Deep Learning for Natural Language Processing: Theory and Practice

Xiaodong He, Jianfeng Gao, Li Deng

Deep Learning Technology Center
Microsoft Research, Redmond, WA

Tutorial presented at CIKM, November 7th, 2014

Tutorial Outline

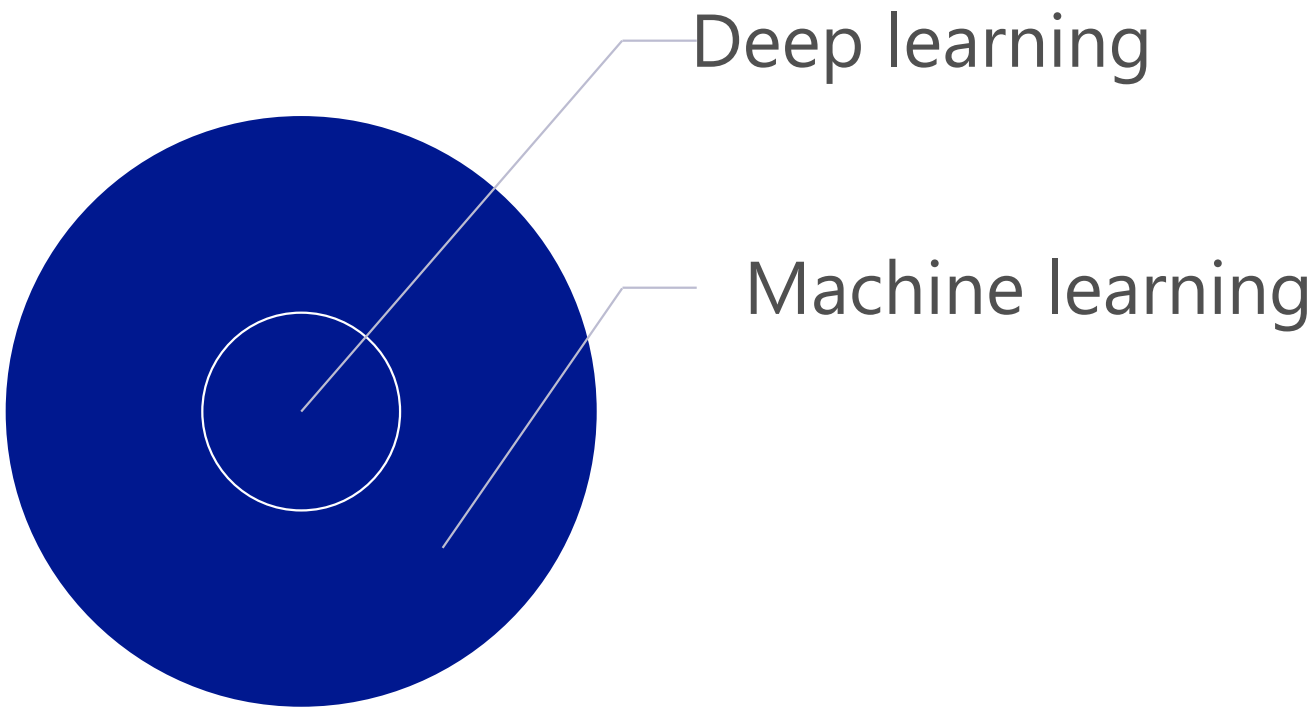
- Part I: Background
- Part II: Deep learning in spoken language understanding
  - Domain & intent detection using DNN
  - Slot filling using RNN
- Part III: Learning semantic embedding
  - Semantic embedding: from words to sentences
  - The Deep Structured Semantic Model (DSSM)
  - DSSM in practice: Word Embedding, Information Retrieval, Question Answering
- Part IV: Deep semantic similarity model for text processing
  - Overview of Semantic Similarity Model
  - DSSM for Web Search
  - DSSM for recommendation
  - DSSM for semantic translation models
- Part V: Conclusion

Part I

Background

Background for deep learning

Machine learning



The Universal Translator … *comes true!*

Deep learning technology enabled speech-to-speech translation

The New York Times

Scientists See Promise in Deep-Learning Programs

John Markoff
November 23, 2012

Rick Rashid in **Tianjin, China**, October, 25, 2012

A voice recognition program translated a speech given by Richard F. Rashid, Microsoft's top scientist, into Mandarin Chinese.

Impact of deep learning in speech technology

Skype to get 'real-time' translator

MIT Technology Review

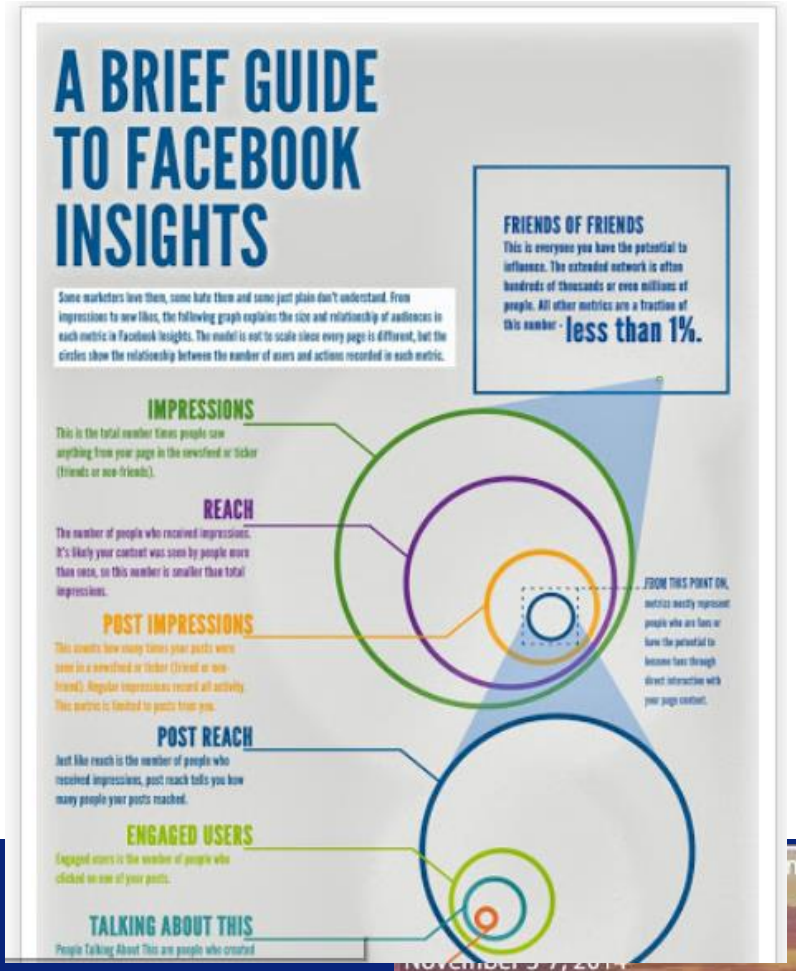
Facebook Launches Advanced AI Effort to Find Meaning in Your Posts

A technique called deep learning could help Facebook understand its users and their data better.

By Tom Simonite on September 20, 2013

September 20, 2013

……Facebook's foray into deep learning sees it following its competitors Google and Microsoft, which have used the approach to impressive effect in the past year. Google has hired and acquired leading talent in the field (see "10 Breakthrough Technologies 2013: Deep Learning"), and last year created software that taught itself to recognize cats and other objects by reviewing stills from YouTube videos. The underlying deep learning technology was later used to slash the error rate of Google's voice recognition services (see "Google's Virtual Brain Goes to Work")….Researchers at Microsoft have used deep learning to build a system that translates speech from English to Mandarin Chinese in real time (see "Microsoft Brings Star Trek's Voice Translator to Life"). Chinese Web giant Baidu also recently established a Silicon Valley research lab to work on deep learning.



MIT Technology Review

BUSINESS NEWS

8 COMMENTS

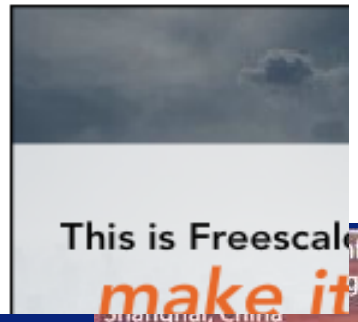
Is Google Cornering the Market on Deep Learning?

A cutting-edge corner of science is being wooed by Silicon Valley, to the dismay of some academics.

By Antonio Regalado on January 29, 2014

How much are a dozen deep-learning researchers worth? Apparently, more than $400 million.

This week, Google reportedly paid that much to acquire DeepMind Technologies, a startup based in



DNN: (Fully-Connected) Deep Neural Networks

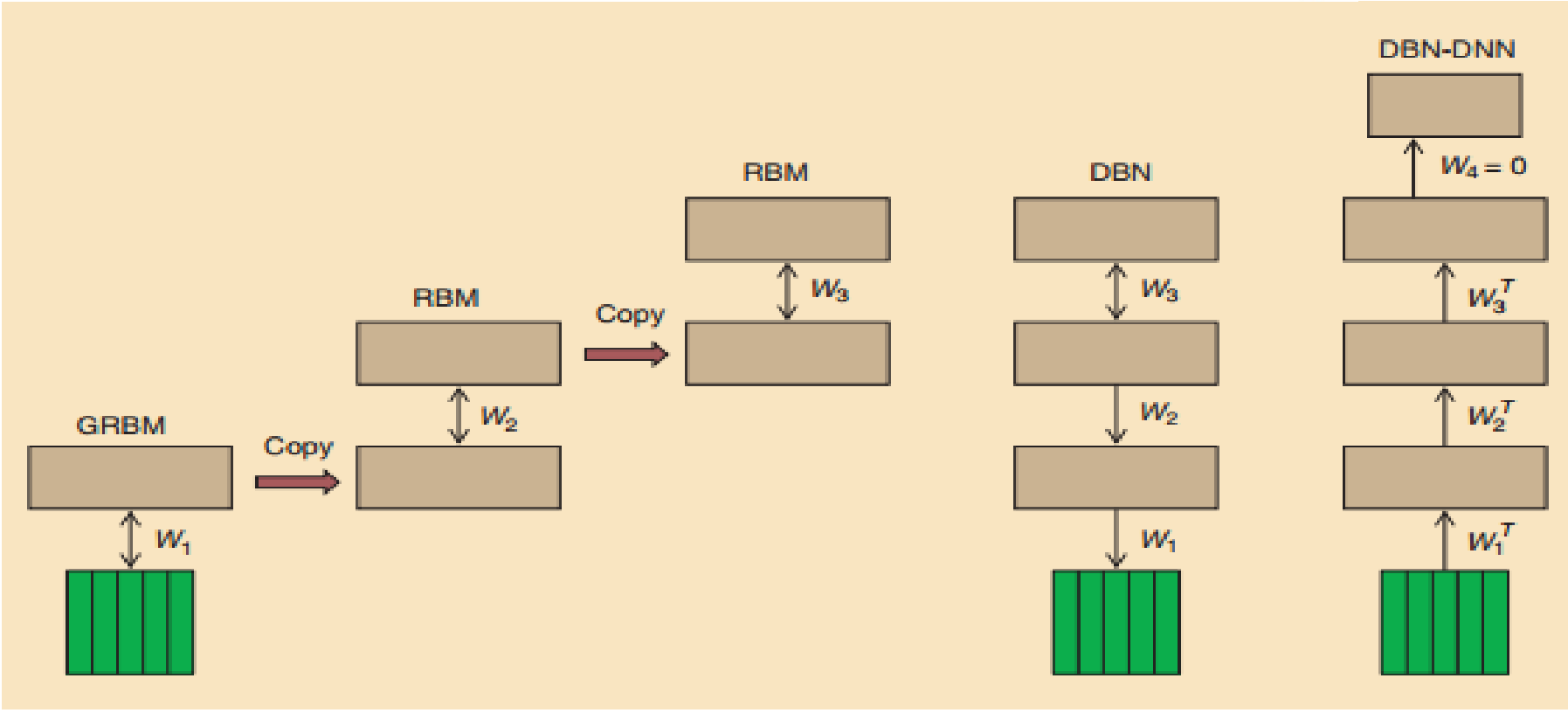
"DNN for acoustic modeling in speech recognition," in *IEEE SPM*, Nov. 2012

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury

Geoff Hinton

Li Deng

Dong Yu

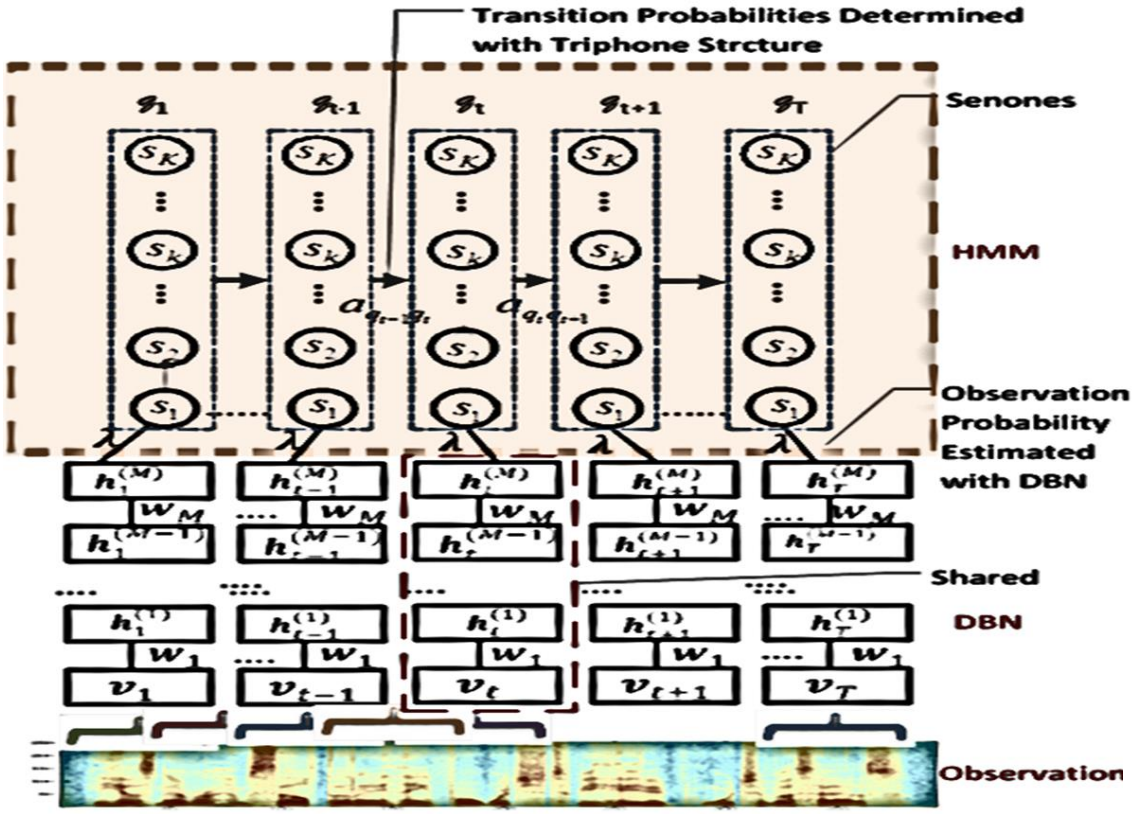


First train a stack of N models each of which has one hidden layer. Each model in the stack treats the hidden variables of the previous model as data.

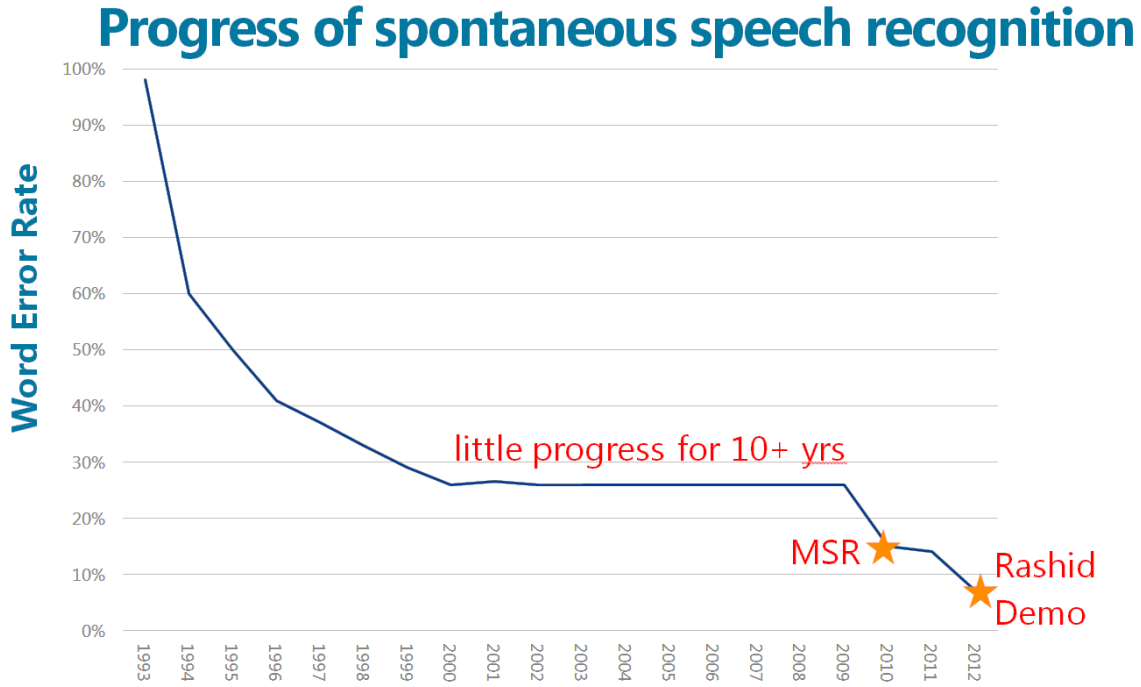
Then compose them into a single Deep Belief Network.

Then add outputs and train the DNN with backprop.

CD-DNN-HMM



Dahl, Yu, Deng, and Acero, "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition," *IEEE Trans. ASLP*, Jan. 2012

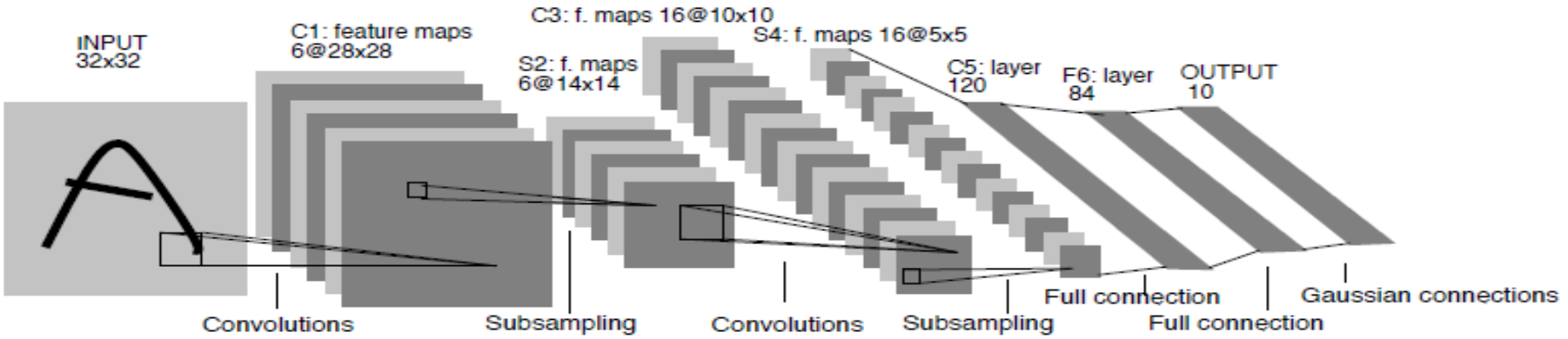


After no improvement for 10+ years by the research community…
…MSR reduced error from **~23%** to **<13%**
(and under 7% for Rick Rashid's S2S demo)!

Deep Convolutional NN for Images

Yann LeCun

CNN: local connections with weight sharing;
pooling for translation invariance

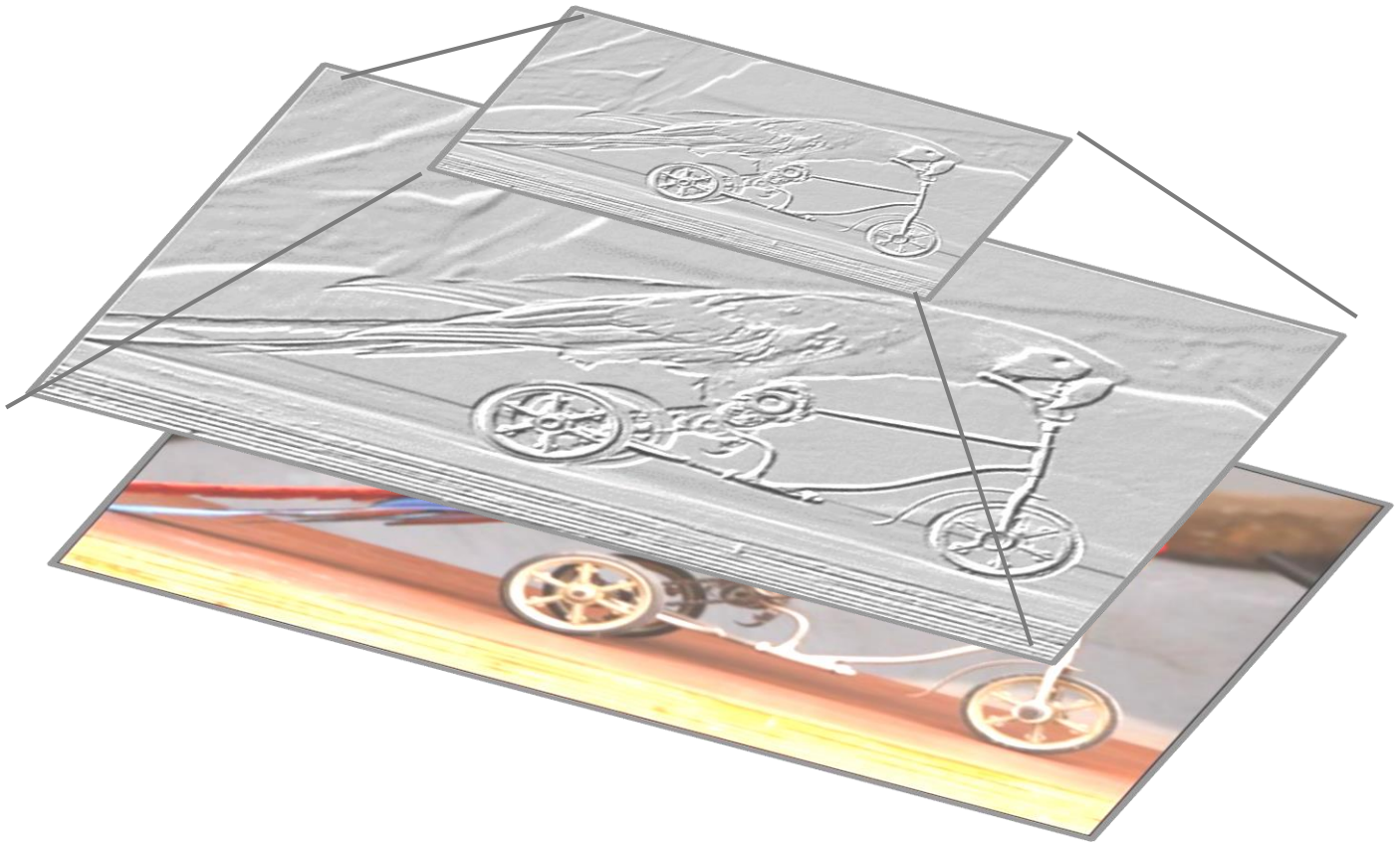


Image

LeCun et al., 1998

Output

A Basic Module of the CNN



Pooling

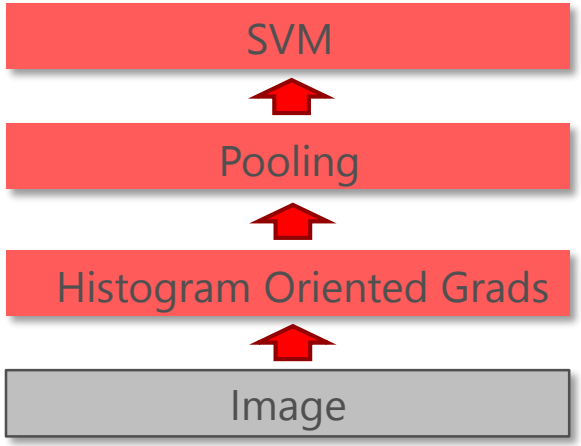
Convolution

Image

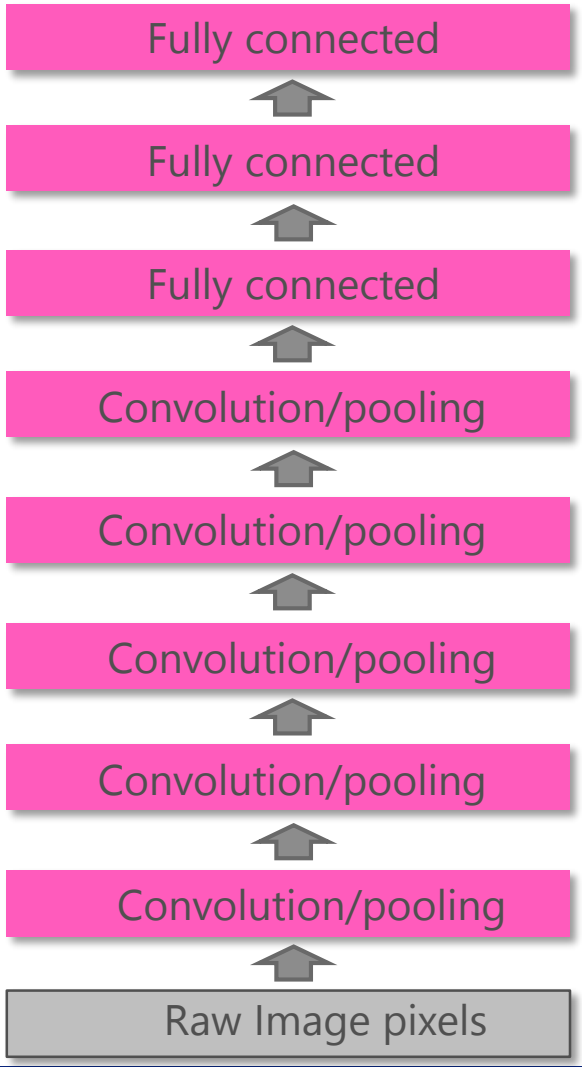
Deep Convolutional NN for Images

A paradigm shift in 2012!

earlier

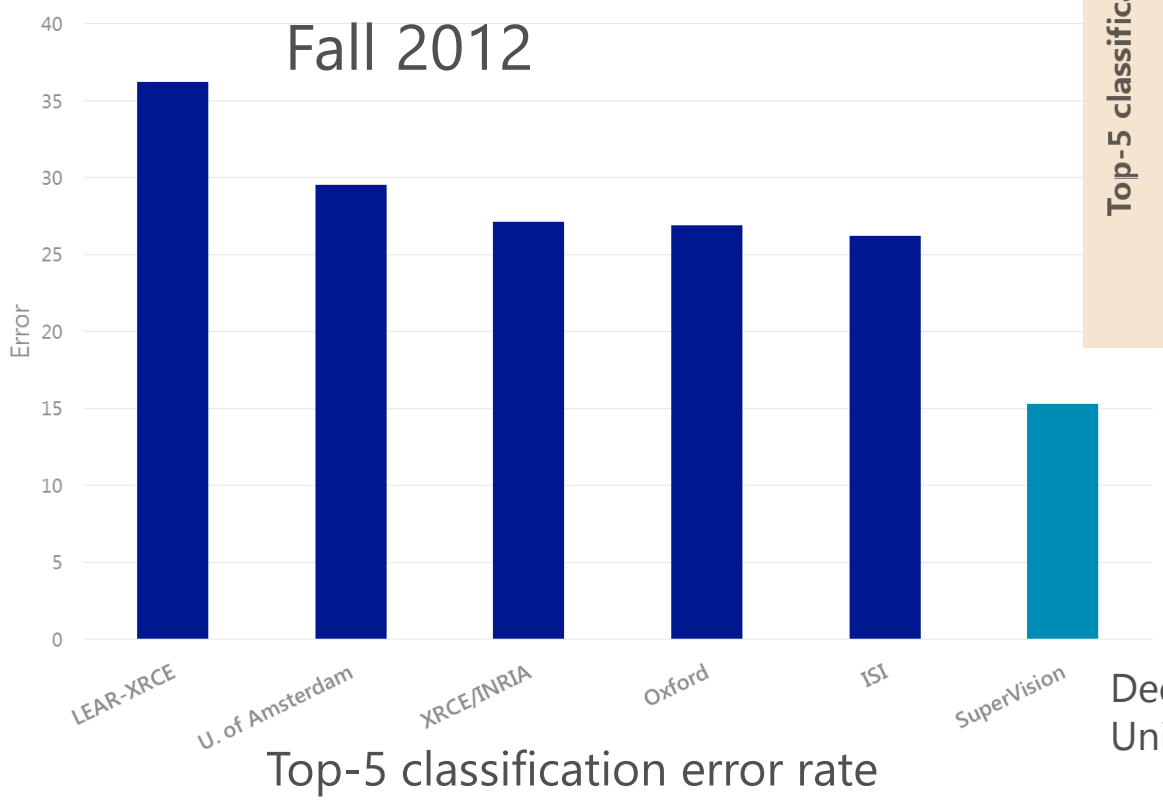


2012

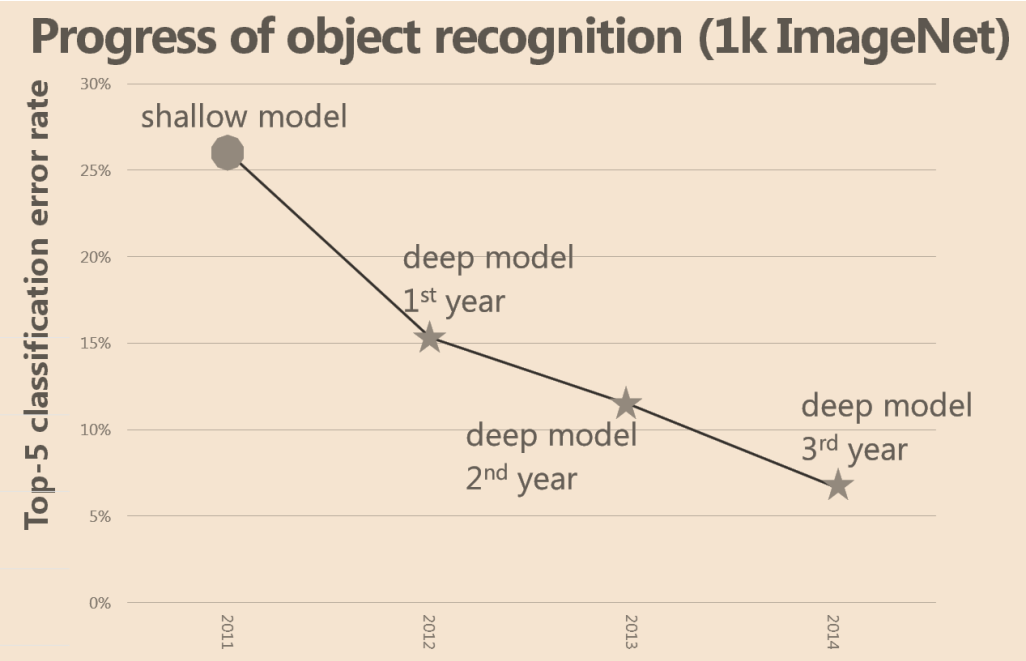


ImageNet 1K Competition

Krizhevsky, Sutskever, Hinton, "ImageNet Classification with Deep Convolutional Neural Networks." *NIPS*, Dec. 2012



