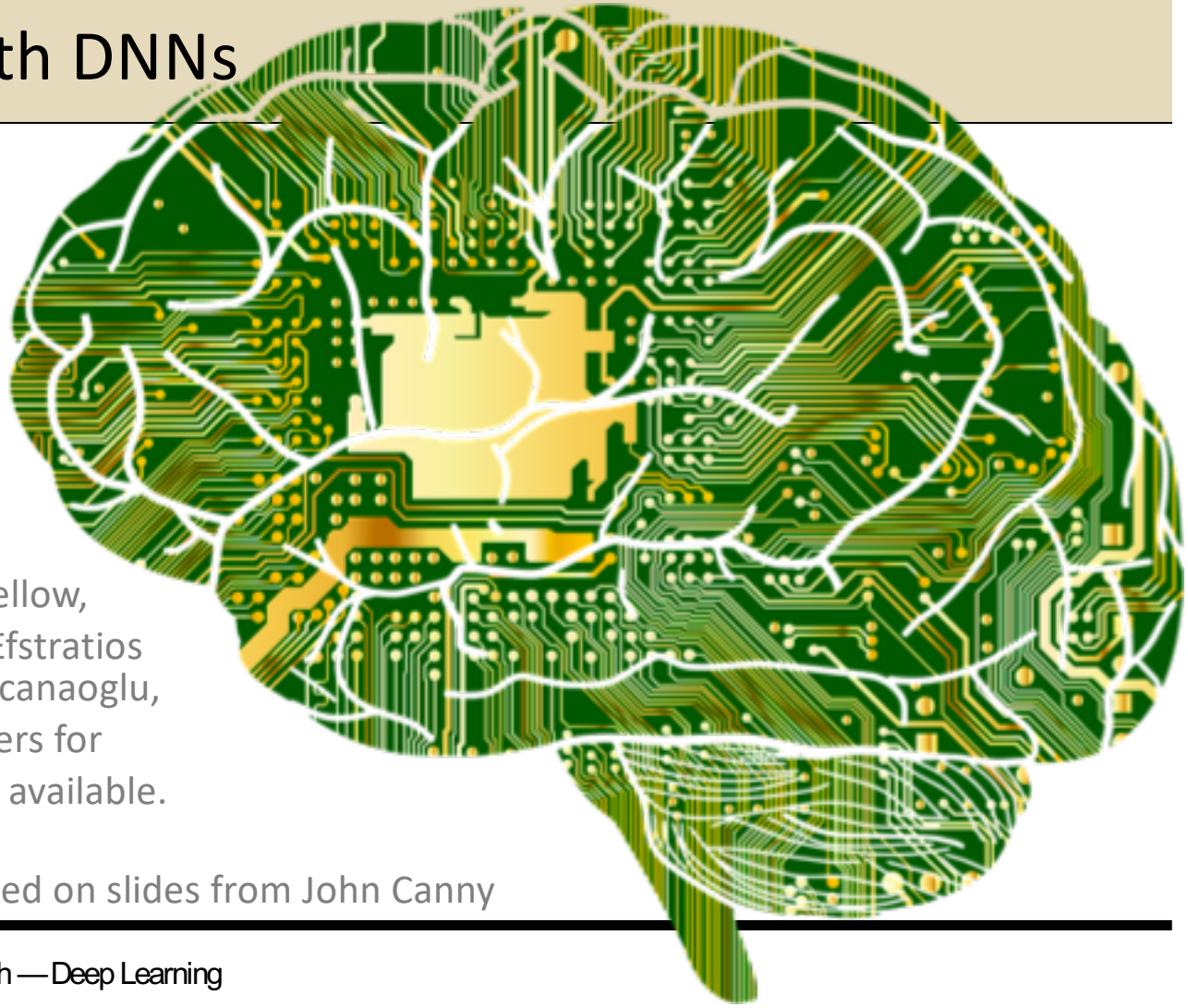
Deep Learning

Architectures and Methods:

Text Processing with DNNs

TECHNISCHE UNIVERSITÄT DARMSTADT



Thanks to John Canny, Ian Goodfellow, Yoshua Bengio, Aaron Courville, Efstratios Gavves, Kirill Gavrilyuk, Berkay Kicanaoglu, and Patrick Putzky and many others for making their materials publically available.

The present slides are mainly based on slides from John Canny

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Semantics

- Propositional models
- Matrix factorization
- Word2vec
- Skip-Thought vectors
- Siamese models

Translation + Structure Extraction

- Translation
- Parsing
- Entity-Relation extraction

Text Semantics

TECHNISCHE UNIVERSITÄT DARMSTADT

- In Natural Language Processing (NLP), ***semantics*** is concerned with the meanings of texts.
- There are two main approaches:
  - ***Propositional or formal semantics:*** A block of text is to converted into a formula in a logical language, e.g. predicate calculus.
  - ***Vector representation.*** Texts are ***embedded*** into a high-dimensional space.

Semantic Approaches

TECHNISCHE UNIVERSITÄT DARMSTADT

Propositional:

- "dog bites man" ➔ bites(dog, man)
- bites(*,*) is a binary relation. man, dog are objects.
- Probabilities can be attached.

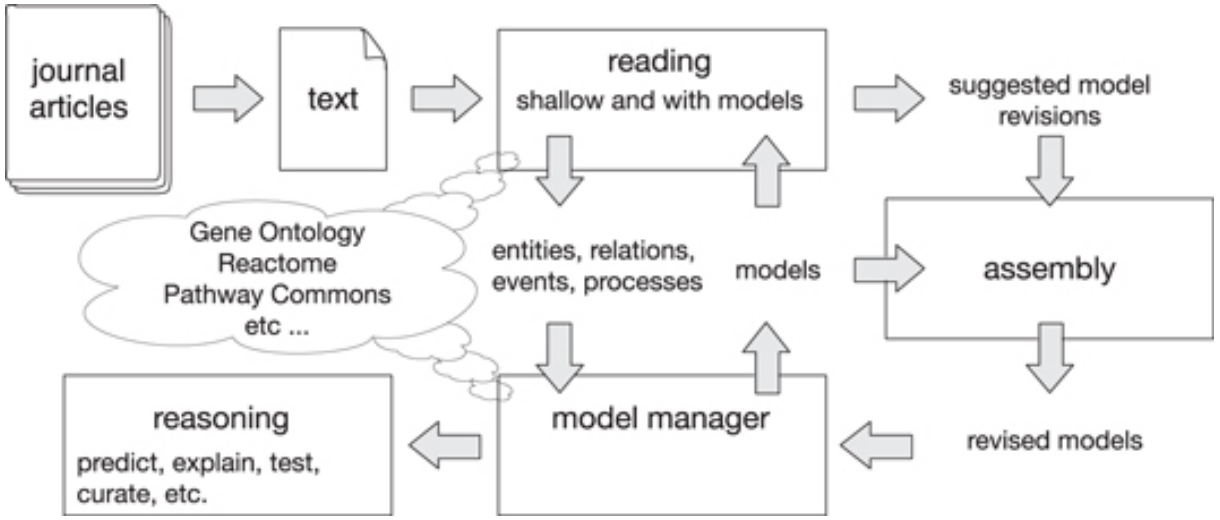
Vector representation:

- vec("dog bites man") = (0.2, -0.3, 1.5,…) $\in \mathfrak{R}^n$
- Sentences similar in meaning should be close to this embedding (e.g. use human judgments)

Propositional Semantics

- Allow logical inferences “Socrates is a man,” + “all men are mortal” → “Socrates is mortal”
- Important for inference in well-defined domains, e.g. inferring ***gene regulation*** from medical journals.

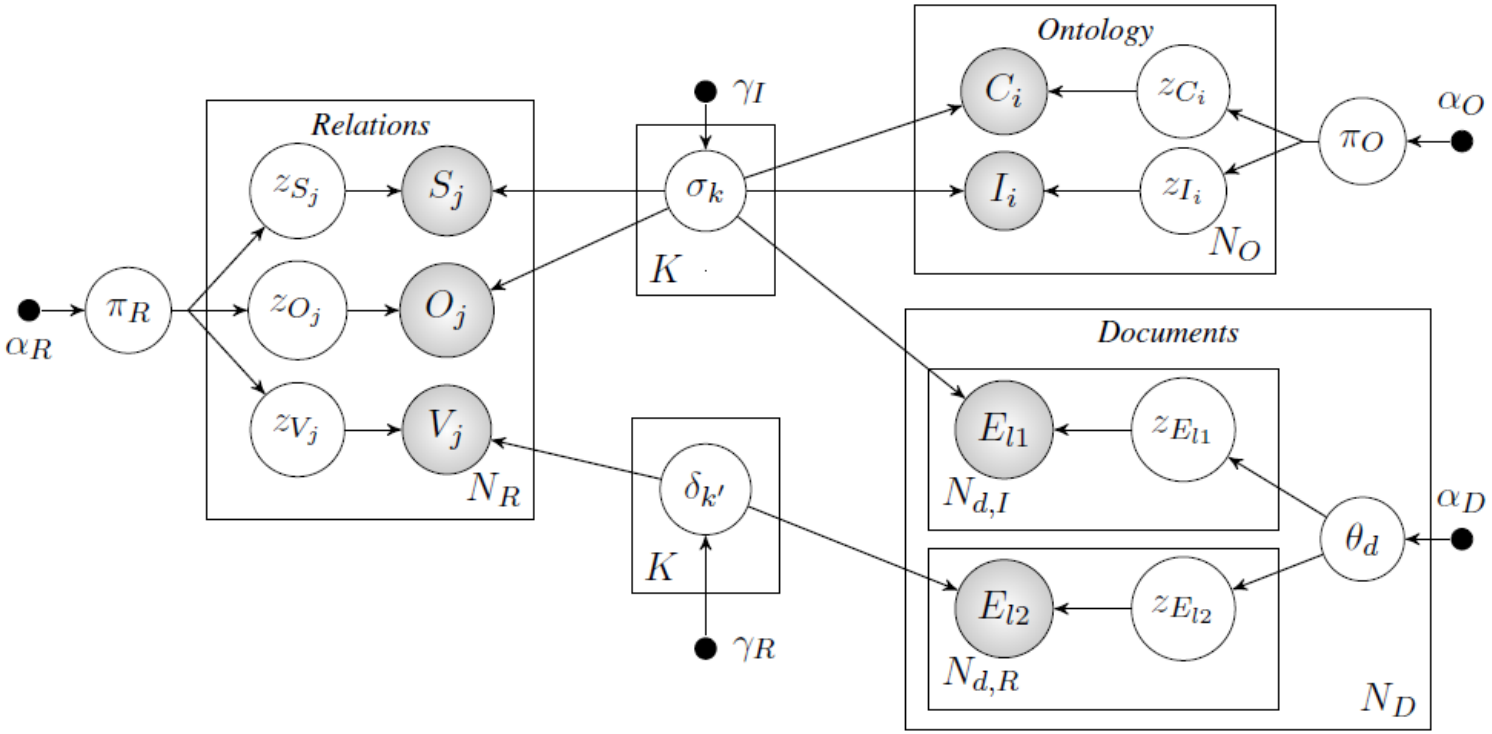
See DARPA’s “Big Mechanism” project



From “DARPA’s Big Mechanism program” Paul R Cohen, Phys. Biol. 12 (2015)

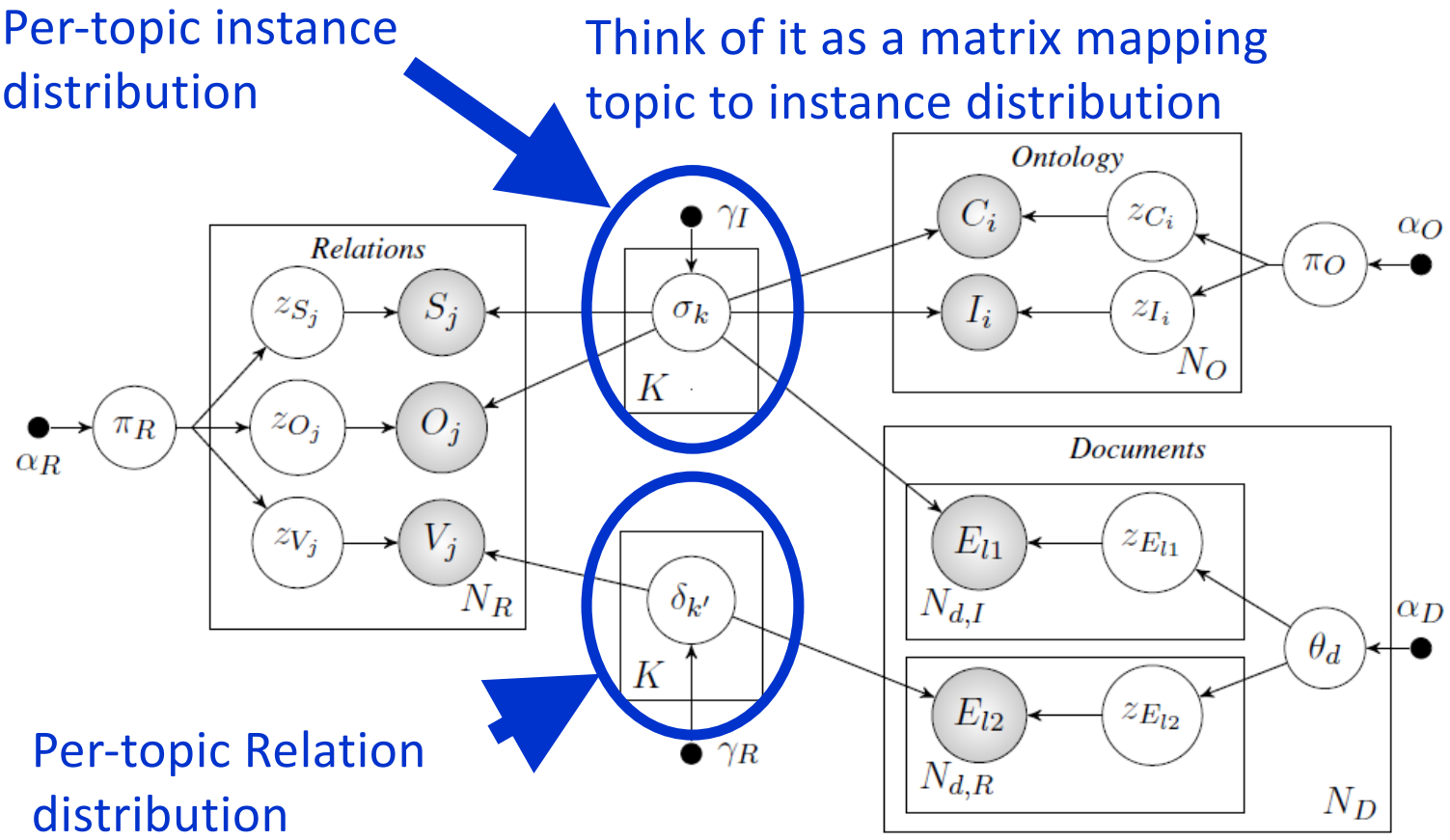
Propositional Semantics

- Contemporary approaches use latent variable models to group entities (objects) and the relations between them in a data-driven way.



