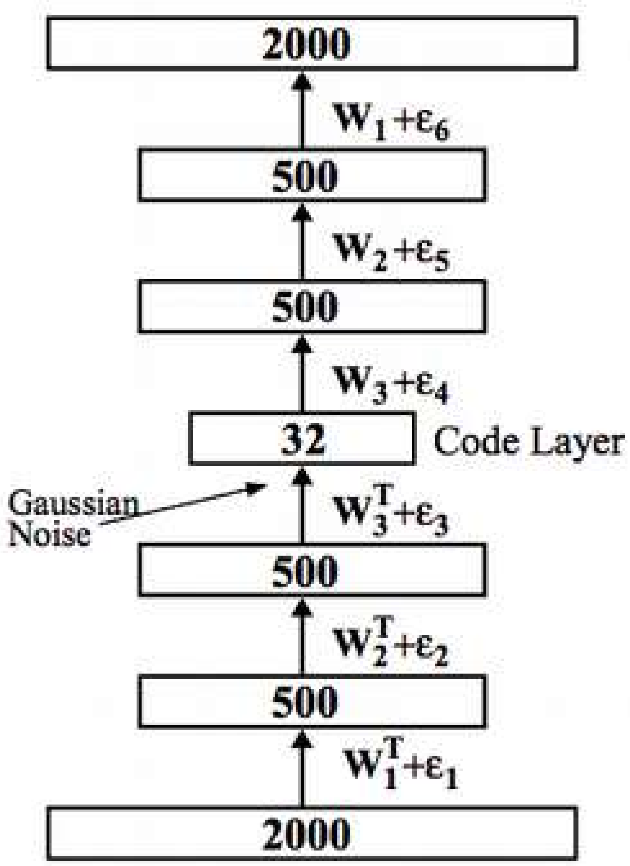
Semantic Hashing



Salakhutdinov & Hinton, 2007

Image: Adapted from Rob Fergus et al.

Semantic Hashing



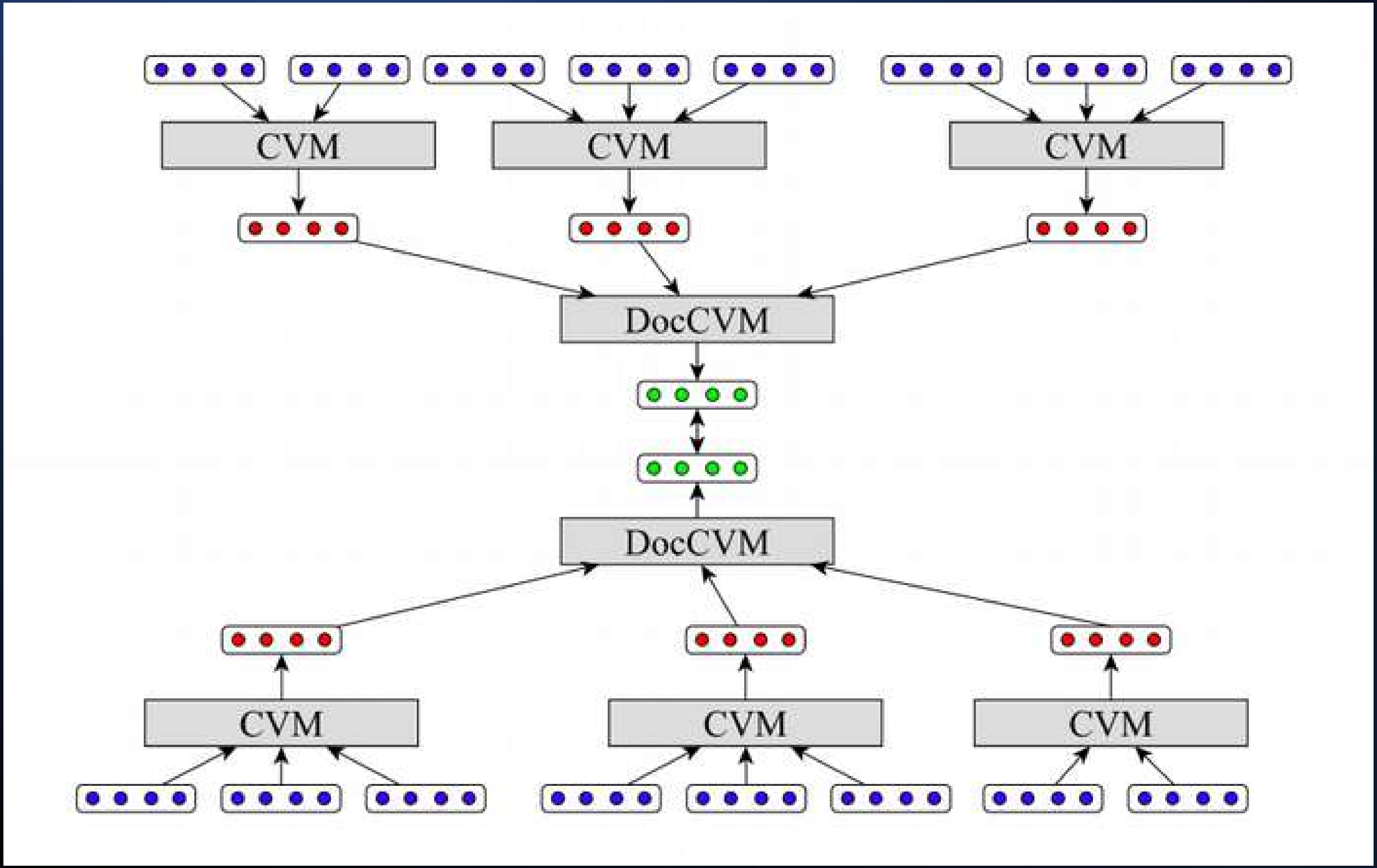
Salakhutdinov & Hinton, 2007

Autoencoder approach

Used RBM-based pretraining

Image: Rob Fergus et al.

Composition from Sentences



Hermann & Blunsom. Multilingual Models for Compositional Distributed Semantics

Paragraph Vectors

(a) Wikipedia nearest neighbours to "Lady Gaga" using Paragraph Vectors. All articles are relevant.

Article	Cosine Similarity
Christina Aguilera	0.674
Beyonce	0.645
Madonna (entertainer)	0.643
Artpop	0.640
Britney Spears	0.640
Cyndi Lauper	0.632
Rihanna	0.631
Pink (singer)	0.628
Born This Way	0.627
The Monster Ball Tour	0.620

(b) Wikipedia nearest neighbours to "Lady Gaga" - "American" + "Japanese" using Paragraph Vectors. Note that Ayumi Hamasaki is one of the most famous singers, and one of the best selling artists in Japan. She also has an album called "Poker Face" in 1998.