Top-5 classification error rate



2012 - 2014

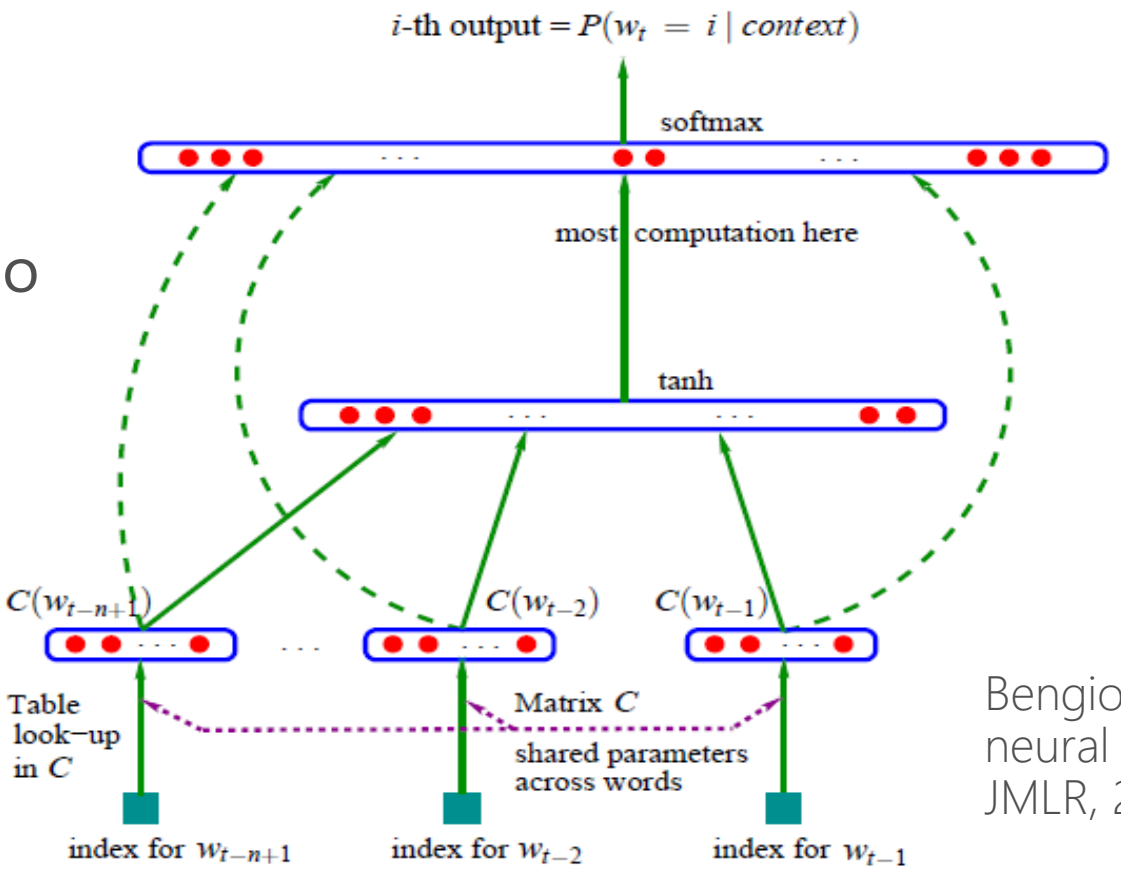
Deep CNN !!!
Univ. Toronto team

Neural network based language model

Yoshua Bengio

LM: predict the next word given the past:

e.g., $p(chases|the\ cat) = ?$, $p(says|the\ cat) = ?$

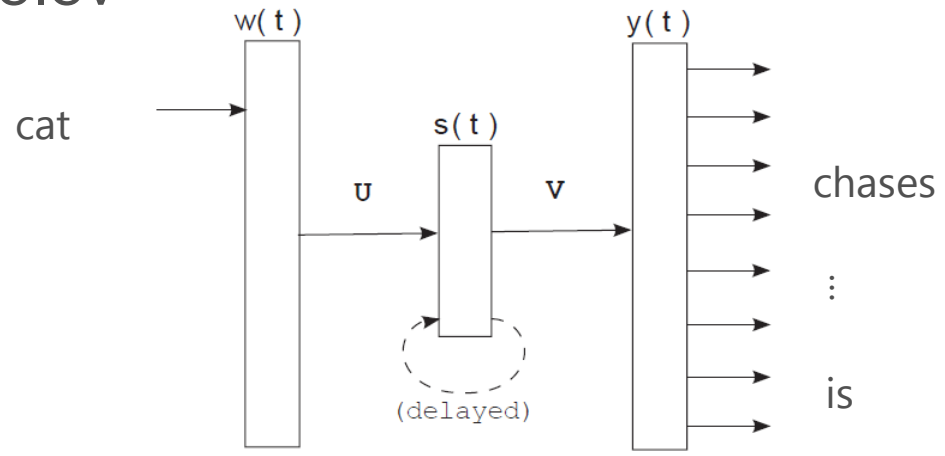


Bengio, Ducharme, Vincent, Jauvin, "A neural probabilistic language model. " JMLR, 2003

Recurrent NN based language model

Tomas Mikolov

Mikolov, Karafiat, Burget, Cernocky, Khudanpur, "Recurrent neural network based language model." Interspeech, 2010



- Large LM perplexity reduction
- Lower ASR WER improvement
- Expensive in learning
- Later turned to FFNN at Google: Word2vec, Skip-gram, etc.
- All UNSUPERVISED

Table 1: *Performance of models on WSJ DEV set when increasing size of training data.*

Model	# words	PPL	WER
KN5 LM	200K	336	16.4
KN5 LM + RNN 90/2	200K	271	15.4
KN5 LM	1M	287	15.1
KN5 LM + RNN 90/2	1M	225	14.0
KN5 LM	6.4M	221	13.5
KN5 LM + RNN 250/5	6.4M	156	11.7

Deep learning demonstrates great success in speech, image, and natural language!

Is **Deep Learning**, the 'holy grail' of big data? - CNBC - Video
video.cnbc.com/gallery/?video=3000192292
Aug 22, 2013
Derrick Harris, GigaOM, explains how "**Deep Learning**" computers are able to process and understand ...

Useful Sites on Deep Learning

- http://www.cs.toronto.edu/~hinton/
- http://ufldl.stanford.edu/wiki/index.php/UFLDL_Recommended_Readings
- http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial (Andrew Ng's group)
- http://deeplearning.net/reading-list/ (Bengio's group)
- http://deeplearning.net/tutorial/
- http://deeplearning.net/deep-learning-research-groups-and-labs/

- Google+ Deep Learning community

Interim Summary

- Deep learning sees great impact in Speech, Image, and Text
- Common deep learning architectures
  - DNN (Deep Neural Nets)
  - CNN (Convolutional Neural Nets)
  - RNN (Recurrent Neural Nets)
- The next 4 parts will elaborate on the learning and applications of deep models in NLP

Part II

Deep learning in spoken language understanding

Deep learning for spoken language processing

The scenarios

- Domain & intent classification
- Semantic slot filling

"Show me flights from Boston to New York today"

Domain: travel

Intent: find_flight

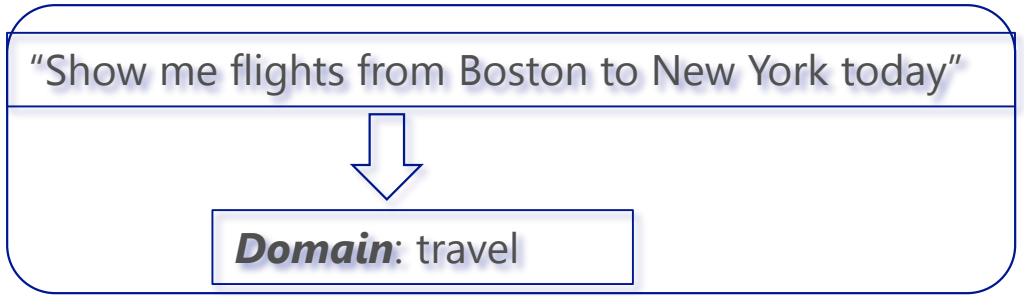
"Show me flights from Boston to New York today"

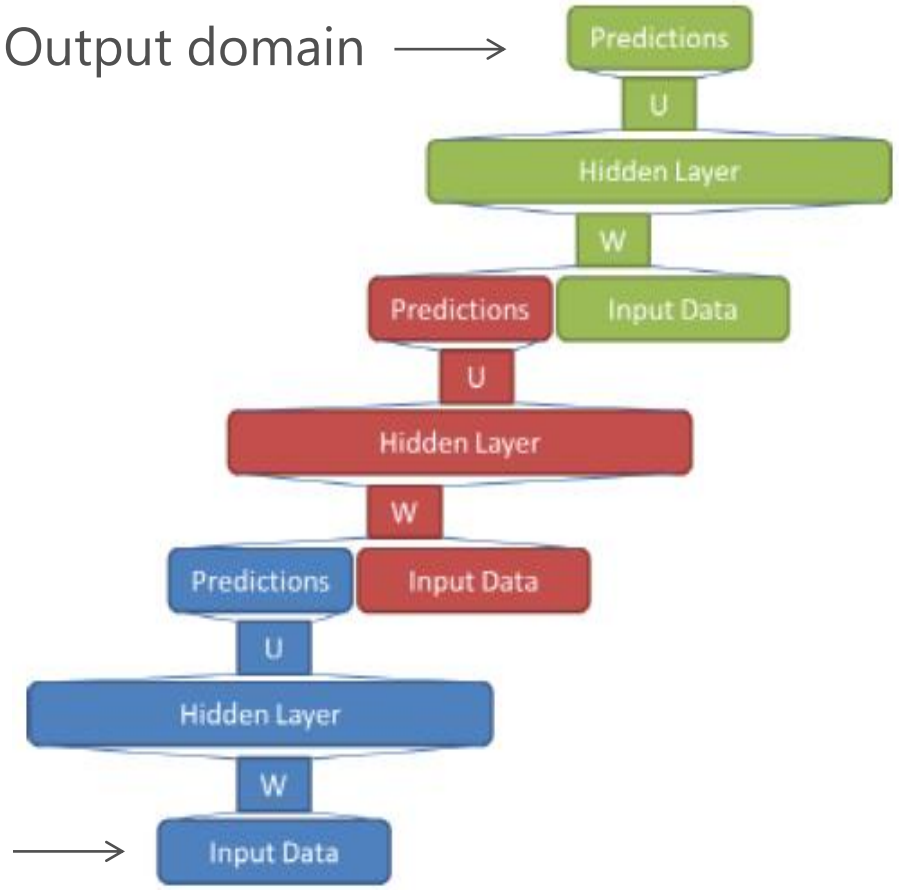
Semantic slots: City-departure City-arrival Date

Deep stack net for domain & intent classification

Deep stack net for semantic utterance classification:

1) A stack of a series of 3-layer perceptron modules
2) Output layer is concatenated with raw input to form input layer of the next module





[Tur, Deng, Hakkani-Tur, He, 2012; Deng, Tur, He, Hakkani-Tur, 2012]

Domain classification results

Table 2. *Comparisons of the domain classification error rates among the boosting-based baseline system, DCN system, and K-DCN system for a domain classification task. Three types of raw features (lexical, query clicks, and name entities) and four ways of their combinations are used for the evaluation as shown in four rows of the table.*

Feature Sets	Baseline	DCN	K-DCN
lexical features	10.40%	10.09%	**9.52%**
lexical features + Named Entities	9.40%	9.32%	**8.88%**
lexical features + Query clicks	8.50%	7.43%	**5.94%**
lexical features + Query clicks + Named Entities	10.10%	7.26%	**5.89%**

30% error reduction over a boosting-based baseline!

Table 3. *More detailed results of K-DCN in Table 2 with Lexical+QueryClick features. Domain classification error rates (percent) on Train set, Dev set, and Test set as a function of the depth of the K-DCN.*

Depth	Train Err%	Dev Error%	Test Err%
1	9.54	12.90	12.20
2	6.36	10.50	9.99
3	4.12	9.25	8.25
4	1.39	7.00	7.20
5	0.28	6.50	5.94
6	0.26	**6.45**	**5.94**
7	0.26	6.55	6.26
8	0.27	6.60	6.20

Error keeps decreasing until up to six layers are added up

Deng, Tur, He, Hakkani-Tur, Use of kernel deep convex networks and end-to-end learning for spoken language understanding, IEEE-SLT 2012

Semantic slot filling

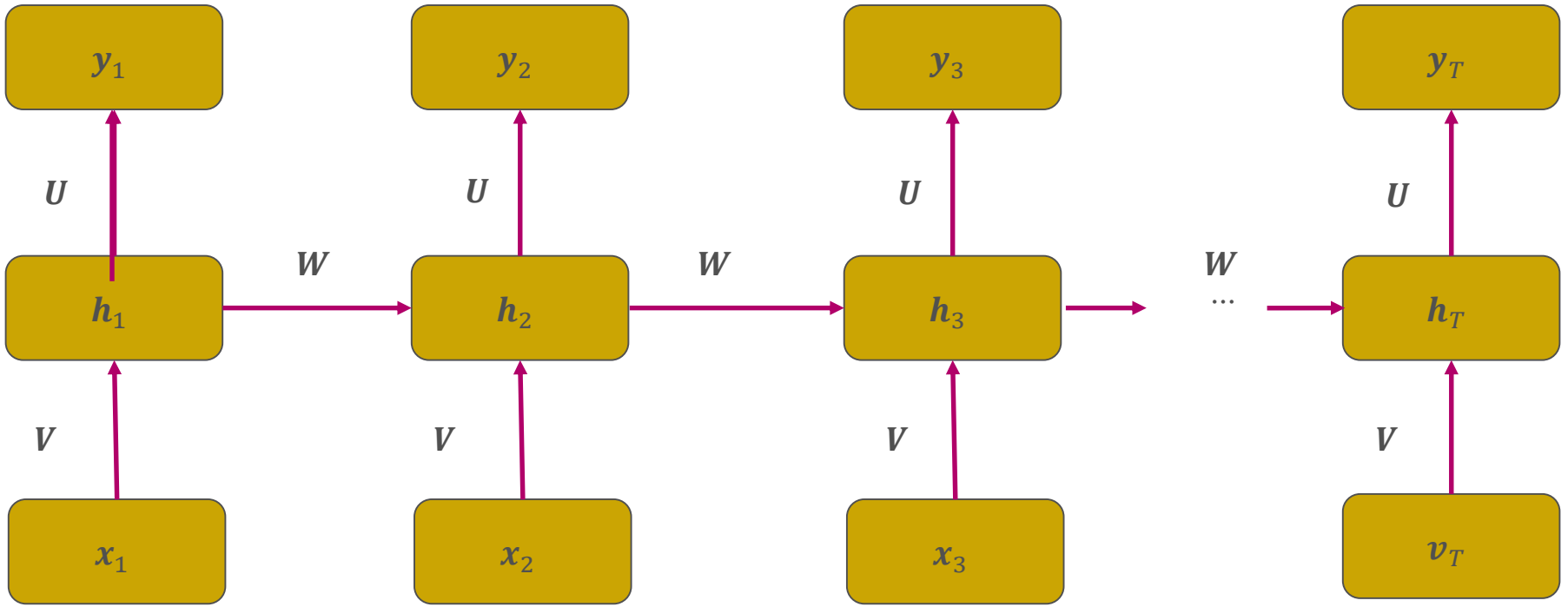
A example in the Airline Travel Information System (ATIS) corpus

	show	*flights*	*from*	*boston*	*to*	*new*	*york*	*today*
Slots	O	O	O	B-dept	O	B-arr	I-arr	B-date

Slot filling can be viewed as a sequential tagging problem

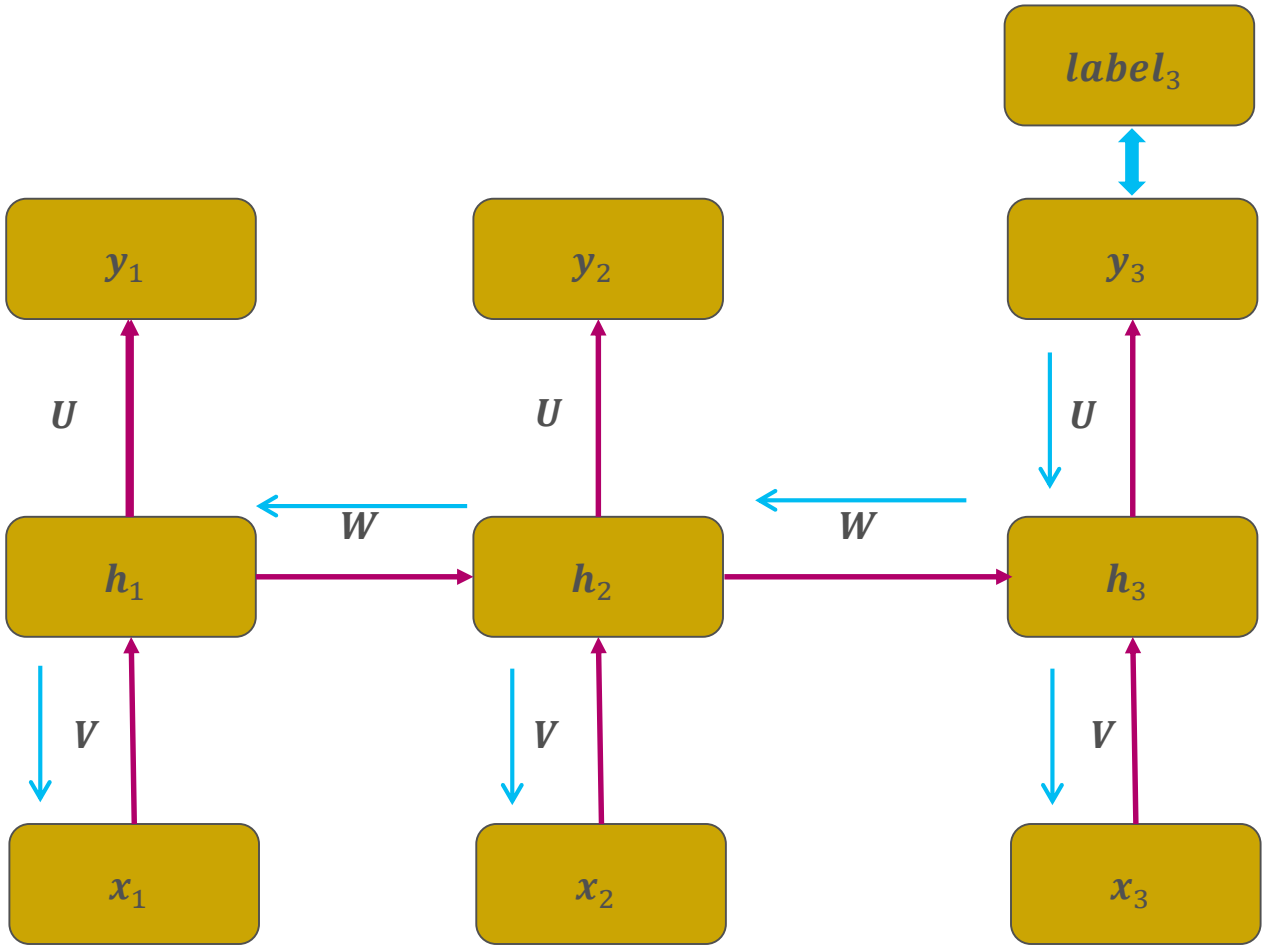
Recurrent neural networks for slot filling

h_t is the hidden layer that carries the information from time $0 \sim t$
where x_t: *the input word* , y_t: *the output tag*
$y_t = SoftMax(U \cdot h_t), where\ h_t = \sigma(W \cdot h_{t-1} + V \cdot x_t)$



[Mesnil, He, Deng, Bengio, 2013; Yao, Zweig, Hwang, Shi, Yu, 2013]

Back-propagation through time (BPTT)



at time $t = 3$

1. Forward propagation
2. Generate output
3. Calculate error
4. Back propagation
5. Back prop. through time

Results

- Evaluated on the ATIS corpus
  - 4978 utterances for training
  - 893 utterances for testing
  - Using word feature only
  - Baseline CRF: 92.94% in F1-measure

~25% error reduction!

SGD vs. minibatch training

With local context window

Model	Elman	Jordan	Hybrid
Stochastic GD	94.55 ±0.51	94.66 ±0.23	94.75 ±0.31
Sentence-minibatch	94.54 ±0.23	94.33 ±0.19	94.25 ±0.28

Left-to-right vs. bi-directional RNN

With local context window

Model	Elman	Jordan
Left-to-right	94.54	94.33
bi-direction	94.73	94.03

Without local context window

Model	Elman	Jordan
Left-to-right	93.15	65.23
bi-direction	93.46	90.31

Interim Summary

- Introduction to SLU
- DNN/DCN/K-DCN for Domain/intent detection
- RNN and its variants for slot filling
- Deep learning models demonstrate superior performances on these tasks

However, understanding human language is more challenging than that …

Part III

Learning Semantic Embedding

Why understanding language is difficult?

Human language has great variability

similar concepts are expressed in different ways, e.g., *kitty* vs. *cat*

Human language has great ambiguity

similar expressions mean different concepts, e.g.,
new york vs. *new york times*

The meaning of text is usually vague and latent

e.g., no clear "supervision" signal to learn from as in speech/image recog.

Learning semantic meaning of texts is a key challenge in NLP

Semantic embedding

Project raw text into a continuous semantic space

e.g., word embedding

Captures the word meaning in a semantic space

$f(\boldsymbol{cat}) =$

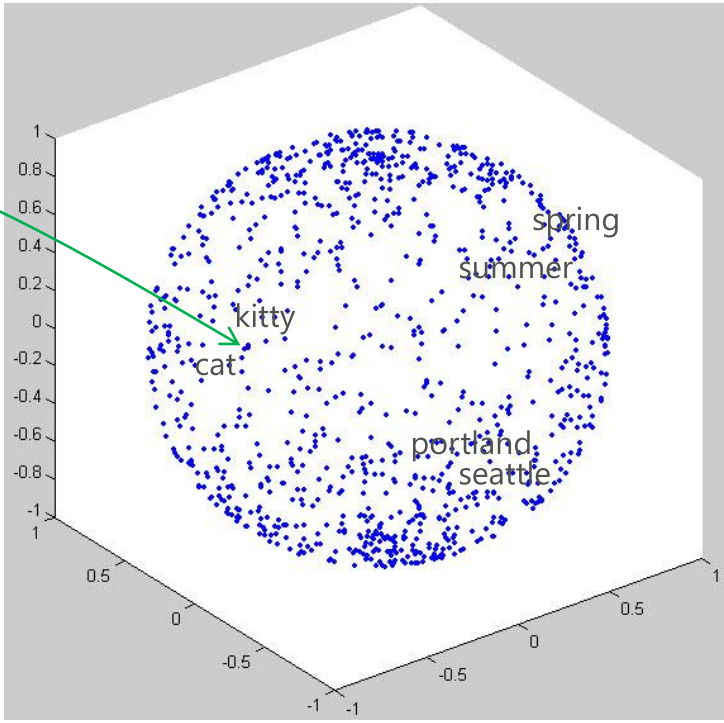
The index of "cat" in the vocabulary

a.k.a the 1-hot word vector

word embedding vector in the semantic space

Dim=100~1000

Dim=|V|=100K~100M



Deerwester, Dumais, Furnas, Landauer, Harshman, "Indexing by latent semantic analysis," JASIS 1990

SENNA word embedding

Scoring:

$$Score(w_1, w_2, w_3, w_4, w_5) = U^T \sigma(W[f_1, f_2, f_3, f_4, f_5] + b)$$

Training:

$$J = \max(0, 1 + S^- - S^+)$$

Update the model until $S^+ > 1 + S^-$

Where

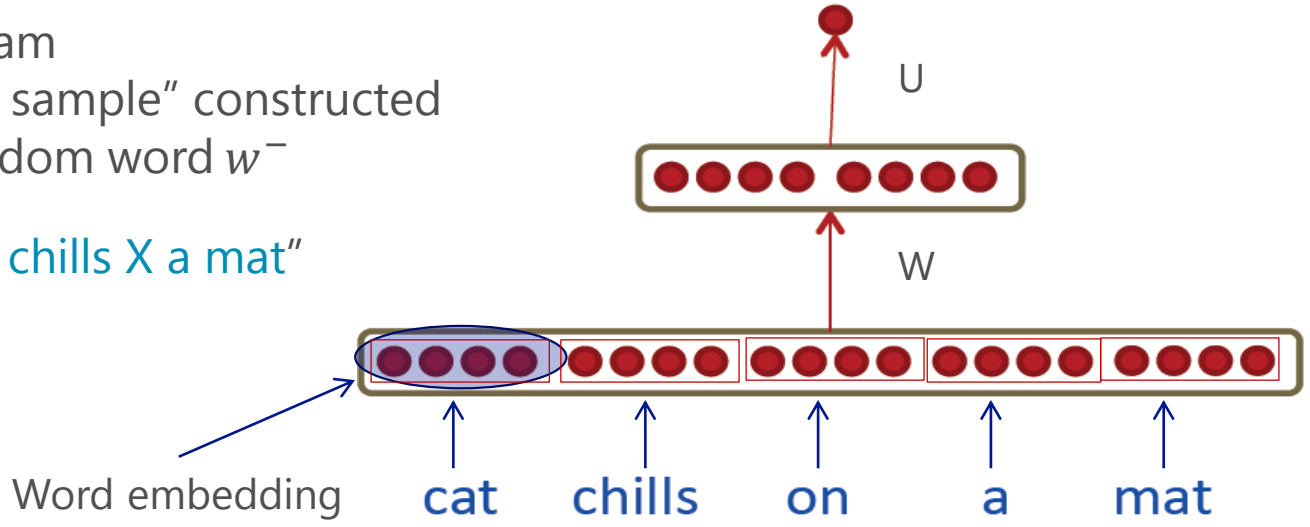
$$S^+ = Score(w_1, w_2, w_3, w_4, w_5)$$
$$S^- = Score(w_1, w_2, w^-, w_4, w_5)$$

And

$< w_1, w_2, w_3, w_4, w_5 >$ is a valid 5-gram

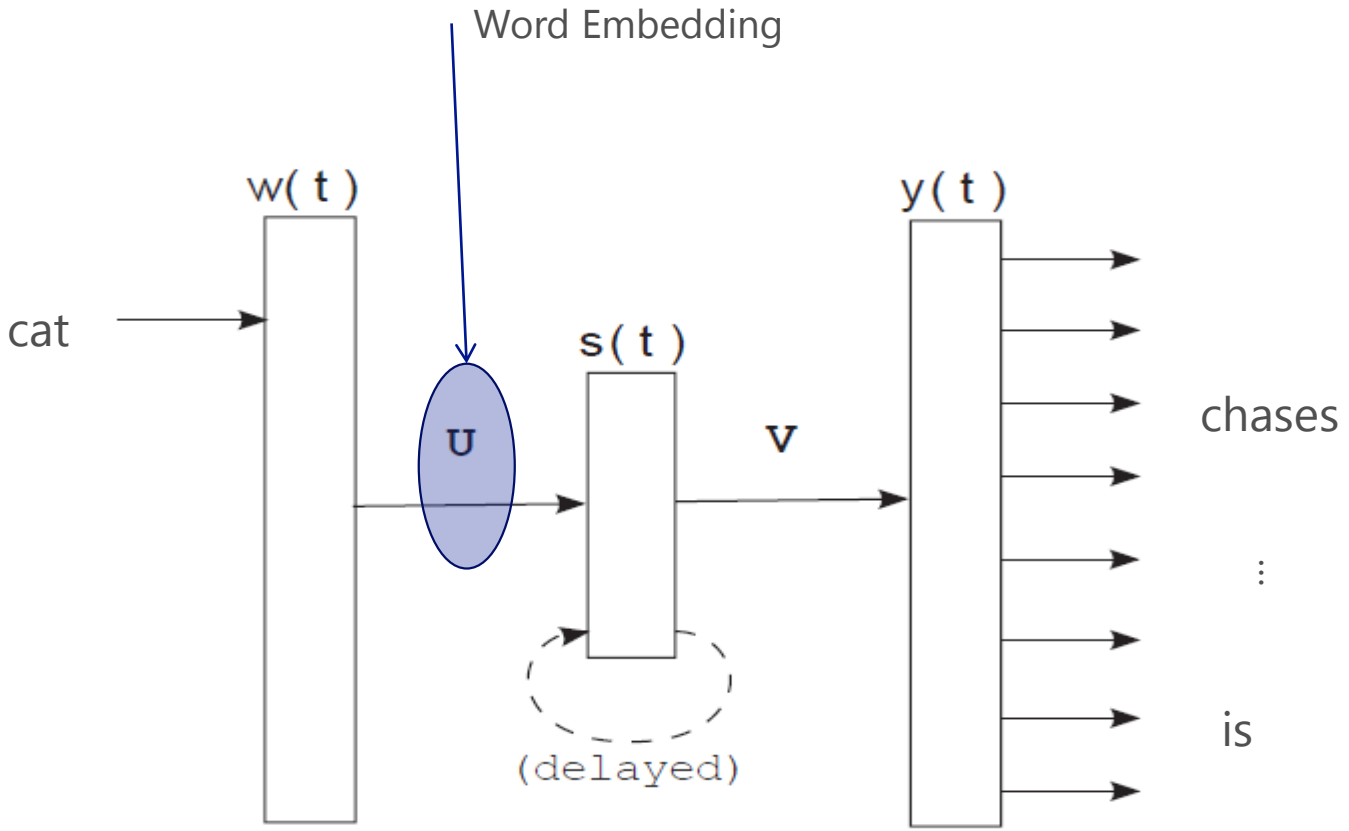
$< w_1, w_2, w^-, w_4, w_5 >$ is a "negative sample" constructed by replacing the word w_3 with a random word w^-

e.g., a negative example: "cat chills X a mat"



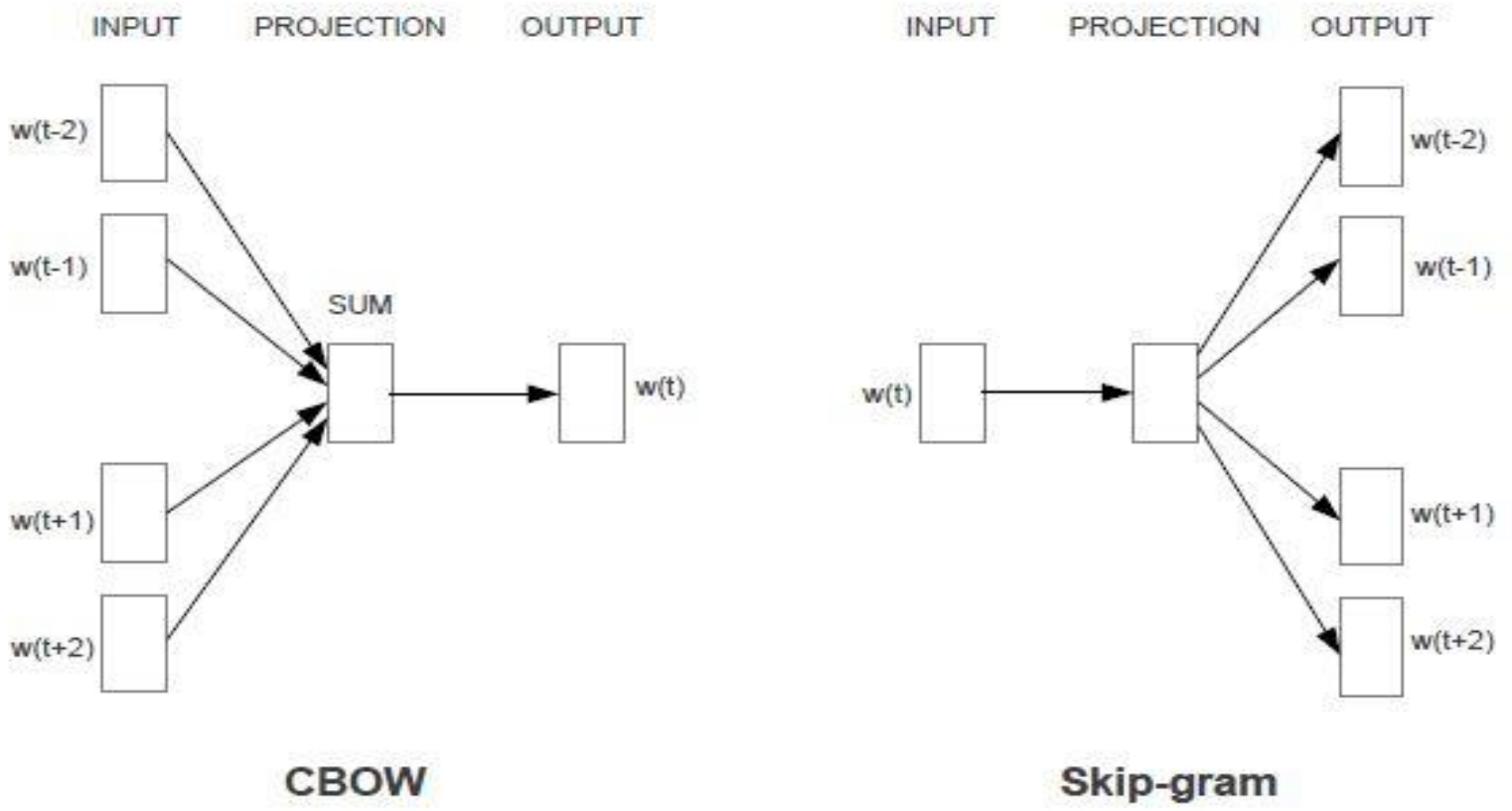
Collobert, Weston, Bottou, Karlen, Kavukcuoglu, Kuksa, "Natural Language Processing (Almost) from Scratch," JMLR 2011

RNN-LM base word embedding



Mikolov, Yih, Zweig, "Linguistic Regularities in Continuous Space Word Representations," NAACL 2013

CBOW/Skip-gram Word Embeddings

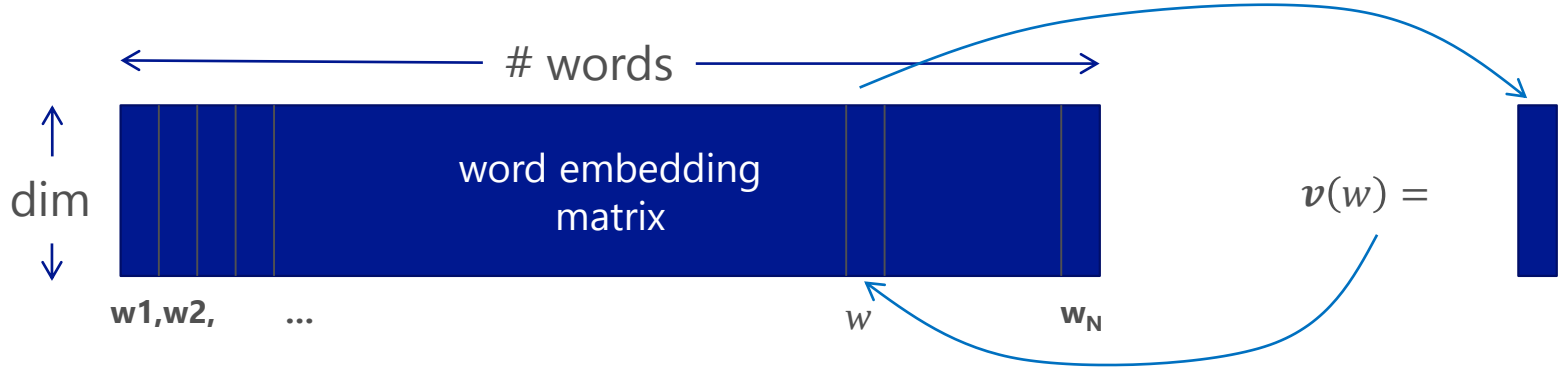


Continuous Bag-of-Words

The CBOW architecture (a) on the left, and the Skip-gram architecture (b) on the right. [Mikolov et al., 2013 ICLR].