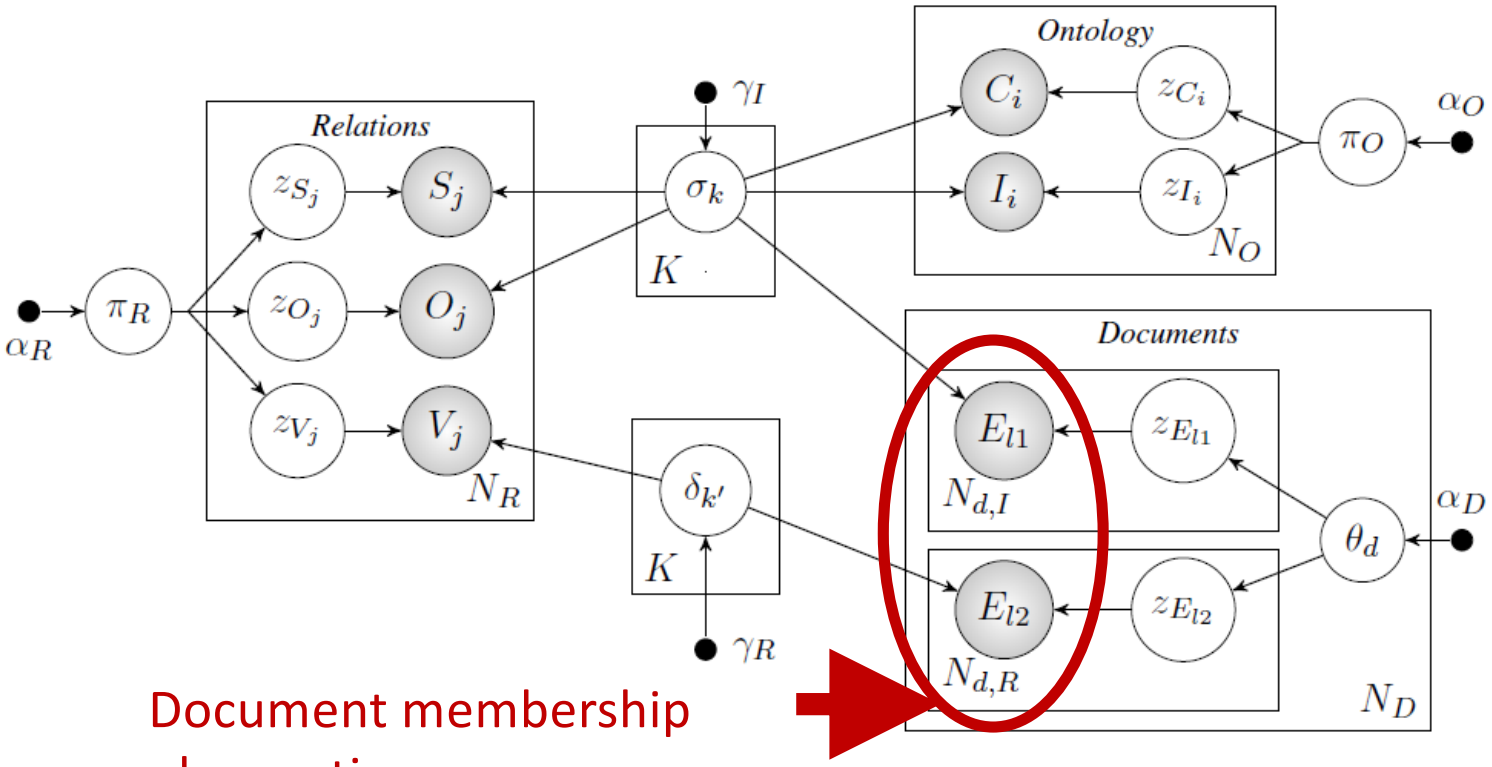
"KB-LDA: Jointly Learning a Knowledge Base of Hierarchy, Relations, and Facts," Dana Movshovitz-Attias. William W. Cohen, ACL 2015

Propositional Semantics



"KB-LDA: Jointly Learning a Knowledge Base of Hierarchy, Relations, and Facts," Dana Movshovitz-Attias. William W. Cohen, ACL 2015

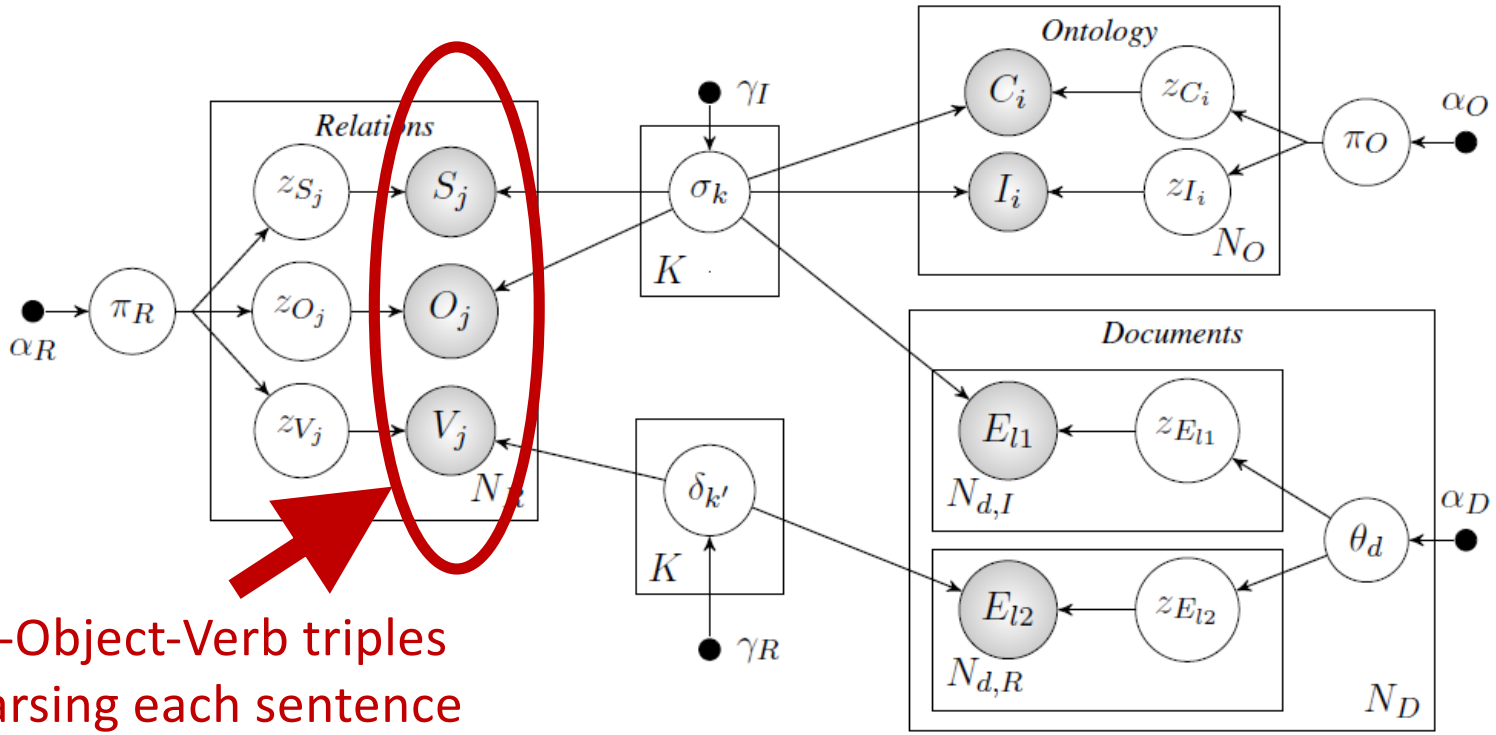
Propositional Semantics



Document membership observations

"KB-LDA: Jointly Learning a Knowledge Base of Hierarchy, Relations, and Facts," Dana Movshovitz-Attias. William W. Cohen, ACL 2015

Propositional Semantics

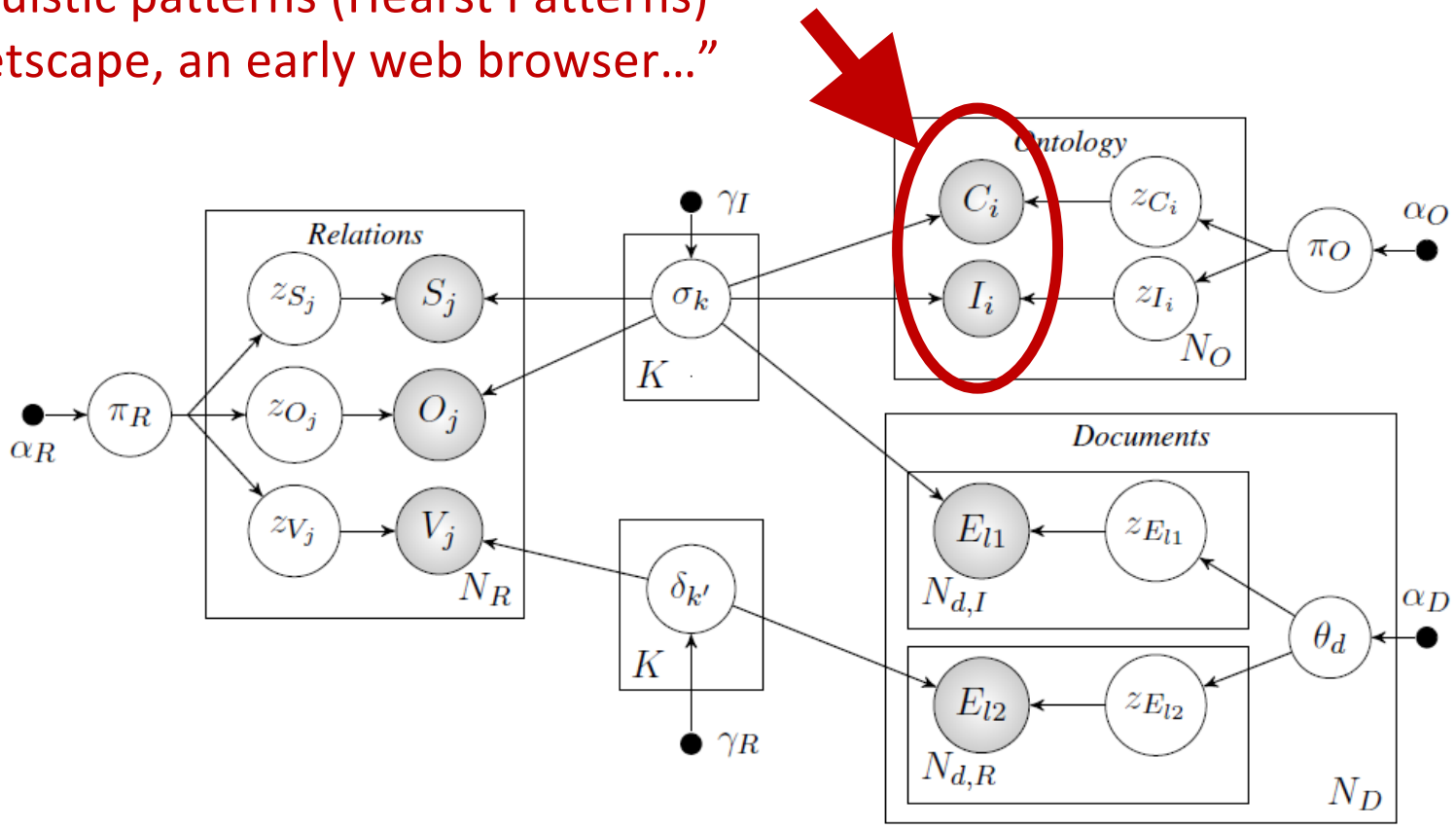


Subject-Object-Verb triples from parsing each sentence

"KB-LDA: Jointly Learning a Knowledge Base of Hierarchy, Relations, and Facts," Dana Movshovitz-Attias. William W. Cohen, ACL 2015

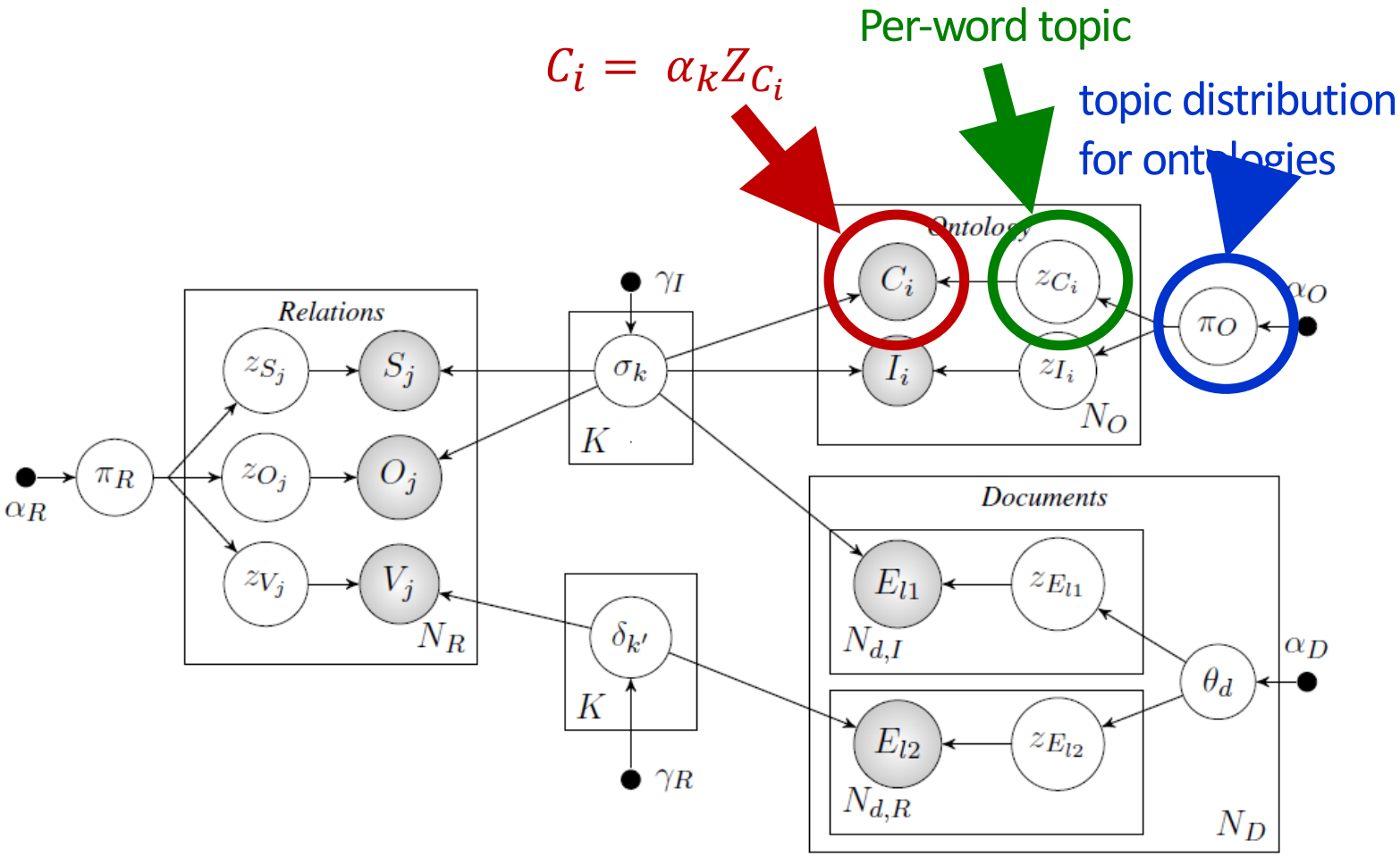
Propositional Semantics

Class-instance relations found from linguistic patterns (Hearst Patterns)
"Netscape, an early web browser…"



"KB-LDA: Jointly Learning a Knowledge Base of Hierarchy, Relations, and Facts," Dana Movshovitz-Attias. William W. Cohen, ACL 2015

Propositional Semantics



"KB-LDA: Jointly Learning a Knowledge Base of Hierarchy, Relations, and Facts," Dana Movshovitz-Attias. William W. Cohen, ACL 2015

KB-LDA

TECHNISCHE UNIVERSITÄT DARMSTADT

Top 2 Topic Concepts	Top 10 Topic Tokens
table, key	table, query, database, sql, column, data, tables, mysql, index, columns
properties, css	image, code, images, problem, point, color, data, size, screen, points
credentials, user information	name, images, id, number, text, password, address, strings, files, string
page, content	page, html, code, file, image, javascript, browser, http, jquery, js
orm tools, orm tool	tomcat, hibernate, server, boost, apache, spring, mongodb, framework, nhibernate, png
clients, apps	app, application, http, android, device, phone, code, api, iphone, google
applications, systems	devices, systems, applications, services, platforms, tools, sites, apps, system, service
systems, platforms	google, windows, linux, facebook, git, ant, database, gmail, android, so
limits, limit	memory, time, thread, code, threads, process, file, program, data, object
data, table	query, table, data, list, example, number, results, search, database, rows

"KB-LDA: Jointly Learning a Knowledge Base of Hierarchy, Relations, and Facts," Dana Movshovitz-Attias. William W. Cohen, ACL 2015

A blueprint for deep semantic network design?