Article	Cosine Similarity
Ayumi Hamasaki	0.539
Shoko Nakagawa	0.531
Izumi Sakai	0.512
Urbangarde	0.505
Ringo Sheena	0.503
Toshiaki Kasuga	0.492
Chihiro Onitsuka	0.487
Namie Amuro	0.485
Yakuza (video game)	0.485
Nozomi Sasaki (model)	0.485

Aka "Doc2Vec"

Andrew M. Dai, Christopher Olah, Quoc V. Le. Document Embedding with Paragraph Vectors. ArXiv 2015

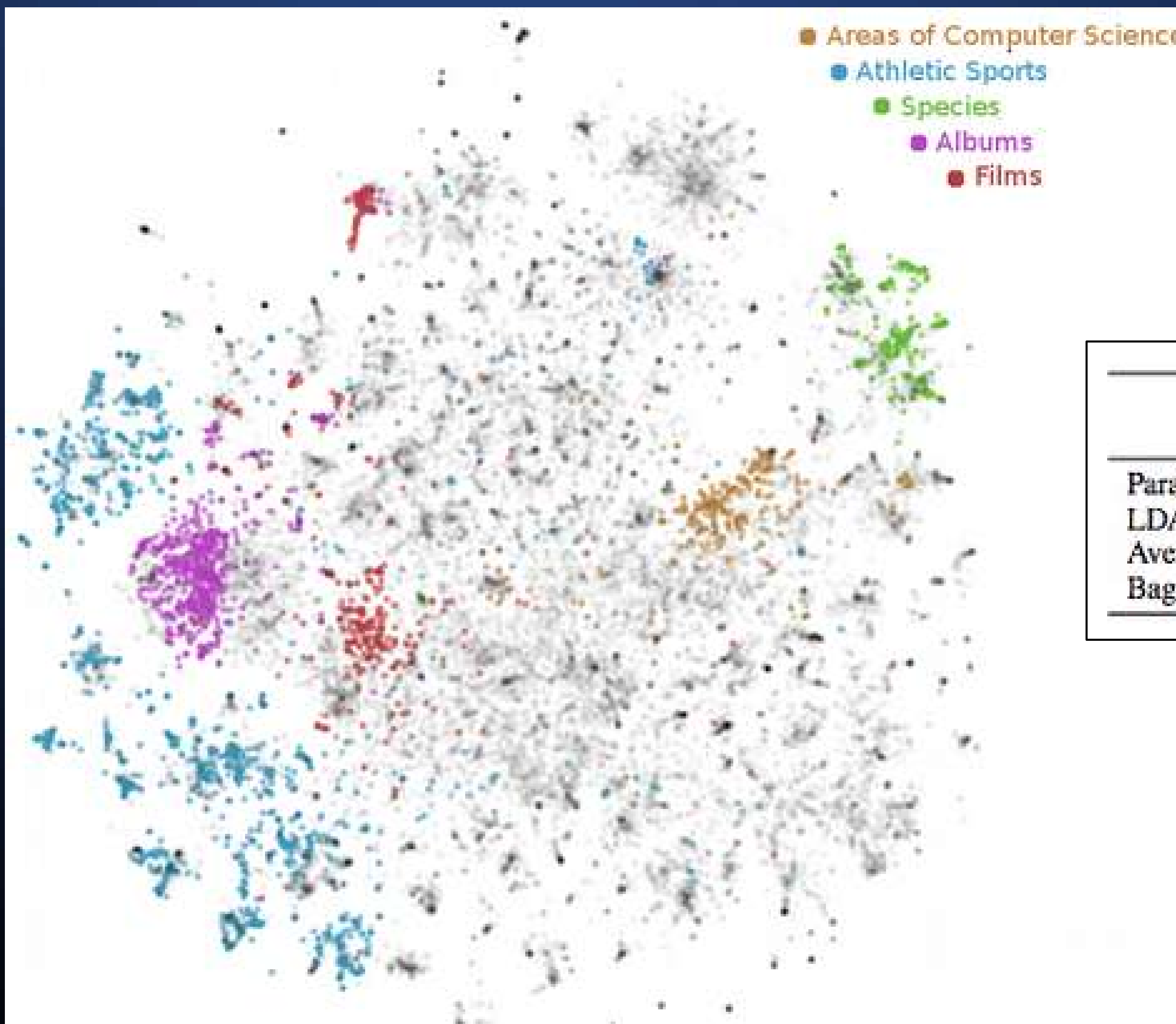
Paragraph Vectors

Title	Cosine Similarity
Evaluating Neural Word Representations in Tensor-Based Compositional Settings	0.771
Polyglot: Distributed Word Representations for Multilingual NLP	0.764
Lexicon Infused Phrase Embeddings for Named Entity Resolution	0.757
A Convolutional Neural Network for Modelling Sentences	0.747
Distributed Representations of Words and Phrases and their Compositionality	0.740
Convolutional Neural Networks for Sentence Classification	0.735
SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation	0.735
Exploiting Similarities among Languages for Machine Translation	0.731
Efficient Estimation of Word Representations in Vector Space	0.727
Multilingual Distributed Representations without Word Alignment	0.721

Nearest neighbours in 886,000 full arXiv papers
For "Distributed Representations of Sentences and Documents"
(original Paragraph Vectors paper)

Andrew M. Dai, Christopher Olah, Quoc V. Le. Document Embedding with Paragraph Vectors. ArXiv 2015

Paragraph Vectors



Wikipedia articles
Embedded using
Paragraph vectors

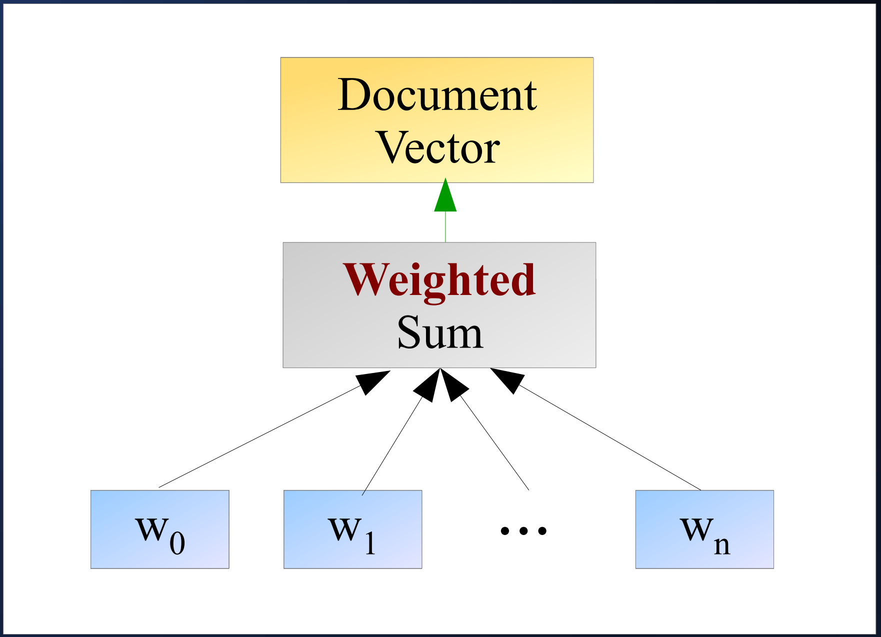
Model	Embedding dimensions/topics	Accuracy
Paragraph vectors	10000	93.0%
LDA	5000	82%
Averaged word embeddings	3000	84.9%
Bag of words		86.0%

Neural models
do outperform
TF-IDF Bag-of-words

Andrew M. Dai, Christopher Olah, Quoc V. Le. Document Embedding with Paragraph Vectors. ArXiv 2015

Word Vector-based Document Vectors

$$\vec{v}_S = \frac{1}{|S|} \sum_{w \in S} \alpha_{S,w} \, \vec{v}_w$$



Additional weights

E.g. 0 for stop words
IDF

Word Vector-based Doc. Vectors: Cross-Lingual Evaluation

Model	Dim	en → de	de → en
Majority class	40	46.8	46.8
MT	40	68.1	67.4
I-Matrix (Klementiev et al., 2012)	40	77.6	71.1
BAE-cr (Sarath Chandar et al., 2014)	40	**91.8**	74.2
CVM-Add (Hermann and Blunsom, 2014)	40	86.4	74.7
DWA (Kočiský et al., 2014)	40	83.1	75.4
BilBOWA (Gouws et al., 2015)	40	86.5	75
UnsupAlign (Luong et al., 2015)	40	87.6	77.8
Trans-gram (Coulmance et al., 2015)	40	87.8	78.7
BRAVE-S(EP)	40	88.1	**78.9**
BRAVE-D(CL-APR)	40	69.4	67.9
CVM-BI (Hermann and Blunsom, 2014)	128	86.1	79.0
UnsupAlign (Luong et al., 2015)	128	88.9	77.4
BRAVE-S(EP)	128	89.7	**80.1**
BRAVE-D(CL-APR)	128	70.4	70.6