Word embedding: rethinking

- Word embedding is a neat and effective representation:



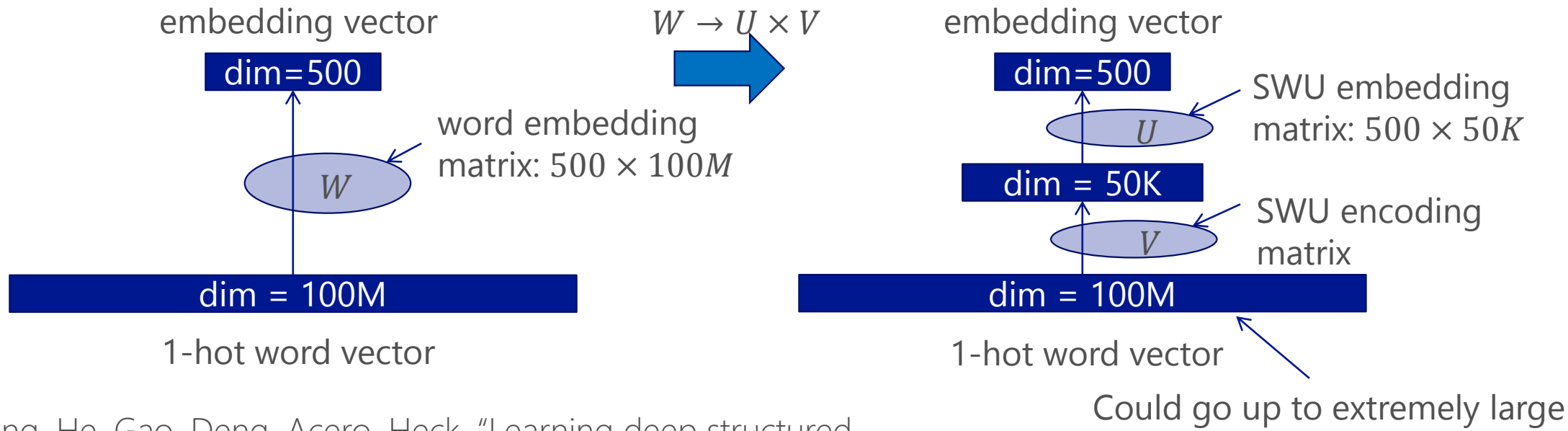
- However, for large scale NL tasks a decomposable, robust representation is preferable
  - Vocabulary of real-world big data tasks could be huge (*scalability*)

    >100M unique words in a modern commercial search engine log, and keeps growing
  - New words, misspellings, and word fragments frequently occur (*generalizability*)

Build semantic embedding on top of sub-word units

Learn semantic embedding on top of sub-word units (SWU)

- Decompose *any* word into sub-word units
- *Scale* the capacity to handle almost unbounded variability (word) based on bounded variability (sub-word)



Huang, He, Gao, Deng, Acero, Heck, "Learning deep structured semantic models for web search using clickthrough data," CIKM, 2013

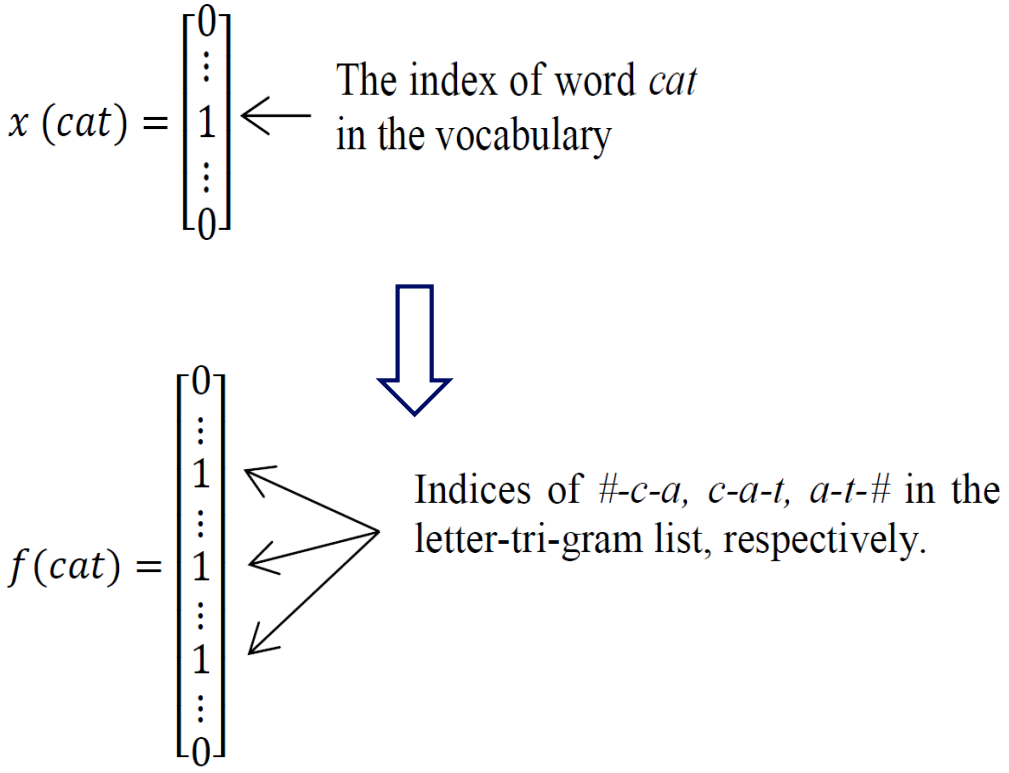
Sub-word unit

- Letters, context-dept letters, positioned-phones, context-dept phones, positioned-roots/morphs, context-dept morphs
- Multi-hashing approach to word input representation

Or random projection (random basis)

Sub-word unit encoding

- E.g., letter-trigram based Word Hashing of "cat"
  - -> #cat#
  - Tri-letters: #-c-a, c-a-t, a-t-#.
- Compact representation
  - |Voc| (500K) → |Letter-trigram| (30K)
- Generalize to unseen words
- Robust to misspelling, inflection, etc.



What if different words have the same word hashing vector (collision)?

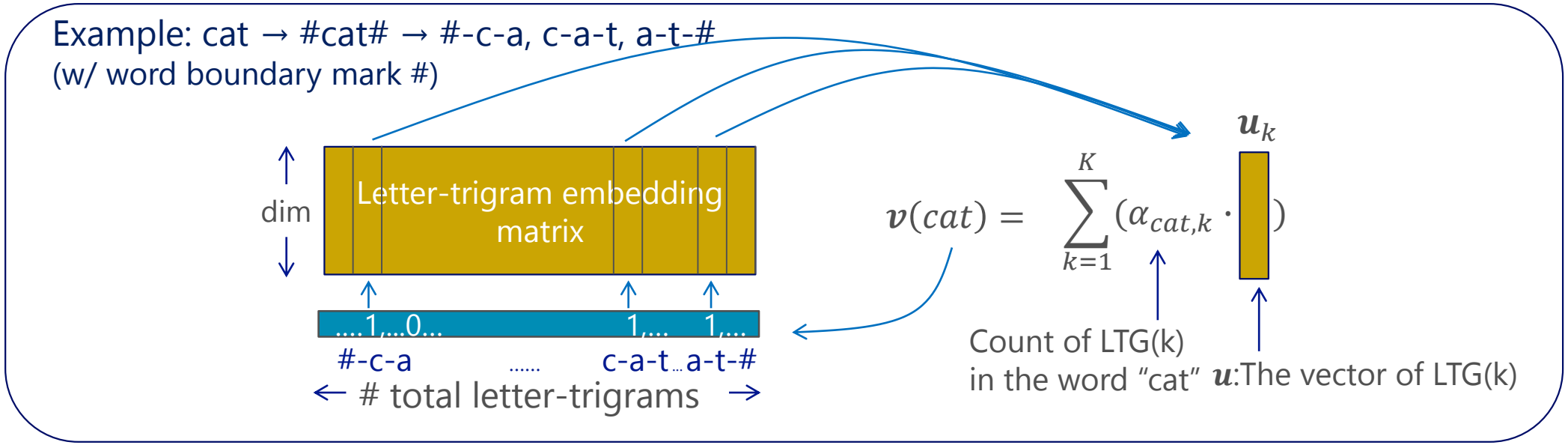
Vocabulary size	Unique letter-tg observed in voc	Number of Collisions
40K	10306	2 (0.005%)
500K	30621	22 (0.004%)

From sub-word unit embedding vectors to word vectors

SWU uses context-dependent letter, e.g., letter-trigram.

Learn **one vector per letter-trigram** (LTG), the encoding matrix is a fixed matrix

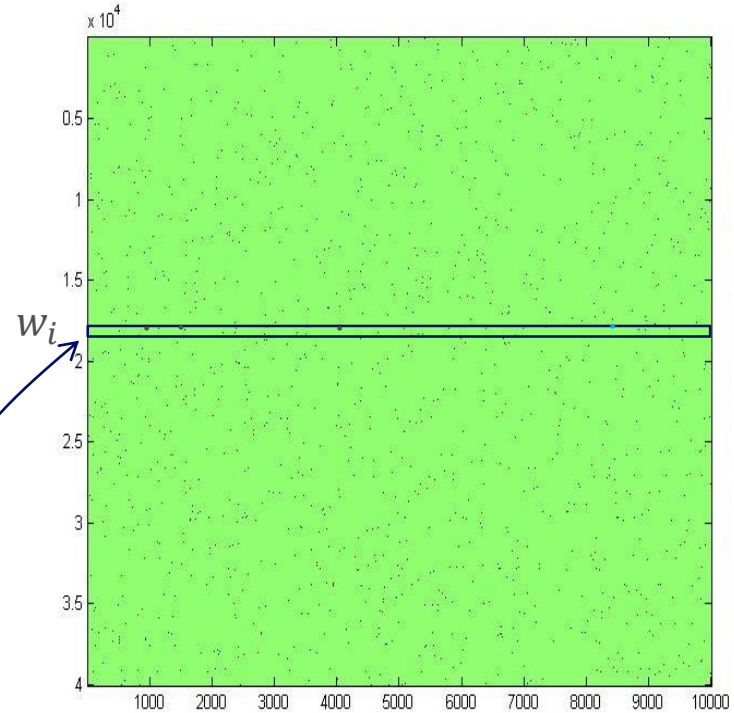
- Use the count of each LTG in the word for encoding



Two words has the same LTG: collision rate ≈ 0.004%

Other representation: random projection

- Sparse random projection matrix R with entries sampled i.i.d. from a distribution over [0, 1, -1]
- Entries of 1 and -1 are equally probable
- $P(R_{ij} = 0) = 1 - \frac{1}{\sqrt{d'}}$, where d is the original input dimensionality.



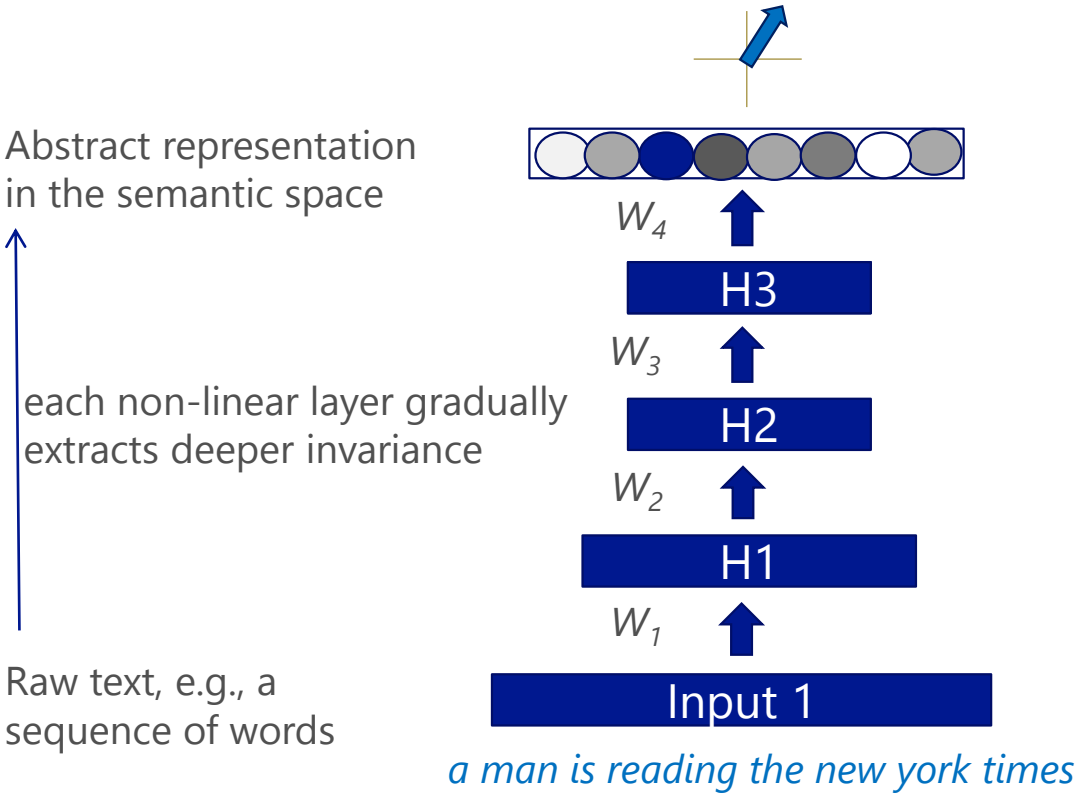
Each word will have a set of sparse random encoding of the 10000 basic units

[Li, Hastie, and Church 2006]

Semantic embedding: from words to sentences

- The semantic intent is better defined at the phrase/sentence level rather than at the word level
  - The meaning of a single word is often ambiguous
  - A phrase/sentence/document contains rich contextual information that could be leveraged

Deep learning for semantic embedding



However

- the semantic meaning of texts – to be learned – is latent
- no clear target for the model to learn
- How to do back-propagation / training?

Fortunately

- we usually know if two texts are "similar" or not.
- That's the training signal for us!

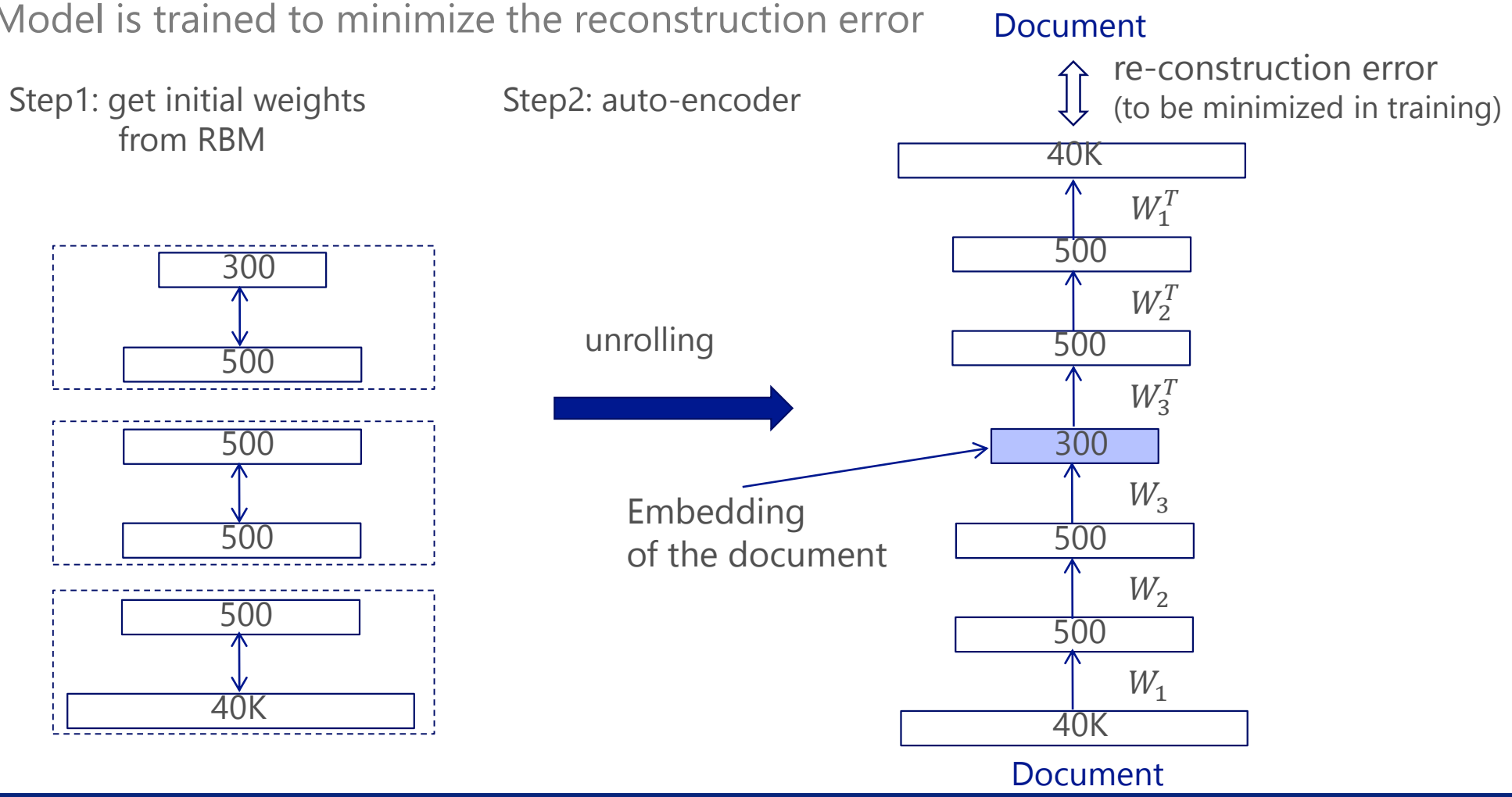
Semantic Hashing

[Salakhutdinov & Hinton 2007, 2010]

1) Single layer learning: Restricted Boltzmann Machine (RBM)

2) Multi-layer training: deep auto-encoder, learn internal representations

Model is trained to minimize the reconstruction error



Auto-encoder: rethinking

- The objective for training the auto-encoder?
  - What is the relation between minimizing re-construction error and good embedding?
- What is a *good* embedding?
  - Good embedding helps end-to-end tasks, so:
    - Optimizing embedding directly instead of minimizing the doc re-construction error
    - Learning the model with end-to-end user behavior log data (weak supervision) beside documents

Deep Structured Semantic Model

Deep Structured Semantic Model/Deep Semantic Similarity Model (DSSM)

the DSSM learns phrase/sentence level semantic vector representation, e.g., query, document

The DSSM is built upon sub-word units for scalability and generalizability

e.g., letter-trigram, phones, roots/morphs

The DSSM is trained by an similarity-driven objective

projecting semantically similar phrases to vectors close to each other

projecting semantically different phrases to vectors far apart

The DSSM is trained using various signals, with or without human labeling effort

semantically-similar text pairs

e.g., user behavior log data, contextual text

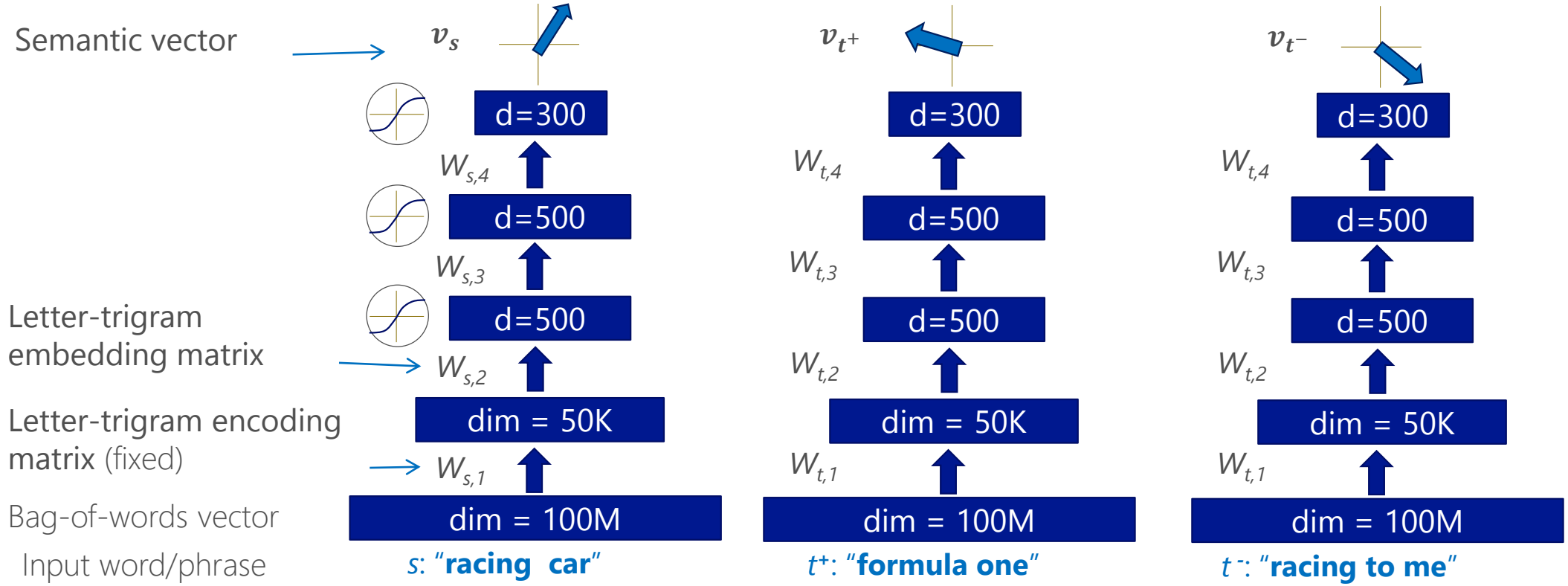
[Huang, He, Gao, Deng, Acero, Heck, CIKM2013]
[Shen, He, Gao, Deng, Mesnil, WWW2014]
[Gao, He, Yih, Deng, ACL2014]
[Yih, He, Meek, ACL2014]
[Song, He, Gao, Deng, Shen, MSR-TR 2014]
[Gao, Pantel, Gamon, He, Deng, Shen, EMNLP2014]
[Shen, He, Gao, Deng, Mesnil, CIKM2014]
[He, Gao, Deng, ICASSP2014 Tutorial]

DSSM for semantic embedding Learning

Initialization:

Neural networks are initialized with random weights

Huang, He, Gao, Deng, Acero, Heck, "Learning deep structured semantic models for web search using clickthrough data," CIKM, 2013

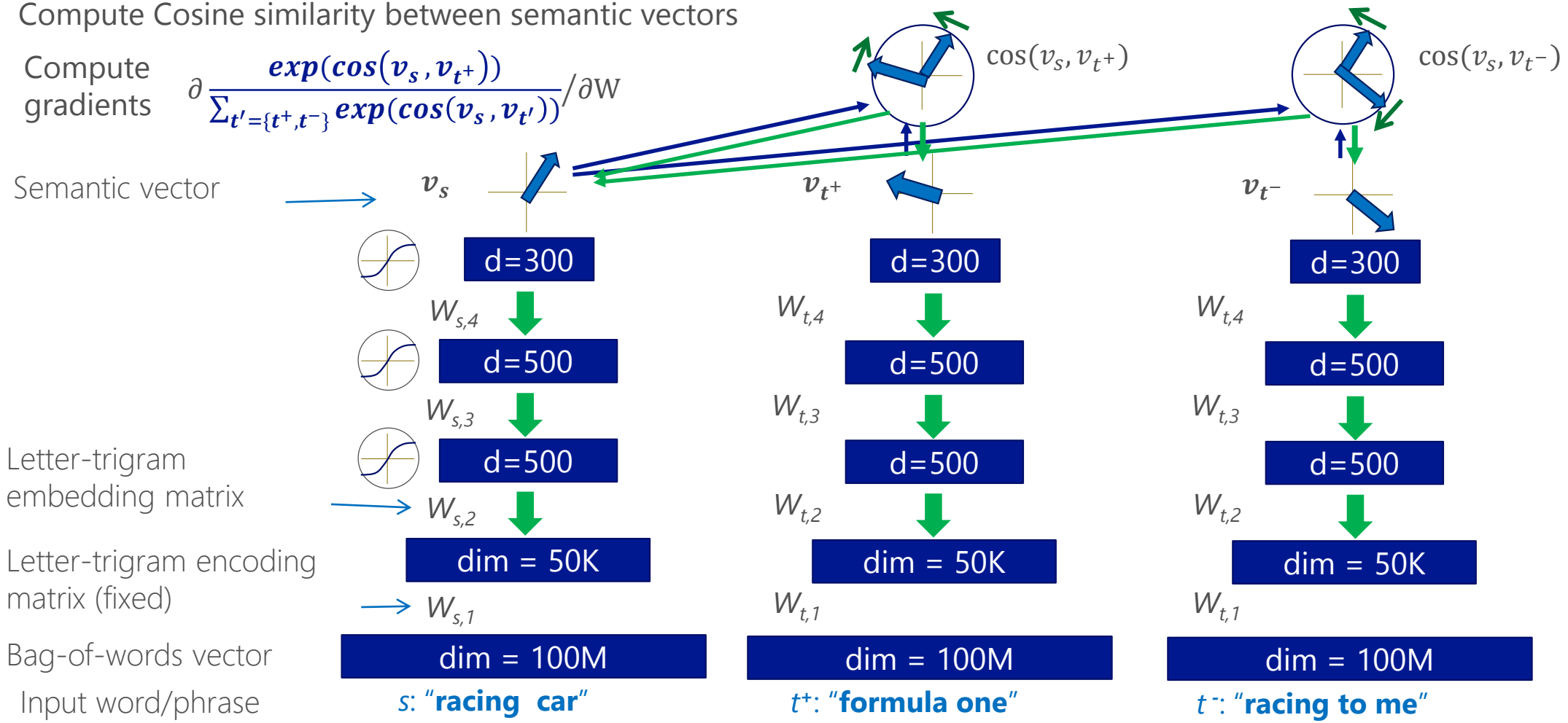


DSSM for semantic embedding learning

Training:

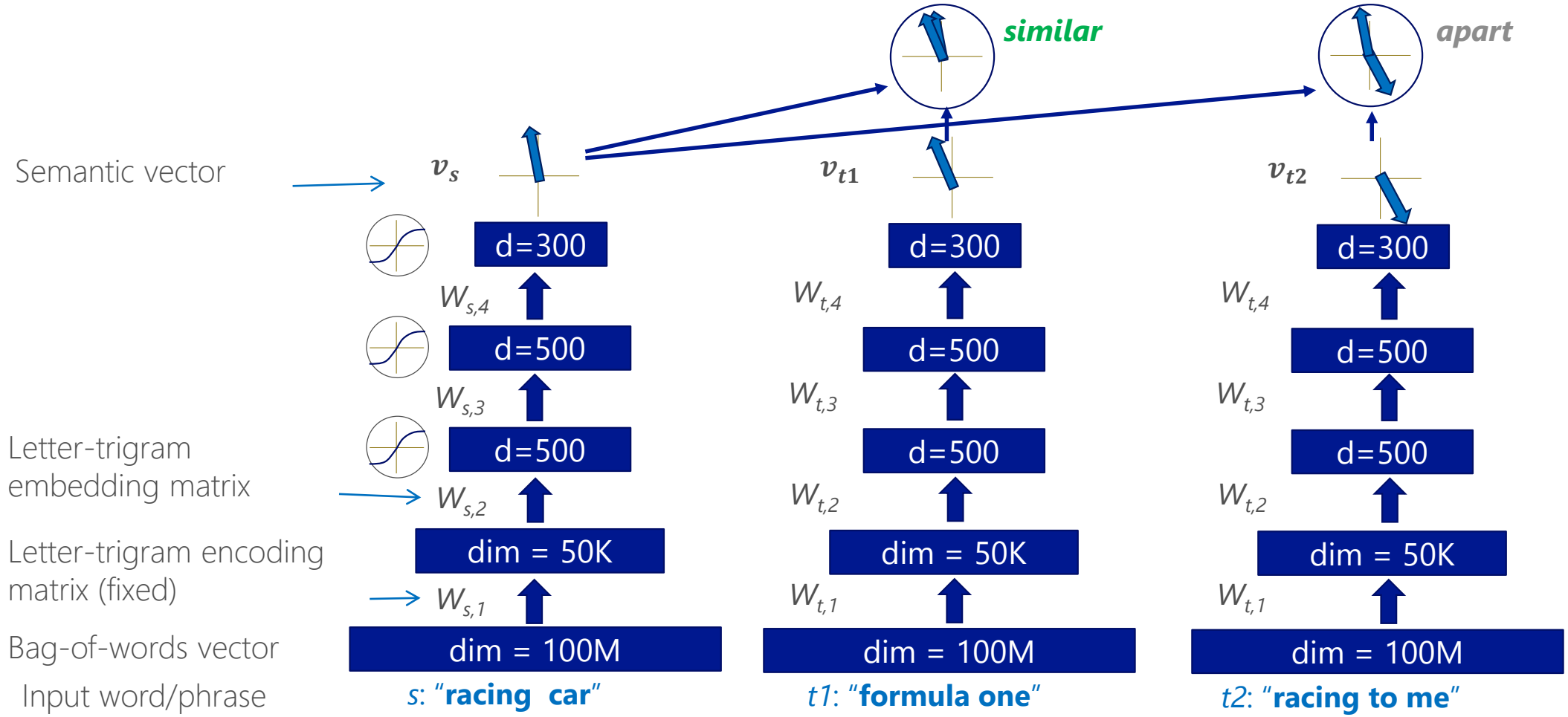
Compute Cosine similarity between semantic vectors

Compute gradients $\partial \frac{\boldsymbol{exp(cos(v_s, v_{t^+}))}}{\sum_{t'=\{t^+,t^-\}} \boldsymbol{exp(cos(v_s, v_{t'}))}} / \partial W$



DSSM for semantic embedding learning

Runtime:



Training of the DSSM

Data: semantically-similar text pairs

e.g., context <-> word in word embedding vector learning

query <-> clicked-doc in Web Search

pattern<-> relationship in Question Answering

Objective: cosine similarity based loss

- Web search as an example: a query $\boldsymbol{q}$ and a list of docs $\boldsymbol{D} = \{\boldsymbol{d}^+, \boldsymbol{d}_1^-, \ldots \boldsymbol{d}_K^-\}$
  - $\boldsymbol{d}^+$ positive doc; $\boldsymbol{d}_1^-, \ldots \boldsymbol{d}_K^-$ are negative docs to $\boldsymbol{q}$ (e.g., sampled from not clicked docs)

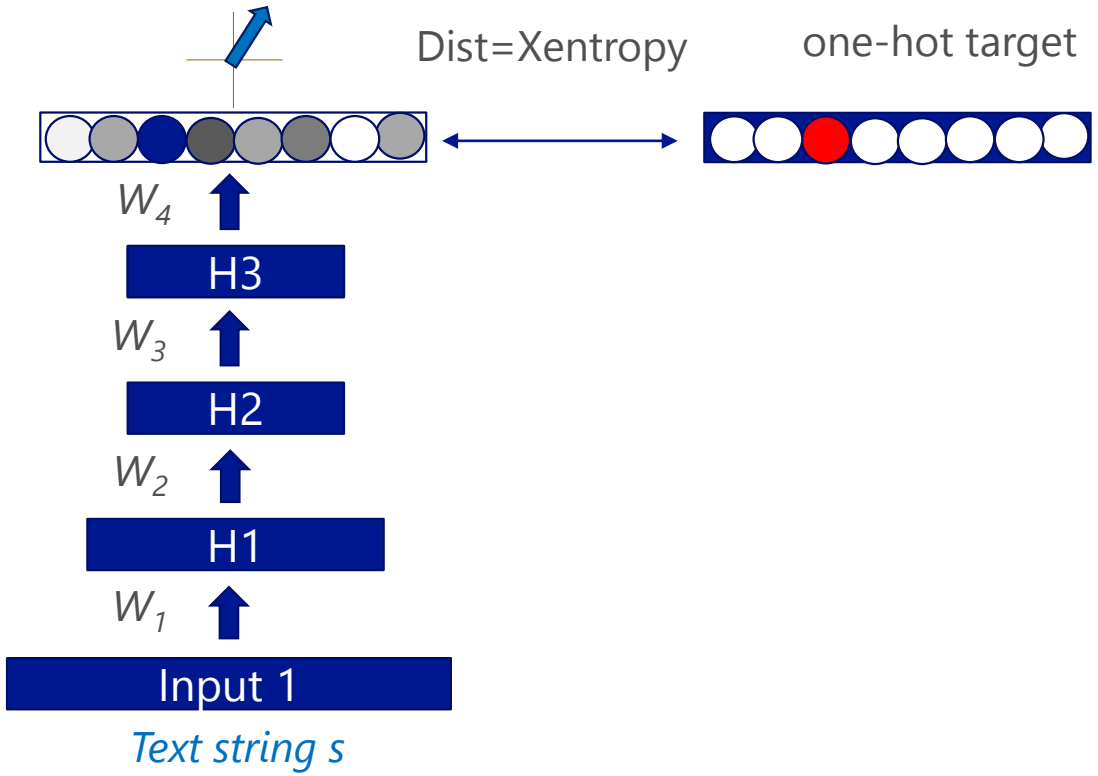
- Objective: the posterior probability of clicked document given query

$$P(d^+|q) = \frac{\exp\left(\gamma\, cos(q, d^+)\right)}{\sum_{d \in \boldsymbol{D}} \exp(\gamma\, cos(q, d))}$$

- Optimize $\boldsymbol{\theta}$ to maximize $\boldsymbol{P}(\boldsymbol{d}^+|\boldsymbol{q})$. SGD training on GPU (NVidia K20x)

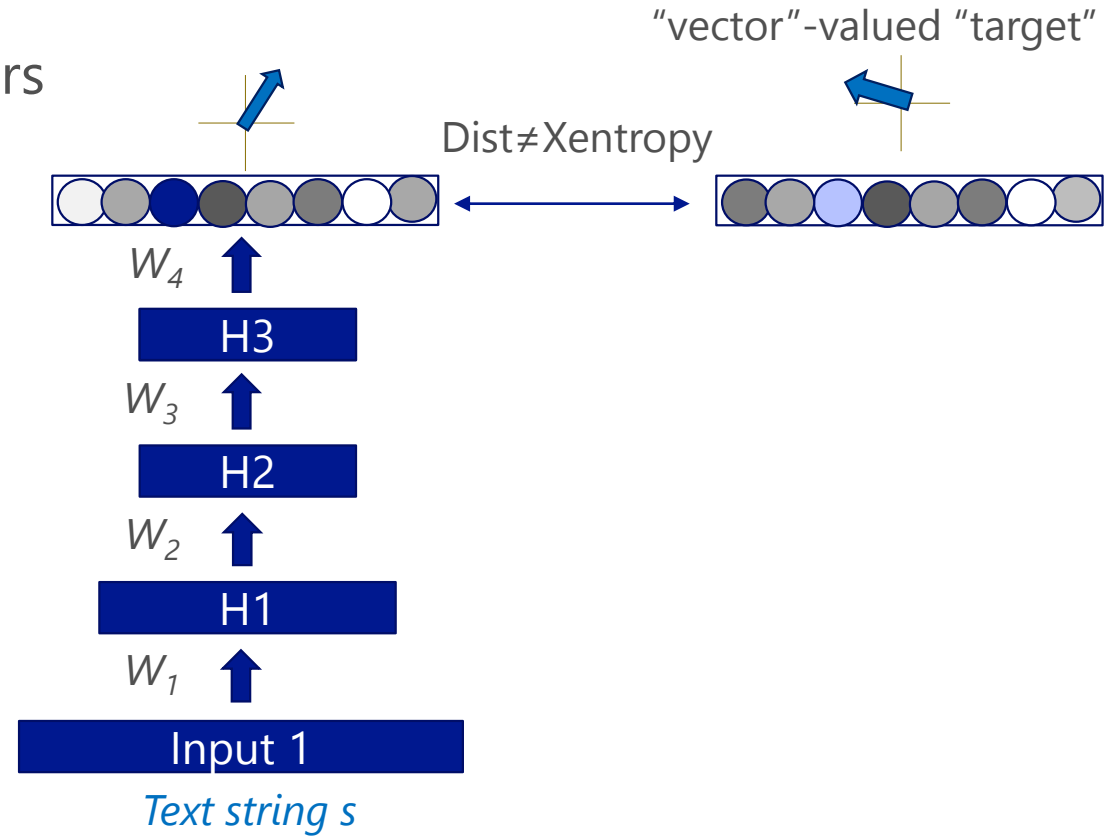
Reflection: from DNN to DSSM

- Common deep models reviewed so far:
  - Mainly for classification (speech, image, LM, SLU)
  - Target: one-hot vector
  - Example of DNN:



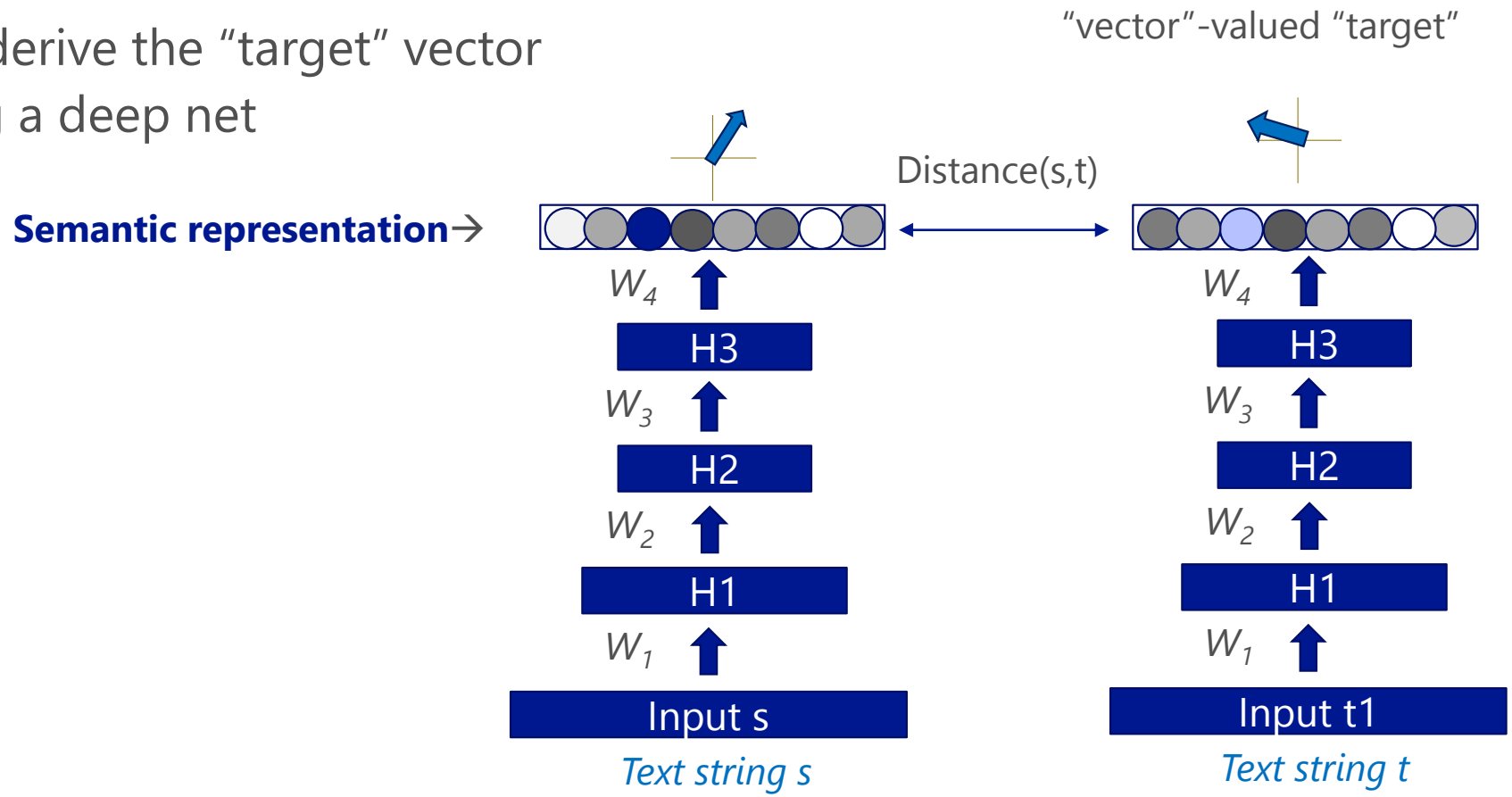
Reflection: from DNN to DSSM

- DSSM
  - Deep Structured Semantic Model or Deep Semantic Similarity Model
  - For semantic matching / ranking (not classification with DNN)
  - Step 1: target from "one-hot"
    to continuous-valued vectors



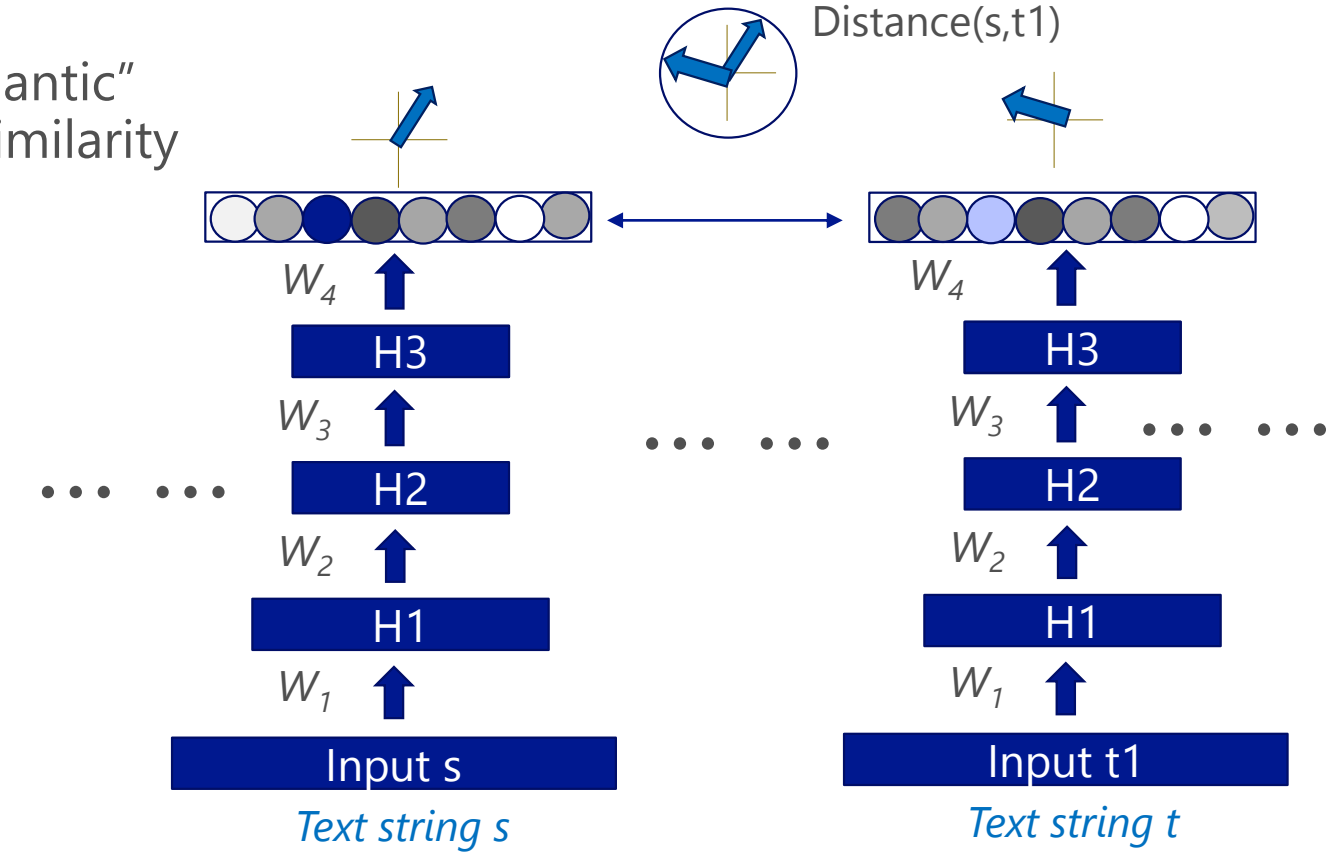
Reflection: from DNN to DSSM

- To construct a DSSM
  - Step 1: target from "one-hot" to continuous-valued vectors
  - Step 2: derive the "target" vector using a deep net



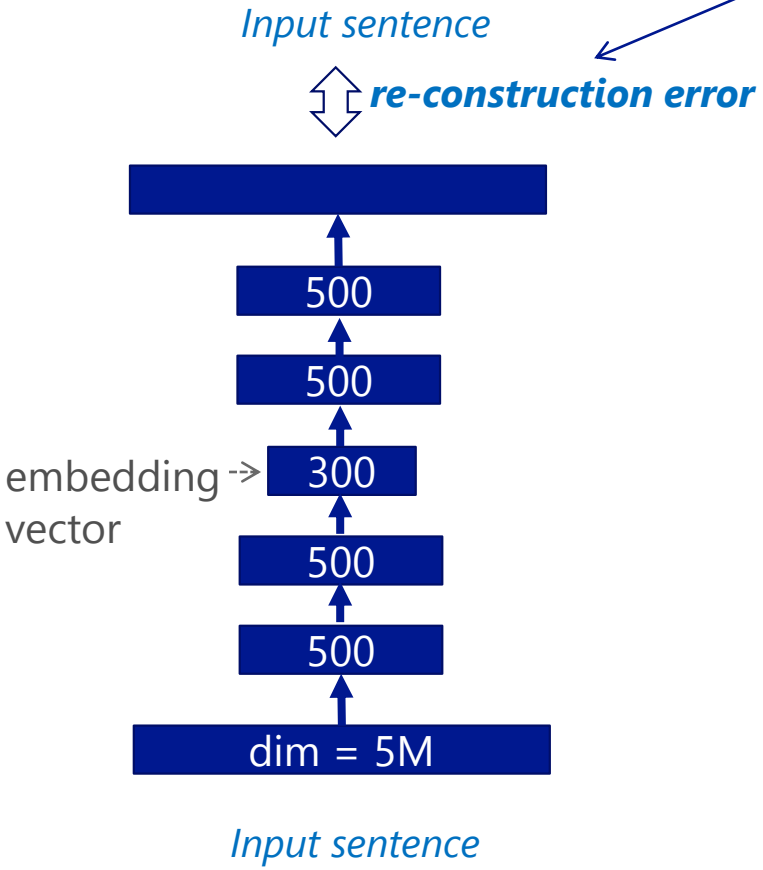
Reflection: from DNN to DSSM

- To construct a DSSM
  - Step 1: target from "one-hot" to a continuous-valued vector
  - Step 2: derive the "target" vector using a deep net
  - Step 3: normalize two "semantic" vectors & computer their similarity



Reflection: from Auto-encoder to DSSM

Auto-encoder

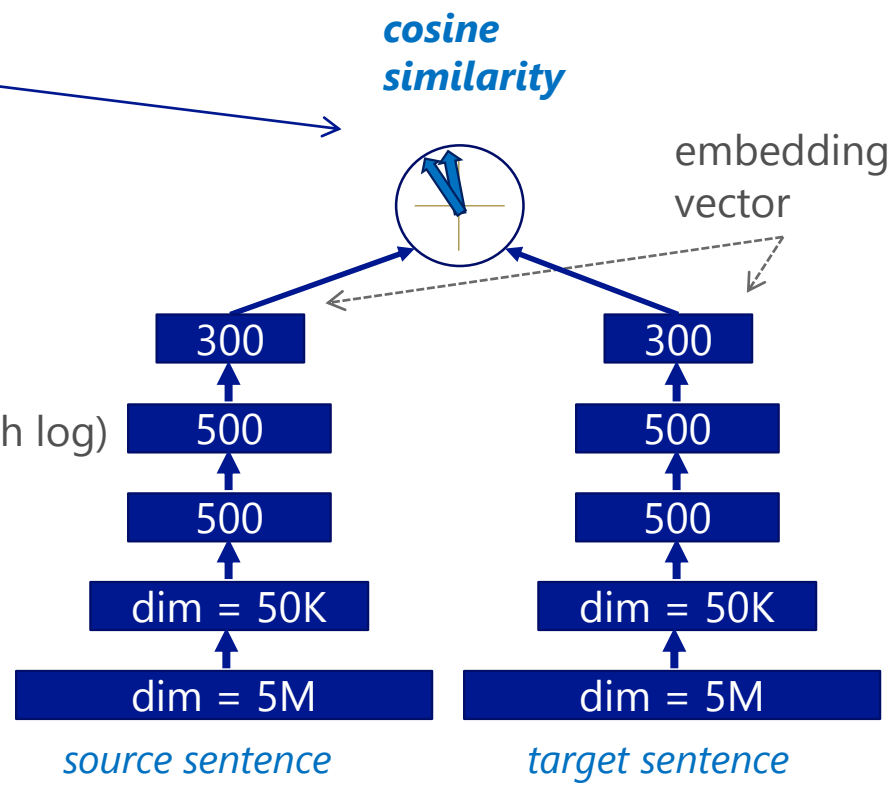


Training loss func.:
AE: reconstruction error of the input
DSSM: distance between embedding vectors

Supervision:
AE: unsupervised (e.g., doc<->doc)
DSSM: weakly supervised (e.g., query<->doc search log)

Input:
AE: 1-hot word vector
DSSM: sub-word unit (e.g., letter-trigram)

DSSM



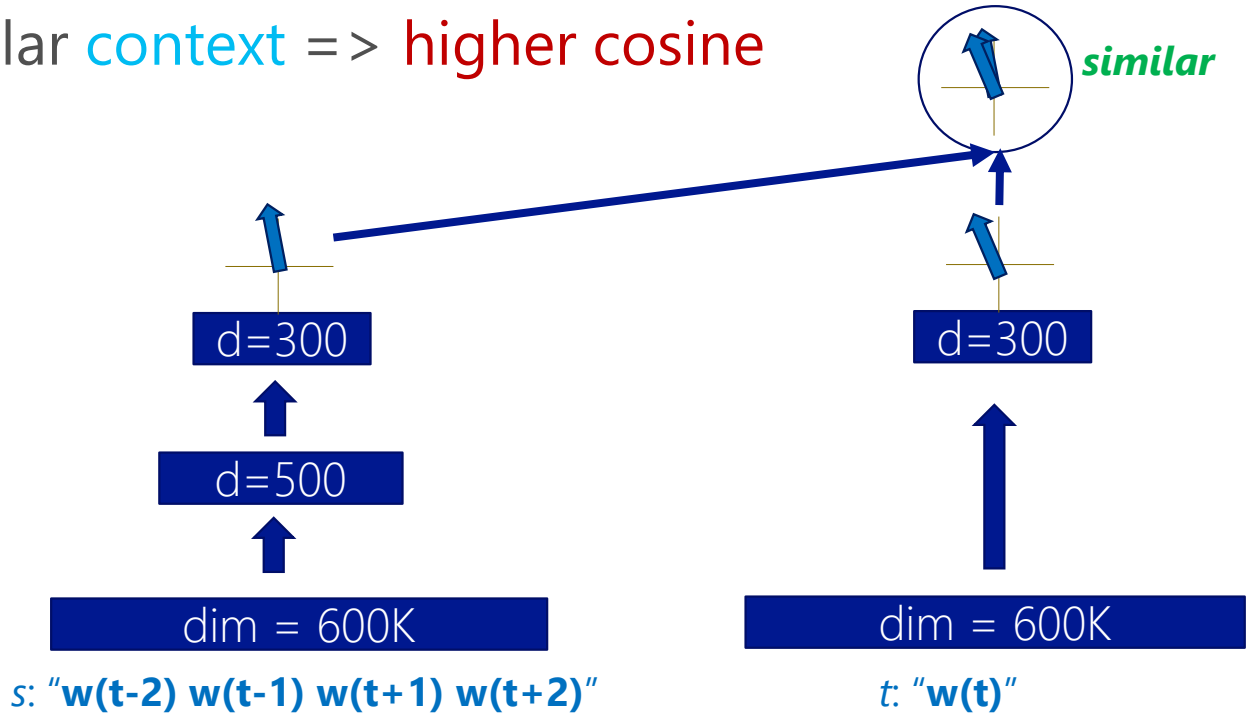
The DSSM can be trained using a variety of weak supervision signals without human labeling effort (e.g., user behavior log data).

Deep Structured Semantic Model (DSSM) in practice

Tasks	*Source*	*Target*
Word semantic embedding	***context***	***word***
Web search	***search query***	***web documents***
Question answering	***pattern / mention*** **(in NL)**	***relation / entity*** **(in KB)**
Recommendation	***doc in reading***	***interesting things / other docs***
Machine translation	***sentence in language a***	***translations in language b***
Text/Image joint learning	***text / image***	***Image / text***
Ad selection	*search query*	*ad keywords*
Entity ranking	*mention (highlighted)*	*entities*
Knowledge-base construction	*entity*	*entity*
…		

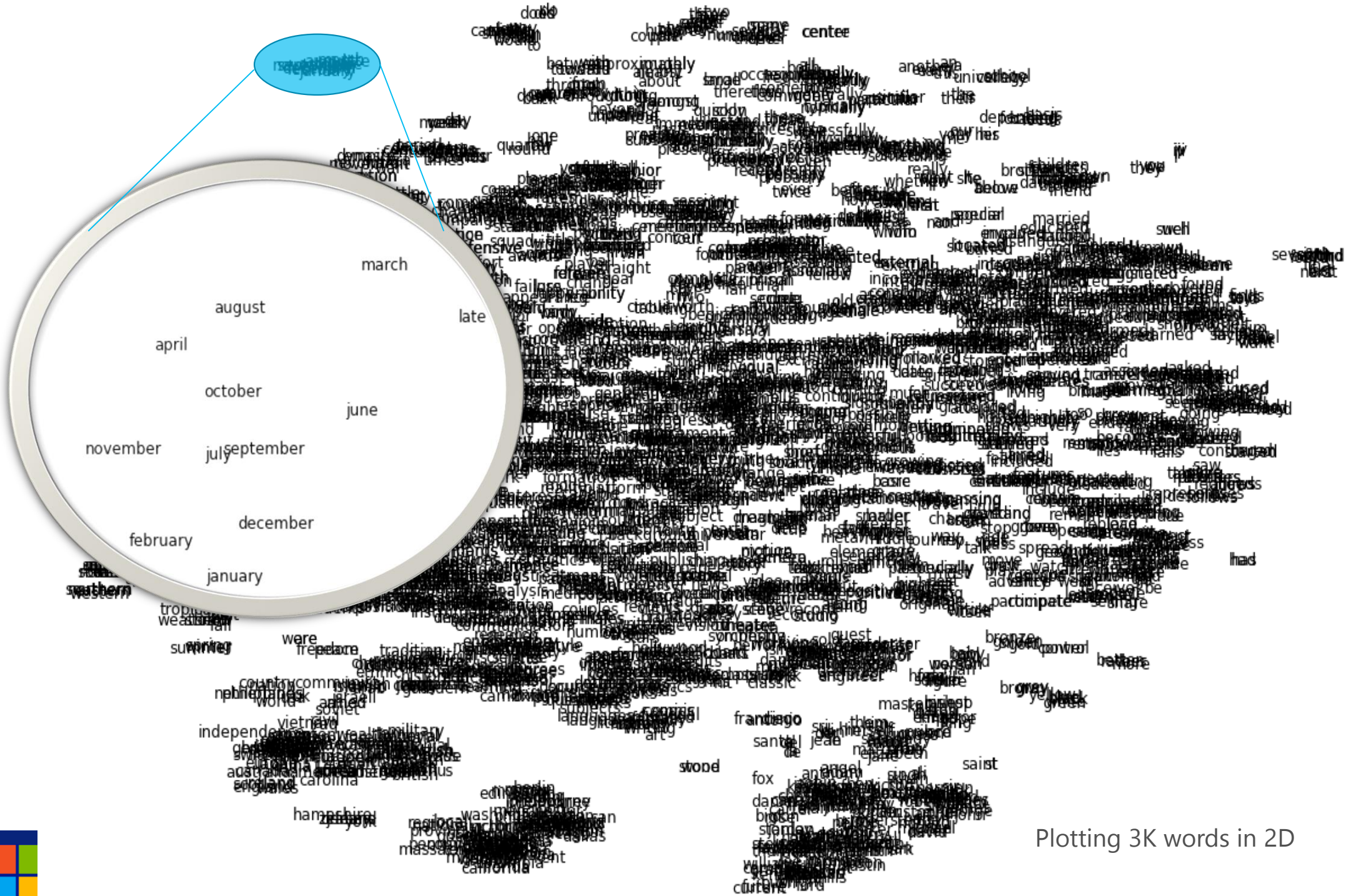
DSSM: learning words' meaning

- Learn a word's semantic meaning by means of its neighbors (context)
  - Construct context <-> word training pair for DSSM
  - Similar words with similar context => higher cosine
- Training Condition:
  - 600K vocabulary size
  - 1B words from Wikipedia
  - 300-dimentional vector

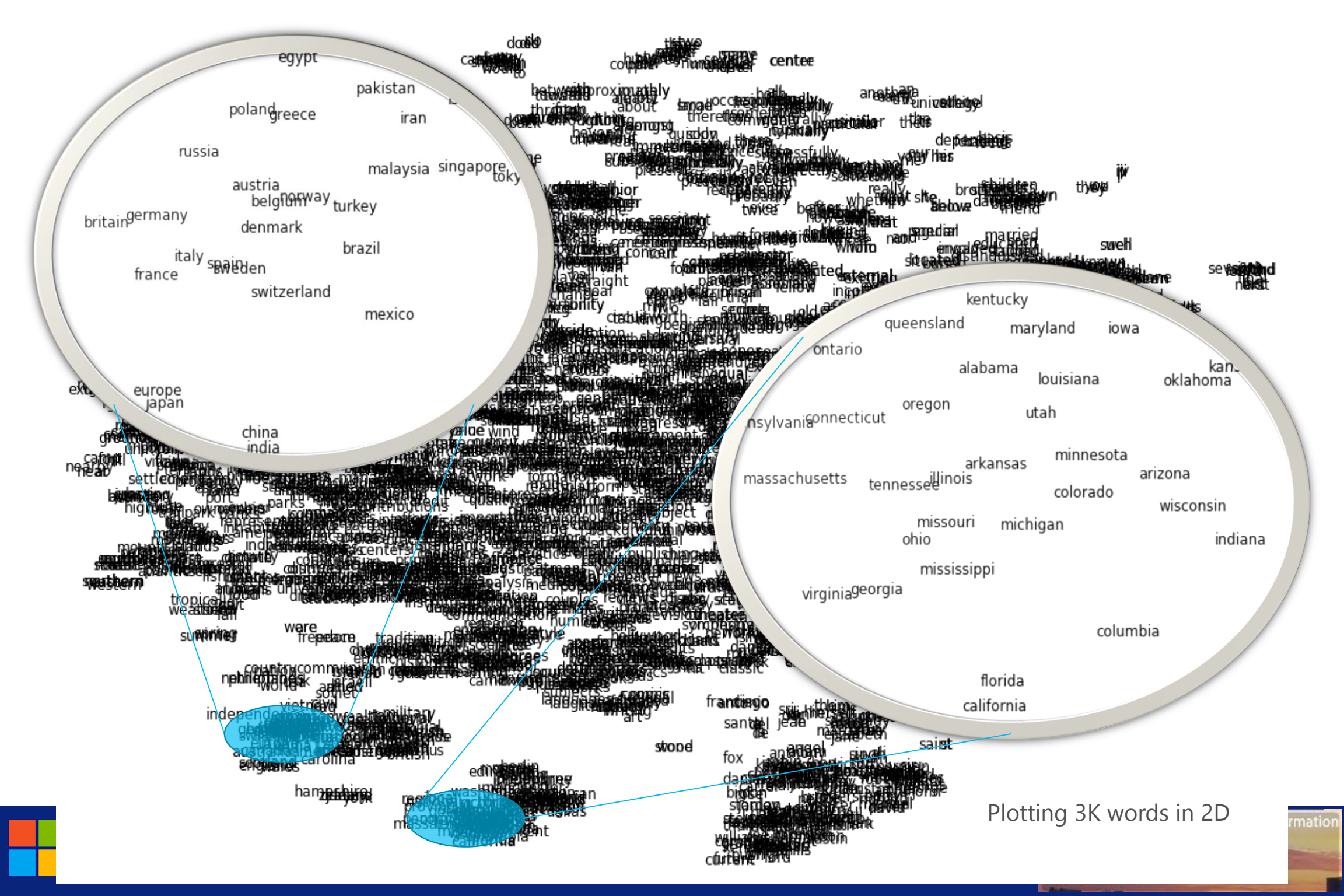


You shall know a word by the company it keeps
(J. R. Firth 1957: 11)