Other Machine Reading Systems

TECHNISCHE UNIVERSITÄT DARMSTADT

Aristo from AI2: Allen Institute for Artificial Intelligence:

http://aristo-demo.allenai.org/

Vector Embedding of Words

Word embeddings depend on a notion of ***word similarity***.

A very useful definition is paradigmatic similarity:

Similar words occur in ***similar contexts***. They are ***exchangable***.

Yesterday { POTUS / The President / Obama } called a press conference

Vector Embedding of Words

Much of the work on text embedding has used word embeddings and bag-of-words representation:

vec("dog") = (0.2, -0.3, 1.5,…)

vec("bites") = (0.5, 1.0, -0.4,…)

vec("man") = (-0.1, 2.3, -1.5,…)

vec("dog bites man") = (0.6, 3.0, -0.4,…)

Vector Embedding: Word Similarity

Word embeddings depend on a notion of ***word similarity***.

A very useful definition is paradigmatic similarity:

Similar words occur in ***similar contexts***. They are ***exchangable***.

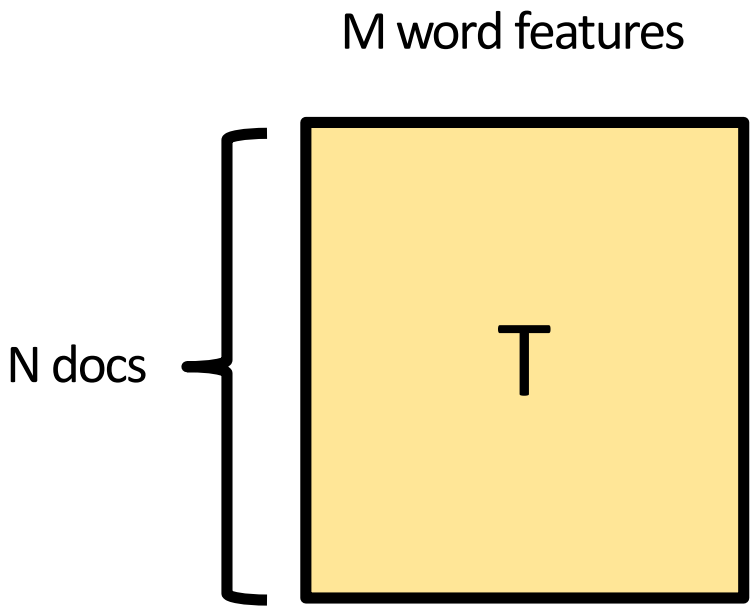
This definition supports unsupervised learning: cluster or embed words according to their contexts.

Embedding: Latent Semantic Analysis

Latent semantic analysis studies documents in **Bag-Of-Words format** (1988).

i.e. given a matrix T encoding some documents:

T_{ij} is the count* of word j in document i. Most entries are 0.



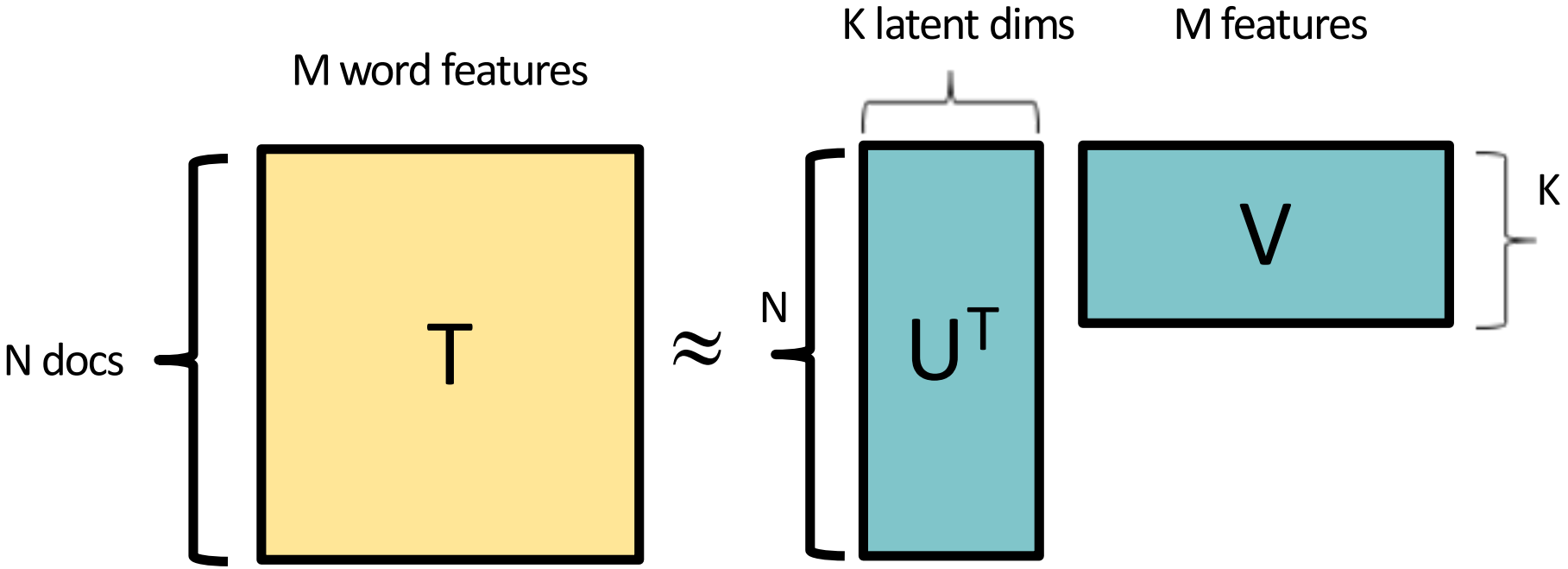
* Often tfidf or other “squashing” functions of the count are used.

Embedding: Latent Semantic Analysis

Given a bag-of-words matrix T, compute a factorization $T \approx U^T * V$ (e.g. a best L_2 approximation to T)

Factors encode similar ***whole document contexts***.

Factors are rows of V.

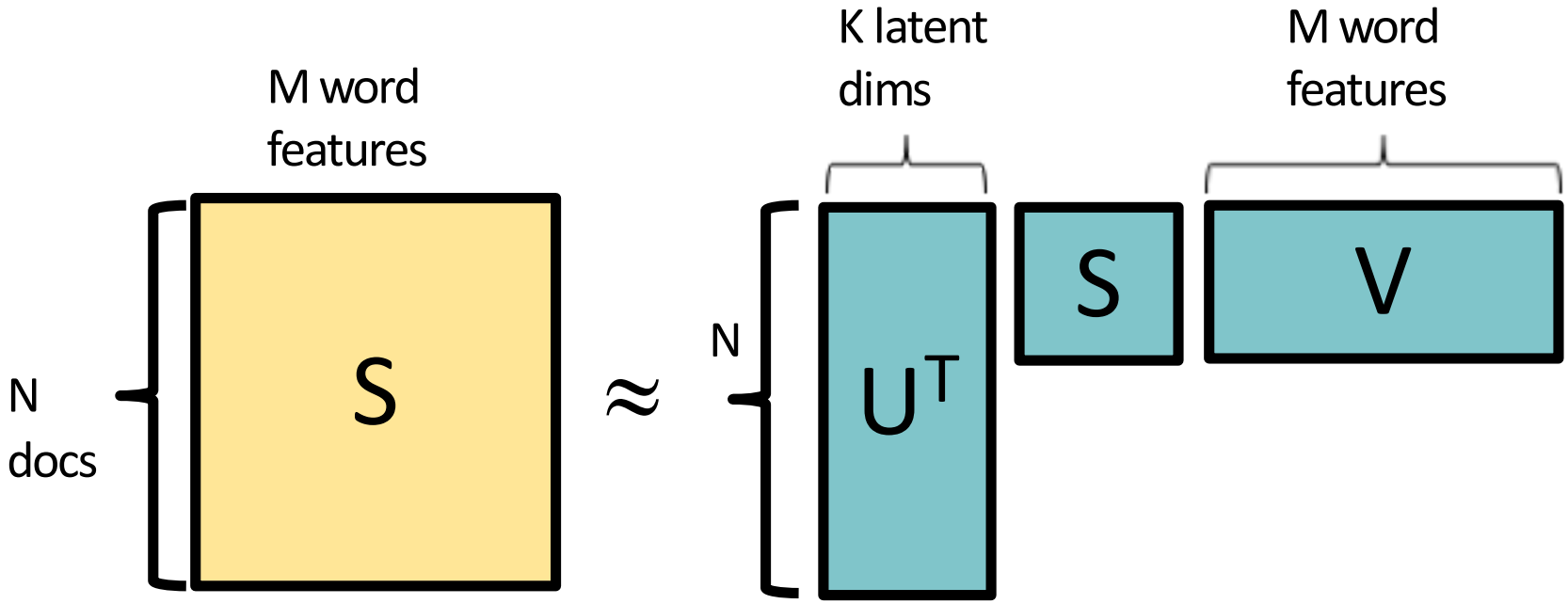


Embedding: Latent Semantic Analysis

If in addition U and V are orthogonal, S a diagonal matrix of singular values. Then if t is a document (row of T):

$v = Vt$ is an embedding of the document in the latent space.

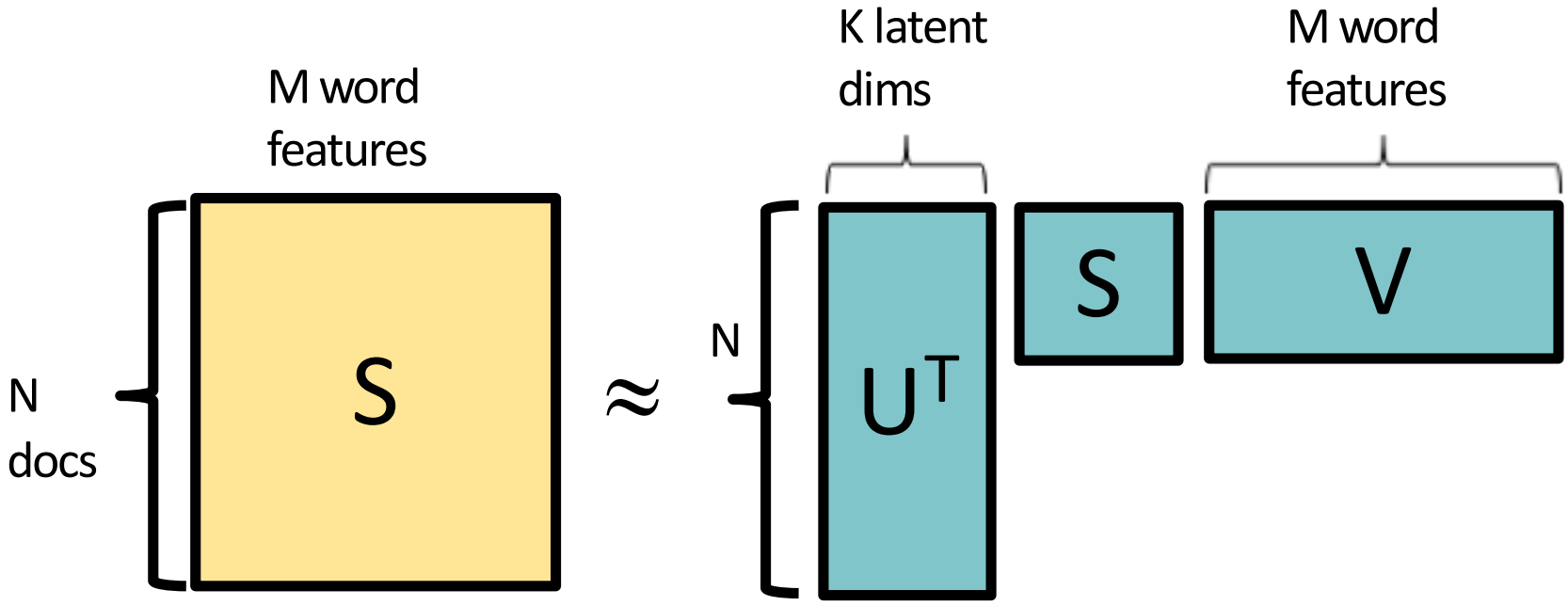
$t' = U^T v = U^T Vt$ is the decoding of the sentence from its embedding.



Embedding: Latent Semantic Analysis

$t' = U^T v = U^T V t$ is the decoding of the sentence from its embedding.