Although neural document models do well (e.g. BRAVE), simply using IDF (or TF-IDF)-weighted word vector sums is almost as good

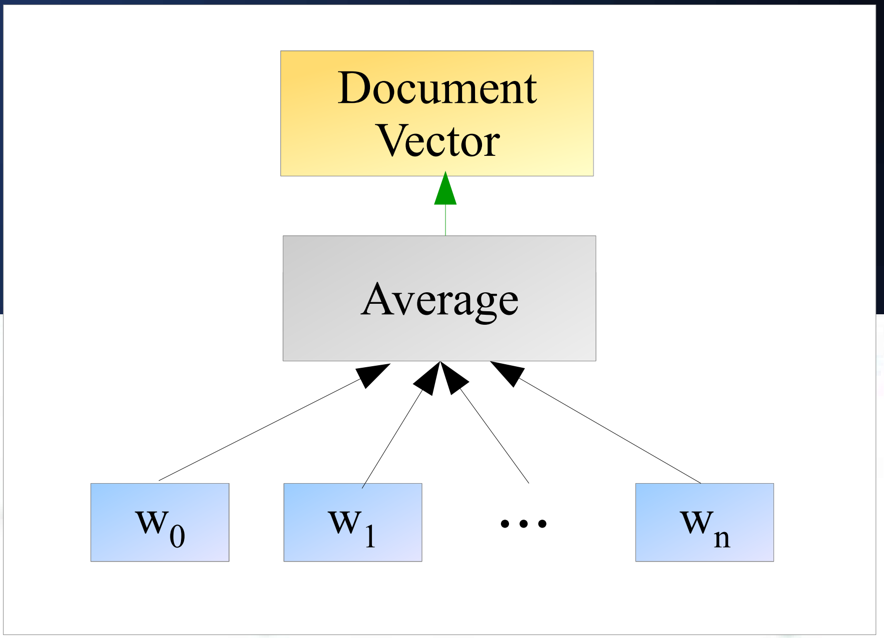
Source: Aditya Mogadala & Achim Rettinger. Bilingual Word Embeddings from Parallel and Non-Parallel Corpora for Cross-Language Text Classification

N-Gram Vector Averaging as in fastText

Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov. Bag of Tricks for Efficient Text Classification. ACL 2017

N-Gram Vector Averaging as in fastText

$$\vec{v}_S = \frac{1}{|S|} \sum_{w \in S} \vec{v}_w$$



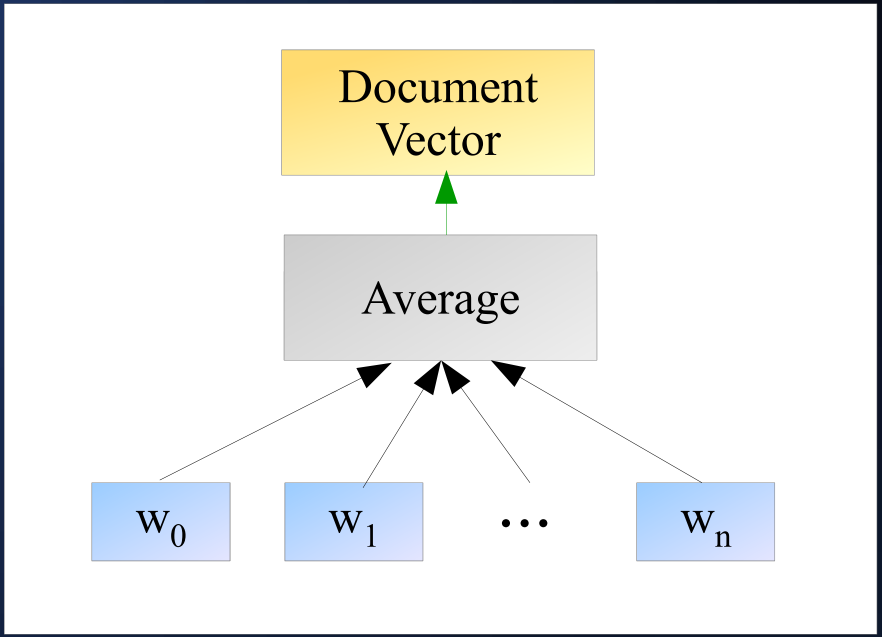
Consider S not as bag of words, but as bag of n-grams

Word vectors for unigrams obtained using regular fastText approach.
For n-grams, use feature hashing with 10M or 100M bins.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov. Bag of Tricks for Efficient Text Classification. ACL 2017