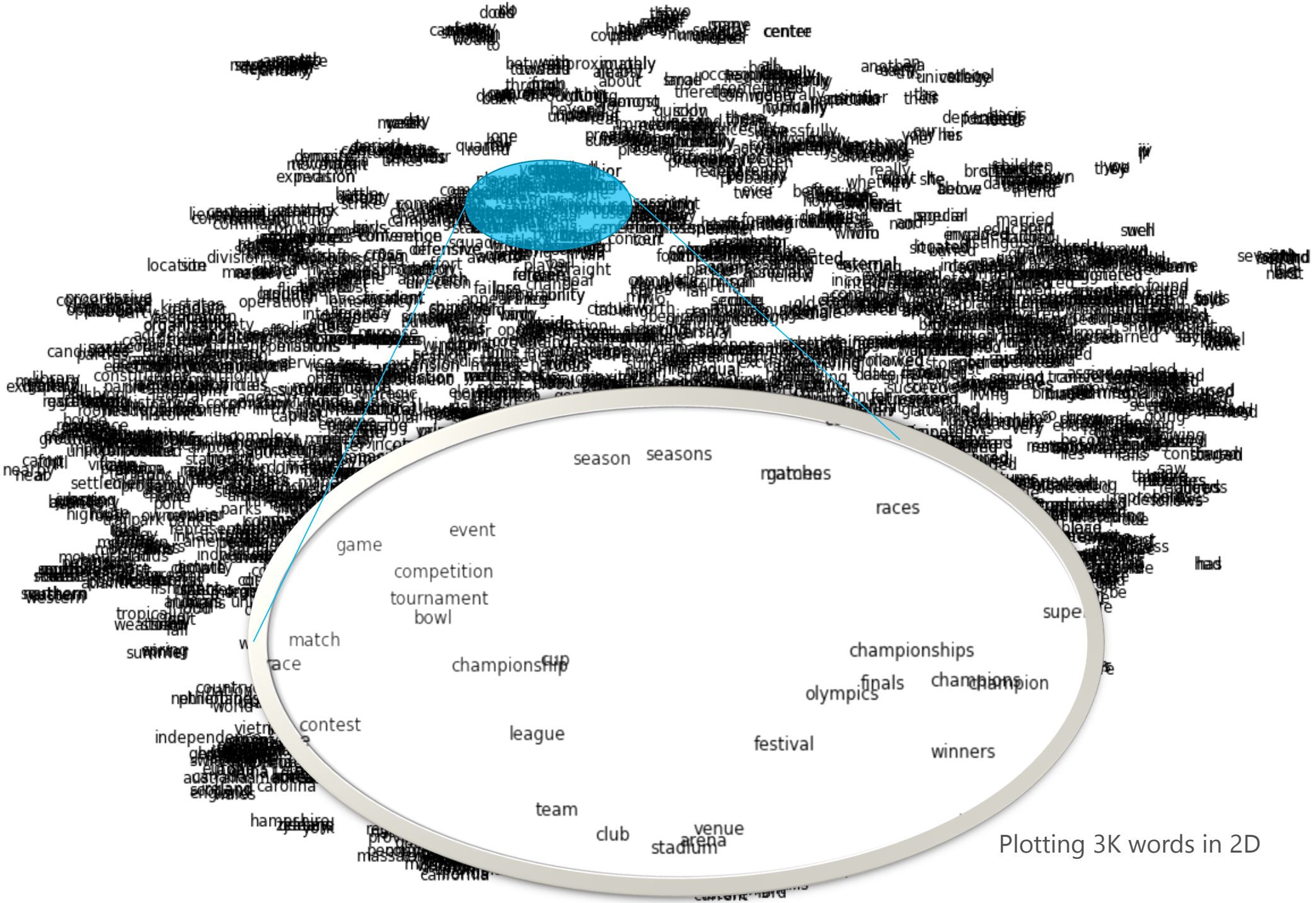
[Song, He, Gao, Deng, Shen, 2014]



Plotting 3K words in 2D



Plotting 3K words in 2D



Plotting 3K words in 2D

Semantic reasoning (as algebra in the semantic space)

Semantic clustering: top 3 neighbors of each word

king	earl (0.77)	pope (0.77)	lord (0.74)
woman	person (0.79)	girl (0.77)	man (0.76)
france	spain (0.94)	italy (0.93)	belgium (0.88)
rome	constantinople (0.81)	paris (0.79)	moscow (0.77)
winter	summer (0.83)	autumn (0.79)	spring (0.74)
rain	rainfall (0.76)	storm (0.73)	wet (0.72)
car	truck (0.8)	driver (0.73)	motorcycle (0.72)

Semantic analogy:

$w_1 : w_2 = w_3 : ? \quad \Rightarrow \quad V_4 = V_3 - V_1 + V_2$ -- retrieve words close to V_4

summer : rain = winter : ?	snow (0.79)	rainfall (0.73)	wet (0.71)
italy : rome = france : ?	paris (0.78)	constantinople (0.74)	egypt (0.73)
man : eye = car : ?	motor (0.64)	brake (0.58)	overhead (0.58)
read : book = listen : ?	sequel (0.65)	tale (0.63)	song (0.60)

Evaluation on the word analogy task

The dataset contains 19,544 word analogy questions:

- Semantic questions, e.g.,: "Athens is to Greece as Berlin is to ?"
- Syntactic questions, e.g.,: "dance is to dancing as fly is to ?"

Model	Dim	Size	Accuracy Avg.(sem+syn)
SG	300	1B	61.0%
CBOW	300	1.6B	36.1%
vLBL	300	1.5B	60.0%
ivLBL	300	1.5B	64.0%
GloVe	300	1.6B	70.3%
DSSM	**300**	**1B**	**71.4%**

(i)vLBL results are from (Mnih et al., 2013); skip-gram (SG) and CBOW results are from (Mikolov et al., 2013a,b); GloVe are from (Pennington, Socher, and Manning, 2014)

DSSM for Web Search

- Training data:
  - 80M query/clicked-doc-title pairs from search log
- Test set:
  - 12,071 English queries sampled from 1-yr. log
  - 5-level relevance label for each query-doc pair
  - Evaluated by NDCG
- Baselines
  - Lexicon matching models: BM25
  - Topic model: PLSA

Results

- Evaluated on a document retrieval task
  - Docs are ranked by the cosine similarity between embedding vectors of the query and the docs

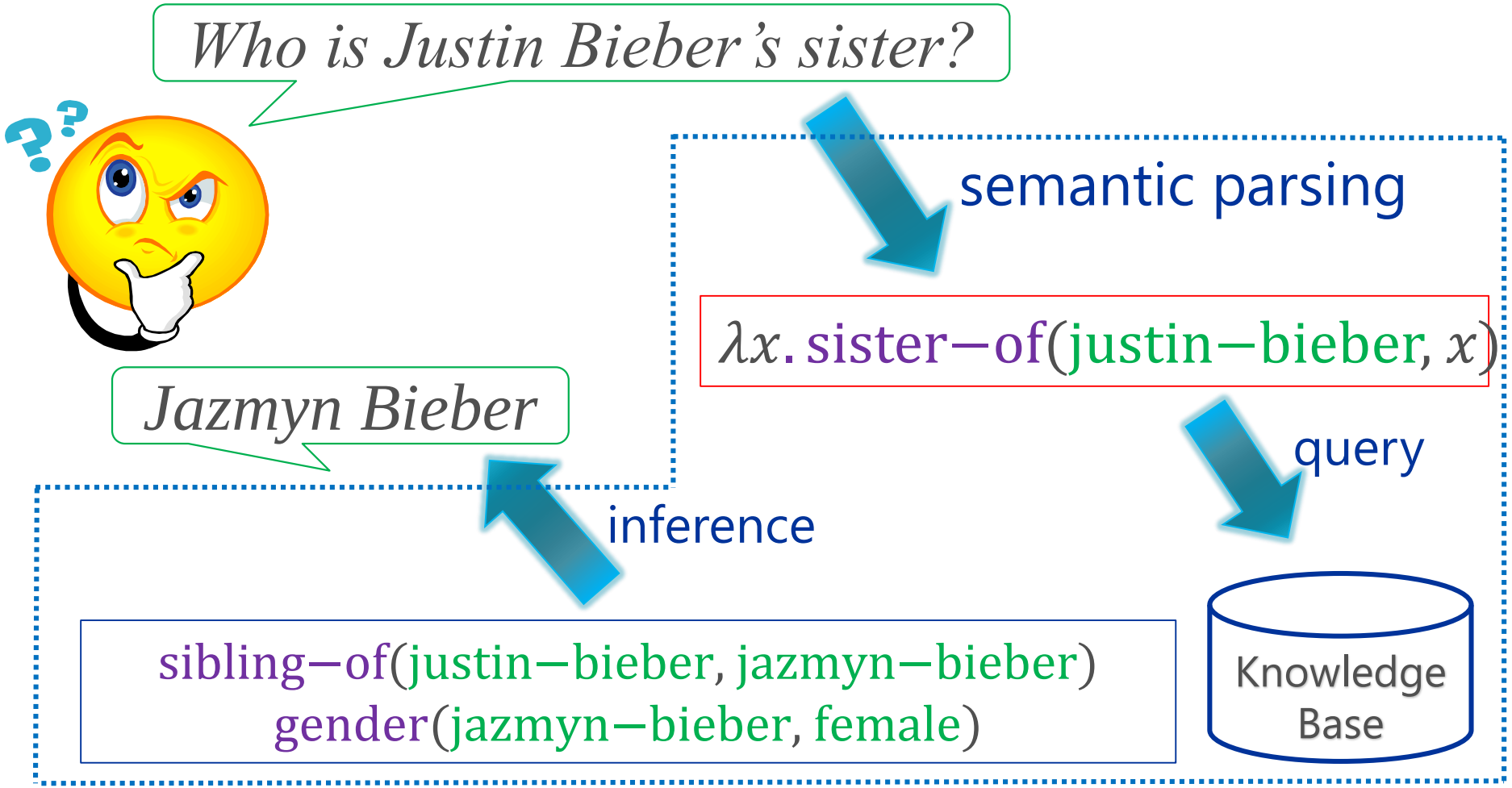
Model	Input dimension	NDCG@1 %
BM25 baseline	--	30.8
Probabilistic LSA (PLSA)		29.5
Auto-Encoder (Word)	40K	31.0 (+0.2)
DSSM (Word)	40K	34.2 (+3.4)
DSSM (Random projection)	30K	35.1 (+4.3)
DSSM (Letter-trigram)	30K	36.2 (+5.4)

The DSSM improves 5~7 pt NDCG over shallow models

The higher the NDCG score the better, 1% NDCG difference is statistically significant.

- The DSSM learns superior semantic embedding
- Letter-trigram + the DSSM gives superior results

Question Answering

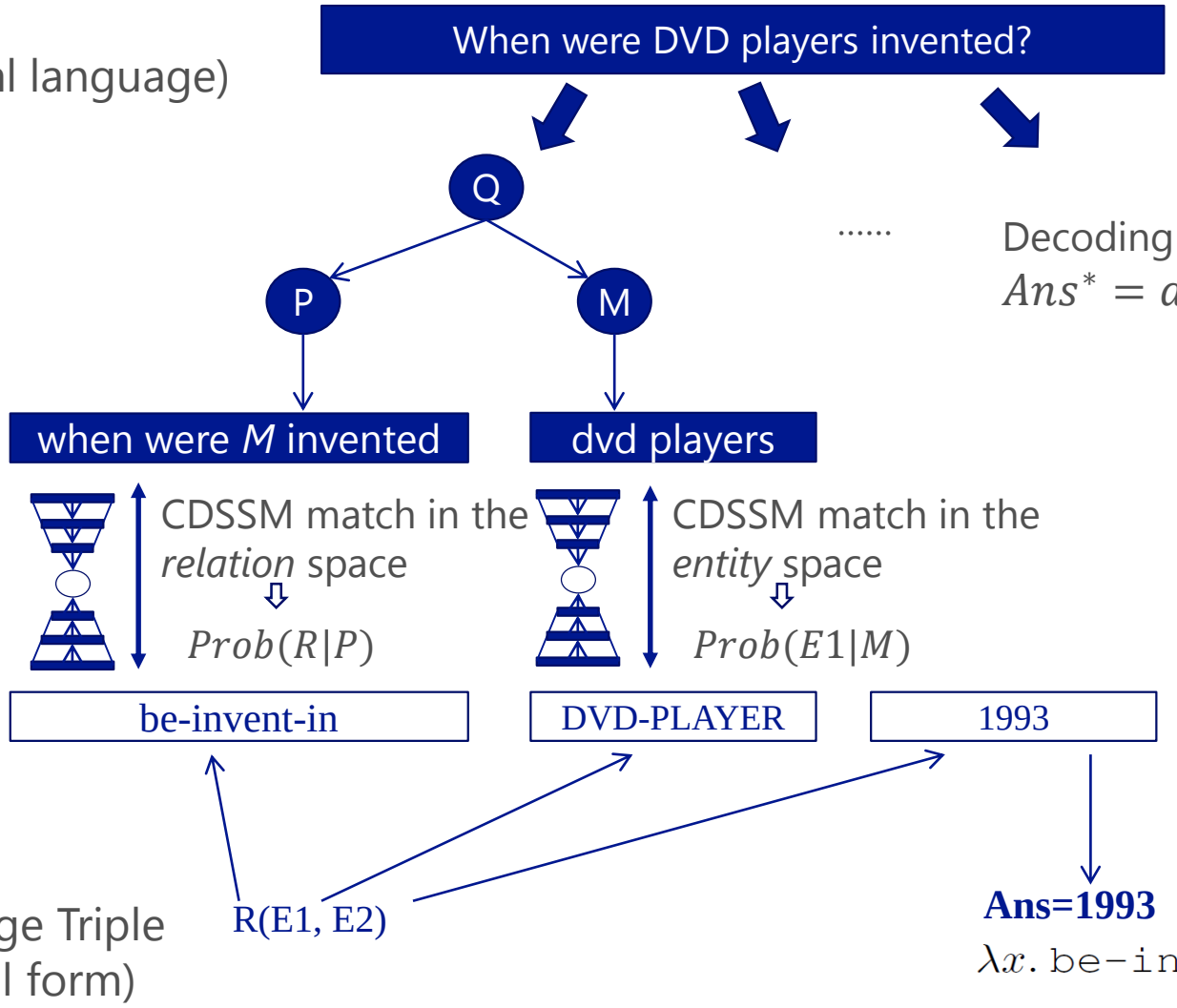


Challenge

- Lots of ways to ask the same question
  - *"What was the date that Minnesota became a state?"*
  - *"Minnesota became a state on?"*
  - *"When was the state Minnesota created?"*
  - *"Minnesota's date it entered the union?"*
  - *"When was Minnesota established as a state?"*
  - *"What day did Minnesota officially become a state?"*
  - *…*

DSSM in question answering

Question
(in natural language)



Knowledge Triple
(in logical form)

Decoding the best answer:

$$Ans^* = argmax_{Ans} P(Ans|KB, Q)$$

$$P(Ans|KB, Q) = \sum_{SP} P(Ans, SP|KB, Q)$$
$$\approx \max_{SP,Triple} P(Ans|SP, KB, Q)P(SP|Q)$$
$$\approx \max_{SP,Triple} Prob(R|P) \times Prob(E1|M)$$

Yih, He, Meek, "Semantic parsing for single-relation question answering," ACL 2014

Experiments: Data

Paralex dataset [Fader et al., 2013]

- 1.8M (question, single-relation queries)

  When were DVD players invented?
  $\lambda x.\,\text{be}-\text{invent}-\text{in}(\text{dvd}-\text{player}, x)$

- 1.2M (relation pattern, relation)

  When were X invented?
  $\text{be}-\text{invent}-\text{in}_2$

- 160k (mention, entity)

  Saint Patrick day
  $\text{st}-\text{patrick}-\text{day}$

Experiments: Task – Question Answering

- Same test questions in the Paralex dataset
- 698 questions from 37 clusters

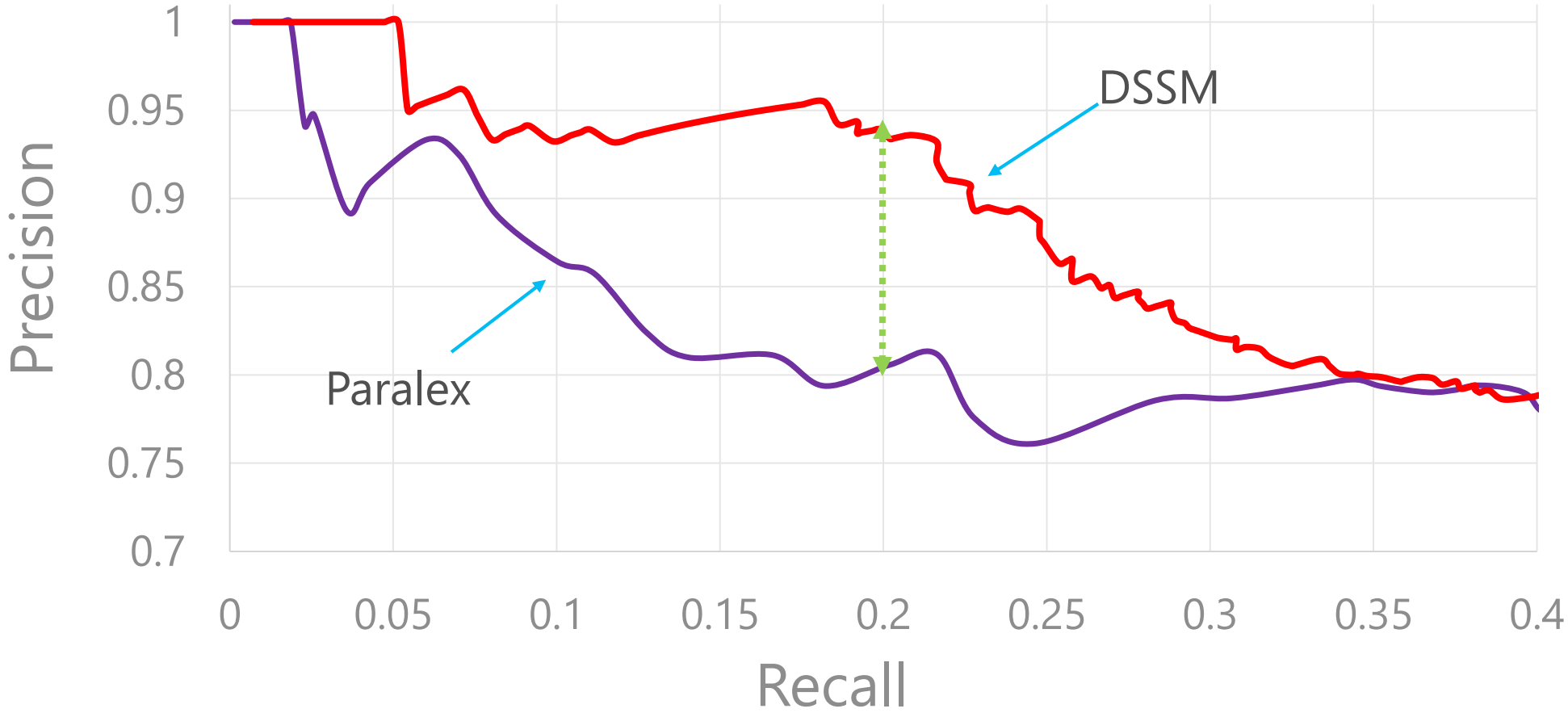
- *What language do people in Hong Kong use?*

  be–speak–in(english, hong–kong)
  be–predominant–language–in
  (cantonese, hong–kong)

- *Where do you find Mt Ararat?*

  be–highest–mountain–in(ararat, turkey)
  be–mountain–in(ararat, armenia)

Experiments: Results



Other relevant work on deep learning in NLP

Long short-term memory RNN (LSTM-RNN)

- Capable to capture long-span dependency in natural language
- LSTM-DSSM for IR (Palangi, et al., "Learning sequential semantic representations," to appear)
- LSTM for MT (Sutskever, et al., "Sequence to sequence learning with neural networks," to appear)

Recursive NN (ReNN)

- Model the hierarchical structure of nature language
- ReNN for parsing (Socher et al., "Parsing natural scenes and natural language with recursive neural networks", 2011)

Tensor product representation (TPR)

- Efficient representation of the structure of natural language
- Smolensky & Legendre: The Harmonic Mind, From Neural Computation to Optimality-Theoretic Grammar, MIT Press, 2006

Interim summary

Exciting advances in learning continuous semantic space

- deep models effectively learn semantic representation vectors
- leads to superior performance in a range of NL tasks
- extend to cross-modality learning – in the next!

Part IV

Deep Semantic Similarity Model For Text Processing

Mission of Machine (Deep) Learning

"Real" world	Data (collected/labeled)
"Artificial" world	Model (architecture)
Link the two worlds	Training (algorithm)

Deep Semantic Similarity Model (DSSM) for Text Processing

- What is DSSM?
- DSSM for web search ranking
- DSSM for recommendation
- DSSM for phrase translation modeling
- DSSM for automatic image captioning

Computing Semantic Similarity

- Fundamental to almost all text processing tasks, e.g.,
  - Machine translation: similarity between sentences in different languages
  - Web search: similarity between queries and documents

- Problems of the existing approaches
  - Lexical matching cannot handle language discrepancy.
  - Unsupervised word embedding or topic models are not optimal for the task of interest.

Deep Semantic Similarity Model (DSSM)

[Huang et al. 2013; Gao et al. 2014a; Gao et al. 2014b; Shen et al. 2014]

- Compute semantic similarity between X and Y
  - Map X and Y to feature vectors in a latent semantic space via deep neural net
  - Compute the cosine similarity between the feature vectors
- DSSM for text processing tasks

Tasks	X	Y
Web search	*Search query*	*Web document*
Automatic highlighting	*Doc in reading*	*Key phrases to be highlighted*
Contextual entity search	*Key phrase and context*	*Entity and its corresponding page*
Machine translation	*Sentence in language A*	*Translations in language B*
Automatic Image captioning	*Image*	*caption*

DSSM for Web Search Ranking

- Task
- Model architecture
- Model training
- Evaluation

[Huang, He, Gao, Deng, Acero, Heck. 2013; Shen, He, Gao, Deng, Mesnil, 2014]

An Example of Web Search

Best Home Remedies for Cold and Flu

Wind Heat External Pathogens

By: Catherine Browne, L.Ac., MH, Dipl. Ac.

In Chinese medicine, colds and flu's are delineated into several different energetic classifications. Here we will outline the different types of cold and flu viruses that you will likely encounter, and then describe the best home remedies for these specific patterns that you can use to treat the cold or influenza virus.

Cold and Flu Basics

The basic pathogenic influences are:

- **Wind**
- **Cold**
- **Heat**
- **Damp**

Wind

Theoretically, wind enters the body through the back of the neck area or nose carrying the pathogen. It first attacks the Lung system (including the sinuses) because the Lung organ system is the most external Yin organ, a thus the most vulnerable to an external invasion. External Wind invasion is marked by acute conditions with a sudden onset of symptoms.

- cold home remedy
- cold remeedy
- flu treatment
- how to deal with stuffy nose

Semantic Matching between Q and D

R&D progress

- Fuzzy keyword matching
  - Q: cold home remedy
  - D: best home remedies for cold and flu

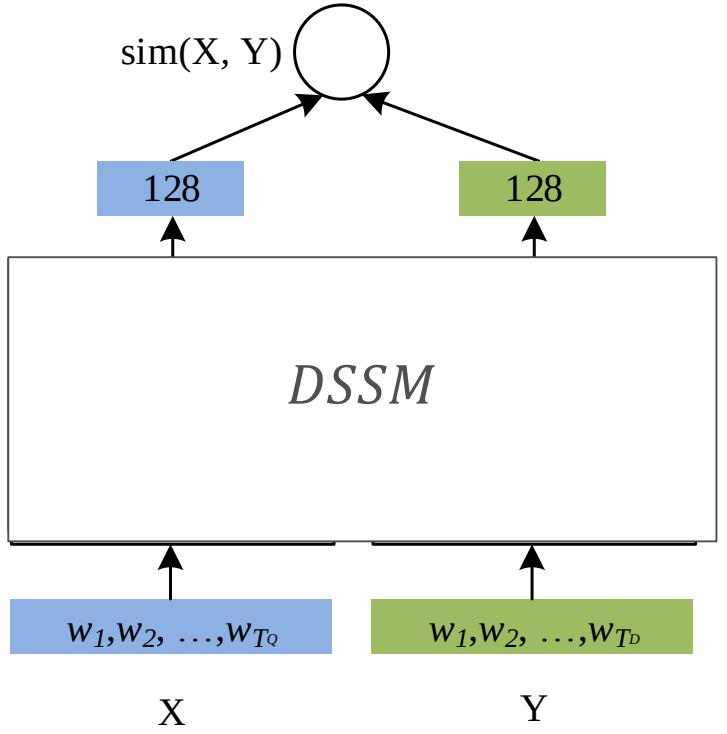
- Spelling correction
  - Q: cold remeedies
  - D: best home remedies for cold and flu

- Query alteration/expansion
  - Q: flu treatment
  - D: best home remedies for cold and flu

- Query/document semantic matching
  - Q: how to deal with stuffy nose
  - D: best home remedies for cold and flu

DSSM: Compute Similarity in Semantic Space

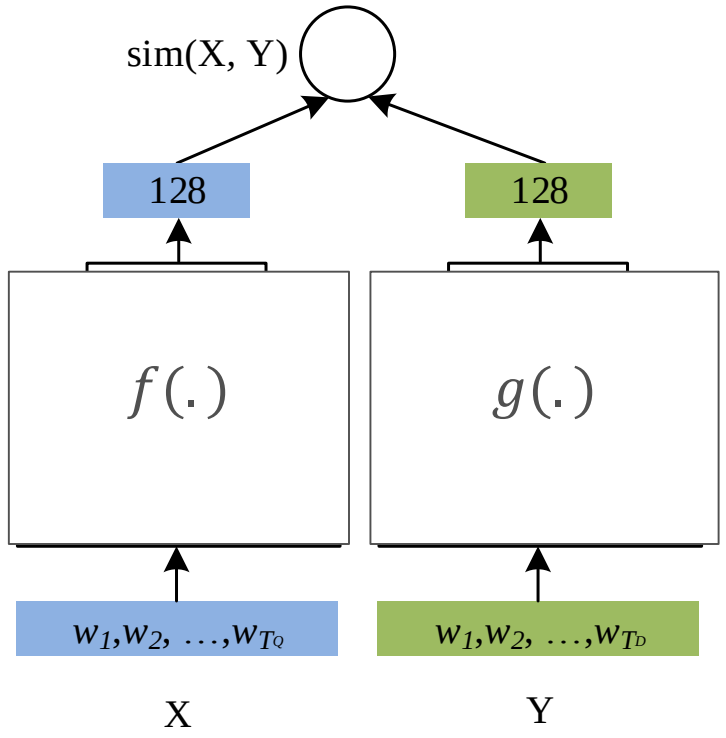
Relevance measured by cosine similarity



Learning: maximize the similarity between X (source) and Y (target)

DSSM: Compute Similarity in Semantic Space

Relevance measured by cosine similarity



Word sequence x_t

Learning: maximize the similarity between X (source) and Y (target)

Representation: use DNN to extract abstract semantic representations

DSSM: Compute Similarity in Semantic Space

Relevance measured by cosine similarity

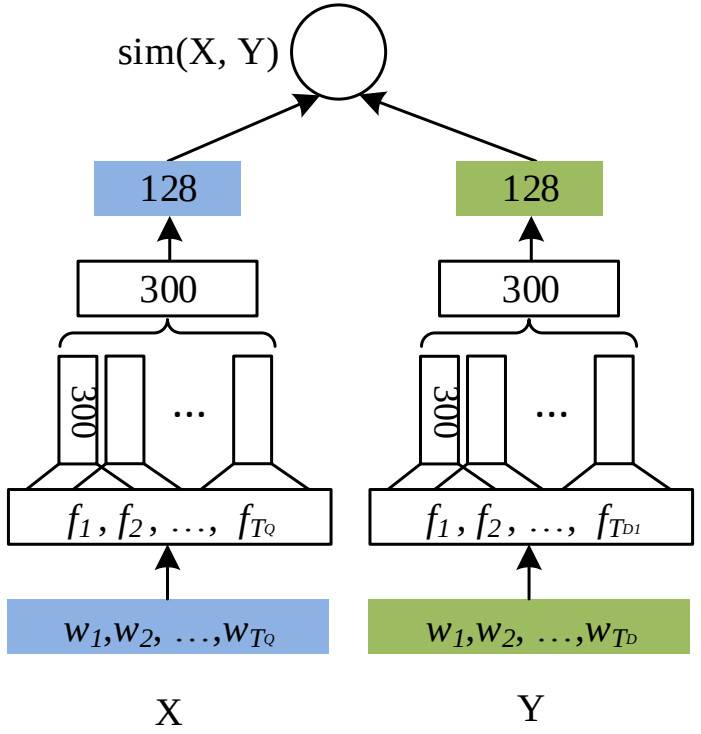
Semantic layer h

Max pooling layer v

Convolutional layer c_t

Word hashing layer f_t

Word sequence x_t



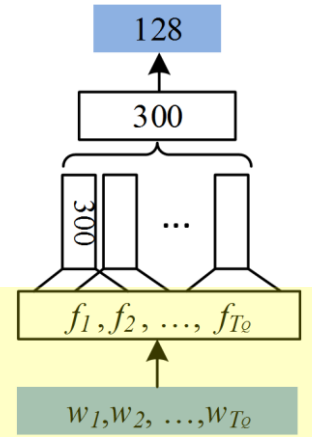
Learning: maximize the similarity between X (source) and Y (target)

Representation: use DNN to extract abstract semantic representations

Convolutional and Max-pooling layer: identify key words/concepts in X and Y

Word hashing: use sub-word unit (e.g., letter n-gram) as raw input to handle very large vocabulary

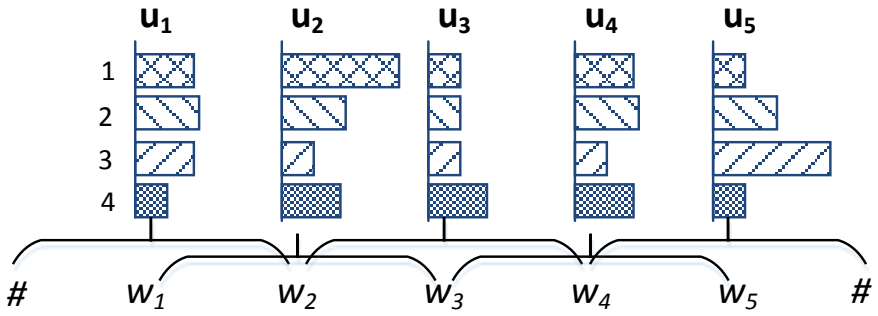
Letter-trigram Representation

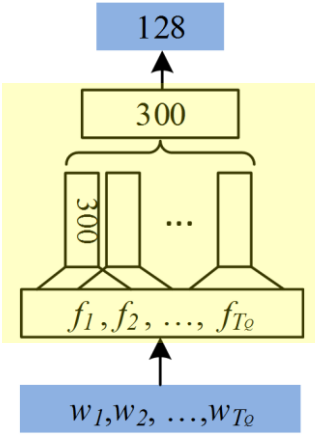


- Control the dimensionality of the input space
  - e.g., cat → #cat# → #-c-a, c-a-t, a-t-#
  - Only ~50K letter-trigrams in English; no OOV issue
- Capture sub-word semantics (e.g., prefix & suffix)
- Words with small typos have similar raw representations
- Collision: different words with same letter-trigram representation?