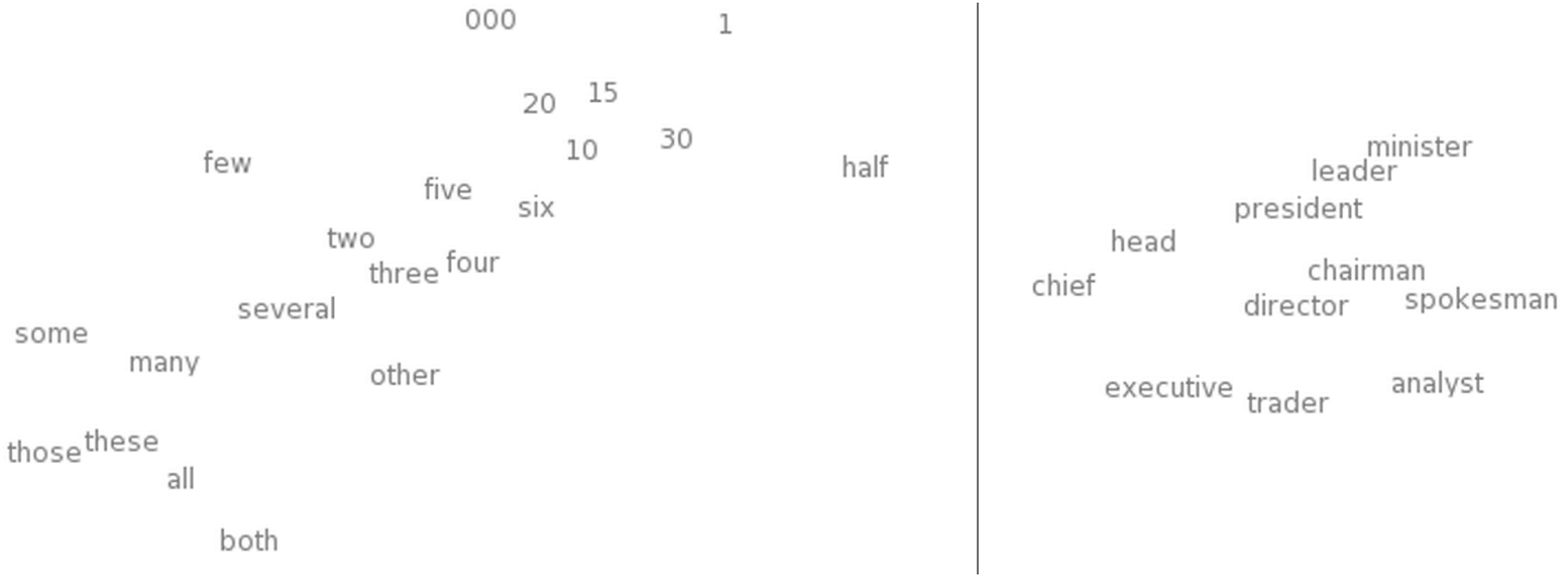
An SVD factorization gives the ***best possible reconstructions*** of the documents t' from their embeddings.



t-SNE of Word Embeddings

From: “Word representations: A simple and general method for semi-supervised learning” Joseph Turian, Lev Ratinov, Yoshua Bengio, ACL 2010.

t-SNE of Word Embeddings



Left: Number Region; Right: Jobs Region

from “Deep Learning, NLP, and Representations” by Chris Olah. See also

http://colah.github.io/posts/2015-01-Visualizing-Representations/

Word2vec: Local contexts

Instead of entire documents, Word2vec uses words a few positions away from each center word. The pairs of center word/context word are called **"skip-grams."**

"It was **a bright cold day in April, and** the clocks were striking"

Center word: red

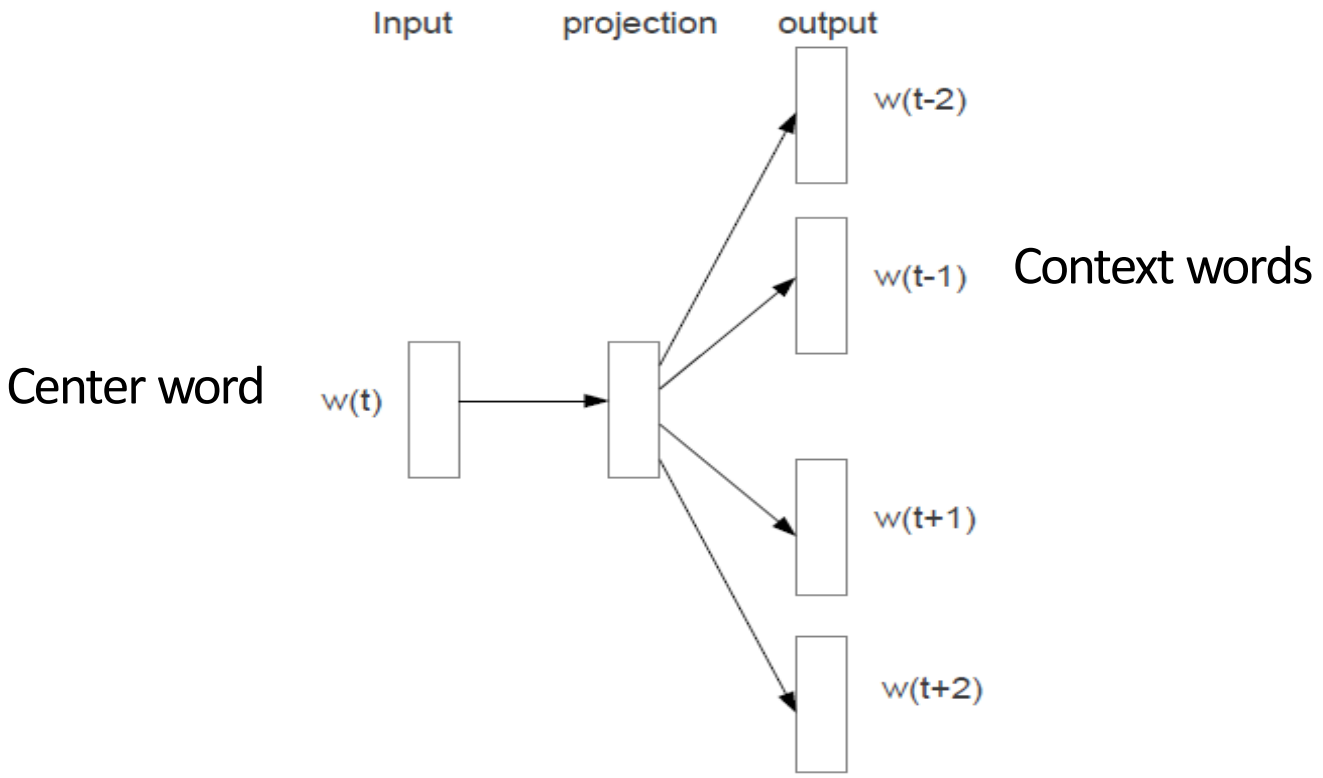
Context words: blue

Word2vec considers all words as center words, and all their context words.

Word2vec: Local contexts

TECHNISCHE UNIVERSITÄT DARMSTADT

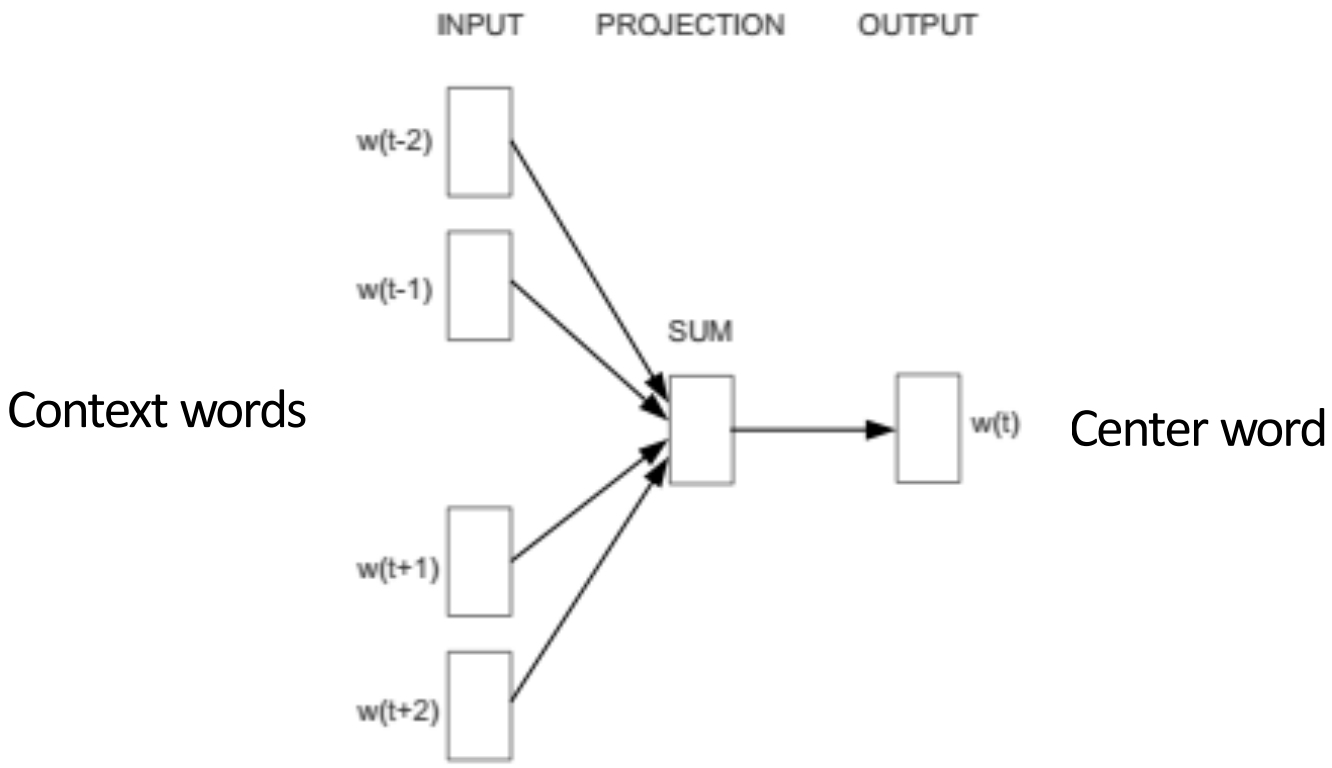
The pairs of center word/context word are called **"skip-grams."**
Typical distances are 3-5 word positions. Skip-gram model:



Distributed Representations of Words and Phrases and their Compositionality
Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, NIPS 2013

Word2vec: Local contexts

Models can also predict center word from context, CBOW model.
Generally, skip-gram performs better.



Distributed Representations of Words and Phrases and their Compositionality
Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, NIPS 2013

Word2vec: Local contexts

Word2vec optimizes a softmax loss for each output word:

$$p(j|i) = \frac{\exp\left(u_j^T v_i\right)}{\sum_{k=1}^{V} \exp(u_k^T v_i)}$$

Where j is the output word, i is the input word. j ranges over a context of $\pm$ 3-5 positions around the input word.

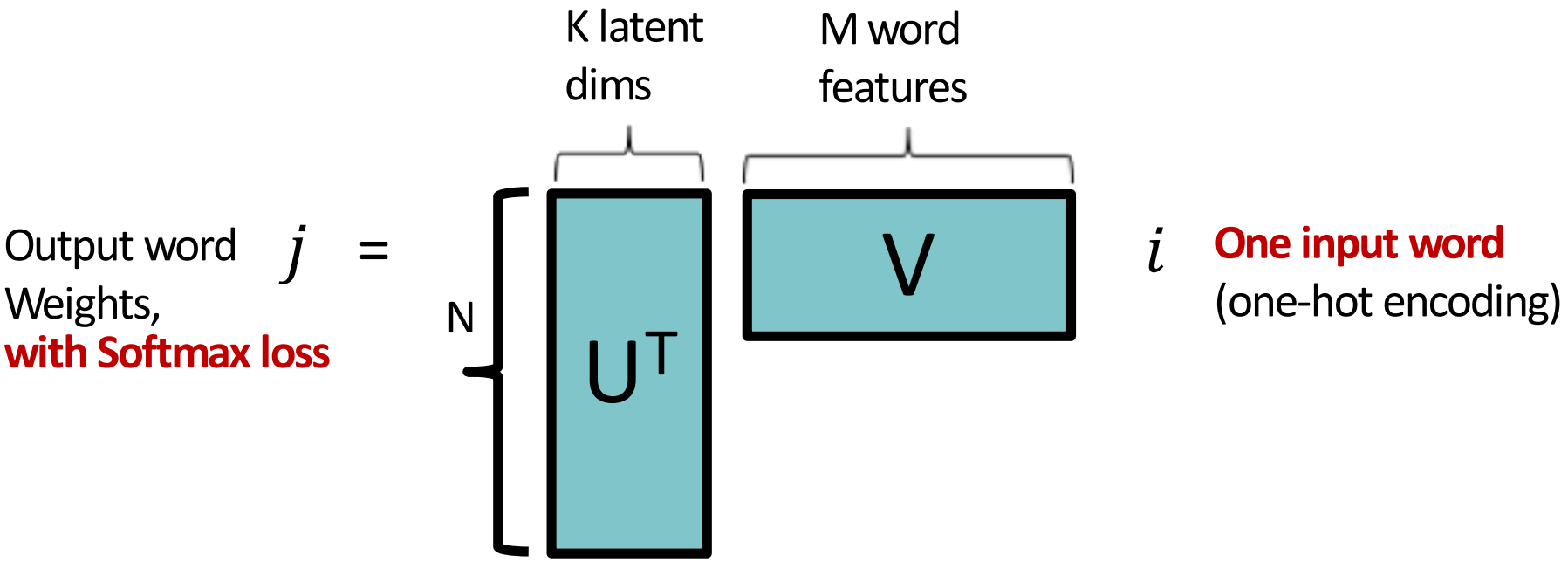
u is an output embedding vector.

v is an input embedding vector.

Word2vec can be implemented with standard DNN toolkits, by backpropagating to optimize u and v.

Matrix perspective

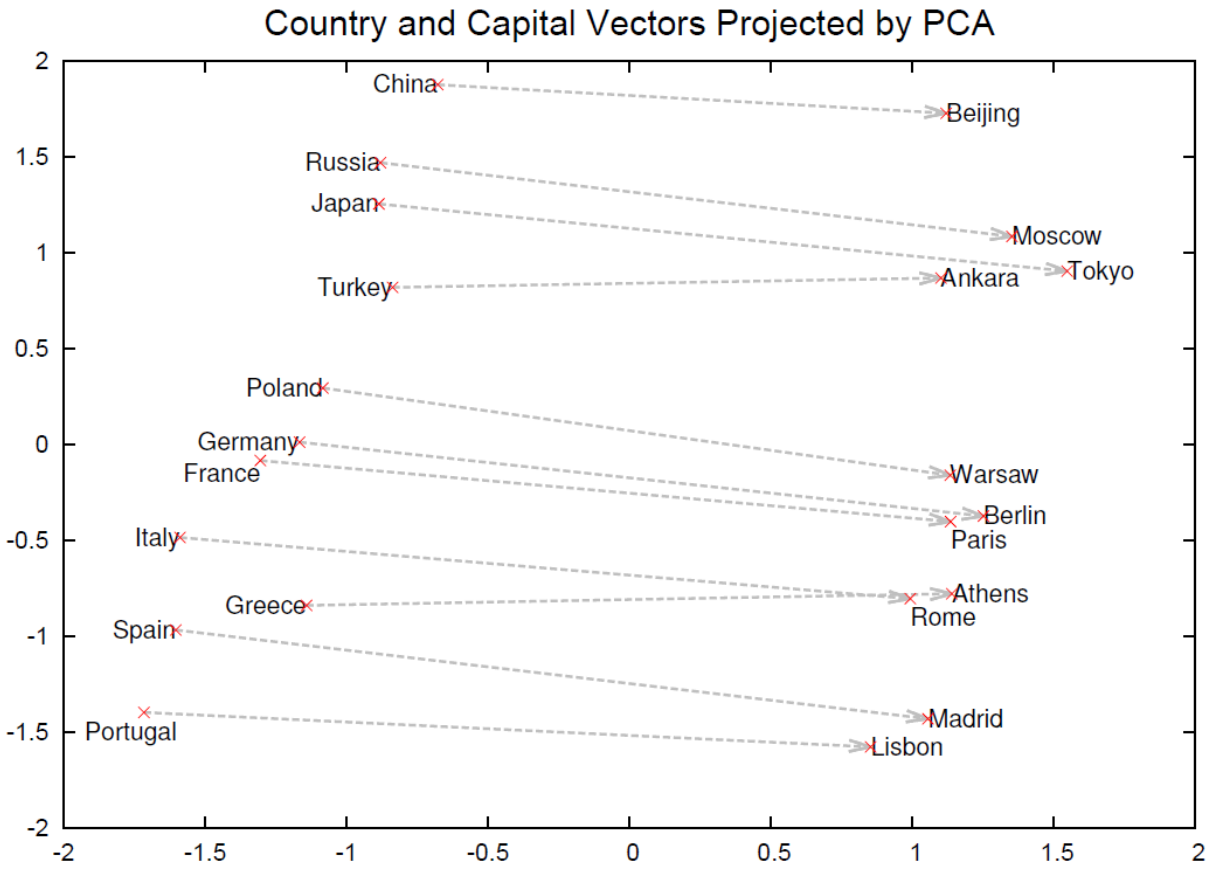
Using matrix representation:



See: “GloVe: Global Vectors for Word Representation” Jeffrey Pennington, Richard Socher, Christopher D. Manning, 2014

Word2vec: Local contexts

Local contexts capture much more information about relations and properties than LSA:

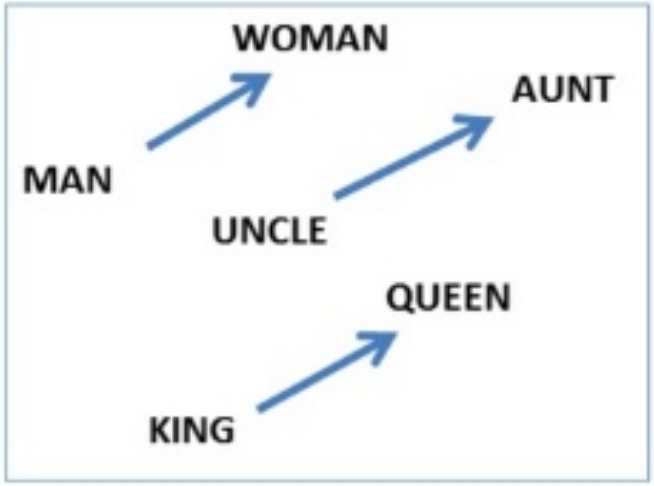


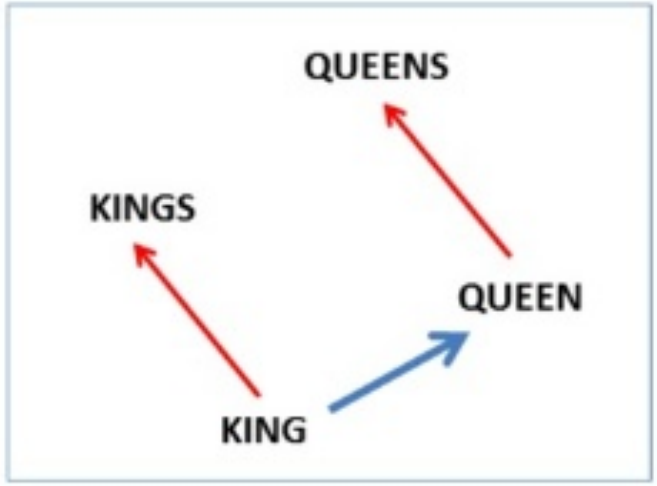
Composition

Algebraic relations:

vec(“woman")−vec(“man") ≃ vec(“aunt")−vec(“uncle")

vec(“woman")−vec(“man") ≃ vec(“queen")−vec(“king")





From “Linguistic Regularities in Continuous Space Word Representations”
Tomas Mikolov , Wen-tau Yih, Geoffrey Zweig, NAACL-HLT 2013

Relations Learned by Word2vec

Word2vec model computed from 6 billion word corpus of news articles

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

"Efficient Estimation of Word Representations in Vector Space" Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Arxiv 2013

Relations Learned by Word2vec

TECHNISCHE UNIVERSITÄT DARMSTADT

A relation is defined by the vector displacement in the first column. For each start word in the other column, the closest displaced word is shown.

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

"Efficient Estimation of Word Representations in Vector Space" Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Arxiv 2013

Relations Learned by Word2vec

TECHNISCHE UNIVERSITÄT DARMSTADT

A relation is defined by the vector displacement in the first column. For each start word in the other column, the closest displaced word is shown.

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

"Efficient Estimation of Word Representations in Vector Space" Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Arxiv 2013

Relations Learned by Word2vec

TECHNISCHE UNIVERSITÄT DARMSTADT

A relation is defined by the vector displacement in the first column. For each start word in the other column, the closest displaced word is shown.

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

"Efficient Estimation of Word Representations in Vector Space" Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Arxiv 2013

Relations Learned by Word2vec

TECHNISCHE UNIVERSITÄT DARMSTADT

A relation is defined by the vector displacement in the first column. For each start word in the other column, the closest displaced word is shown.

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

"Efficient Estimation of Word Representations in Vector Space" Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Arxiv 2013

Relations Learned by Word2vec

TECHNISCHE UNIVERSITÄT DARMSTADT

A relation is defined by the vector displacement in the first column. For each start word in the other column, the closest displaced word is shown.

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

"Efficient Estimation of Word Representations in Vector Space" Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Arxiv 2013

Relations Learned by Word2vec

TECHNISCHE UNIVERSITÄT DARMSTADT

A relation is defined by the vector displacement in the first column. For each start word in the other column, the closest displaced word is shown.

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

"Efficient Estimation of Word Representations in Vector Space" Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Arxiv 2013

Relations Learned by Word2vec

TECHNISCHE UNIVERSITÄT DARMSTADT

A relation is defined by the vector displacement in the first column. For each start word in the other column, the closest displaced word is shown.

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany	bratwurst	France	tapas	USA	pizza

"Efficient Estimation of Word Representations in Vector Space" Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Arxiv 2013

Lexical and Compositional Semantics

TECHNISCHE UNIVERSITÄT DARMSTADT

Lexical Semantics: focuses on the meaning of individual words.

Compositional Semantics: meaning depends on the words, and on how they are combined.

Beyond Bag-Of-Words: Skip-Thought Vectors

The models we discussed so far embed texts as **the sum of their words** (lexical semantics).

Clearly there is a lot missing from these representations:

“man bites dog” = “dog bites man”

“the quick, brown fox jumps over the lazy dog” =

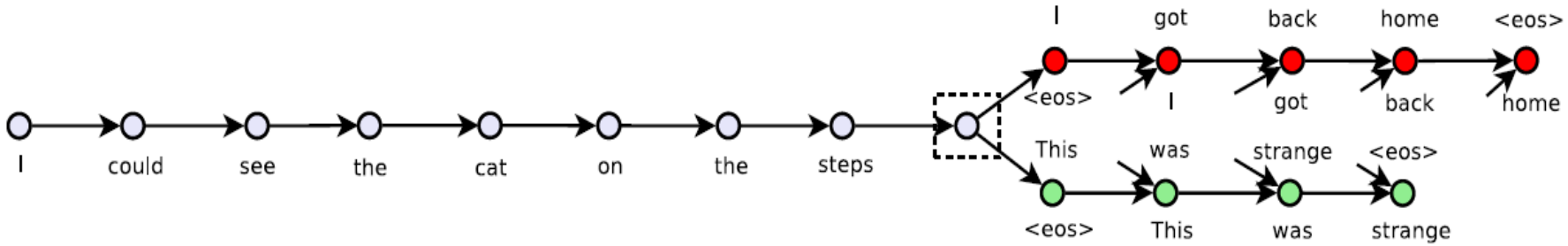
“the lazy fox over the brown dog jumps quick”

…

How can we model text structure as well as word meanings?

Beyond Bag-Of-Words: Skip-Thought Vectors

Skip-thought embeddings use sequence-to-sequence RNNs to predict the next and previous ***sentences***.



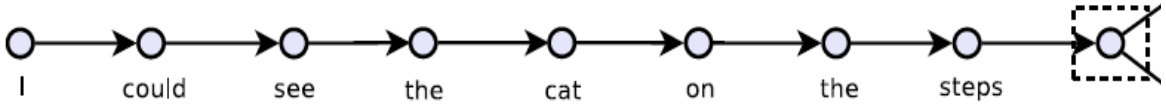
The output state vector of the boundary layer (dotted box) forms the embedding. RNN units are GRU units.

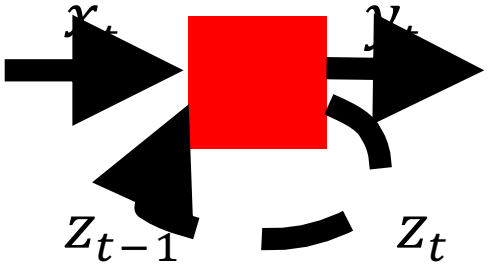
Once the network is trained, we can discard the red and green sections of the network.

From “Skip-Thought Vectors,” Ryan Kiros et al., Arxiv 2015.

Beyond Bag-Of-Words: Skip-Thought Vectors

Skip-thought embeddings use sequence-to-sequence RNNs to predict the next and previous ***sentences***.



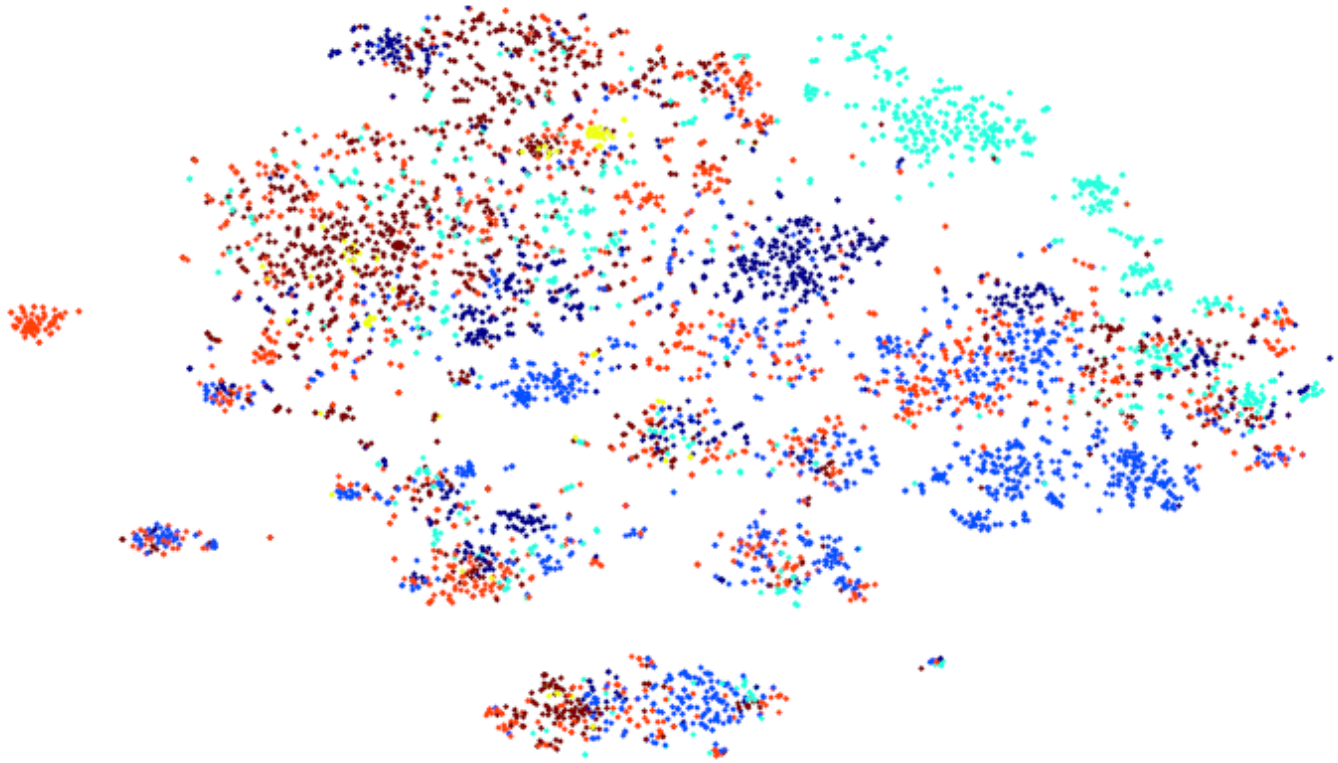


Encoding doesn't require backpropagation, so we can represent the encoder as a (truly) recurrent network.

Thus we can encode longer units of text: sentences or paragraphs.

Embedding of TREC queries

Points are colored by query type (t-SNE embedding):



From “Skip-Thought Vectors,” Ryan Kiros et al., Arxiv 2015.

Sentence Similarity

TECHNISCHE UNIVERSITÄT DARMSTADT

Query and nearest sentence
he ran his hand inside his coat , double-checking that the unopened letter was still there . he slipped his hand between his coat and his shirt , where the folded copies lay in a brown envelope .
im sure youll have a glamorous evening , she said , giving an exaggerated wink . im really glad you came to the party tonight , he said , turning to her .
although she could tell he had n't been too invested in any of their other chitchat , he seemed genuinely curious about this . although he had n't been following her career with a microscope , he 'd definitely taken notice of her appearances .
an annoying buzz started to ring in my ears , becoming louder and louder as my vision began to swim . a weighty pressure landed on my lungs and my vision blurred at the edges , threatening my consciousness altogether .
if he had a weapon , he could maybe take out their last imp , and then beat up errol and vanessa . if he could ram them from behind , send them sailing over the far side of the levee , he had a chance of stopping them .
then , with a stroke of luck , they saw the pair head together towards the portaloos . then , from out back of the house , they heard a horse scream probably in answer to a pair of sharp spurs digging deep into its flanks .
" i 'll take care of it , " goodman said , taking the phonebook . " i 'll do that , " julia said , coming in .
he finished rolling up scrolls and , placing them to one side , began the more urgent task of finding ale and tankards . he righted the table , set the candle on a piece of broken plate , and reached for his flint , steel , and tinder .

Approximately two weeks of training on a billion-word Books corpus

Semantic Relatedness Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

SICK semantic relatedness task: score sentences for semantic similarity from 1 to 5 (average of 10 human ratings)

Sentence A: A man is jumping into an empty pool
Sentence B: There is no biker jumping in the air
Relatedness score: 1.6

Sentence A: Two children are lying in the snow and are making snow angels
Sentence B: Two angels are making snow on the lying children
Relatedness score: 2.9

Sentence A: The young boys are playing outdoors and the man is smiling nearby
Sentence B: There is no boy playing outdoors and there is no man smiling
Relatedness score: 3.6

Sentence A: A person in a black jacket is doing tricks on a motorbike
Sentence B: A man in a black jacket is doing tricks on a motorbike
Relatedness score: 4.9

Semantic Relatedness Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

Note: a separate model is trained to predict the scores from pairs of embedded sentences.