Doc2VecC: Doc2Vec with Corruption

$$\vec{v}_S = \frac{1}{|S|} \sum_{w \in S} \vec{v}_w$$

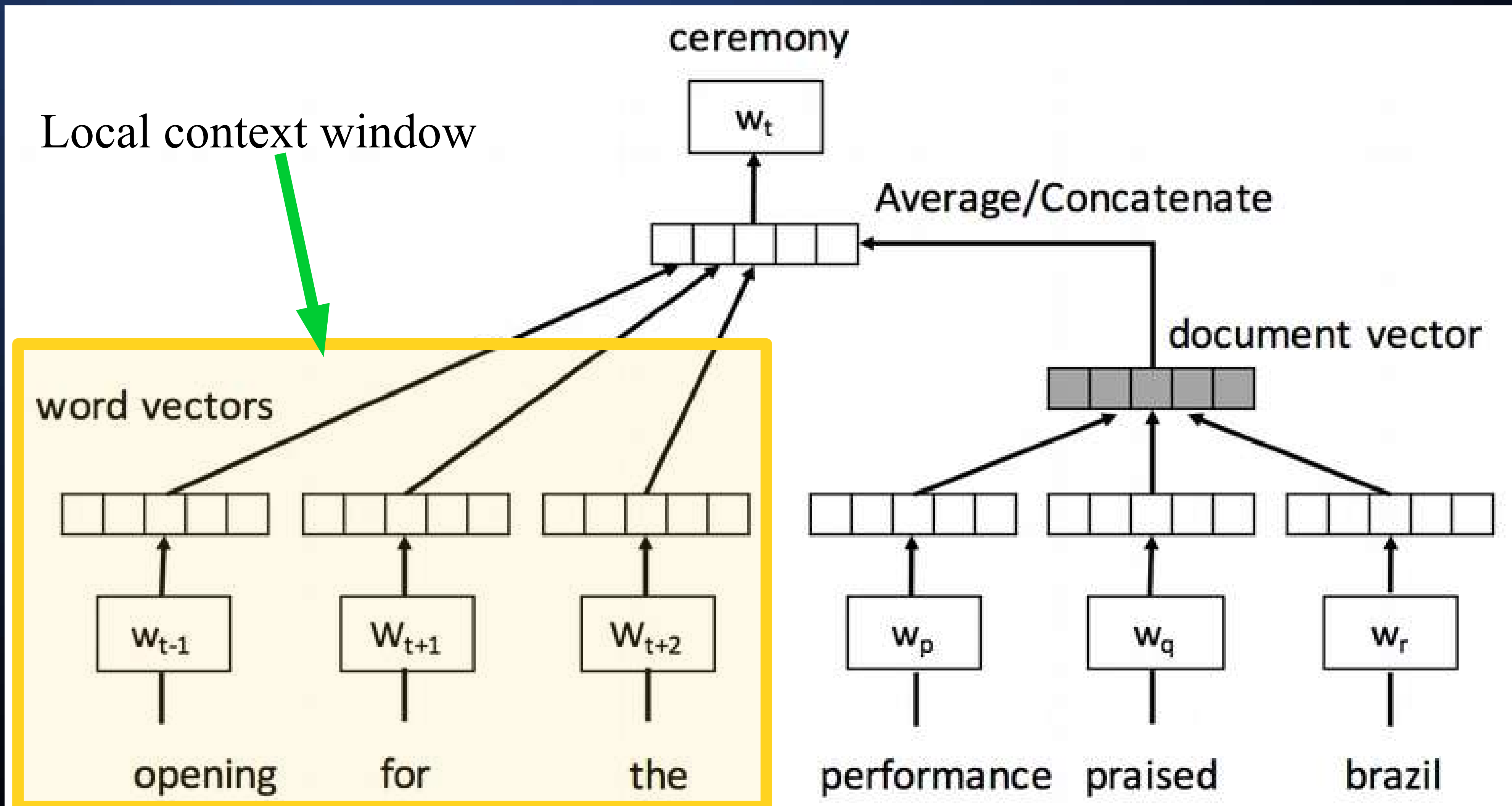


Simple averaging

However, the word vectors are learned using a drop-out like procedure.

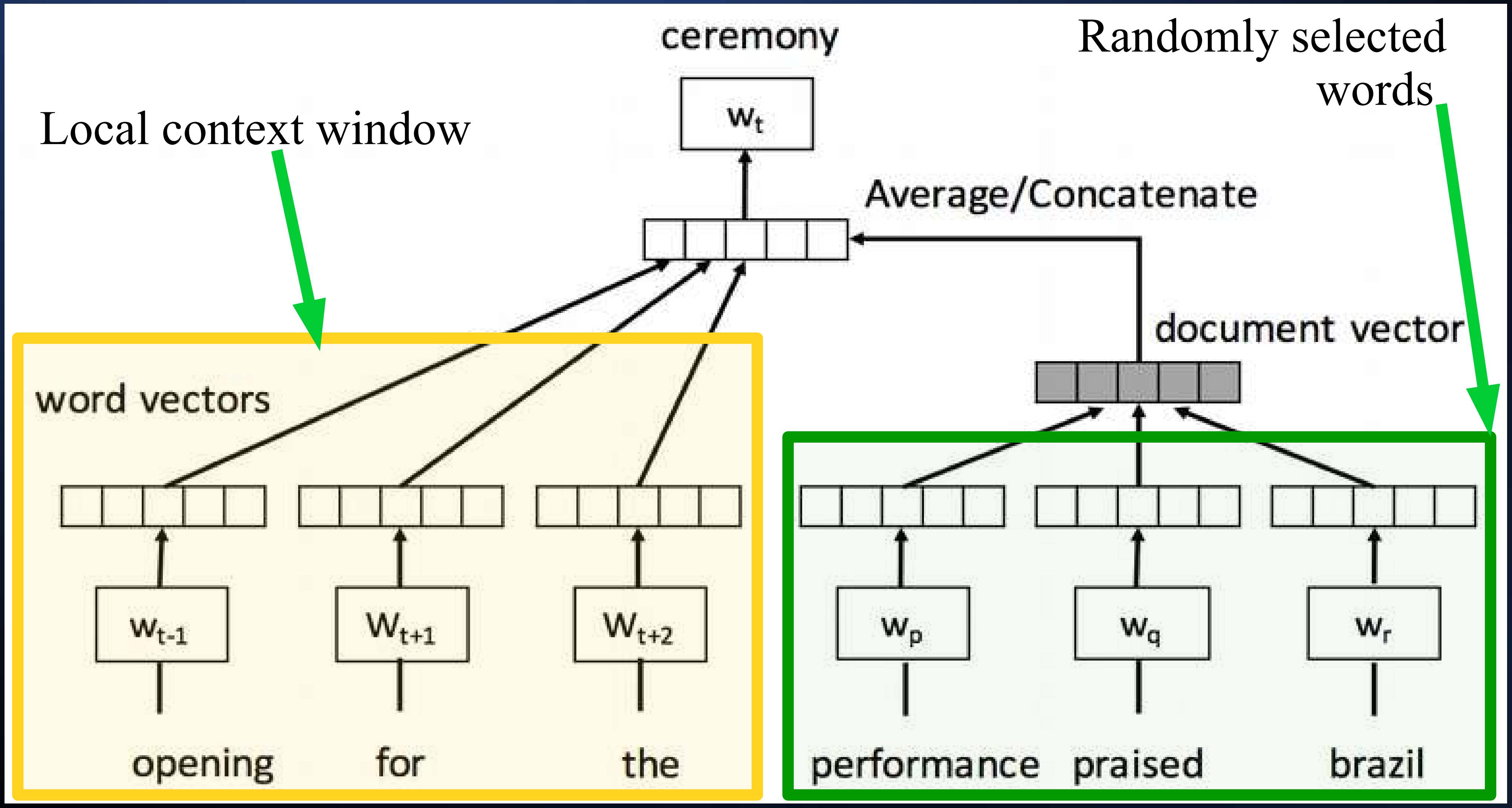
Minmin Chen. Efficient Vector Representation for Documents Through Corruption. ICLR 2017

Doc2VecC: Doc2Vec with Corruption



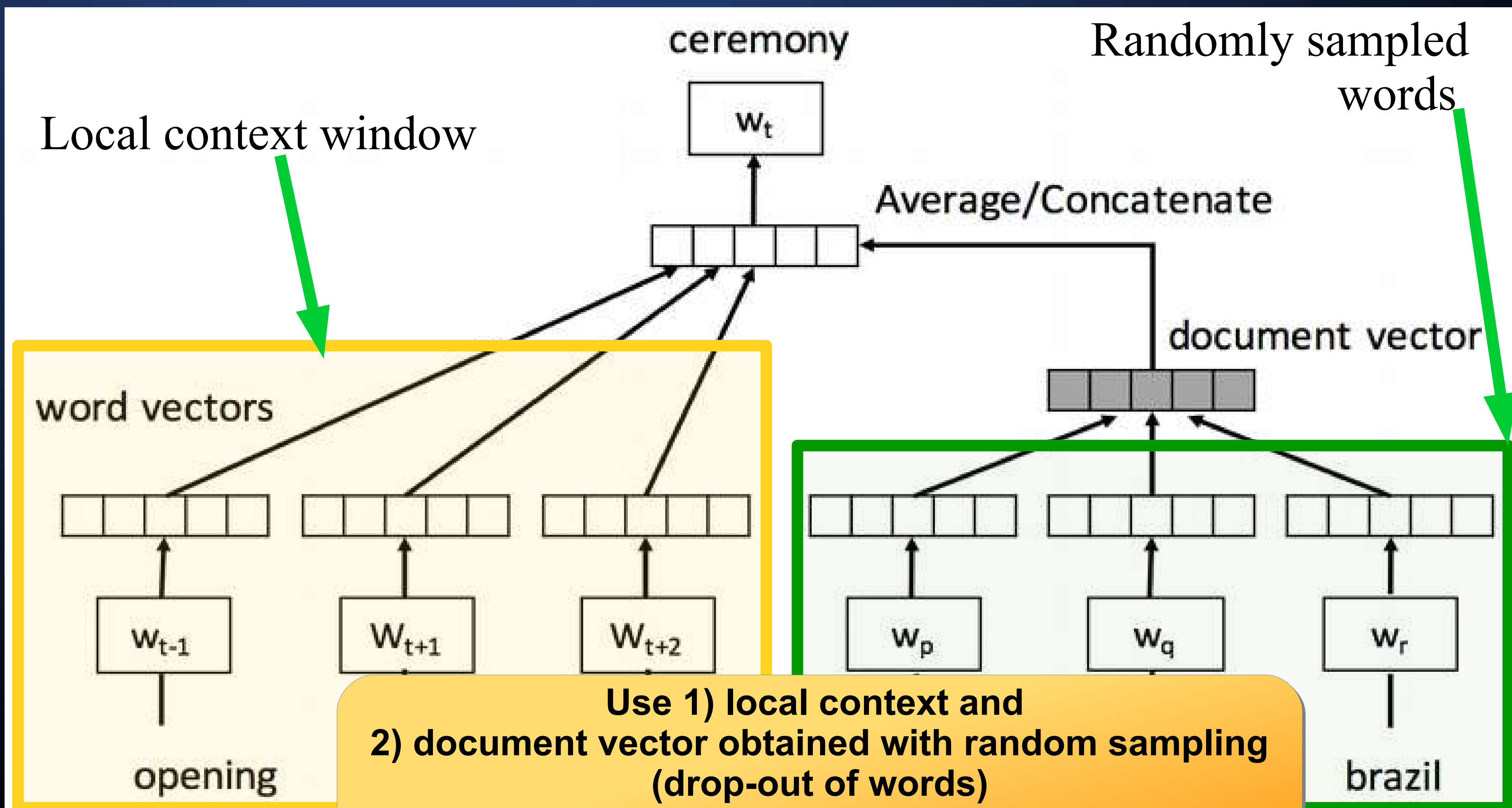
Minmin Chen. Efficient Vector Representation for Documents Through Corruption. ICLR 2017