Vocabulary size	# of unique letter-trigrams	# of Collisions	Collision rate
40K	10,306	2	0.0050%
500K	30,621	22	0.0044%
5M	49,292	179	0.0036%

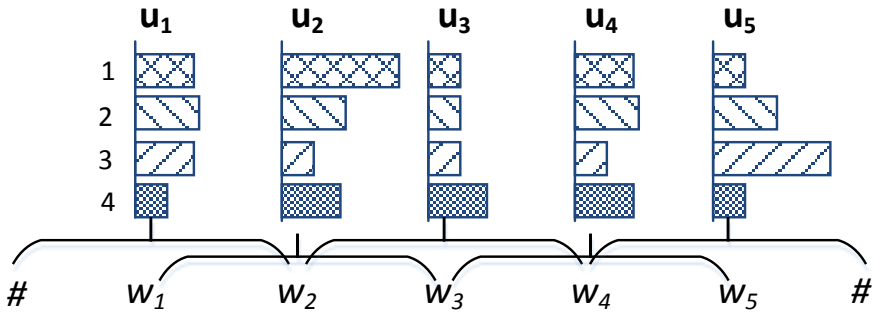
Convolutional Layer

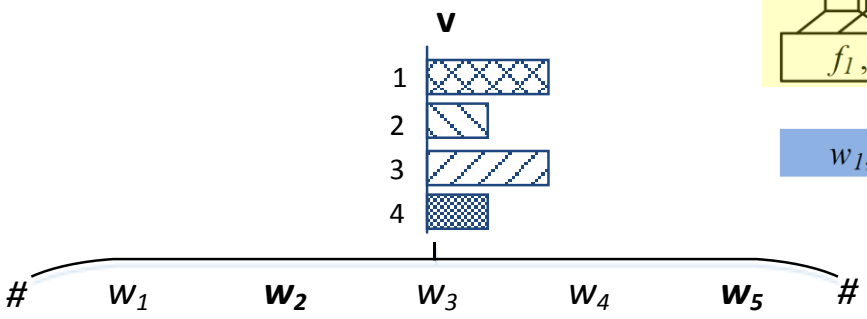


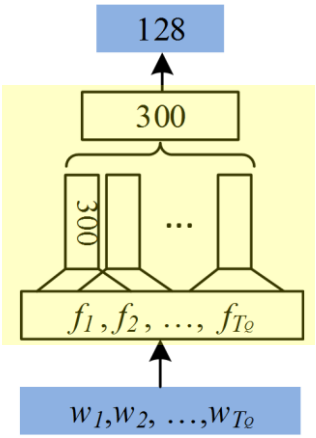


- Extract local features using convolutional layer
  - {w1, w2, w3} → topic 1
  - {w2, w3, w4} → topic 4

Max-pooling Layer

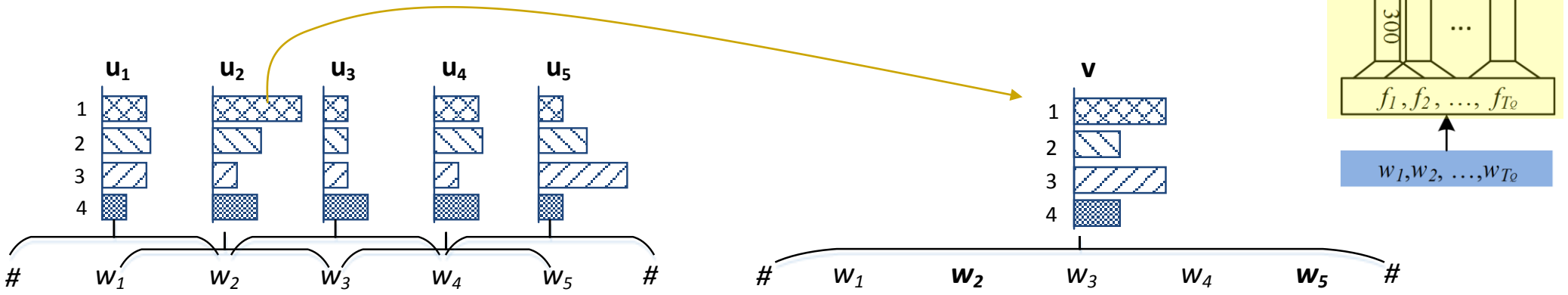






- Extract local features using convolutional layer
  - {w1, w2, w3} → topic 1
  - {w2, w3, w4} → topic 4
- Generate global features using max-pooling
  - Key topics of the text → topics 1 and 3
  - keywords of the text: w2 and w5

Max-pooling Layer

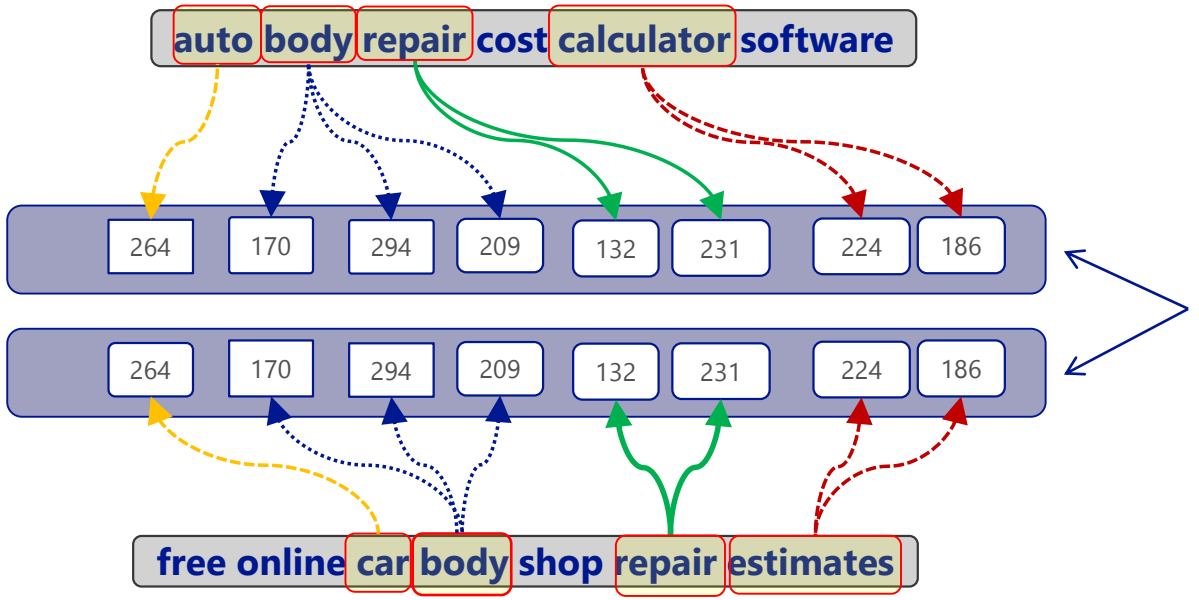


- Extract local features using convolutional layer
  - {w1, w2, w3} → topic 1
  - {w2, w3, w4} → topic 4
- Generate global features using max-pooling
  - Key topics of the text → topics 1 and 3
  - keywords of the text: w2 and w5

… the **comedy festival** formerly known as the us **comedy arts** festival is a comedy festival held each year in **las vegas nevada** from its 1985 inception to 2008 . it was held annually at the **wheeler opera house** and other venues in **aspen colorado** . the primary sponsor of the festival was hbo with co-sponsorship by caesars palace . the primary venue tbs **geico insurance** twix candy bars and **smirnoff vodka** **hbo** exited the festival business in 2007 …

Intent Matching via Convolutional-Pooling

- Semantic matching of query and document



More Examples

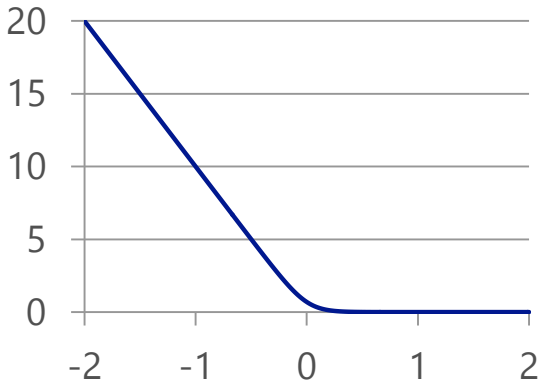
Query	Title of the top-1 returned document retrieved by CLSM
warm environment arterioles do what	**thermoregulation** wikipedia the free encyclopedia
auto body repair cost **calculator** software	free online **car body shop** repair **estimates**
what happens if our body **absorbs excessive** amount **vitamin d**	**calcium** supplements and **vitamin d** discussion stop **sarcoidosis**
how do **camera** use **ultrasound focus** automatically	**wikianswers how** does a **camera focus**
how to **change font excel office 2013**	**change font** default styles in **excel 2013**
where do i get my **federal tax return transcript**	how to get **trasncripts** of **federal** income **tax returns fast** ehow
12 **fishing boats trailers**	**trailer** kits and accessories **motorcycle** utility **boat snowmobile**
acp **ariakon combat pistol** 2.0	paintball acp **combat** pistol paintball **gun paintball pistol** package deal marker and gun

Learning DSSM from Labeled X-Y Pairs

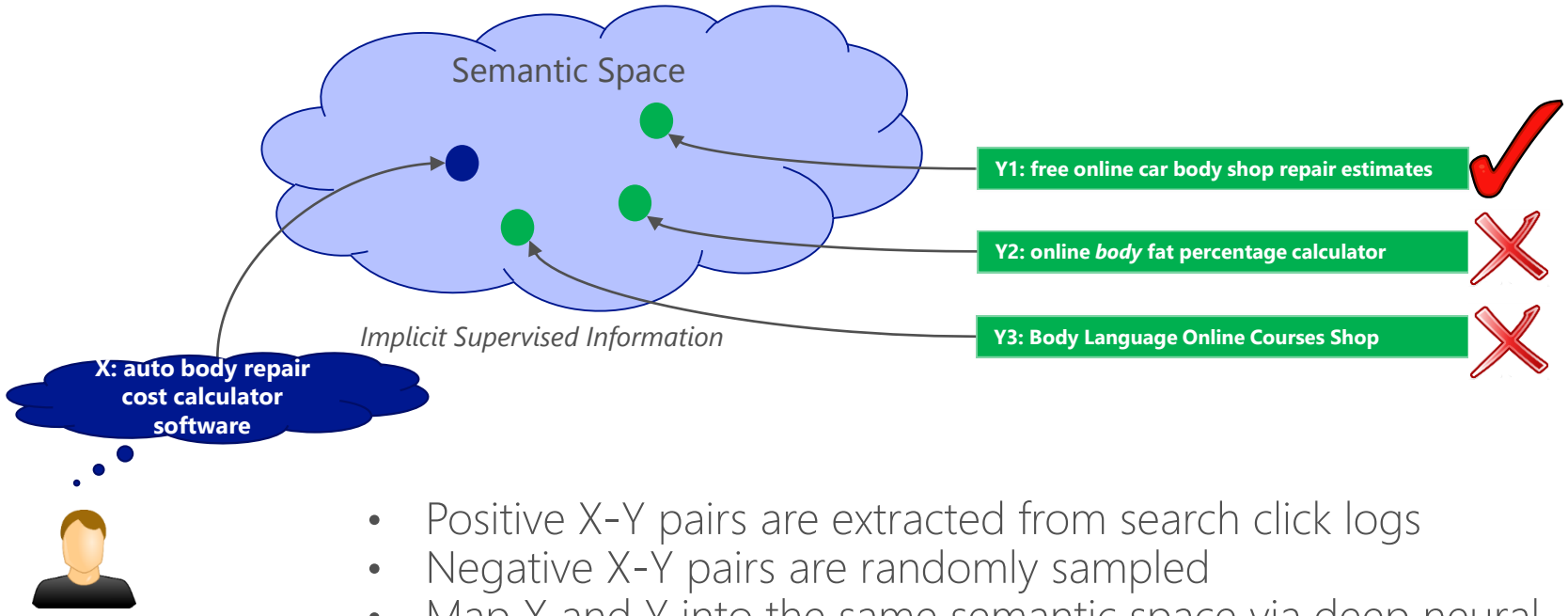
- Consider a query X and two docs Y^+ and Y^-
  - Assume Y^+ is more relevant than Y^- with respect to X
- $\mathbf{sim_\theta}(X, Y)$ is the cosine similarity of X and Y in semantic space, mapped by DSSM parameterized by $\boldsymbol{\theta}$

Learning DSSM from Labeled X-Y Pairs

- Consider a query X and two docs Y^+ and Y^-
  - Assume Y^+ is more relevant than Y^- with respect to X
- $\mathbf{sim_\theta}(X, Y)$ is the cosine similarity of X and Y in semantic space, mapped by DSSM parameterized by $\boldsymbol{\theta}$
- $\Delta = \mathbf{sim_\theta}(X, Y^+) - \mathbf{sim_\theta}(X, Y^-)$
  - We want to maximize Δ
- $Loss(\Delta; \boldsymbol{\theta}) = \log(1 + \exp(-\gamma\Delta))$
- Optimize $\boldsymbol{\theta}$ using mini-batch SGD on GPU

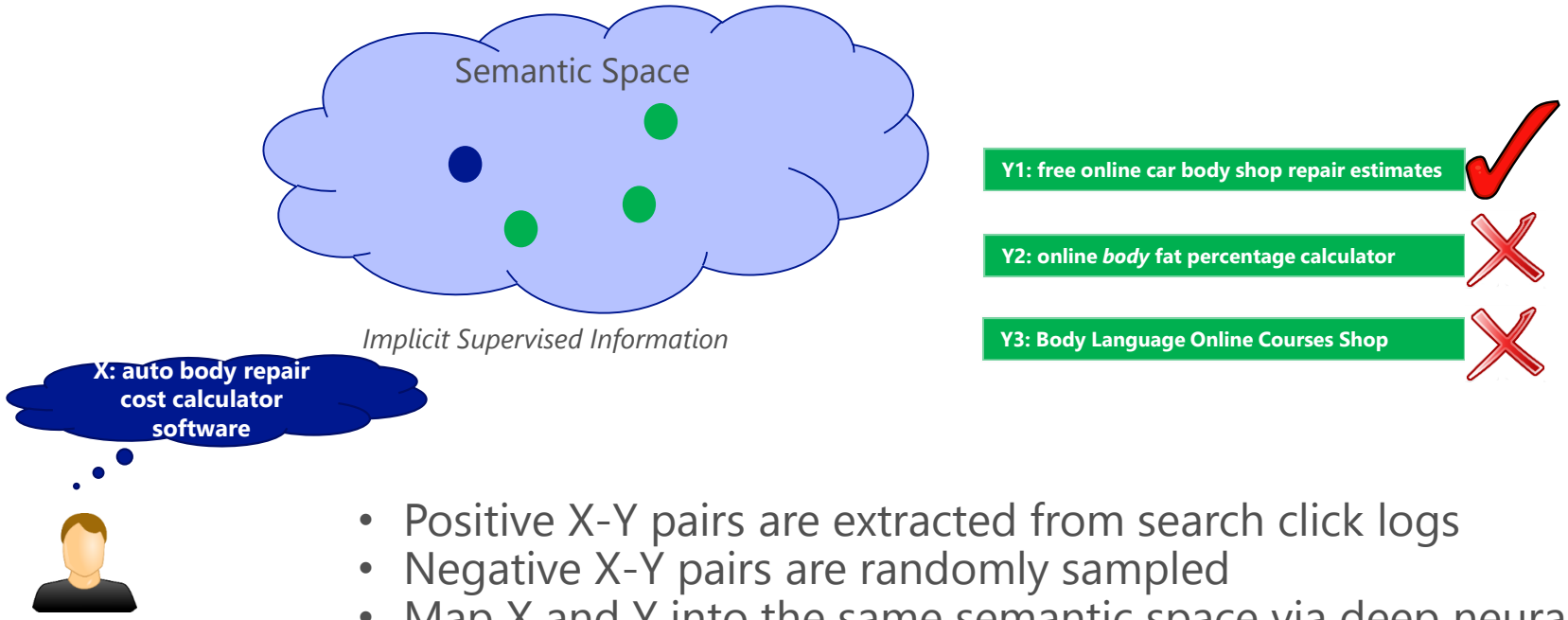


Learning DSSM from Labeled X-Y Pairs



- Positive X-Y pairs are extracted from search click logs
- Negative X-Y pairs are randomly sampled
- Map X and Y into the same semantic space via deep neural net

Learning DSSM from Labeled X-Y Pairs



- Positive X-Y pairs are extracted from search click logs
- Negative X-Y pairs are randomly sampled
- Map X and Y into the same semantic space via deep neural net
- Positive Y are closer to X than negative Y in that space

Mine "Labeled" X-Y Pairs from Search Logs

how to deal with stuffy nose?	⟷	NO CLICK
stuffy nose treatment	⟷	NO CLICK
cold home remedies	⟷	http://www.agelessherbs.com/BestHomeRemediesColdFlu.html

[Gao, He, Nie, 2010]

Mine "Labeled" X-Y Pairs from Search Logs

how to deal with stuffy nose? ⟷

stuffy nose treatment ⟷

cold home remedies ⟷

Best Home Remedies for Cold and Flu

Wind Heat External Pathogens

By: Catherine Browne, L.Ac., MH, Dipl. Ac.

In Chinese medicine, colds and flu's are delineated into several different energetic classifications. Here we will outline the different types of cold and flu viruses that you will likely encounter, and then describe the best home remedies for these specific patterns that you can use to treat the cold or influenza virus.

Cold and Flu Basics

The basic pathogenic influences are:

- **Wind**
- **Cold**
- **Heat**
- **Damp**

Wind

Theoretically, wind enters the body through the back of the neck area or nose carrying the pathogen. It first attacks the Lung system (including the sinuses) because the Lung organ system is the most external Yin organ, a thus the most vulnerable to an external invasion. External Wind invasion is marked by acute conditions with a sudden onset of symptoms.

[Gao, He, Nie, 2010]

Mine "Labeled" X-Y Pairs from Search Logs

QUERY (Q)	Title (T)
how to deal with stuffy nose	best home remedies for cold and flu
stuffy nose treatment	best home remedies for cold and flu
cold home remedies	best home remedies for cold and flu
... ...	... ...
go israel	forums goisrael community
skate at wholesale at pr	wholesale skates southeastern skate supply
breastfeeding nursing blister baby	clogged milk ducts babycenter
thank you teacher song	lyrics for teaching educational children s music
immigration canada lacolle	cbsa office detailed information

[Gao, He, Nie, 2010]

Evaluation Methodology

- Measurement: NDCG, t-test
- Test set:
  - 12,071 English queries sampled from 1-y log
  - 5-level relevance label for each query-doc pair
- Training data for translation models:
  - 82,834,648 query-title pairs
- Baselines
  - Lexicon matching models: BM25, ULM
  - Translation models
  - Topic models

Translation Models for Web Search

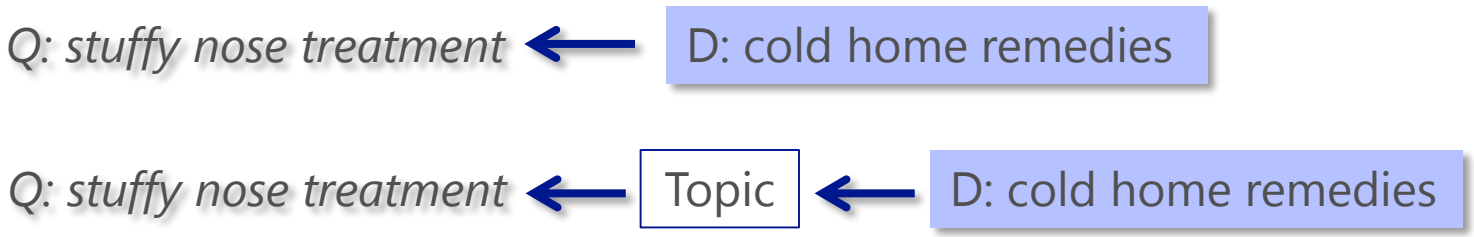
D: best home remedies for cold and flu

Q: how to deal with stuffy nose

- Leverage statistical machine translation (SMT) tech to improve search relevance
- Model docs and queries as different languages
- Cast mapping queries to docs as bridging the language gap via translation
- Given a Q, D can be ranked by how likely it is that Q is "translated" from D, $P(\mathbf{Q}|\mathbf{D})$

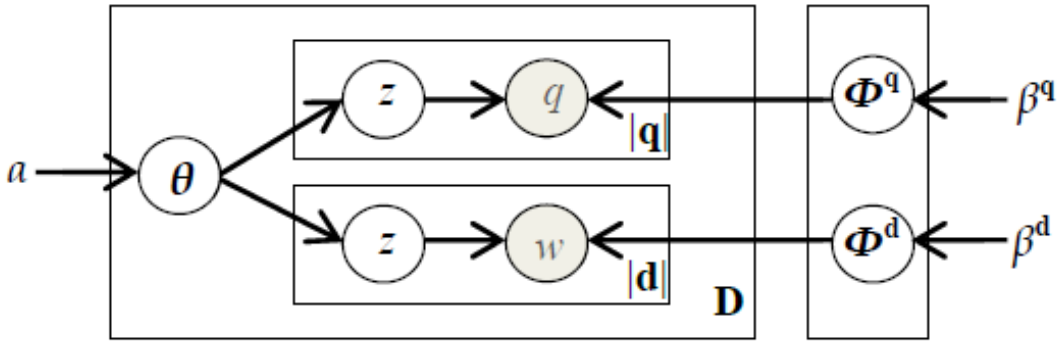
[Gao, He, Nie, 2010]

Generative Topic Models



- Probabilistic latent Semantic Analysis (PLSA)
  - $P(\mathrm{Q}|\mathrm{D}) = \prod_{q\in\mathrm{Q}} \sum_z P(q|\boldsymbol{\phi}_z) P(z|\mathrm{D},\boldsymbol{\theta})$
  - D is assigned a single most likely topic vector
  - Q is generated from the topic vectors
- Latent Dirichlet Allocation (LDA) generalizes PLSA
  - a posterior distribution over topic vectors is used
  - PLSA = LDA with MAP inference

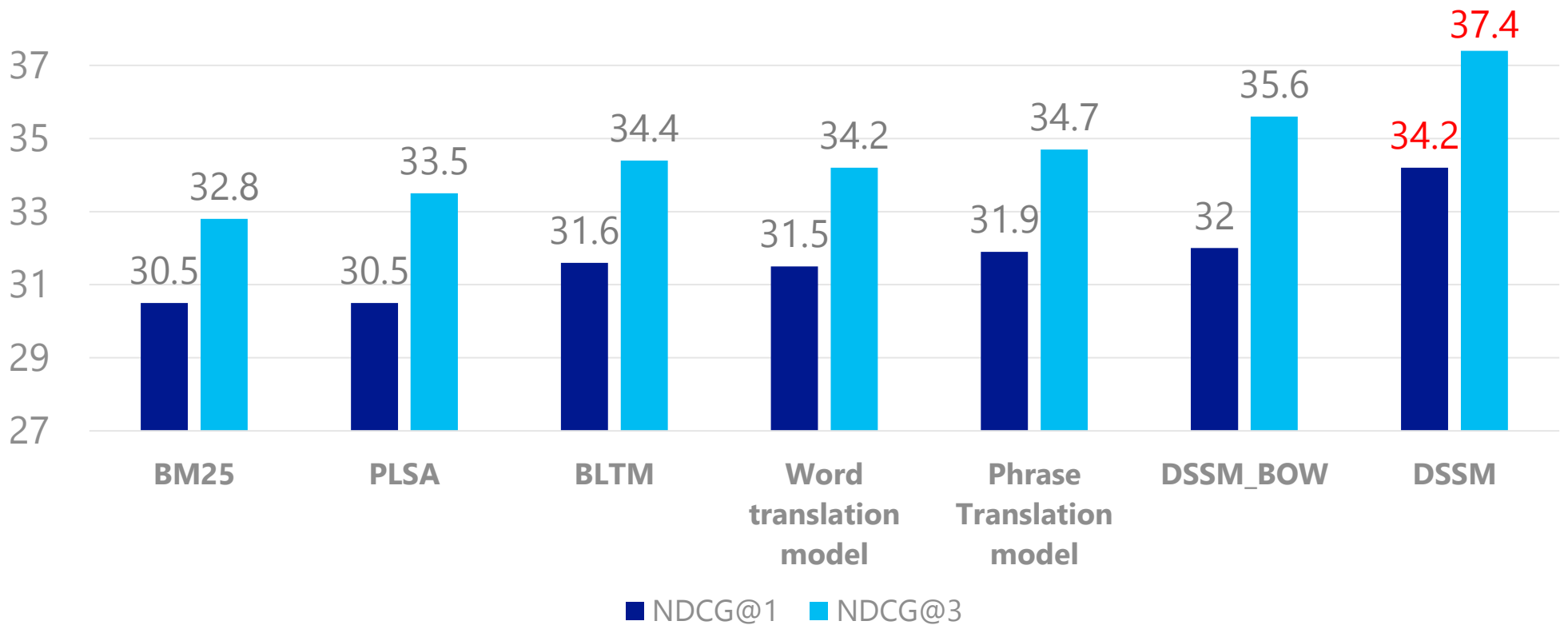
Bilingual Topic Model for Web Search



- For each topic z: $\left(\boldsymbol{\phi}_z^{\mathrm{Q}}, \boldsymbol{\phi}_z^{\mathrm{D}}\right) \sim \mathbf{Dir}(\boldsymbol{\beta})$
- For each Q-D pair: $\boldsymbol{\theta} \sim \mathbf{Dir}(\boldsymbol{\alpha})$
- Each q is generated by $z \sim \boldsymbol{\theta}$ and $q \sim \boldsymbol{\phi}_z^{\mathrm{Q}}$
- Each w is generated by $z \sim \boldsymbol{\theta}$ and $w \sim \boldsymbol{\phi}_z^{\mathrm{D}}$

[Gao, Toutanova, Yih, 2011]

Web Doc Ranking Results



Summary

- Map the queries and documents into the same latent semantic space
- Doc ranking score is the cosine distance of Q/D vectors in that space
- DSSM outperforms all the competing models

DSSM for Recommendation

- Two interestingness tasks for recommendation
- Modeling interestingness via DSSM
- Training data acquisition
- Evaluation
- Summary

[Gao, Pantel, Gamon, He, Deng, Shen, 2014]

Two Tasks of Modeling Interestingness

- Automatic highlighting
  - Highlight the key phrases which represent the entities (person/loc/org) that interest a user when reading a document
  - Doc semantics influences what is perceived as interesting to the user
  - e.g., article about movie → articles about an actor/character
- Contextual entity search
  - Given the highlighted key phrases, recommend new, interesting documents by searching the Web for supplementary info about the entities
  - A key phrase may refer to different entities; need to use the contextual information to disambiguate

The Einstein Theory of Relativity

(1) The perihelion of Mercury shows a discrepancy which has long puzzled astronomers. This discrepancy is fully accounted for by Einstein. At the time when he published his theory, this was its only experimental verification.

(2) Modern physicists were willing to suppose that light might be subject to gravitation—i.e., that a ray of light passing near a great mass like the sun might be deflected to the extent to which a particle moving with the same velocity would be deflected according to the orthodox theory of gravitation. But Einstein's theory required that the light should be deflected just twice as much as this. The matter could only be tested during an eclipse among a number of bright stars. Fortunately a peculiarly favourable eclipse occurred last year. The results of the observations

The Einstein Theory of Relativity

(1) The perihelion of Mercury shows a discrepancy which has long puzzled astronomers. This discrepancy is fully accounted for by Einstein. At the time when he published his theory, this was its only experimental verification.

(2) Modern physicists were willing to suppose that light might be subject to gravitation—i.e., that a ray of light passing near a great mass like the sun might be deflected to the extent to which a particle moving with the same velocity would be deflected according to the orthodox theory of gravitation. But Einstein's theory required that the light should be deflected just twice as much as this. The matter could only be tested during an eclipse among a number of bright stars. Fortunately a peculiarly favourable eclipse occurred last year. The results of the observations

The Einstein Theory of Relativity

(1) The perihelion of Mercury shows a discrepancy which has long puzzled astronomers. This discrepancy is fully accounted for by Einstein. At the time when he published his theory, this was its only experimental verification.

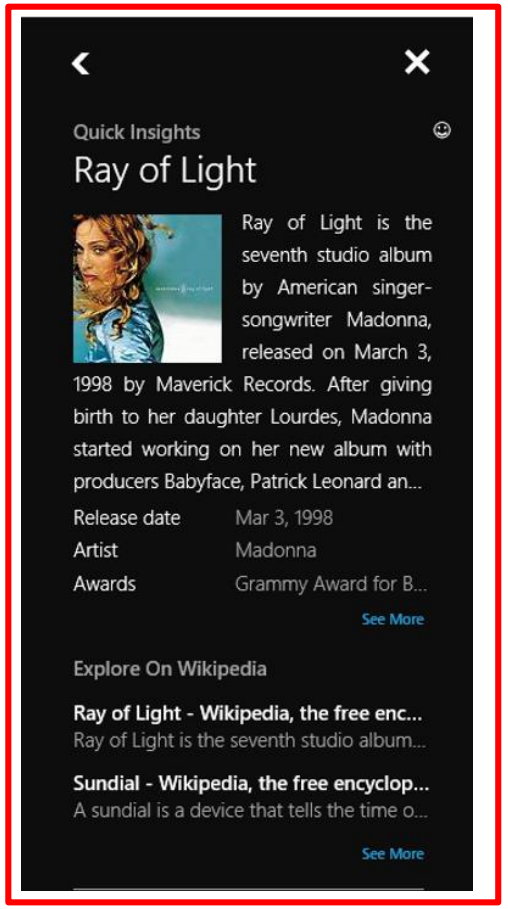
(2) Modern physicists were willing to suppose that light might be subject to gravitation—i.e., that a ray of light passing near a great mass like the sun might be deflected to the extent to which a particle moving with the same velocity would be deflected according to the orthodox theory of gravitation. But Einstein's theory required that the light should be deflected just twice as much as this. The matter could only be tested during an eclipse among a number of bright stars. Fortunately a peculiarly favourable eclipse occurred last year. The results of the observations

The Einstein Theory of Relativity

(1) The perihelion of Mercury shows a discrepancy which has long puzzled astronomers. This discrepancy is fully accounted for by Einstein. At the time when he published his theory, this was its only experimental verification.