Sentence A: A man is jumping into an empty pool
Sentence B: There is no biker jumping in the air
Relatedness score: 1.6

Sentence A: Two children are lying in the snow and are making snow angels
Sentence B: Two angels are making snow on the lying children
Relatedness score: 2.9

Sentence A: The young boys are playing outdoors and the man is smiling nearby
Sentence B: There is no boy playing outdoors and there is no man smiling
Relatedness score: 3.6

Sentence A: A person in a black jacket is doing tricks on a motorbike
Sentence B: A man in a black jacket is doing tricks on a motorbike
Relatedness score: 4.9

Semantic Relatedness Evaluation

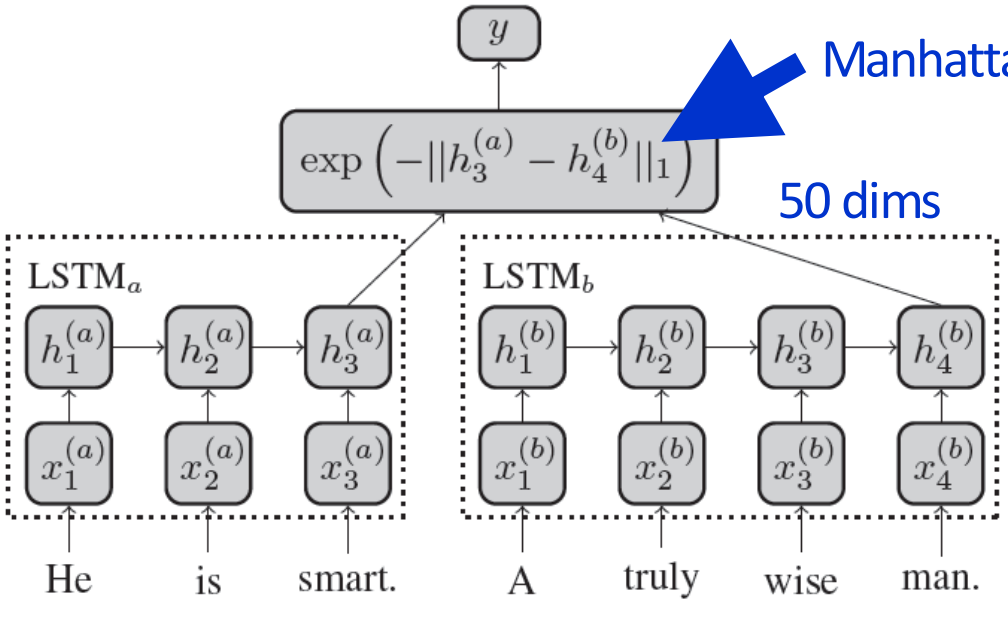
TECHNISCHE UNIVERSITÄT DARMSTADT

SICK semantic relatedness scores for skip-thought methods:

Method	r	ρ	MSE
Illinois-LH [18]	0.7993	0.7538	0.3692
UNAL-NLP [19]	0.8070	0.7489	0.3550
Meaning Factory [20]	0.8268	0.7721	0.3224
ECNU [21]	0.8414	–	–
Mean vectors [22]	0.7577	0.6738	0.4557
DT-RNN [23]	0.7923	0.7319	0.3822
SDT-RNN [23]	0.7900	0.7304	0.3848
LSTM [22]	0.8528	0.7911	0.2831
Bidirectional LSTM [22]	0.8567	0.7966	0.2736
Dependency Tree-LSTM [22]	**0.8676**	**0.8083**	**0.2532**
uni-skip	0.8477	0.7780	0.2872
bi-skip	0.8405	0.7696	0.2995
combine-skip	0.8584	0.7916	0.2687
combine-skip+COCO	0.8655	0.7995	0.2561

A Siamese Network for Semantic Relatedness

This network is trained on pairs of sentences a, b with a similarity label y.



Parameters are shared between the two networks.

From “Siamese Recurrent Architectures for Learning Sentence Similarity”
Jonas Mueller, Aditya Thyagarajan, AAAI-2016

A Siamese Network for Semantic Relatedness

The network is trained on Semeval similar sentence pairs, expanded by substituting for random words using WordNet (a dataset of synonyms). Results:

Method	r	ρ	MSE
Illinois-LH (Lai and Hockenmaier 2014)	0.7993	0.7538	0.3692
UNAL-NLP (Jimenez et al. 2014)	0.8070	0.7489	0.3550
Meaning Factory (Bjerva et al. 2014)	0.8268	0.7721	0.3224
ECNU (Zhao, Zhu, and Lan 2014)	0.8414	–	–
Skip-thought+COCO (Kiros et al. 2015)	0.8655	0.7995	0.2561
Dependency Tree-LSTM (Tai, Socher, and Manning 2015)	0.8676	0.8083	0.2532
ConvNet (He, Gimpel, and Lin 2015)	0.8686	0.8047	0.2606
MaLSTM	**0.8822**	**0.8345**	**0.2286**

From “Siamese Recurrent Architectures for Learning Sentence Similarity” Jonas Mueller, Aditya Thyagarajan, AAAI-2016

A Siamese Network for Semantic Relatedness

The network is trained on Semeval similar sentence pairs, expanded by substituting for random words using WordNet (a dataset of synonyms). Results:

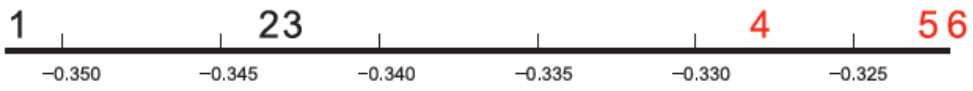
Method	r	ρ	MSE
Illinois-LH (Lai and Hockenmaier 2014)	0.7993	0.7538	0.3692
UNAL-NLP (Jimenez et al. 2014)	0.8070	0.7489	0.3550
Meaning Factory (Bjerva et al. 2014)	0.8268	0.7721	0.3224
ECNU (Zhao, Zhu, and Lan 2014)	0.8414	–	–
Skip-thought+COCO (Kiros et al. 2015)	0.8655	0.7995	0.2561
Dependency Tree-LSTM (Tai, Socher, and Manning 2015)	0.8676	0.8083	0.2532
ConvNet (He, Gimpel, and Lin 2015)	0.8686	0.8047	0.2606
MaLSTM	**0.8822**	**0.8345**	**0.2286**

r = Pearson correlation, ρ = Spearman's rank correlation.

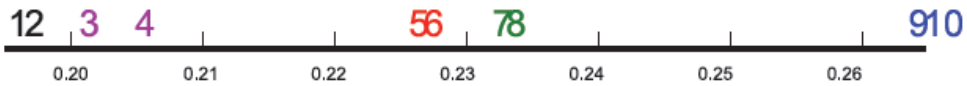
A Siamese Network for Semantic Relatedness

Moral: Train for your evaluation metric!

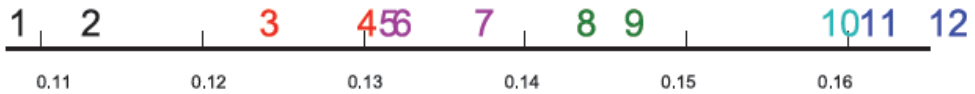
Hidden Unit Factors 1,2, and 6



1 There is no man pointing at a car
2 The woman is not playing the flute
3 The man is not riding a horse
4 A man is pointing at a silver sedan
5 The woman is playing the flute
6 A man is riding a horse



1 Two kids are bouncing on colorful balls
2 Two children are bouncing on colorful balls
3 The golden dog is running through a field of tall grass
4 A brown dog is running through tall green grass
5 A woman is putting on makeup carefully
6 A woman is carefully removing her makeup
7 A woman is applying cosmetics to her eyelid
8 A woman is carefully applying cosmetics to her eyelid
9 There is no woman cutting potatoes
10 A woman is slicing carrots



1 The cat is running across the gravel
2 A cat is playing a keyboard
3 The brown animal is jumping in the air
4 The animal with big eyes is eating
5 A dog is bouncing on a trampoline
6 A dog is running on the ground
7 A dog is running on the road
8 Several boys are jumping on a trampoline
9 A little boy is running on the ground and playing with a little girl
10 Someone is playing a piano
11 A man is running on the road
12 A man is playing an electronic keyboard

Semantic Entailment Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

SICK semantic entailment task: score sentences for relations: ENTAILMENT, CONTRADICTION, NEUTRAL:

Sentence A: Two teams are competing in a football match
Sentence B: Two groups of people are playing football
Entailment judgment: ENTAILMENT

Sentence A: The brown horse is near a red barrel at the rodeo
Sentence B: The brown horse is far from a red barrel at the rodeo
Entailment judgment: CONTRADICTION

Sentence A: A man in a black jacket is doing tricks on a motorbike
Sentence B: A person is riding the bicycle on one wheel
Entailment judgment: NEUTRAL

Semantic Entailment for MaLSTM

TECHNISCHE UNIVERSITÄT DARMSTADT

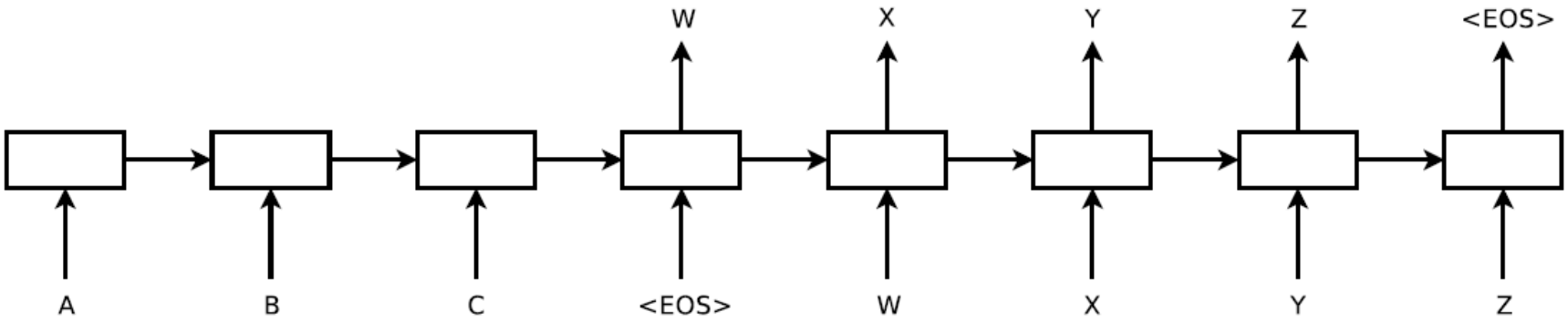
Method	Accuracy
Illinois-LH (Lai and Hockenmaier 2014)	**84.6**
ECNU (Zhao, Zhu, and Lan 2014)	83.6
UNAL-NLP (Jimenez et al. 2014)	83.1
Meaning Factory (Bjerva et al. 2014)	81.6
Reasoning-based n-best (Lien and Kouylekov 2015)	80.4
LangPro Hybrid-800 (Abzianidze 2015)	81.4
SNLI-transfer 3-class LSTM (Bowman et al. 2015)	80.8
MaLSTM features + SVM	84.2

Translation Models

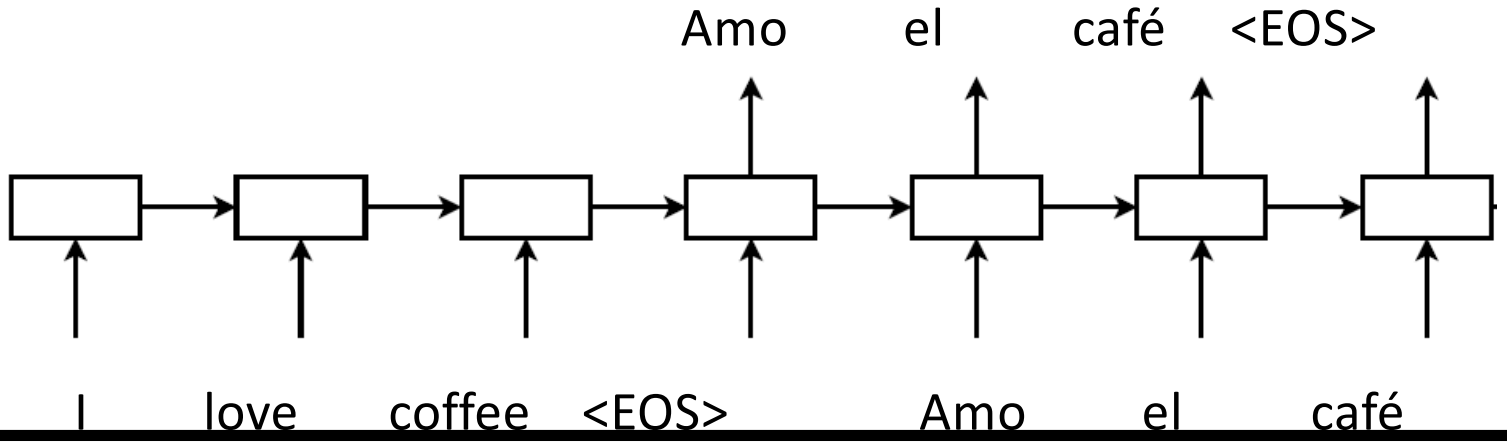
Sequence-To-Sequence RNNs

TECHNISCHE UNIVERSITÄT DARMSTADT

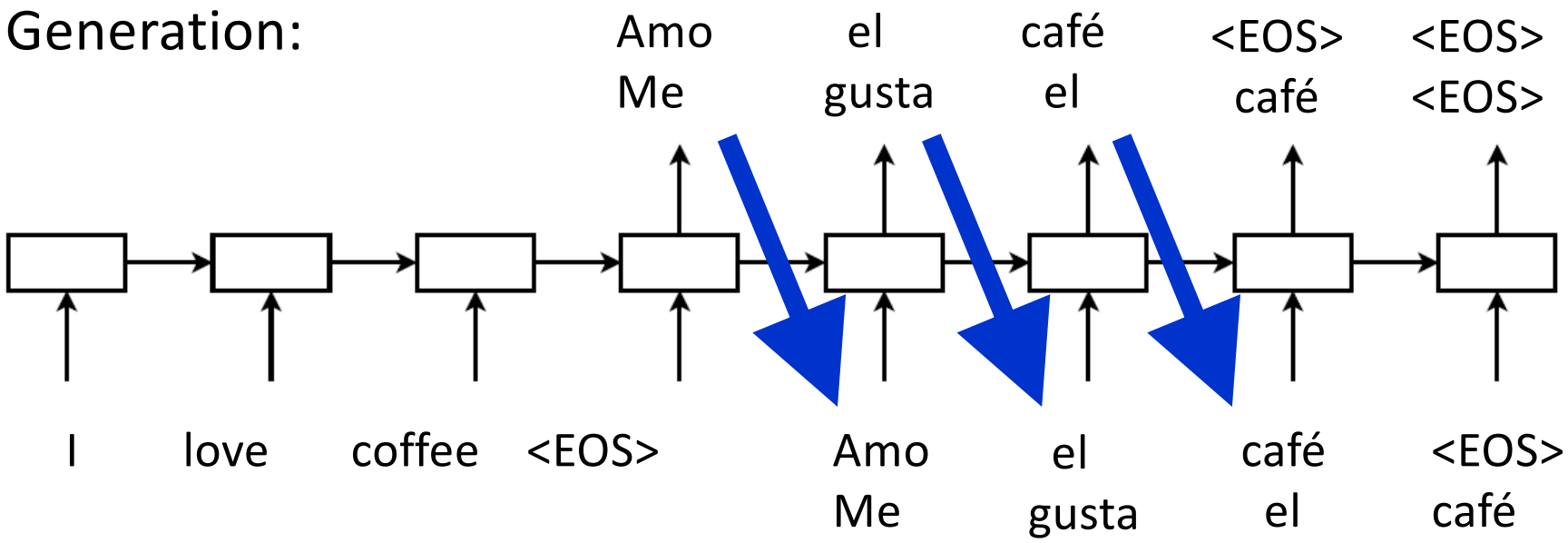
An input sequence is fed to the left array, output sentence to the right array for training:



For translation:



Sequence-To-Sequence RNNs



Keep an n-best list of partial sentences, along with their partial softmax scores.

Bleu Scores for Translation

The goal of bleu scores is to compare machine translations against human-generated translations, allowing for variation.

Consider these translations for a Chinese sentence:

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Bleu Scores for Translation

TECHNISCHE UNIVERSITÄT DARMSTADT

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

Bleu Scores for Translation

TECHNISCHE UNIVERSITÄT DARMSTADT

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

Bleu Scores for Translation

TECHNISCHE UNIVERSITÄT DARMSTADT

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

Bleu Scores for Translation

Unigram precision:

$$\frac{\text{correct unigrams occuring in reference sentence}}{\text{unigrams occuring in test sentence}}$$

Modified unigram precision: clip counts by maximum occurrence in any reference sentence:

Candidate: the the the the the the the.
Reference 1: The cat is on the mat.
Reference 2: There is a cat on the mat.

Modified precision is 2/7.