Doc2VecC: Doc2Vec with Corruption



Minmin Chen. Efficient Vector Representation for Documents Through Corruption. ICLR 2017

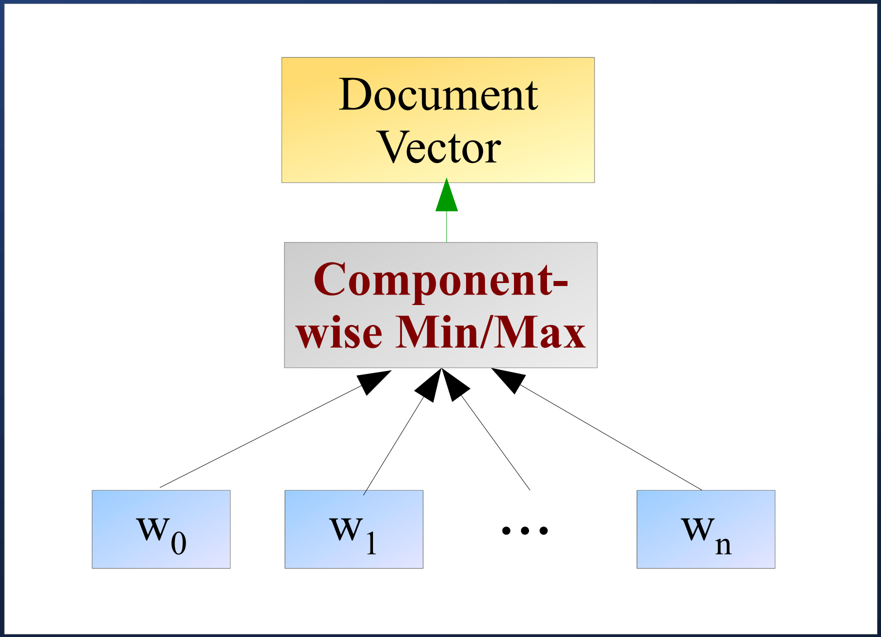
Doc2VecC: Doc2Vec with Corruption



Use 1) local context and
2) document vector obtained with random sampling (drop-out of words)
in order to predict target word.
Back-propagation to update embeddings.

Minmin Chen. Efficient Vector Representation for Documents Through Corruption. ICLR 2017

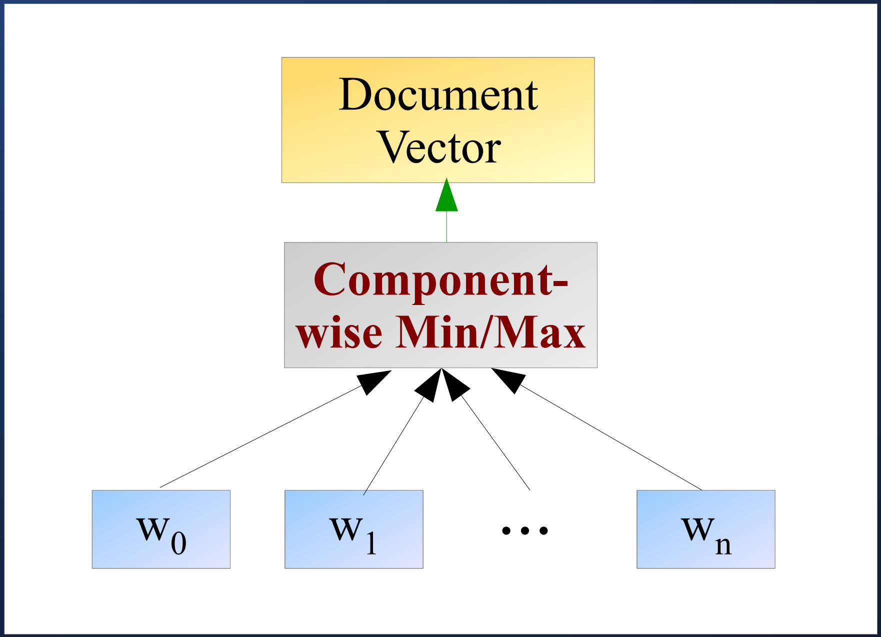
Component-Wise Aggregation



Min. Aggregation:	For each dimension, take the min. value across all word vectors
Max. Aggregation:	For each dimension, take the max. value across all word vectors
Min/Max Aggregation:	Concatenate Min./Max. Aggregation Vectors

Cedric De Boom, Steven Van Canneyt, Thomas Demeester, Bart Dhoedt (2016).
Representation learning for very short texts using weighted word embedding aggregation. Pattern Recognition Letters

Component-Wise Aggregation



Min. Aggregation:	For each dimension, take the min. value across all word vectors
Max. Aggregation:	For each dimension, take the max. value across all word vectors
Min/Max Aggregation:	Concatenate Min./Max. Aggregation Vectors

Performed best

Cedric De Boom, Steven Van Canneyt, Thomas Demeester, Bart Dhoedt (2016).
Representation learning for very short texts using weighted word embedding aggregation. Pattern Recognition Letters