(2) Modern physicists were willing to suppose that light might be subject to gravitation—i.e., that a ray of light passing near a great mass like the sun might be deflected to the extent to which a particle moving with the same velocity would be deflected according to the orthodox theory of gravitation. But Einstein's theory required that the light should be deflected just twice as much as this. The matter could only be tested during an eclipse among a number of bright stars. Fortunately a peculiarly favourable eclipse occurred last year. The results of the observations

Entity



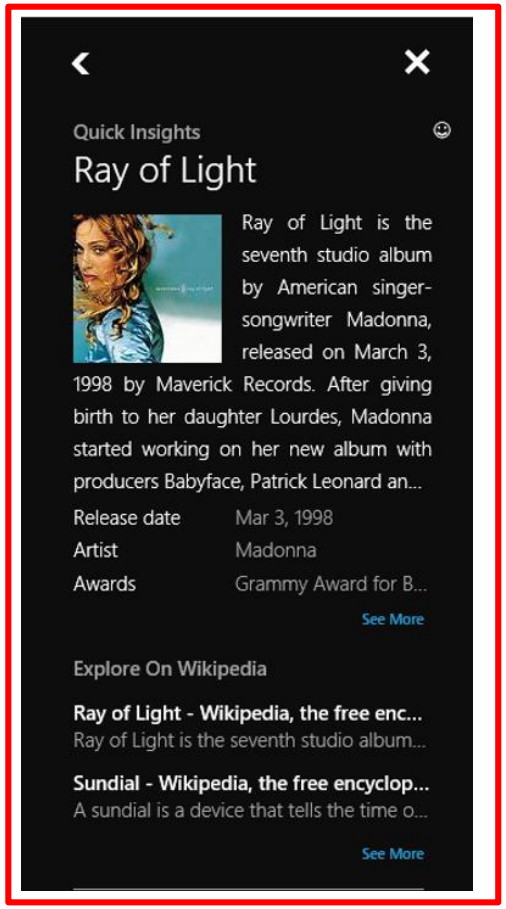
The Einstein Theory of Relativity

Context

(1) The perihelion of Mercury shows a discrepancy which has long puzzled astronomers. This discrepancy is fully accounted for by Einstein. At the time when he published his theory, this was its only experimental verification.

(2) Modern physicists were willing to suppose that light might be subject to gravitation—i.e., that a ray of light passing near a great mass like the sun might be deflected to the extent to which a particle moving with the same velocity would be deflected according to the orthodox theory of gravitation. But Einstein's theory required that the light should be deflected just twice as much as this. The matter could only be tested during an eclipse among a number of bright stars. Fortunately a peculiarly favourable eclipse occurred last year. The results of the observations

Entity



The Einstein Theory of Relativity

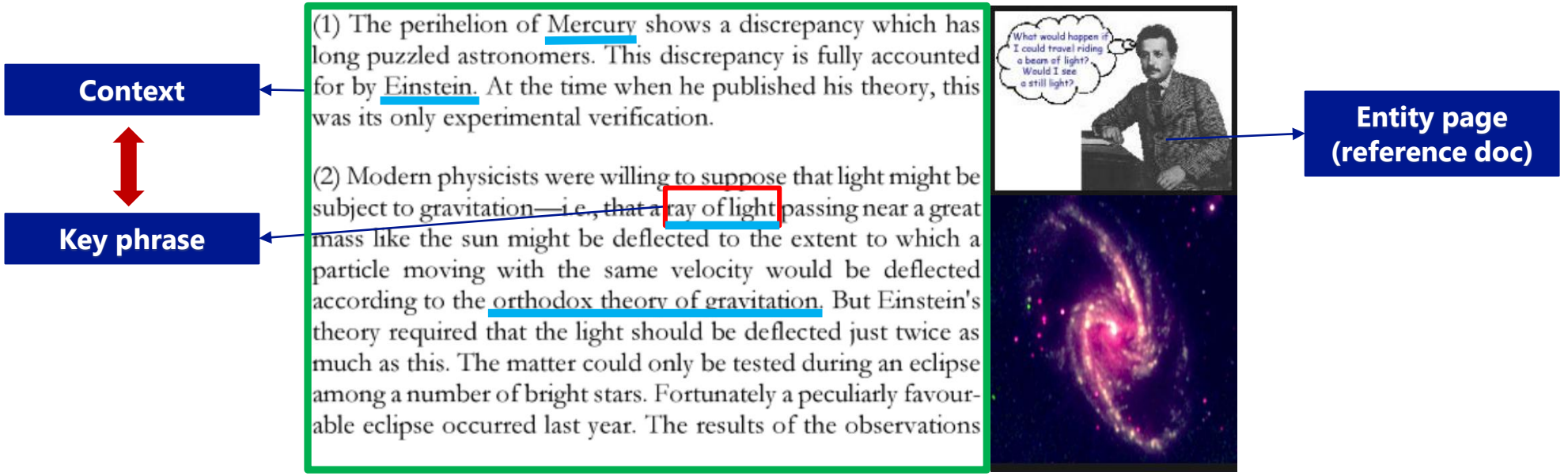
Context

(1) The perihelion of Mercury shows a discrepancy which has long puzzled astronomers. This discrepancy is fully accounted for by Einstein. At the time when he published his theory, this was its only experimental verification.

(2) Modern physicists were willing to suppose that light might be subject to gravitation—i.e., that a ray of light passing near a great mass like the sun might be deflected to the extent to which a particle moving with the same velocity would be deflected according to the orthodox theory of gravitation. But Einstein's theory required that the light should be deflected just twice as much as this. The matter could only be tested during an eclipse among a number of bright stars. Fortunately a peculiarly favourable eclipse occurred last year. The results of the observations

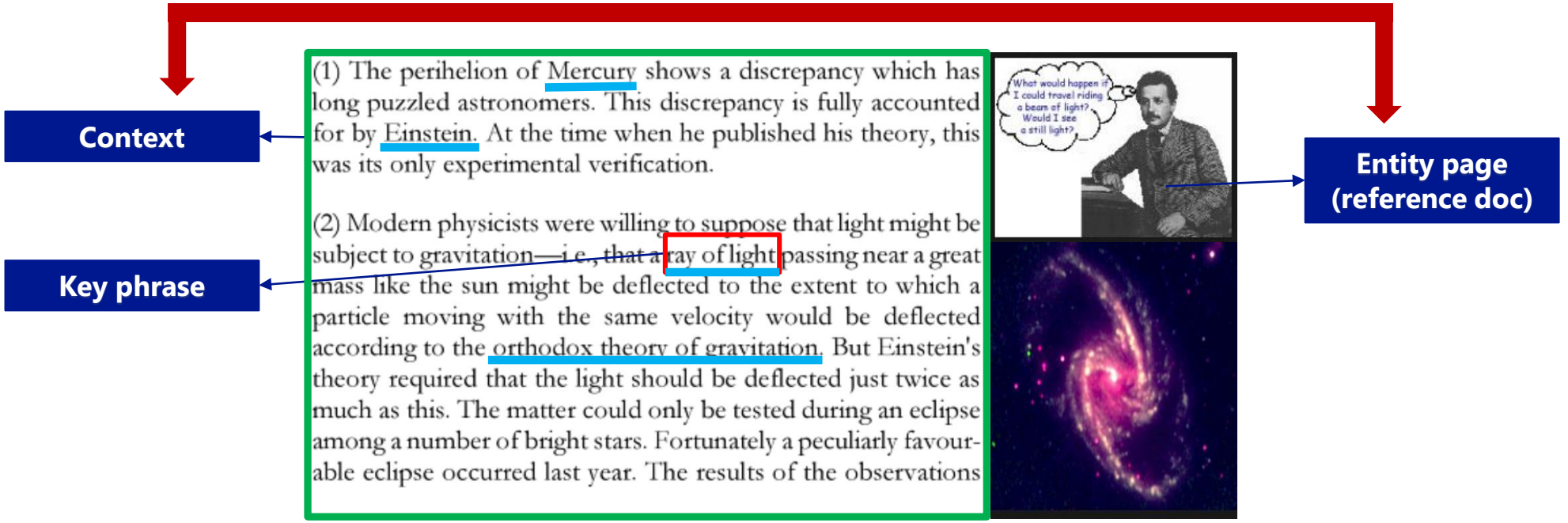
Entity

DSSM for Modeling Interestingness



Tasks	*X (source text)*	*Y (target text)*
Automatic highlighting	***Doc in reading***	***Key phrases to be highlighted***
Contextual entity search	*Key phrase and context*	*Entity and its corresponding (wiki) page*

DSSM for Modeling Interestingness



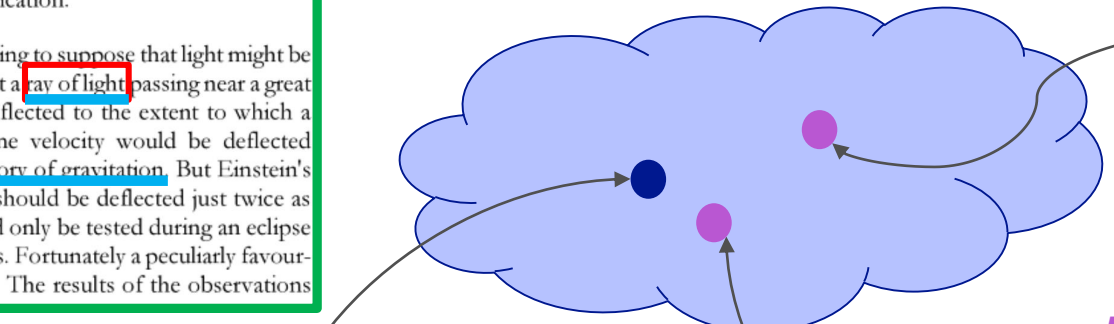
Tasks	*X (source text)*	*Y (target text)*
Automatic highlighting	*Doc in reading*	*Key phrases to be highlighted*
Contextual entity search	***Key phrase and context***	***Entity and its corresponding (wiki) page***

Learning DSSM from Labeled X-Y Pairs

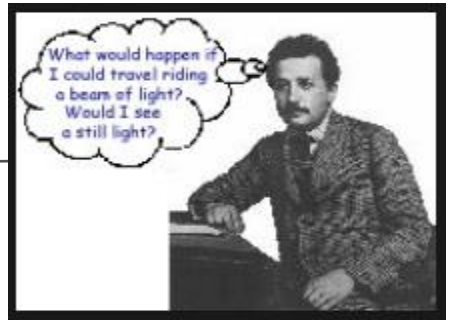
The Einstein Theory of Relativity

(1) The perihelion of Mercury shows a discrepancy which has long puzzled astronomers. This discrepancy is fully accounted for by Einstein. At the time when he published his theory, this was its only experimental verification.

(2) Modern physicists were willing to suppose that light might be subject to gravitation—i.e., that a ray of light passing near a great mass like the sun might be deflected to the extent to which a particle moving with the same velocity would be deflected according to the orthodox theory of gravitation. But Einstein's theory required that the light should be deflected just twice as much as this. The matter could only be tested during an eclipse among a number of bright stars. Fortunately a peculiarly favourable eclipse occurred last year. The results of the observations



Ray of Light (Experiment)



ray of light

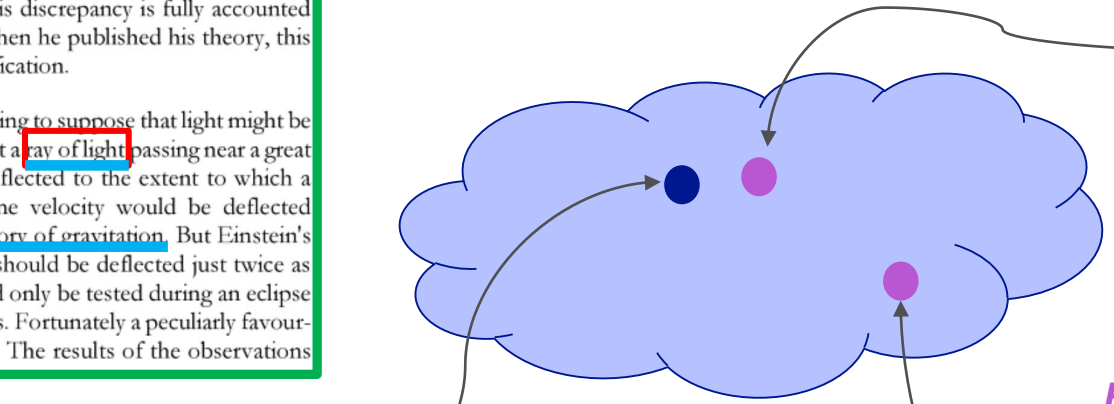
Ray of Light (Song)

Learning DSSM from Labeled X-Y Pairs

The Einstein Theory of Relativity

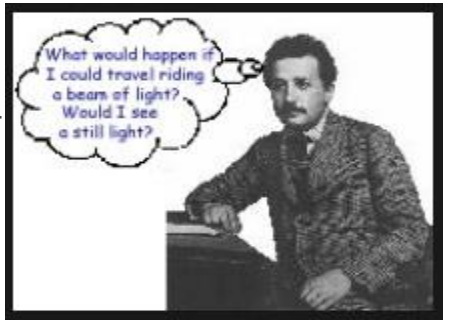
(1) The perihelion of Mercury shows a discrepancy which has long puzzled astronomers. This discrepancy is fully accounted for by Einstein. At the time when he published his theory, this was its only experimental verification.

(2) Modern physicists were willing to suppose that light might be subject to gravitation—i.e., that a ray of light passing near a great mass like the sun might be deflected to the extent to which a particle moving with the same velocity would be deflected according to the orthodox theory of gravitation. But Einstein's theory required that the light should be deflected just twice as much as this. The matter could only be tested during an eclipse among a number of bright stars. Fortunately a peculiarly favourable eclipse occurred last year. The results of the observations



ray of light

Ray of Light (Experiment)



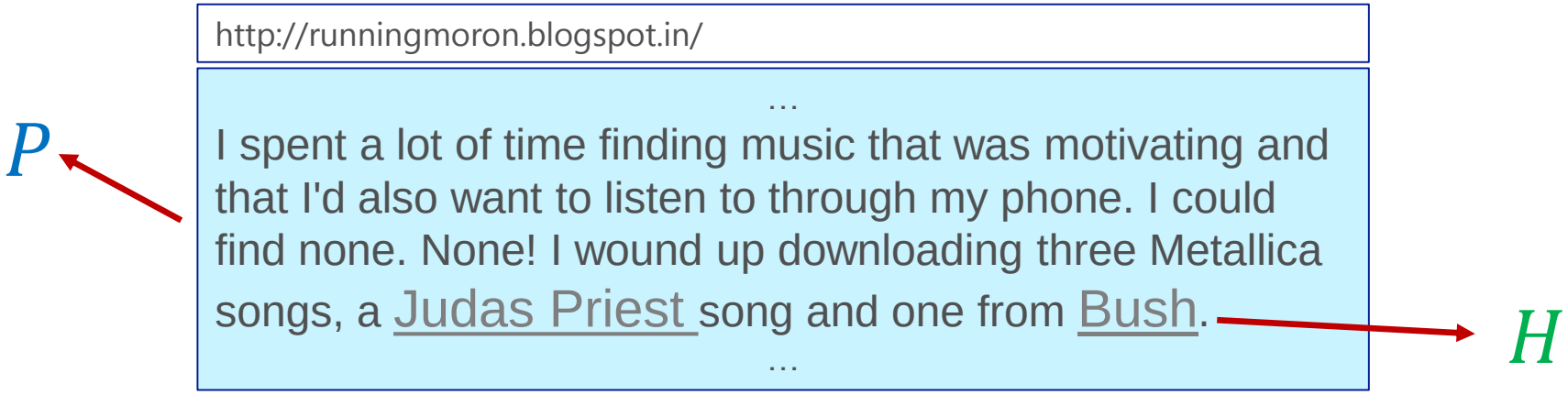
Ray of Light (Song)

DSSM for recommendation

- Two interestingness tasks for recommendation
- Modeling interestingness via DSSM
- Training data acquisition
- Evaluation
- Summary

Extract Labeled Pairs from Web Browsing Logs Automatic Highlighting

- When reading a page P, the user *clicks* a hyperlink H



- (text in P, anchor text of H)

Extract Labeled Pairs from Web Browsing Logs Contextual Entity Search

- When a hyperlink H points to a Wikipedia P'

http://runningmoron.blogspot.in/

…

I spent a lot of time finding music that was motivating and that I'd also want to listen to through my phone. I could find none. None! I wound up downloading three Metallica songs, a Judas Priest song and one from Bush.

…

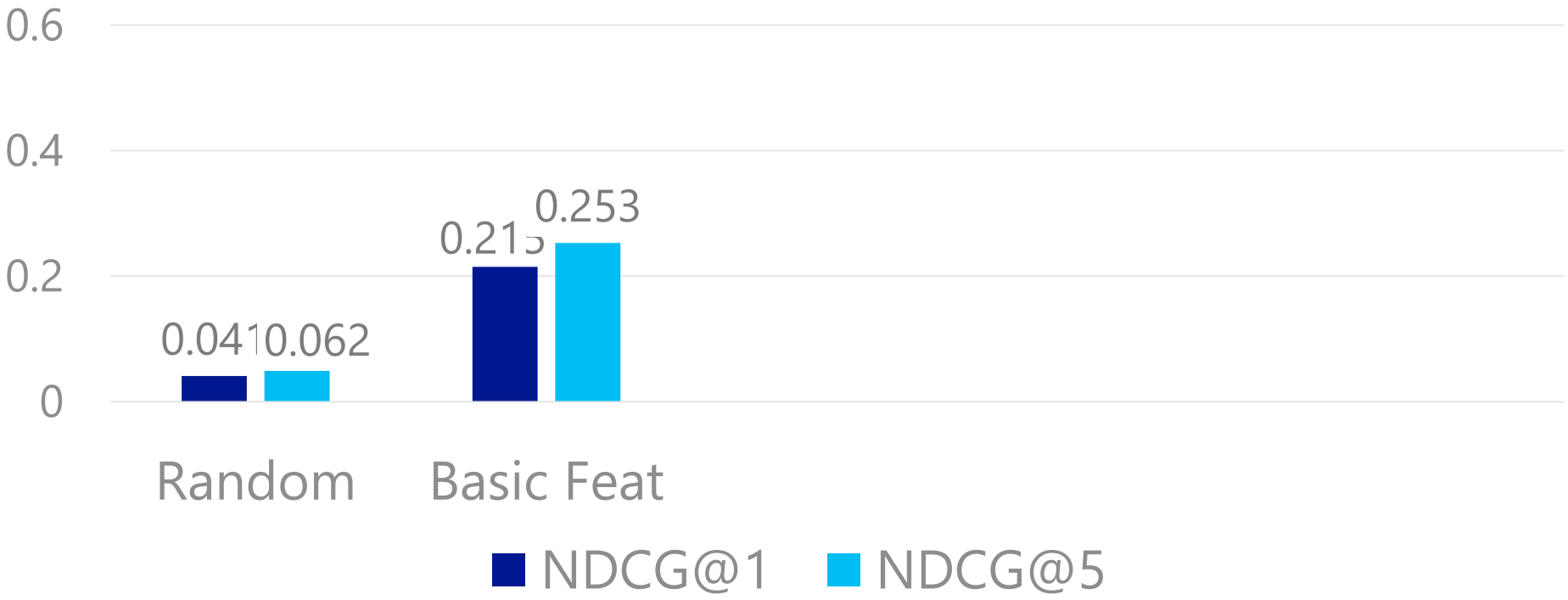
http://en.wikipedia.org/wiki/Bush_(band)

- (anchor text of H & surrounding words, text in P')

Automatic Highlighting: Settings

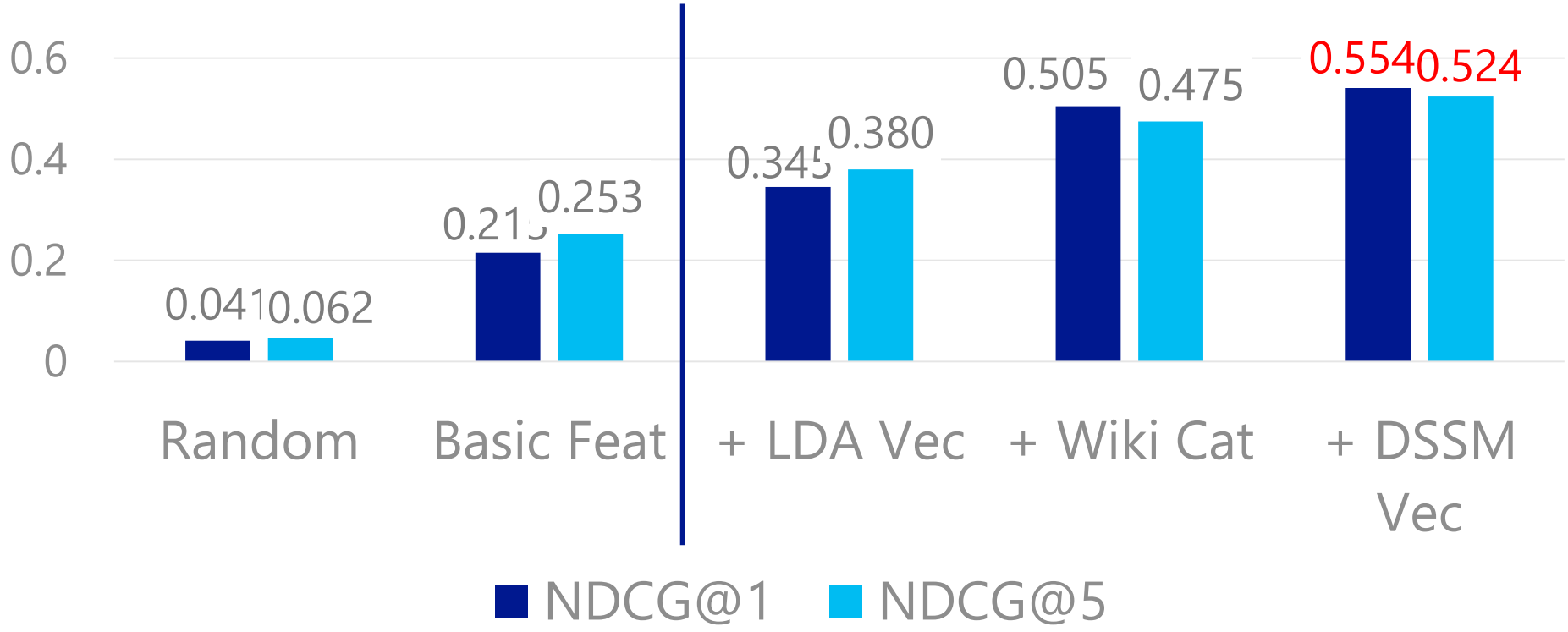
- Simulation
  - Use a set of anchors as candidate key phrases to be highlighted
  - Gold standard rank of key phrases – determined by # user clicks
  - Model picks top-k keywords from the candidates
  - Evaluation metric: NDCG
- Data
  - 18 million occurrences of user clicks from a Wiki page to another, collected from 1-year Web browsing logs
  - 60/20/20 split for training/validation/evaluation

Automatic Highlighting Results: Baselines



- Random: Random baseline
- Basic Feat: Boosted decision tree learner with document features, such as anchor position, freq. of anchor, anchor density, etc.

Automatic Highlighting Results: Semantic Features

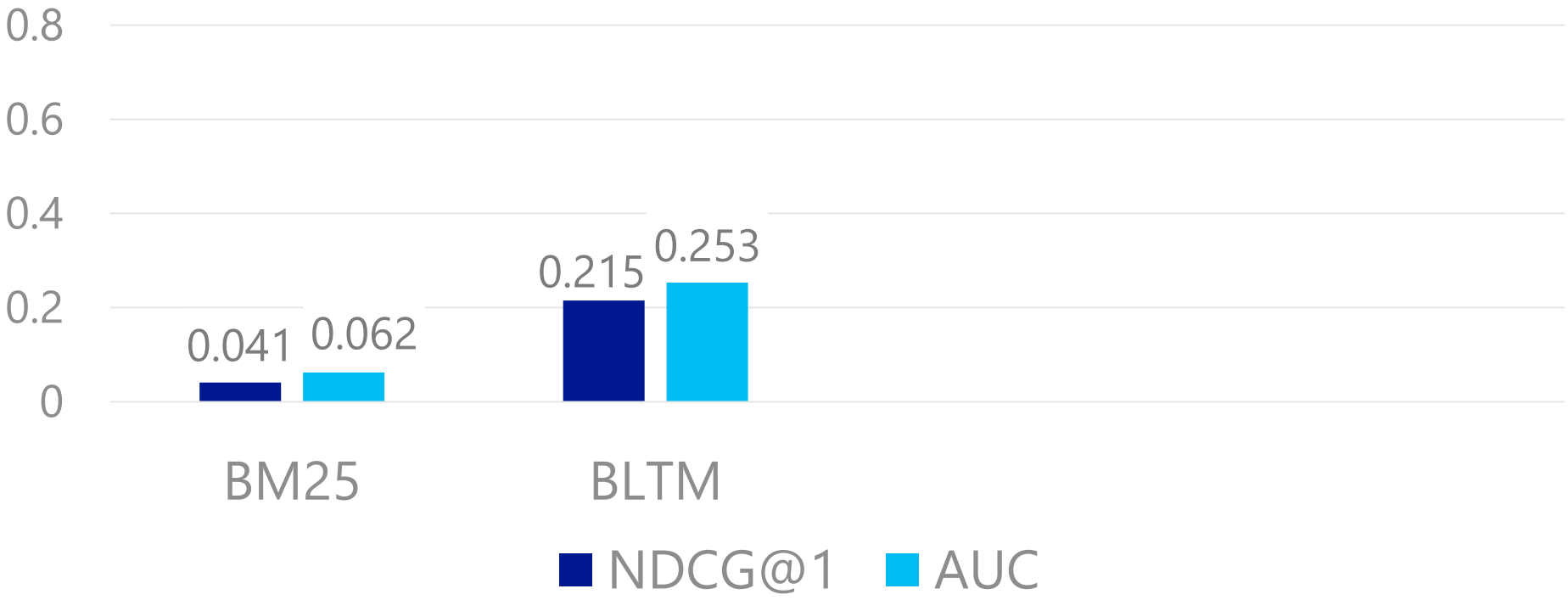


- + LDA Vec: Basic + Topic model (LDA) vectors [Gamon+ 2013]
- + Wiki Cat: Basic + Wikipedia categories (do not apply to general documents)
- + DSSM Vec: Basic + DSSM vectors

Contextual Entity Search: Settings

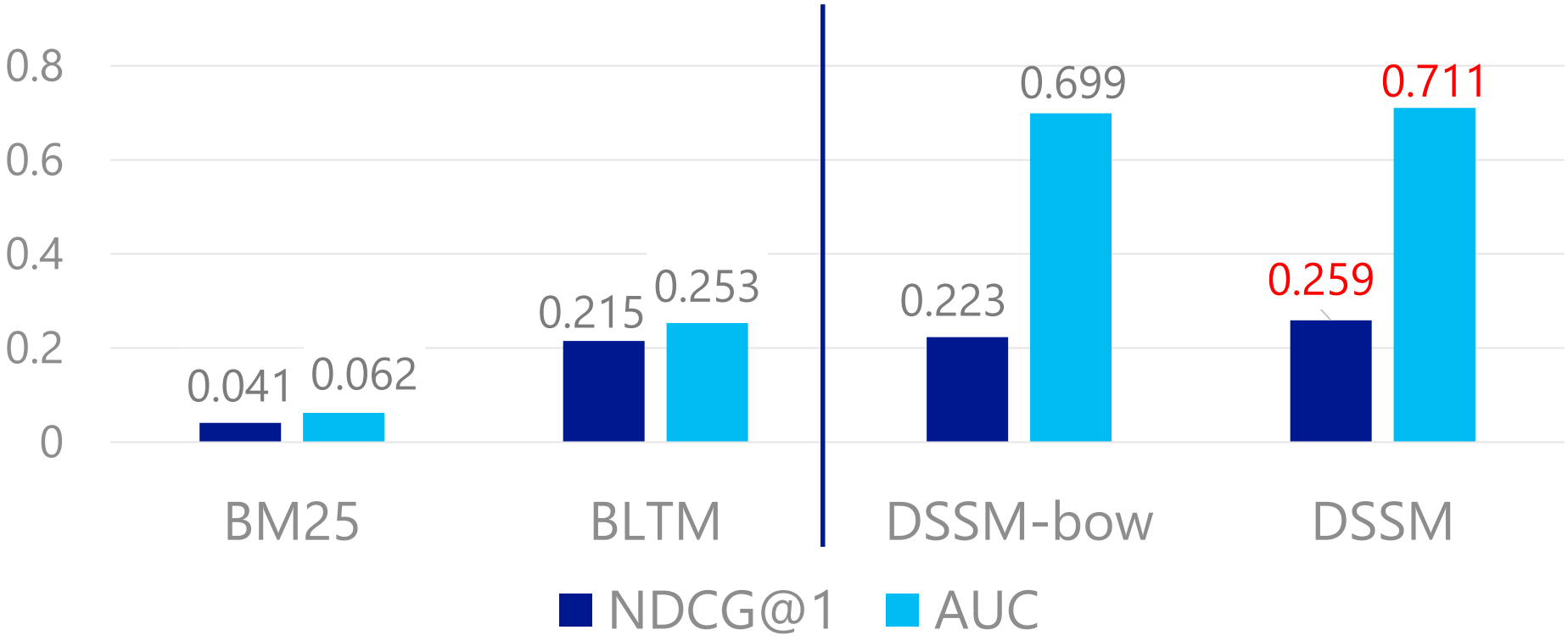
- Training/validation data: same as in *automatic highlighting*
- Evaluation data
  - Sample 10k Web documents as the source documents
  - Use named entities in the doc as query; retain up to 100 returned documents as target documents
  - Manually label whether each target document is a good page describing the entity
  - 870k labeled pairs in total
- Evaluation metric: NDCG and AUC

Contextual Entity Search Results: Baselines



- BM25: The classical document model in IR [Robertson+ 1994]
- BLTM: Bilingual Topic Model [Gao+ 2011]

Contextual Entity Search Results: DSSM



- DSSM-bow: DSSM without convolutional layer and max-pooling structure
- DSSM outperforms classic doc model and state-of-the-art topic model

Summary

- Extract labeled pairs from Web browsing logs
- DSSM outperforms state-of-the-art topic models
- DSSM learned semantic features outperform the thousands of features coming from the manually assigned semantic labels

DSSM for Phrase Translation Modeling

- Introduction
- DSSM for phrase translation modeling
- Model training using Expected-BLEU objective
- Evaluation on WMT
- Summary

[Gao, He, Yih, Deng, 2014]

Statistical Machine Translation (SMT)

C: 救援 人员 在 倒塌的 房屋 里 寻找 生还者
E: Rescue workers search for survivors in collapsed houses