Bleu Scores for Translation

TECHNISCHE UNIVERSITÄT DARMSTADT

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party. **unigram precision 17/18**

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct. **unigram precision 8/14**

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

Bleu Scores for Translation

TECHNISCHE UNIVERSITÄT DARMSTADT

N-gram precision is defined similarly:

$$\frac{\text{correct ngrams occuring in reference sentence}}{\text{ngrams occuring in test sentence}}$$

Modified ngram precision: clip counts by maximum occurrence in any reference sentence.

Unigram scores tend to capture ***adequacy***

Ngram scores tend to capture ***fluency***

Bleu Scores for Translation

TECHNISCHE UNIVERSITÄT DARMSTADT

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party. **bigram precision 10/17**

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct. **bigram precision 1/13**

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

Bleu Scores for Translation

TECHNISCHE UNIVERSITÄT DARMSTADT

How to combine scores for different n-grams?
Averaging sounds good, but precisions are very different for different n (unigrams have much higher scores).

BLEU Score: Take a weighted geometric mean of the logs of n-gram precisions up to some length (usually 4). Add a penalty for too-short predictions.

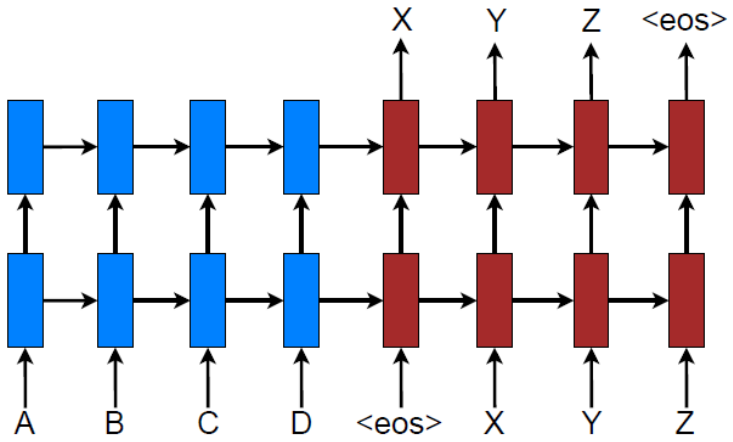
$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

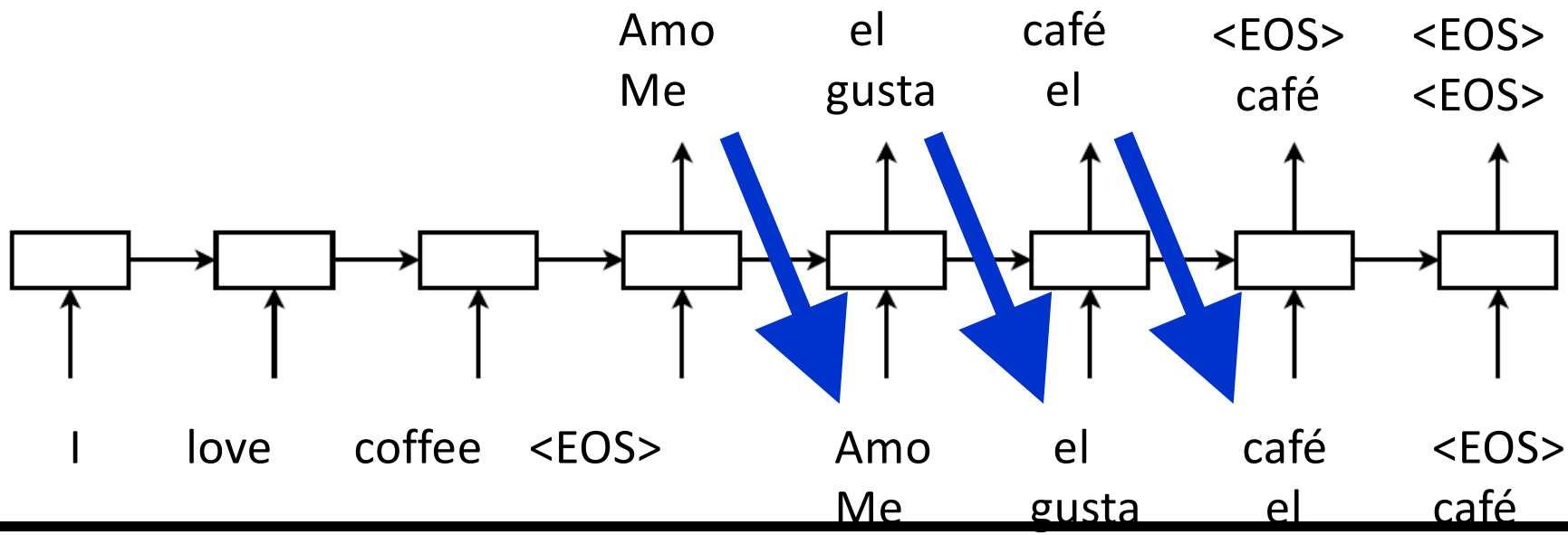
$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

Candidate length c shorter than reference r translation

Sequence-To-Sequence Model Translation

Raw scores for French-English Translation, depth = 4 for these results.





Sequence-To-Sequence Model Translation

TECHNISCHE UNIVERSITÄT DARMSTADT

Raw scores for French-English Translation

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

Sequence-To-Sequence Model Translation

Scores using the LSTM model to rerank 1000-best sentences from a baseline Machine Translation system:

Method	test BLEU score (ntst14)
Baseline System [29]	33.30
Cho et al. [5]	34.54
Best WMT'14 result [9]	**37.0**
Rescoring the baseline 1000-best with a single forward LSTM	35.61
Rescoring the baseline 1000-best with a single reversed LSTM	35.85
Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs	**36.5**
Oracle Rescoring of the Baseline 1000-best lists	~45

Sequence-To-Sequence Model Translation

TECHNISCHE UNIVERSITÄT DARMSTADT

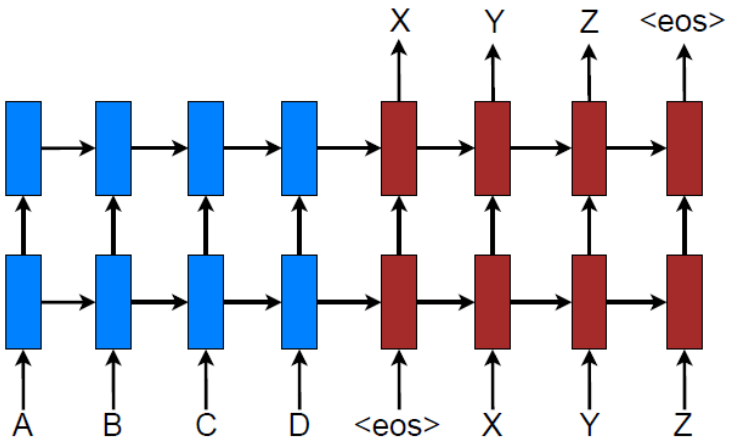
Training details:

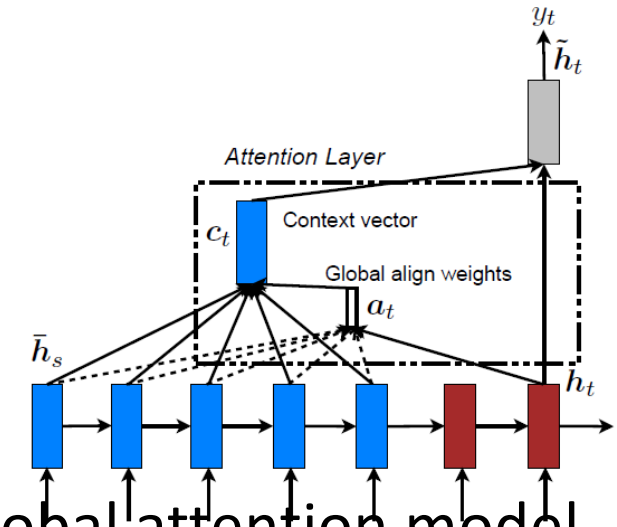
- **LSTM Array depth = 4**. Deeper is better.
- LSTM params initialized from uniform distribution [-0.8,0.8]
- Stochastic gradient descent w/o momentum, fixed learning rate of 0.7.
- After 5 epochs, learning rate was halved every half epoch.
- Models trained for a total of 7.5 epochs.
- Batch size of 128 sequences.
- Gradient clipping at $\|g\| = 5$.
- Sentences were grouped into minibatches of approximately the same size.

State-of-the-Art Neural Machine Translation

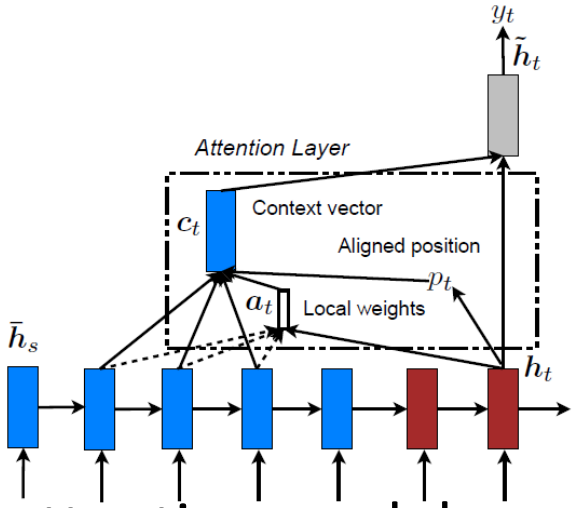
TECHNISCHE UNIVERSITÄT DARMSTADT

A RNN array with an attention network to regulate information flow from the source network.





Global attention model



Local attention model

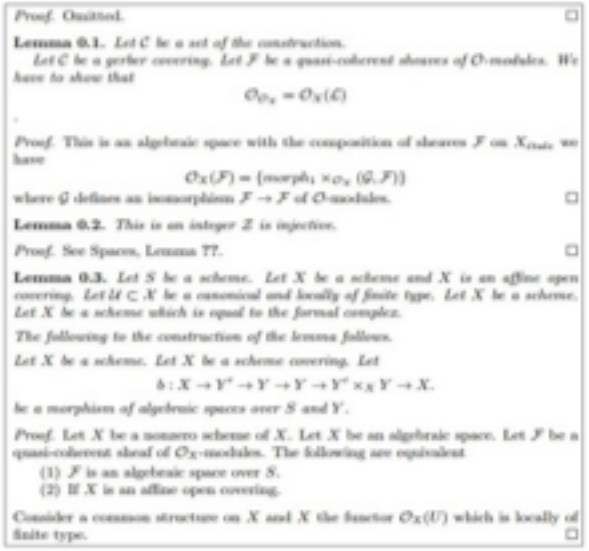
State-of-the-Art Neural Machine Translation

English-German translation (WMT 14 results):

System	Ppl	BLEU
Winning WMT'14 system – *phrase-based* + *large LM* (Buck et al., 2014)		20.7
Existing NMT systems		
RNNsearch (Jean et al., 2015)		16.5
RNNsearch + unk replace (Jean et al., 2015)		19.0
RNNsearch + unk replace + large vocab + *ensemble* 8 models (Jean et al., 2015)		**21.6**
Our NMT systems		
Base	10.6	11.3
Base + reverse	9.9	12.6 (+*1.3*)
Base + reverse + dropout	8.1	14.0 (+*1.4*)
Base + reverse + dropout + global attention (*location*)	7.3	16.8 (+*2.8*)
Base + reverse + dropout + global attention (*location*) + feed input	6.4	18.1 (+*1.3*)
Base + reverse + dropout + local-p attention (*general*) + feed input	5.9	19.0 (+*0.9*)
Base + reverse + dropout + local-p attention (*general*) + feed input + unk replace		20.9 (+*1.9*)
Ensemble 8 models + unk replace		**23.0 (+*2.1*)**

Parsing

Recall RNNs ability to generate Latex, C code:



Proof. Omitted. □

Lemma 0.1. *Let $\mathcal{C}$ be a set of the construction.*

Let $\mathcal{C}$ be a gerbe covering. Let $\mathcal{F}$ be a quasi-coherent sheaves of $\mathcal{O}$-modules. We have to show that

$$\mathcal{O}_{\mathcal{O}_X} = \mathcal{O}_X(\mathcal{L})$$

.

Proof. This is an algebraic space with the composition of sheaves $\mathcal{F}$ on $X_{\acute{e}tale}$ we have

$$\mathcal{O}_X(\mathcal{F}) = \{morphisms \times_{\mathcal{O}_X} (\mathcal{G}, \mathcal{F})\}$$

where $\mathcal{G}$ defines an isomorphism $\mathcal{F} \to \mathcal{F}$ of $\mathcal{O}$-modules. □

Lemma 0.2. *This is an integer $\mathcal{Z}$ is injective.*

Proof. See Spaces, Lemma ??. □

Lemma 0.3. *Let S be a scheme. Let X be a scheme and X is an affine open covering. Let $U \subset X$ be a canonical and locally of finite type. Let X be a scheme. Let X be a scheme which is equal to the formal complex.*

The following to the construction of the lemma follows.

Let X be a scheme. Let X be a scheme covering. Let

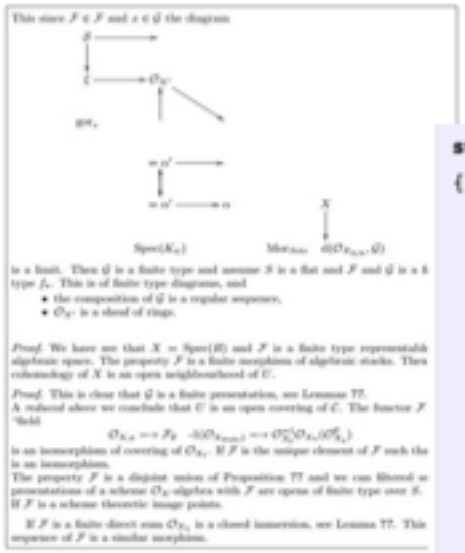
$$b : X \to Y' \to Y \to Y \to Y' \times_Y Y \to X.$$

be a morphism of algebraic spaces over S and Y.

Proof. Let X be a nonzero scheme of X. Let X be an algebraic space. Let $\mathcal{F}$ be a quasi-coherent sheaf of $\mathcal{O}_X$-modules. The following are equivalent

(1) $\mathcal{F}$ is an algebraic space over S.
(2) If X is an affine open covering.

Consider a common structure on X and X the functor $\mathcal{O}_X(U)$ which is locally of finite type. □



```
static void do_command(struct seq_file *m, void *v)
{
  int column = 32 << (cmd[2] & 0x80);
  if (state)
    cmd = (int)(int_state ^ (in_8(&ch->ch_flags) & Cmd) ? 2 : 1);
  else
    seq = 1;
  for (i = 0; i < 16; i++) {
    if (k & (1 << 1))
      pipe = (in_use & UMXTHREAD_UNCCA) +
        ((count & 0x00000000fffffff8) & 0x000000f) << 8;
    if (count == 0)
      sub(pid, ppc_md.kexec_handle, 0x20000000);
    pipe_set_bytes(i, 0);
  }
  /* Free our user pages pointer to place camera if all dash */
  subsystem_info = &of_changes[PAGE_SIZE];
  rek_controls(offset, idx, &soffset);
  /* Now we want to deliberately put it to device */
  control_check_polarity(&context, val, 0);
  for (i = 0; i < COUNTER; i++)
    seq_puts(s, "policy ");
}
```

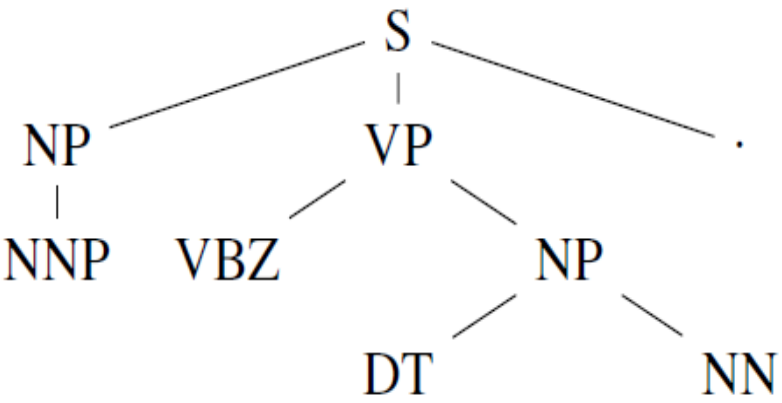
They seem to do well with tree-structured data.