Exploit Document Labels

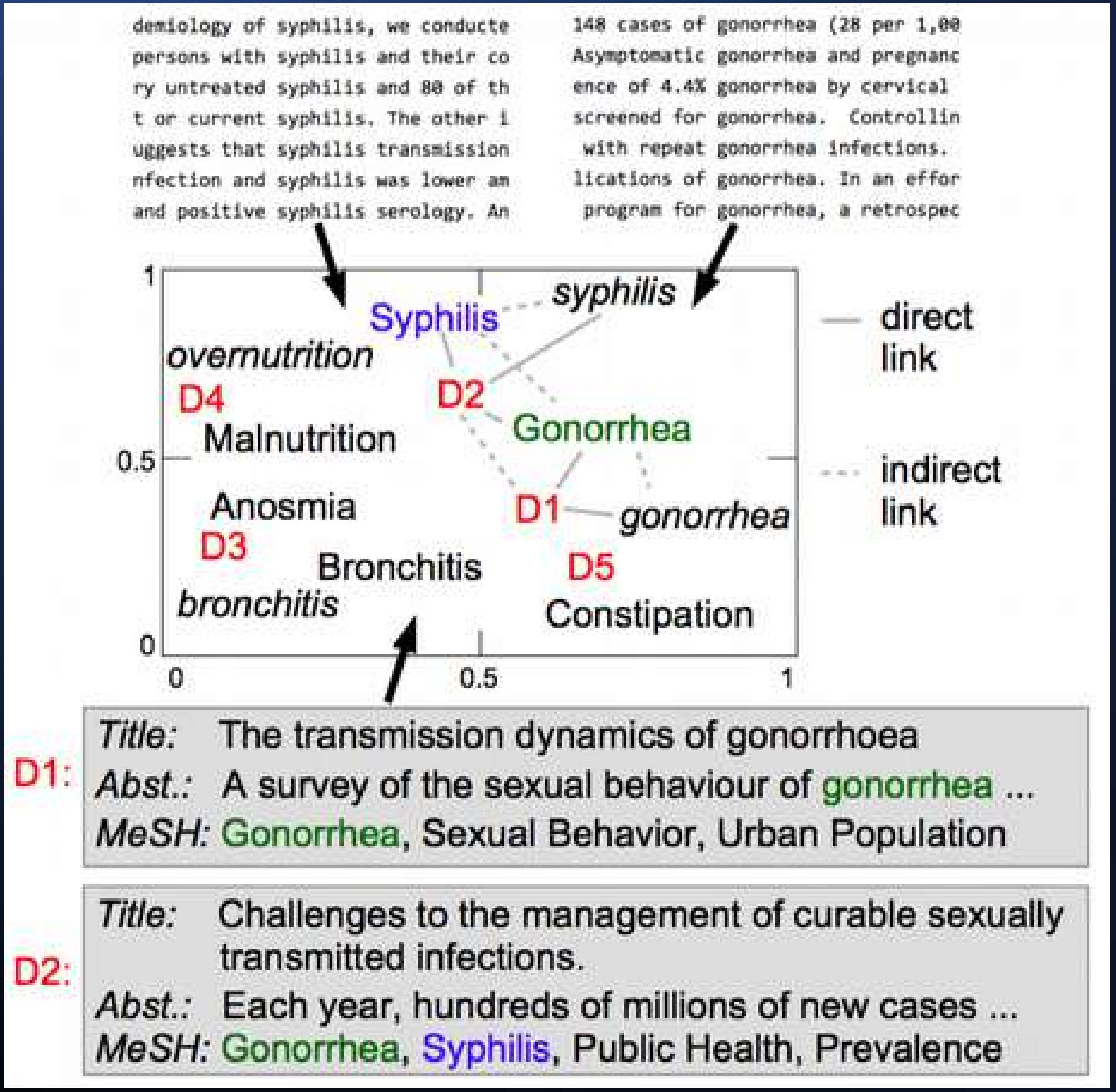
Title: Spotlight on Global Malnutrition: A Continuing Challenge in the 21st Century.
Abstract: Malnutrition as undernutrition, overnutrition, or an imbalance of specific nutrients, can be found in all countries and in both community and hospital settings around the world. The prevalence of malnutrition is unacceptably high . . .
MeSH terms: Acute Disease, Chronic Disease, Food Habits, Global Health, Humans, Malnutrition, Nutritional, Support, Overnutrition, Risk Factors, Socioeconomic Factors

Title: Fetal and early-postnatal developmental patterns of obese-genotype piglets exposed to prenatal programming by maternal over- and undernutrition.
Abstract: The present study evaluated the effect of nutritional imbalances during pregnancy, either by excess or deficiency, on fertility and conceptus development in obese-genotype swine (Iberian pig). Twenty-five multiparous sows were . . .
MeSH terms: Animals, Newborn Animals, Body Weight, Fetal Development, Genotype, Malnutrition, Obesity, Overnutrition, Pregnancy, Prenatal Exposure Delayed Effects, Swine

Title: Predictors of maternal and child double burden of malnutrition in rural Indonesia and Bangladesh
Abstract: BACKGROUND: Many developing countries now face the double burden of malnutrition, defined as the coexistence of a stunted child and overweight mother within the same household. OBJECTIVE: This study sought to . . .
MeSH terms: Adult, Body Mass Index, Preschool Child, Cost of Illness, Cross-Sectional Studies, Developing Countries, Family Characteristics, Humans, Indonesia, Infant, Logistic Models, Malnutrition, Mothers, Overnutrition, Population Surveillance, Prevalence Risk Factors, Rural Health, Urban Health

Loza Mencía, de Melo, Nam (2016). Medical Concept Embeddings via Labeled Background Corpora

Exploit Document Labels



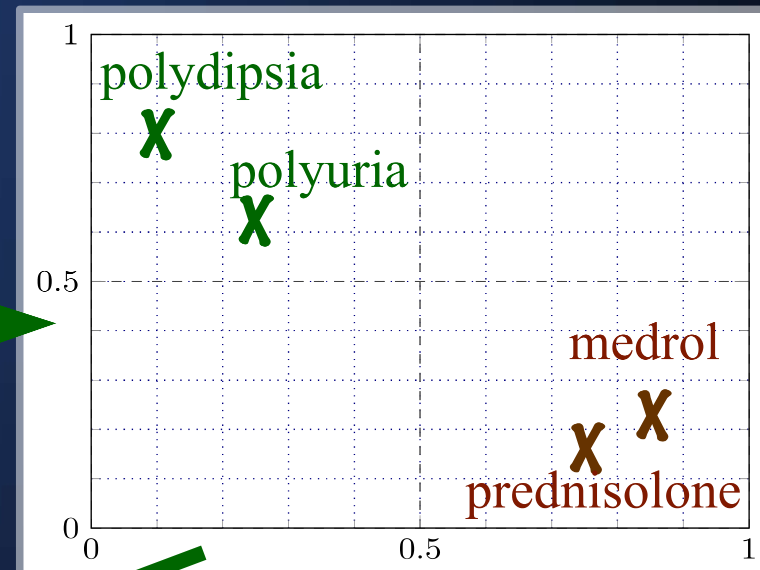
Loza Mencía, de Melo, Nam (2016). Medical Concept Embeddings via Labeled Background Corpora

Exploit Document Labels: Biomedical Representations

Title: Spotlight on Global Malnutrition: A Continuing Challenge in the 21st Century.
Abstract: Malnutrition as undernutrition, overnutrition, or an imbalance of specific nutrients, can be found in all countries and in both community and hospital settings around the world. The prevalence of malnutrition is unacceptably high …
MeSH terms: Acute Disease, Chronic Disease, Food Habits, Global Health, Humans, Malnutrition, Nutritional, Support, Overnutrition, Risk Factors, Socioeconomic Factors

Title: Fetal and early-postnatal developmental patterns of obese-genotype piglets exposed to prenatal programming by maternal over- and undernutrition.
Abstract: The present study evaluated the effect of nutritional imbalances during pregnancy, either by excess or deficiency, on fertility and conceptus development in obese-genotype swine (Iberian pig). Twenty-five multiparous sows were …
MeSH terms: Animals, Newborn Animals, Body Weight, Fetal Development, Genotype, Malnutrition, Obesity, Overnutrition, Pregnancy, Prenatal Exposure Delayed Effects, Swine

Title: Predictors of maternal and child double burden of malnutrition in rural Indonesia and Bangladesh
Abstract: BACKGROUND: Many developing countries now face the double burden of malnutrition, defined as the coexistence of a stunted child and overweight mother within the same household. OBJECTIVE: This study sought to …
MeSH terms: Adult, Body Mass Index, Preschool Child, Cost of Illness, Cross-Sectional Studies, Developing Countries, Family Characteristics, Humans, Indonesia, Infant, Logistic Models, Malnutrition, Mothers, Overnutrition, Population Surveillance, Prevalence Risk Factors, Rural Health, Urban Health



Angina	Dyspnea
Xanax	Ativan
Hernias	Earache
Ataxia	Ethanol
Overnutrition	Malnutrition
Cirrhosis	Hematemesis
Anosmia	Constipation
Pallor	Iron
Starvation	Anorexia

Loza Mencía, de Melo, Nam (2016). Medical Concept Embeddings via Labeled Background Corpora

Application: Cross-Lingual Text Classification

- Given: training documents with class labels
- Goal: guess class labels for test documents in some other language
- better than plain machine translation

Map to concepts

	Reuters Spanish			
	Topics		Geography	
	F_1	error rate	F_1	error rate
B	80.97	18.61 ±0.30	81.86	18.12 ±0.30
CM	89.23	10.49 ±0.24	85.74	14.58 ±0.28
ORM	89.53	10.36 ±0.24	87.33	12.97 ±0.26
ORM+B	91.88	8.04 ±0.21	91.92	8.22 ±0.21
T	90.96	8.80 ±0.22	88.76	11.43 ±0.25
TCM	90.75	9.06 ±0.22	91.12	9.16 ±0.23
TORM	91.12	8.74 ±0.22	93.89	6.28 ±0.19
TORM+T	92.46	7.43 ±0.20	94.44	5.68 ±0.18

	Wikipedia Japanese	
	F_1	error rate
T	86.26	14.00 ±0.38
TCM	85.38	15.10 ±0.40
TORM	86.67	13.52 ±0.38
TORM+T	87.29	12.86 ±0.37

Gerard de Melo, Stefan Siersdorfer. Multilingual Text Classification using Ontologies

Application: Cross-Lingual Text Classification

- Given: training documents with class labels
- Goal: guess class labels for test documents in some other language
- better than plain machine translation

Expand concepts

	Reuters Spanish						
	Topics			Geography			
	F_1	error rate		F_1	error rate		
B	80.97	18.61	±0.30	81.86	18.12	±0.30	
CM	89.23	10.49	±0.24	85.74	14.58	±0.28	
ORM	89.53	10.36	±0.24	87.33	12.97	±0.26	
ORM+B	91.88	8.04	±0.21	91.92	8.22	±0.21	
T	90.96	8.80	±0.22	88.76	11.43	±0.25	
TCM	90.75	9.06	±0.22	91.12	9.16	±0.23	
TORM	91.12	8.74	±0.22	93.89	6.28	±0.19	
TORM+T	92.46	7.43	±0.20	94.44	5.68	±0.18	

	Wikipedia Japanese		
	F_1	error rate	
T	86.26	14.00	±0.38
TCM	85.38	15.10	±0.40
TORM	86.67	13.52	±0.38
TORM+T	87.29	12.86	±0.37

Gerard de Melo, Stefan Siersdorfer. Multilingual Text Classification using Ontologies

Application: Dataless Text Classification

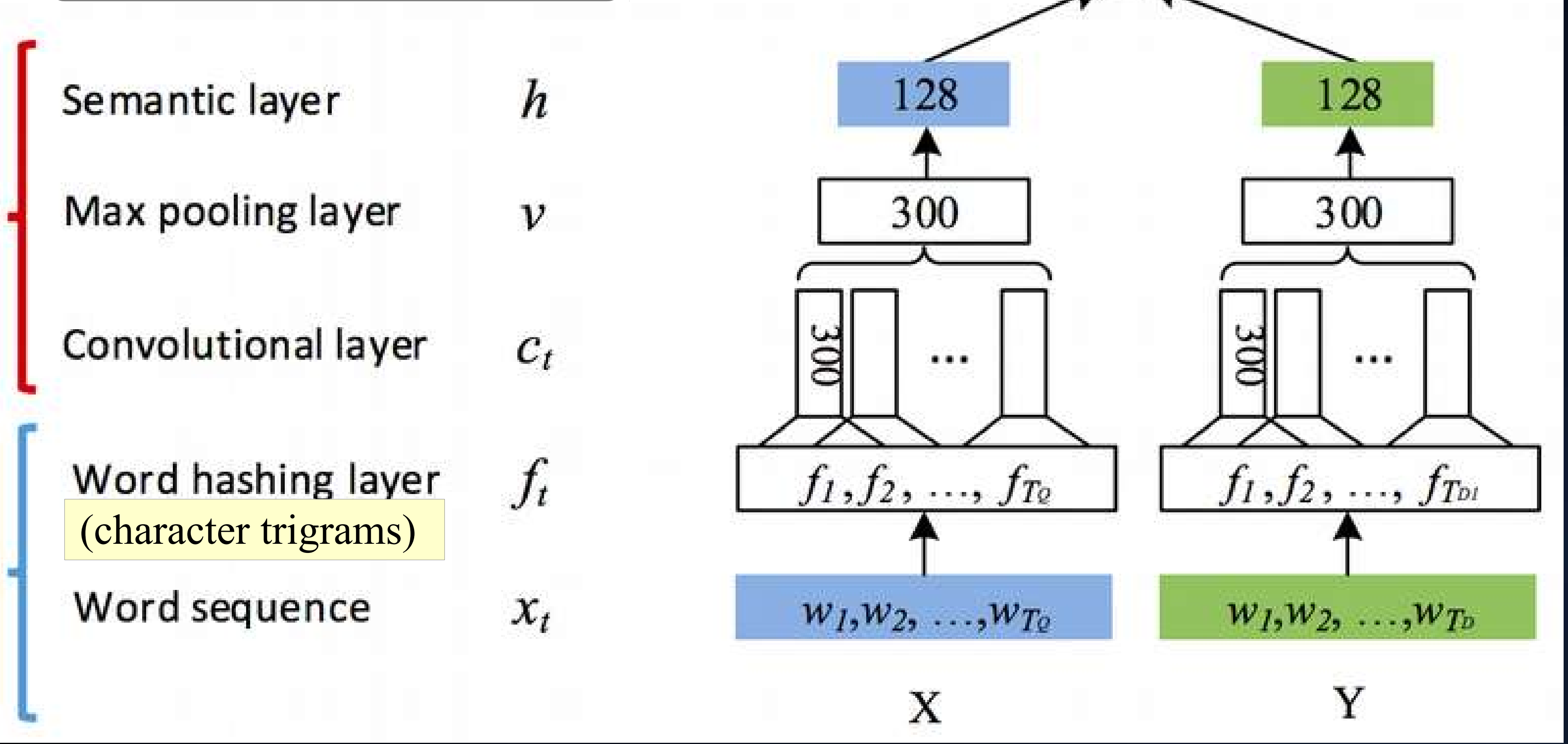
Newsgroup Name	Expanded Label
talk.politics.guns	politics guns
talk.politics.mideast	politics mideast
talk.politics.misc	politics
alt.atheism	atheism
soc.religion.christian	society religion christianity christian
talk.religion.misc	religion
comp.sys.ibm.pc.hardware	computer systems ibm pc hardware
comp.sys.mac.hardware	computer systems mac macintosh apple hardware
sci.electronics	science electronics
comp.graphics	computer graphics
comp.windows.x	computer windows x windowsx
comp.os.ms-windows.misc	computer os operating system microsoft windows
misc.forsale	for sale discount
rec.autos	cars
rec.motorcycles	motorcycles
rec.sport.baseball	baseball
rec.sport.hockey	hockey
sci.crypt	science cryptography
sci.med	science medicine
sci.space	science space