What about natural language parsing?

Parsing

Sequence models generate linear structures, but these can easily encode trees by “closing parens” (prefix tree notation):

John has a dog . →

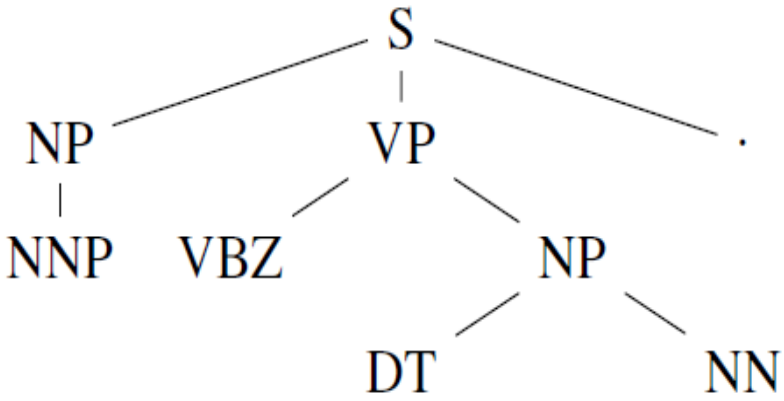


John has a dog . → $(\text{S } (\text{NP NNP })_{\text{NP}} (\text{VP VBZ } (\text{NP DT NN })_{\text{NP}})_{\text{VP}} \text{ . })_{\text{S}}$

Parsing Cheat Sheet

TECHNISCHE UNIVERSITÄT DARMSTADT

John has a dog . →



John has a dog . → $(\text{S } (\text{NP NNP })_{\text{NP}} (\text{VP VBZ } (\text{NP DT NN })_{\text{NP}})_{\text{VP}} \text{ . })_{\text{S}}$

S = Sentence

NP = Noun Phrase

VP = Verb Phrase

NNP = Proper Noun (“John”)

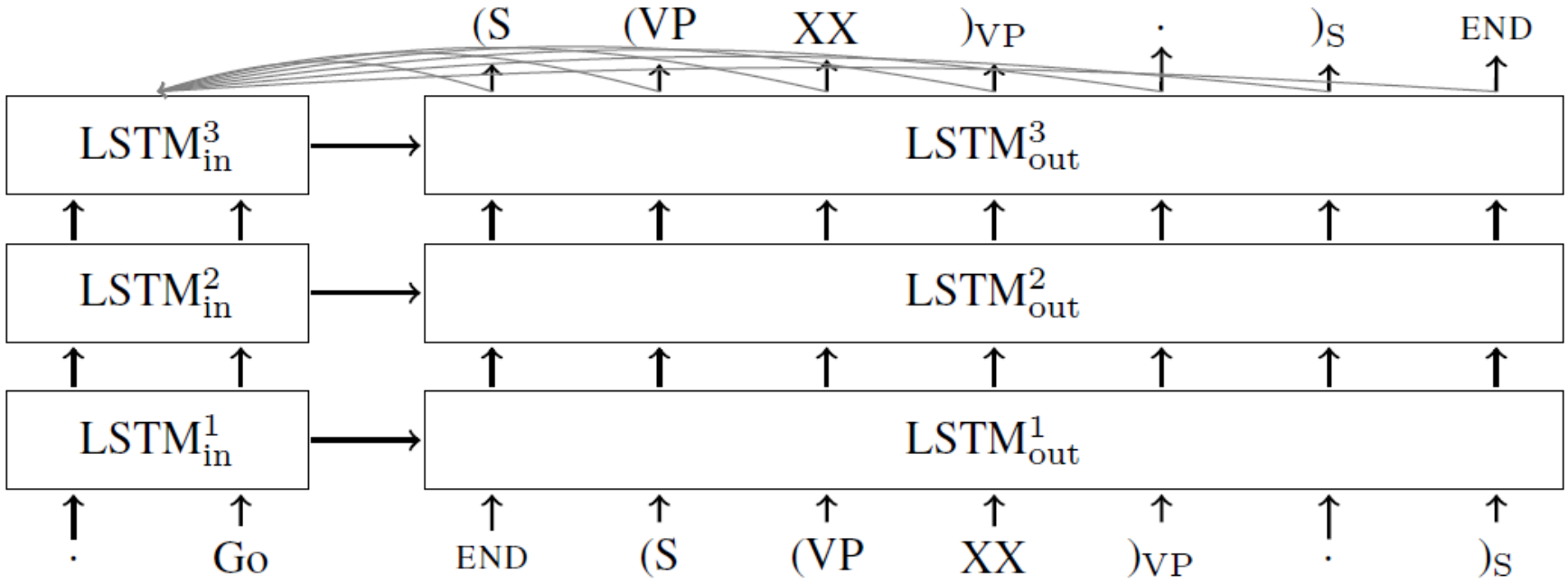
VBZ = Verb, 3rd person, singular (“has”)

DT = Determiner (“a”)

NN = Noun, singular (“dog”)

A Sequence-To-Sequence Parser

The model is a depth-3 sequence-to-sequence predictor, augmented with the attention model of Bahdanau 2014.



Grammar as a Foreign Language Oriol Vinyals, Google, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, Geoffrey Hinton, NIPS 2015

"Neural machine translation by jointly learning to align and translate." Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. arXiv 2014.

A Sequence-To-Sequence Parser

TECHNISCHE UNIVERSITÄT DARMSTADT

Chronology:

- First tried training a basic sequence-to-sequence model on human-annotated training treebanks. **Poor results.**
- Then training on parse trees **generated by the Berkeley Parser**, achieved similar performance (90.5 F1 score) to it.
- Next added the attention model, trained **on human treebank data**, also achieved 90.5 F1.
- Finally, created a synthetic dataset of **high-confidence parse trees** (agreed on by two parsers). Achieved a new state-of-the-art of 92.5 F1 score (WSJ dataset).

F1 is a widely-used accuracy measure that combines precision and recall

A Sequence-To-Sequence Parser

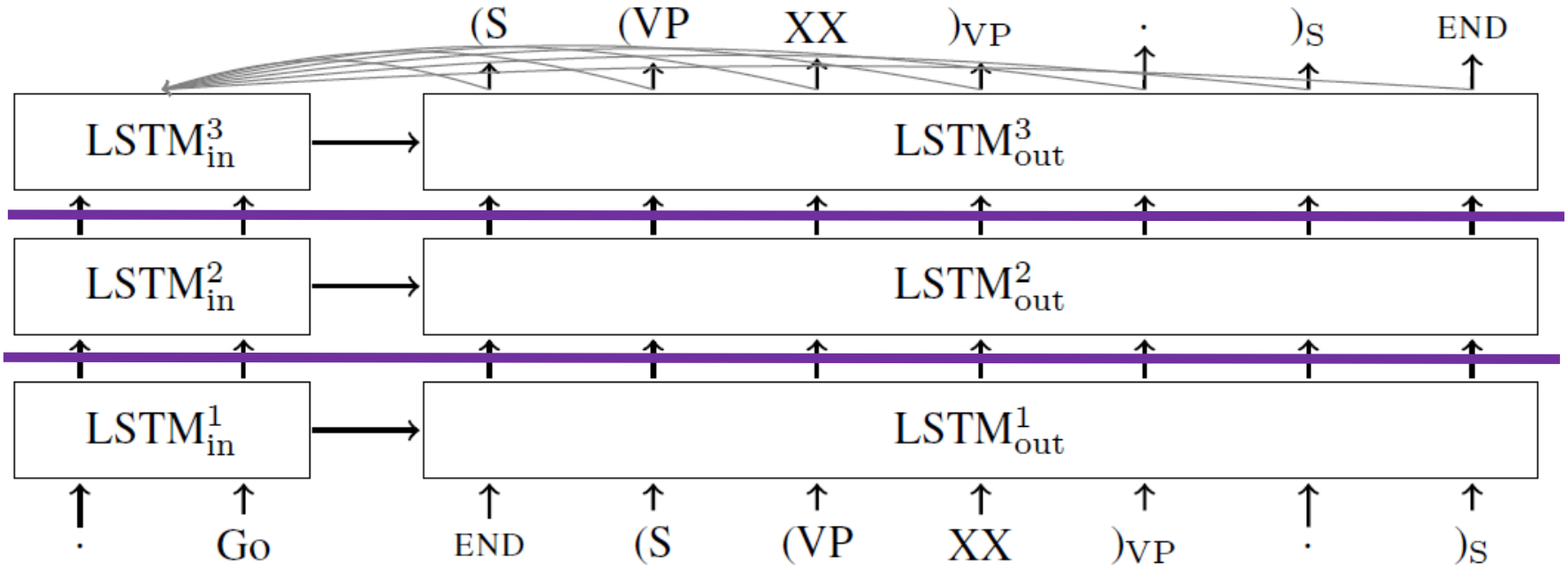
TECHNISCHE UNIVERSITÄT DARMSTADT

Quick Training Details:

- Depth = 3, layer dimension = 256.
- **Dropout** between layers 1 and 2, and 2 and 3.
- **No POS tags!!** Improved by F1 1 point by leaving them out.
- Input reversing.

A Sequence-To-Sequence Parser

Dropout layers shown in purple:



This use of dropout in LSTM arrays is now widely used.

Neural Entity-Relation Extraction

TECHNISCHE UNIVERSITÄT DARMSTADT

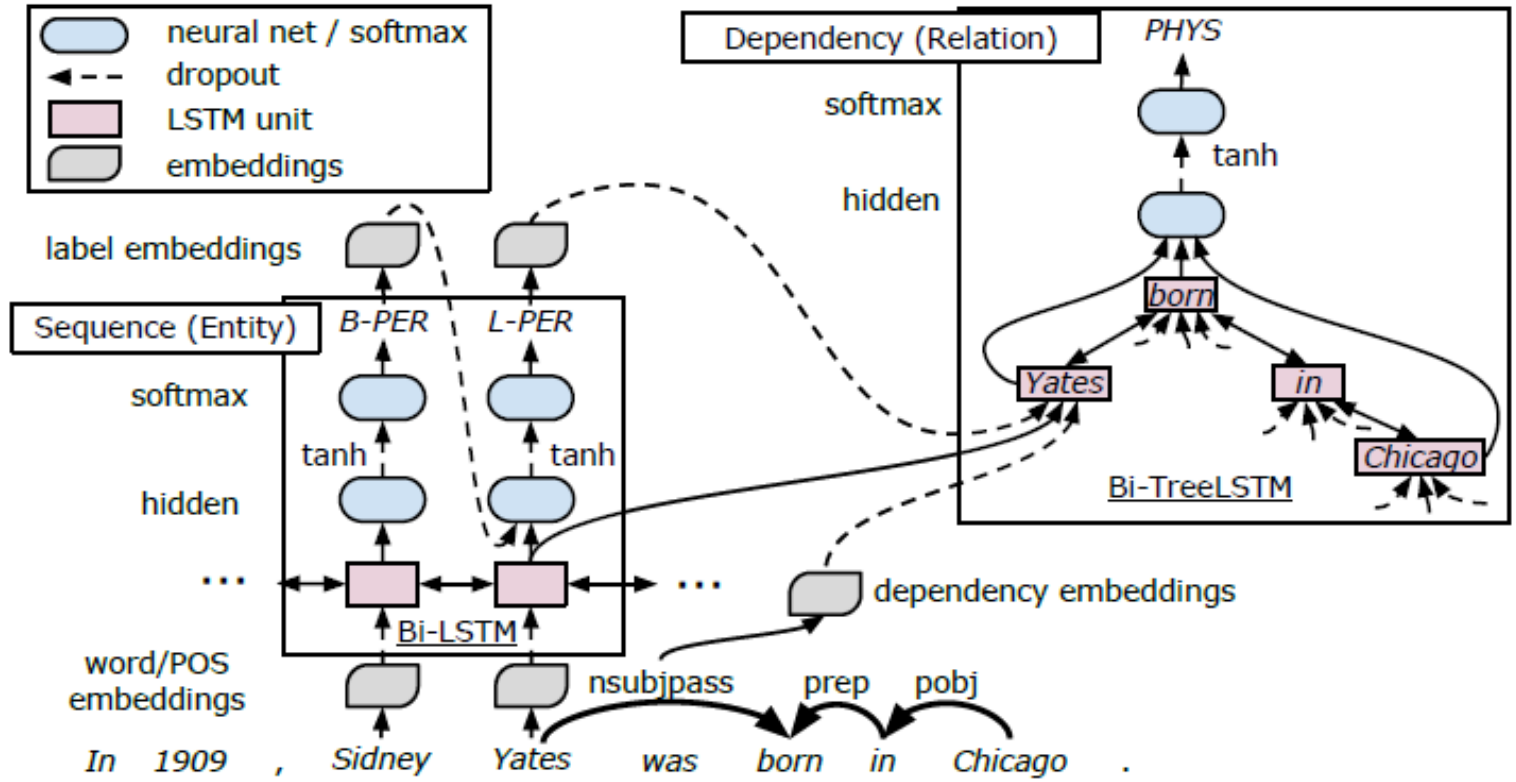
Several approaches tried already:

- RNNs running on the raw text only have trouble getting the relation structure right.
- Adding dependency tree information helps dramatically.

Recent approach: run separate RNNs on the text and dependency parse tree data.

Neural Entity-Relation Extraction

Recent approach: run separate RNNs on the text and dependency parse tree data.



Neural Entity-Relation Extraction

TECHNISCHE UNIVERSITÄT DARMSTADT

Equals previous best scores on an entity-relation benchmark

Corpus	Settings	Entity			Relation		
		P	R	F1	P	R	F1
ACE05	Our Model (SPTree)	0.829	**0.839**	**0.834**	0.572	**0.540**	**0.556**
	Li and Ji (2014)	**0.852**	0.769	0.808	**0.654**	0.398	0.495
ACE04	Our Model (SPTree)	0.808	**0.829**	**0.818**	0.487	**0.481**	**0.484**
	Li and Ji (2014)	**0.835**	0.762	0.797	**0.608**	0.361	0.453

Wrapup

TECHNISCHE UNIVERSITÄT DARMSTADT

Semantics

- Propositional models, entity-relation extraction
- Matrix factorization
- Word2vec
- Skip-Thought vectors
- Siamese models

Translation + Structure Extraction

- Translation
- Parsing
- Entity-Relation extraction

Take-Aways

TECHNISCHE UNIVERSITÄT DARMSTADT

Training data quality (consistency) matters!

- DNNs can model anything, but it shouldn't be human inconsistency.

DNNs need good advice (hints)! c.f. resNets

- DNNs are capable of state-of-the-art parsing, but need a parser to do good ER-extraction now.

Depth matters – Deeper is better

- Between-level dropout is a good regularization scheme