Instead of supervision from labeled data, use proximity to representation of the label In concept-based representation space for classification

Chang et al. Importance of Semantic Representation: Dataless Classification
Song et al. Cross-Lingual Dataless Classification for Many Languages

Application: Information Retrieval via Siamese Models

Convolutional Deep Structured Semantic Model (CDSSM)

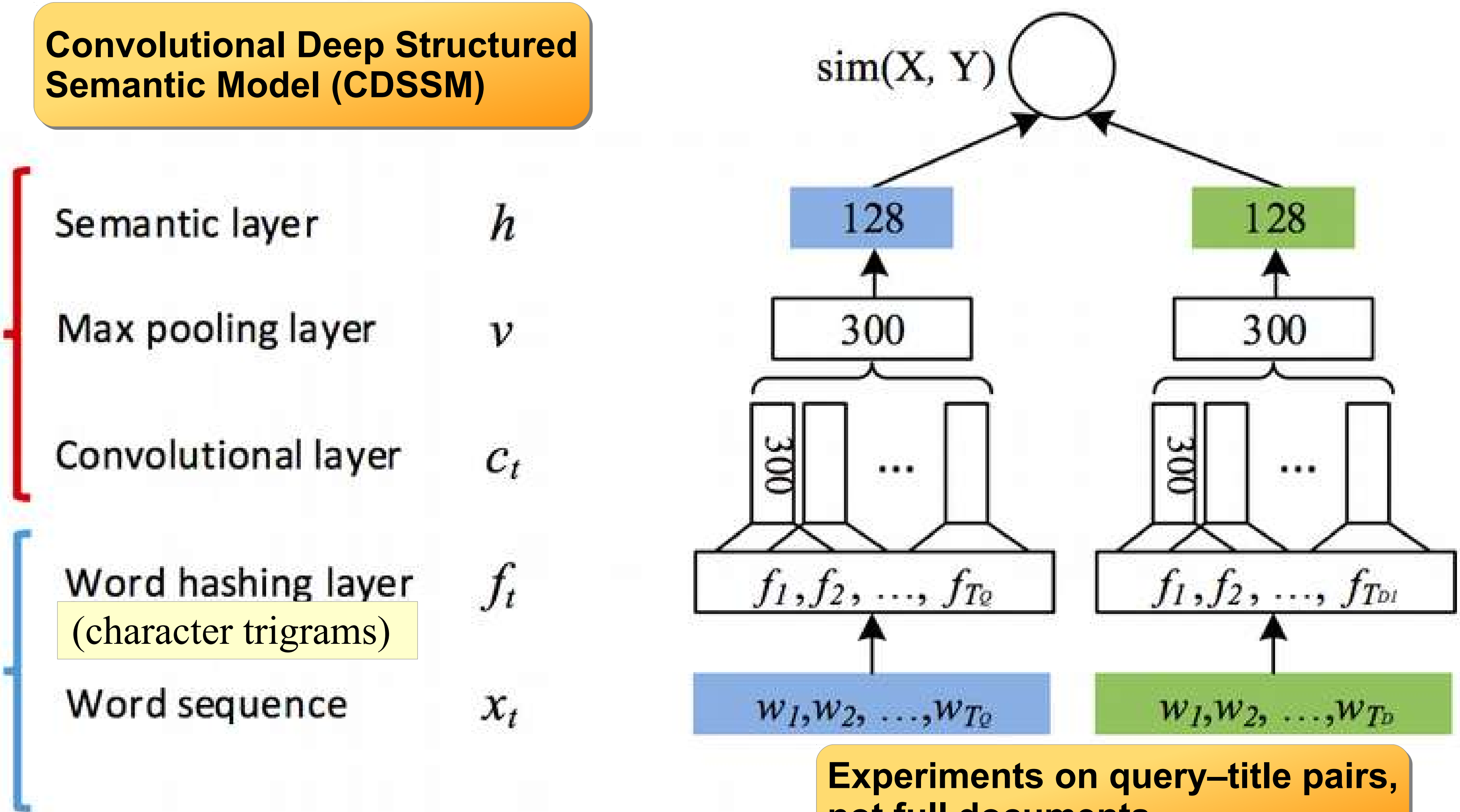


Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, Gregoire Mesnil. A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval

Image: Microsoft

Application: Information Retrieval via Siamese Models

Convolutional Deep Structured Semantic Model (CDSSM)

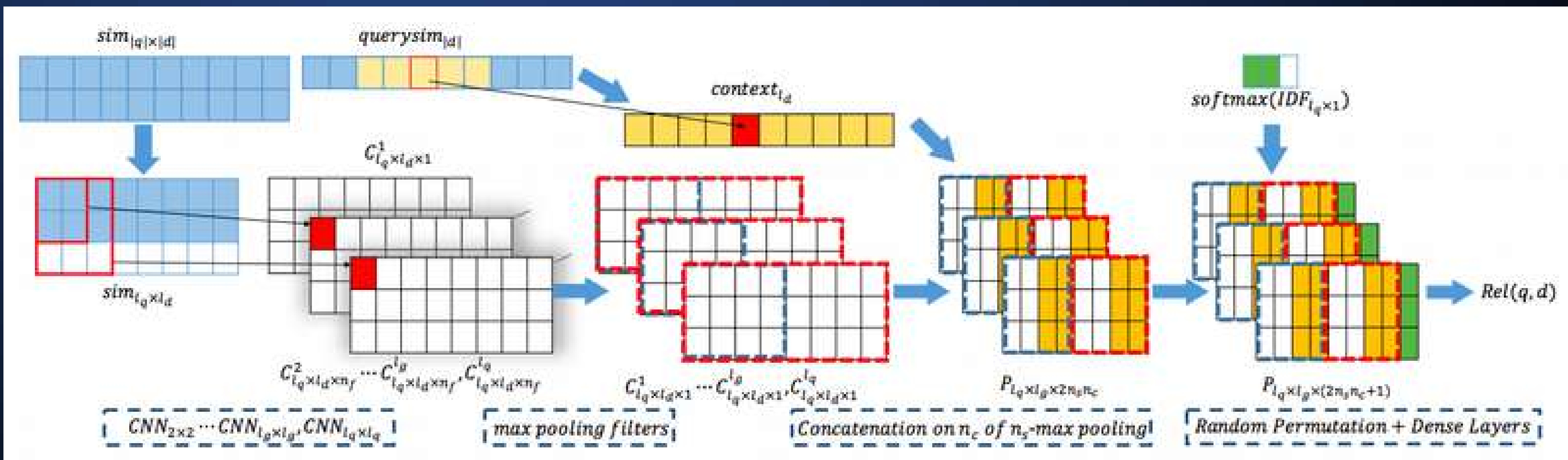


Experiments on query–title pairs, not full documents

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, Gregoire Mesnil. A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval

Image: Microsoft

Application: Information Retrieval via Relevance Matching (Co-PACRR)



<u>Asymmetry of Query and Document:</u>
Support for proximity of query term matches within document, but permutation of query term order as regularization

Kai Hui, Andrew Yates, Klaus Berberich, Gerard de Melo. PACRR: A Position-Aware Neural IR Model for Relevance Matching. EMNLP 2017.

Kai Hui, Andrew Yates, Klaus Berberich, Gerard de Melo. Co-PACRR: A Context-Aware Neural IR Model for Ad-hoc Retrieval. WSDM 2018