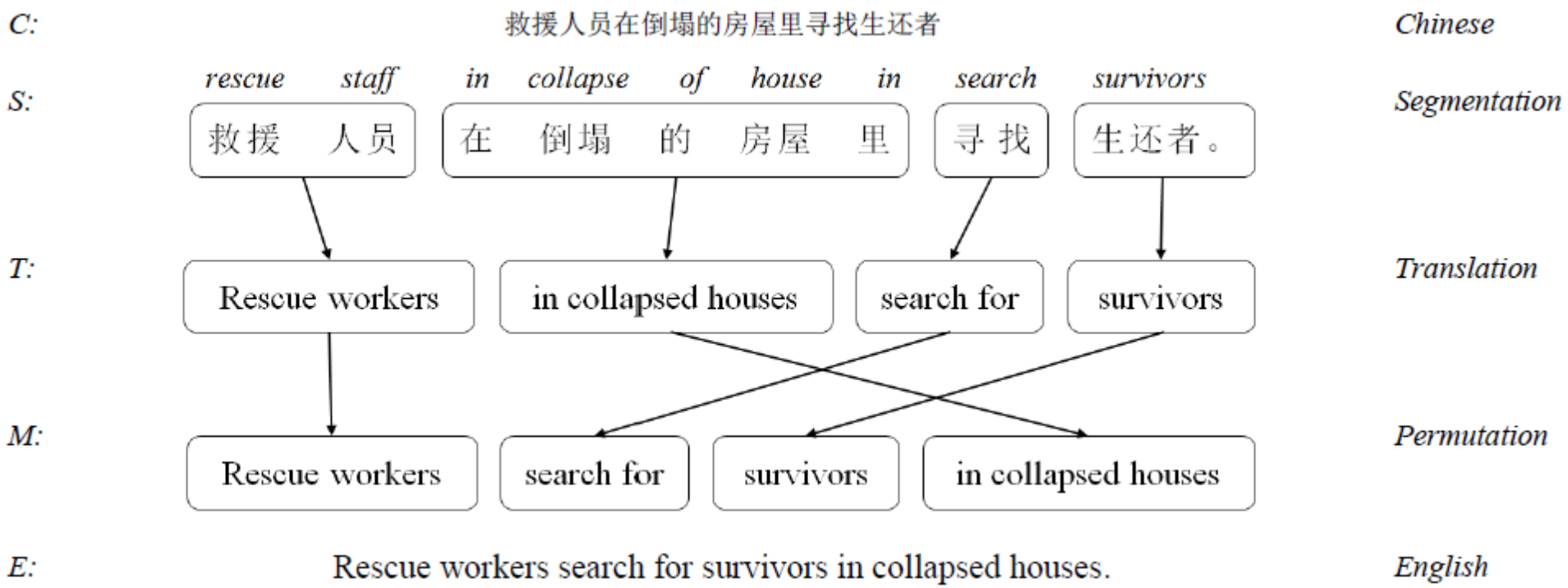
- Statistical decision: $E^* = \underset{E}{\operatorname{argmax}}\, P(E|C)$
- Source-channel model: $E^* = \underset{E}{\operatorname{argmax}}\, P(C|E)P(E)$
- Translation models: $P(C|E)$ and $P(E|C)$
- Log-linear model: $P(E|C) = \frac{1}{Z(C,E)} \exp \sum_i \lambda_i h_i(C,E)$
- Evaluation metric: BLEU score (higher is better)

[Koehn 2009]

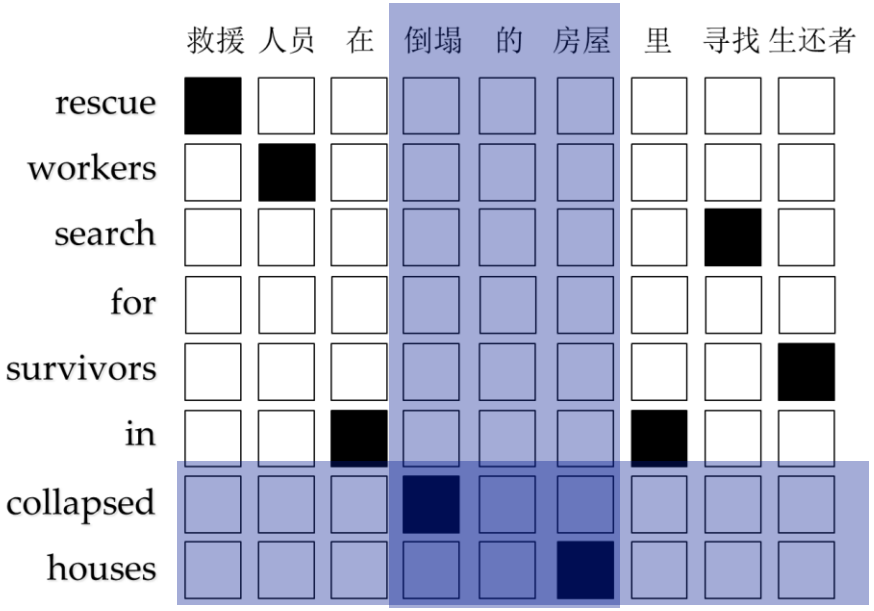
Phrase-based Models

C:	救援人员在倒塌的房屋里寻找生还者	*Chinese*

Phrase-based Models



Parameter Estimation

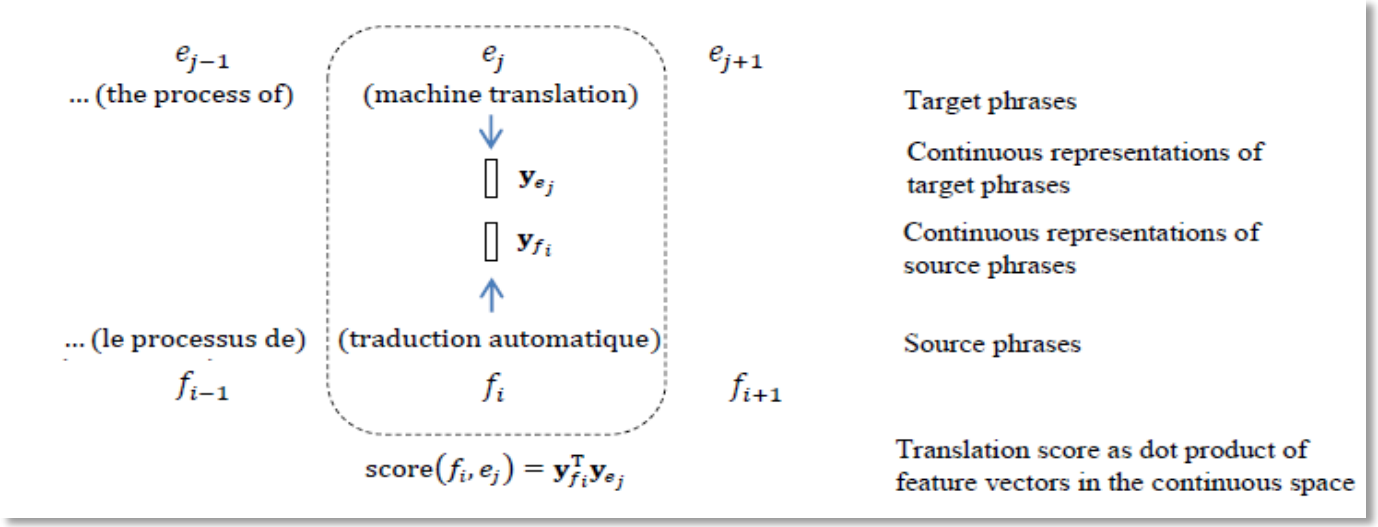


(救援, rescue)
(人员, workers)
(在, in)
(倒塌, collapsed)
(房屋, house)
(里, in)
(寻找, search)
(生还者, survivors)
(救援 人员, rescue workers)
(在 倒塌, in collapsed)
(倒塌 的, collapsed)
(的 房屋, house)
(寻找, search for)
(寻找 生还者, search for survivors)
(生还者, for survivors)
(倒塌 的 房屋, collapsed house)

MLE: $P(\mathbf{e}|\mathbf{c}) = \frac{N(\mathbf{c},\mathbf{e})}{\sum_{\mathbf{e}'} N(\mathbf{c},\mathbf{e}')}$

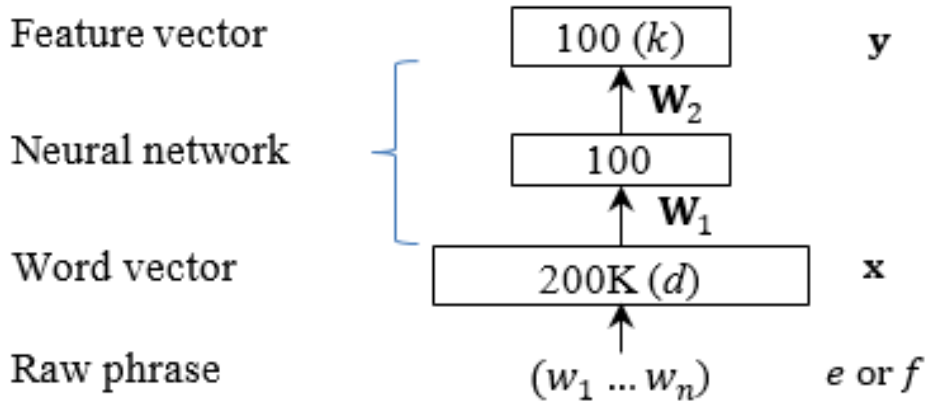
with smoothing

DSSM for Phrase Translation Modeling



- Follows the "story" of phrase translation models, but
- Uses different parameter estimation method
  - Map source/target phrases into the same semantic space
  - Phrase translation score == similarity btw their feature vectors in semantic space

A Closer Look at the *Mapping*



- Bag-of-words representation of a phrase: $\mathbf{x}$
- Map $\mathbf{x}$ to a low-dim semantic space: $\boldsymbol{\phi}(\mathbf{x})\mathbf{:}\ \mathbb{R}^d \rightarrow \mathbb{R}^k$
- Mapping is performed using a neural net:

$$\mathbf{y} \equiv \phi(\mathbf{x}) = \tanh\left(\mathbf{W}_2{}^{\mathrm{T}}\left(\tanh\left(\mathbf{W}_1{}^{\mathrm{T}}\mathbf{x}\right)\right)\right)$$

- Translation score as similarity between feature vectors

$$\mathbf{score}(f, e) \equiv \mathbf{sim}_{\boldsymbol{\theta}}(\mathbf{x}_f, \mathbf{x}_e) = \mathbf{y}_f^{\mathrm{T}}\mathbf{y}_e$$

Using the DSSM for SMT

- Define a new translation feature:

$$h_{M+1}(F_i, E, \boldsymbol{\theta}) = \sum_{(f,e)\in A} \mathrm{sim}_{\boldsymbol{\theta}}(\mathbf{x}_f, \mathbf{x}_e)$$

- Integrate into the log-linear model for SMT:

$$P(E|F) = \frac{1}{Z(F,E)} \exp \sum_i \lambda_i h_i(F,E)$$

$$E^* = \underset{E}{\mathrm{argmax}} \sum_i \lambda_i h_i(F,E)$$

Parameter Estimation

- Parameters $(\boldsymbol{\lambda}, \boldsymbol{\theta})$
  - $\boldsymbol{\lambda}$: a handful of parameters in log-linear model.
  - $\boldsymbol{\theta}$: projection matrices of the DSSM.
- Take three steps to learn $(\boldsymbol{\lambda}, \boldsymbol{\theta})$:
  - Generate N-best lists using a baseline SMT system
  - Fix $\boldsymbol{\lambda}$, and optimize $\boldsymbol{\theta}$ w.r.t. a loss function on the N-best lists of training data.
  - Fix $\boldsymbol{\theta}$, and optimize $\boldsymbol{\lambda}$ to maximize BLEU on development data.

How to Learn DSSM Parameters?

- Problem 1: can we optimize translation quality (BLEU) directly?
- Solution: end2end optimization based on a task-specific objective

- Problem 2: we have sentence pairs, but not phrase pairs, for training.
- Solution: use the chain rule to decompose the sentence error to phrase errors

Training DSSM Parameters, $\boldsymbol{\theta}$

- Define a loss function $\mathcal{L}(\boldsymbol{\theta})$, which is
  - Friendly to optimizer: differentiable/convex
  - Aiming the right target: closely related to task-specific metric (BLEU)
- Update $\boldsymbol{\theta}$ with gradient descent

$$\boldsymbol{\theta}^{new} = \boldsymbol{\theta} - \eta \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

- Algorithms
  - Batch training, L-BFGS
  - Stochastic Gradient Descent (SGD)

Loss Function: $\mathcal{L}(\boldsymbol{\theta})$

- Expected BLEU based on n-best list
  - $\mathrm{xBleu}(\boldsymbol{\theta}) = \sum_{E \in \mathrm{GEN}(F_i)} P(E|F_i)\mathrm{sBleu}(E_i, E)$
  - $P(E|F_i) = \frac{\exp\left(\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{h}(F_i,E,A)+\lambda_{M+1}h_{M+1}(F_i,E,\boldsymbol{\theta})\right)}{\sum_{E \in \mathrm{GEN}(F_i)} \exp(\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{h}(F_i,E,A)+\lambda_{M+1}h_{M+1}(F_i,E,\boldsymbol{\theta}))}$
- Friendly to optimizer?
  - Differentiable but non-convex
- Aiming the right target?
  - Closely related to BLEU

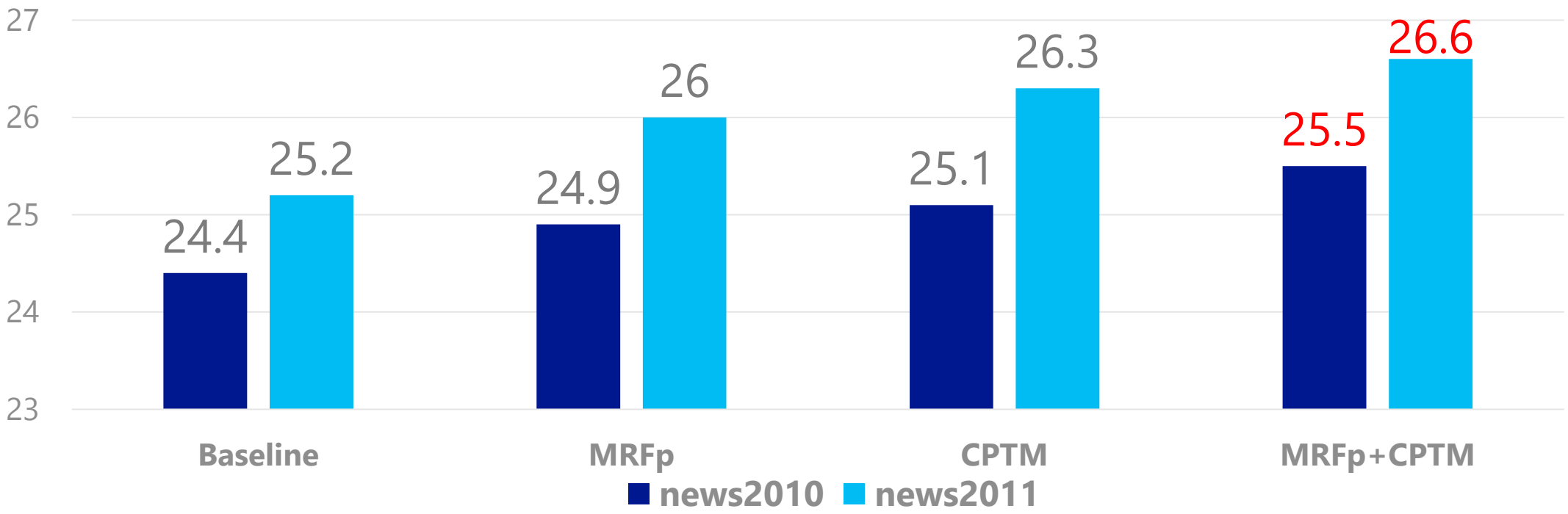
Gradient: $\partial \mathcal{L}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$

- $\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{(f,e)} \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \mathrm{sim}_{\boldsymbol{\theta}}(\mathbf{x}_f, \mathbf{x}_e)} \frac{\partial \mathrm{sim}_{\boldsymbol{\theta}}(\mathbf{x}_f, \mathbf{x}_e)}{\partial \boldsymbol{\theta}}$
- Error term: $-\partial \mathcal{L}(\boldsymbol{\theta}) / \partial \mathrm{sim}_{\boldsymbol{\theta}}(\mathbf{x}_f, \mathbf{x}_e)$
  - how the overall loss changes with the translation score of the phrase pair
- $\partial sim_{\boldsymbol{\theta}}(\mathbf{x}_f, \mathbf{x}_e) / \partial \boldsymbol{\theta}$ can be computed via Back Propagation (BP)
  - Similar to DSSM for web search ranking

Evaluation

- Two Europarl translation tasks
  - English-to-French (EN-FR)
  - German-to-English (DE-EN)
- Baseline
  - A state-of-the-art phrase-based SMT system, i.e., Moses
- Evaluation metric
  - case insensitive BLEU score
  - 1 reference

Results on WMT2012 datasets



- MRF: Markov Random Fields with xBleu [Gao and He 2013]
- CPTM: DSSM with xBleu
- **Up to 1.3 BLEU point improvement over the baseline**

Summary

- Map the sentences in source/target languages into the same, language-independent semantic space
- The DSSM-based semantic translation model leads up to 1.3 BLEU improvement
- DSSM training: end2end optimization based on a task-specific objective
- Other DNNs for SMT
  - [Auli et al. 2013; Auli and Gao, 2014; Hu et al. 2014; Devlin et al. 2014]

Deep Semantic Similarity Model (DSSM): learning semantic similarity between *X* and *Y*

Tasks	*X*	*Y*
Web search	***Search query***	***Web documents***
Ad selection	***Search query***	***Ad keywords***
Entity ranking	***Mention (highlighted)***	***Entities***
Recommendation	***Doc in reading***	***Interesting things in doc or other docs***
Machine translation	***Sentence in language A***	***Translations in language B***
Nature User Interface	*Command (text/speech)*	*Action*
Summarization	*Document*	*Summary*
Query rewriting	*Query*	*Rewrite*
Image retrieval	*Text string*	*Images*
…	…	…

[Huang et al. 2013; Shen et al. 2014; Gao et al. 2014a; Gao et al. 2014b]

Mission of Machine (Deep) Learning

"Real" world	Data (collected/labeled)
"Artificial" world	Model (architecture)
Link the two worlds	Training (algorithm)

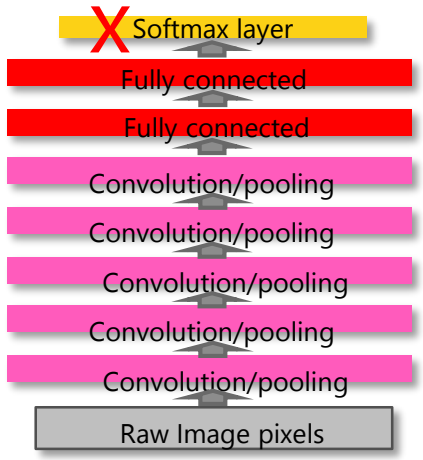
Mission of Machine Deep Learning

"Real" world	*Labeled* Data
"Artificial" world	*Deep* Neural network
Link the two worlds	*Stochastic* Gradient Descent

Text is boring, let's have some fun!

DSSM for Text-Image Joint Representation Learning

- Recall DSSM for text inputs: *s, t1, t2, t3, …*
- Now: replace text s by image **s**
- Using DNN/CNN features of image
- Can rank/generate text's given image or can rank images given text.



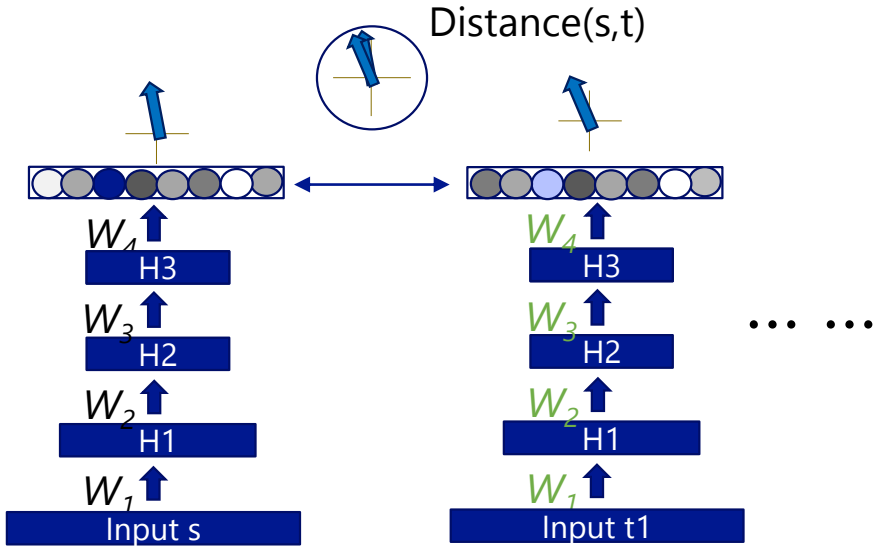


Image features s

Text: *a parrot rides a tricycle*

Automatic Image Captioning

Detector Models, Deep Neural Net Features

Computer Vision System

street
signs
light
under
on
pole
red
stop
sign
building
bus
city
traffic

Language Model

Caption Generation System

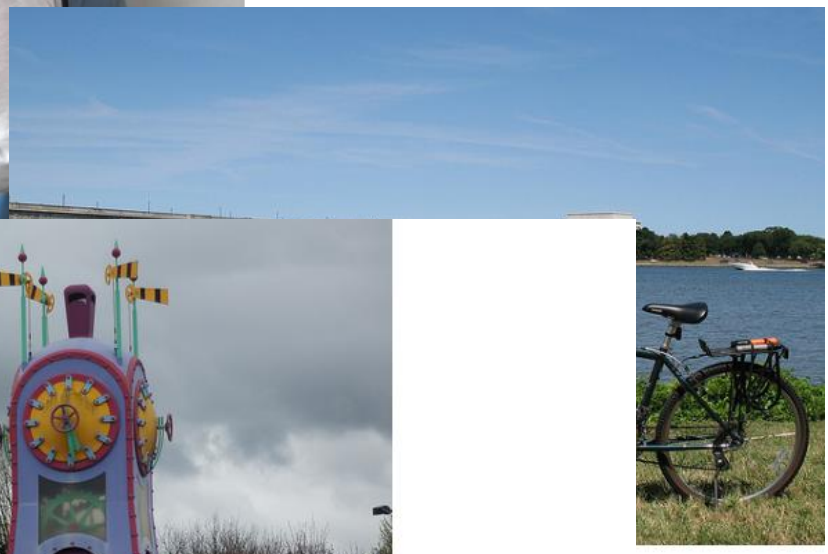
a red stop sign sitting under a traffic light on a city street
a stop sign at an intersection on a street
a stop sign with two street signs on a pole on a sidewalk
a stop sign at an intersection on a city street
...
a stop sign
a red traffic light

a stop sign at an intersection on a city street

DSSM Model

Global Semantic Ranking System

Fang, Gupta, Iandola, Srivastava, Deng, Dollar, Gao, He, Mitchell, Platt, Zitnick, Zweig, "From captions to visual concepts and back," on arXiv



Machine-generated (but turker prefered)	a group of motorc to a motorcycle
Human-annotated (but turker not prefered)	two girls wearing skirts and one of motorcycle while nearby

Machine-generated (but turker prefered)	a man holding a tennis racquet on a tennis court
Human-annotated (but turker not prefered)	the man is on the tennis court playing a game

next to a

next to a

rated (but)	a clock tower in the middle of the street
Human-annotated (but turker not prefered)	a statue with a clock on it near a parking lot

References

- Auli, M., Galley, M., Quirk, C. and Zweig, G., 2013. Joint language and translation modeling with recurrent neural networks. In EMNLP.
- Auli, M., and Gao, J., 2014. Decoder integration and expected bleu training for recurrent neural network language models. In ACL.
- Bengio, Y., 2009. Learning deep architectures for AI. Foundumental Trends Machine Learning, vol. 2.
- Bengio, Y., Courville, A., and Vincent, P. 2013. Representation learning: A review and new perspectives. IEEE Trans. PAMI, vol. 38, pp. 1798-1828.
- Bengio, Y., Ducharme, R., and Vincent, P., 2000. A Neural Probabilistic Language Model, in NIPS.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P., 2011. Natural language processing (almost) from scratch. in JMLR, vol. 12.
- Dahl, G., Yu, D., Deng, L., and Acero, 2012. A. Context-dependent, pre-trained deep neural networks for large vocabulary speech recognition, IEEE Trans. Audio, Speech, & Language Proc., Vol. 20 (1), pp. 30-42.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T., and Harshman, R. 1990. Indexing by latent semantic analysis. J. American Society for Information Science, 41(6): 391-407
- Deng, L., Seltzer, M., Yu, D., Acero, A., Mohamed, A., and Hinton, G., 2010. Binary Coding of Speech Spectrograms Using a Deep Auto-encoder, in Interspeech.
- Deng, L., Tur, G, He, X, and Hakkani-Tur, D. 2012. Use of kernel deep convex networks and end-to-end learning for spoken language understanding, Proc. IEEE Workshop on Spoken Language Technologies.
- Deng, L., Yu, D. and Acero, A. 2006. Structured speech modeling, IEEE Trans. on Audio, Speech and Language Processing, vol. 14, no. 5, pp. 1492-1504.
- Deng, L., Yu, D., and Platt, J. 2012. Scalable stacking and learning for building deep architectures, Proc. ICASSP.
- Deoras, A., and Sarikaya, R., 2013. Deep belief network based semantic taggers for spoken language understanding, in INTERSPEECH.
- Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., and Makhoul, J., 2014. Fast and Robust Neural Network Joint Models for Statistical Machine Translation, ACL.
- Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T., 2013. DeViSE: A Deep Visual-Semantic Embedding Model, Proc. NIPS.
- Gao, J., He, X., Yih, W-t., and Deng, L. 2014a. Learning continuous phrase representations for translation modeling. In ACL.
- Gao, J., He, X., and Nie, J-Y. 2010. Clickthrough-based translation models for web search: from word models to phrase models. In CIKM.
- Gao, J., Pantel, P., Gamon, M., He, X., and Deng, L. 2014. Modeling interestingness with deep neural networks. In EMNLP
- Gao, J., Toutanova, K., Yih., W-T. 2011. Clickthrough-based latent semantic models for web search. In SIGIR.
- Gao, J., Yuan, W., Li, X., Deng, K., and Nie, J-Y. 2009. Smoothing clickthrough data for web search ranking. In SIGIR.
- Gao, J., and He, X. 2013. Training MRF-based translation models using gradient ascent. In NAACL-HLT.
- Graves, A., Jaitly, N., and Mohamed, A., 2013a. Hybrid speech recognition with deep bidirectional LSTM, Proc. ASRU.
- Graves, A., Mohamed, A., and Hinton, G., 2013. Speech recognition with deep recurrent neural networks, Proc. ICASSP.

References

- He, X. and Deng, L., 2013. Speech-Centric Information Processing: An Optimization-Oriented Approach, in Proceedings of the IEEE.
- He, X. and Deng, L., 2012. Maximum Expected BLEU Training of Phrase and Lexicon Translation Models , ACL.
- He, X., Deng, L., and Chou, W., 2008. Discriminative learning in sequential pattern recognition, Sept. IEEE Sig. Proc. Mag.
- Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B., 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition, IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97.
- Hinton, G., and Salakhutdinov, R., 2010. Discovering binary codes for documents by learning deep generative models. Topics in Cognitive Science.
- Hu, Y., Auli, M., Gao, Q., and Gao, J. 2014. Minimum translation modeling with recurrent neural networks. In EACL.
- Huang, E., Socher, R., Manning, C, and Ng, A. 2012. Improving word representations via global context and multiple word prototypes, Proc. ACL.
- Huang, P., He, X., Gao, J., Deng, L., Acero, A., and Heck, L. 2013. Learning deep structured semantic models for web search using clickthrough data. In CIKM.
- Hutchinson, B., Deng, L., and Yu, D., 2012. A deep architecture with bilinear modeling of hidden representations: Applications to phonetic recognition, Proc. ICASSP.
- Hutchinson, B., Deng, L., and Yu, D., 2013. Tensor deep stacking networks, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 35, pp. 1944 - 1957.
- Kiros, R., Zemel, R., and Salakhutdinov, R. 2013. Multimodal Neural Language Models, Proc. NIPS Deep Learning Workshop.
- Koehn, P. 2009. Statistical Machine Translation. Cambridge University Press.
- Krizhevsky, A., Sutskever, I, and Hinton, G., 2012. ImageNet Classification with Deep Convolutional Neural Networks, NIPS.
- Le, H-S, Oparin, I., Allauzen, A., Gauvain, J-L., Yvon, F., 2013. Structured output layer neural network language models for speech recognition, IEEE Transactions on Audio, Speech and Language Processing.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. 1998. Gradient-based learning applied to document recognition, Proceedings of the IEEE, Vol. 86, pp. 2278-2324.
- Li, P., Hastie, T., and Church, K.. 2006. Very sparse random projections, in Proc. SIGKDD.
- Mesnil, G., He, X., Deng, L., and Bengio, Y., 2013. Investigation of Recurrent-Neural-Network Architectures and Learning Methods for Spoken Language Understanding, in Interspeech.
- Mikolov, T. 2012. Statistical Language Models based on Neural Networks, Ph.D. thesis, Brno University of Technology.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficient estimation of word representations in vector space, Proc. ICLR.
- Mikolov, T., Kombrink,. S., Burget, L., Cernocky, J., Khudanpur, S., 2011. Extensions of Recurrent Neural Network LM. ICASSP.
- Mikolov, T., Yih, W., Zweig, G., 2013. Linguistic Regularities in Continuous Space Word Representations. In NAACL-HLT.
- Mohamed, A., Yu, D., and Deng, L. 2010. Investigation of full-sequence training of deep belief networks for speech recognition, Proc. Interspeech.
- Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. 2011. Multimodal deep learning, Proc. ICML.

References

- Sainath, T., Mohamed, A., Kingsbury, B., and Ramabhadran, B. 2013. Convolutional neural networks for LVCSR, Proc. ICASSP.
- Salakhutdinov R., and Hinton, G., 2007 Semantic hashing. in Proc. SIGIR Workshop Information Retrieval and Applications of Graphical Models
- Sarikaya, R., Hinton, G., and Ramabhadran, B., 2011. Deep belief nets for natural language call-routing, in Proceedings of the ICASSP.
- Schwenk, H., Dchelotte, D., Gauvain, J-L., 2006. Continuous space language models for statistical machine translation, in COLING-ACL
- Seide, F., Li, G., and Yu, D. 2011. Conversational speech transcription using context-dependent deep neural networks, Proc. Interspeech
- Shen, Y., He, X., Gao, J., Deng, L., and Mesnil, G. 2014. Learning Semantic Representations Using Convolutional Neural Networks for Web Search, in Proceedings of WWW.
- Shen, Y., He, X., Gao, J., Deng, L., and Mesnil, G. 2014. A convolutional latent semantic model for web search. CIKM
- Socher, R., Huval, B., Manning, C., Ng, A., 2012. Semantic compositionality through recursive matrix-vector spaces. In EMNLP.
- Socher, R., Lin, C., Ng, A., and Manning, C. 2011. Learning continuous phrase representations and syntactic parsing with recursive neural networks, Proc. ICML.
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng A., and Potts. C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, Proc. EMNLP
- Song, X. He, X., Gao. J., and Deng, L. 2014. Learning Word Embedding Using the DSSM. MSR Tech Report.
- Song, Y., Wang, H., and He, X., 2014. Adapting Deep RankNet for Personalized Search. Proc. WSDM.
- Tur, G., Deng, L., Hakkani-Tur, D., and He, X., 2012. Towards Deeper Understanding Deep Convex Networks for Semantic Utterance Classification, in ICASSP.
- Wright, S., Kanevsky, D., Deng, L., He, X., Heigold, G., and Li, H., 2013. Optimization Algorithms and Applications for Speech and Language Processing, in IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 11.
- Xu, P., and Sarikaya, R., 2013. Convolutional neural network based triangular crf for joint intent detection and slot filling, in IEEE ASRU.
- Yao, K., Zweig, G., Hwang, M-Y. , Shi, Y., Yu, D., 2013. Recurrent neural networks for language understanding, submitted to Interspeech.
- Yann, D., Tur, G., Hakkani-Tur, D., Heck, L., 2014. Zero-Shot Learning and Clustering for Semantic Utterance Classification Using Deep Learning, in ICLR.
- Yih, W., Toutanova, K., Platt, J., and Meek, C. 2011. Learning discriminative projections for text similarity measures. In CoNLL.
- Yih, W., He, X., Meek, C. 2014. Semantic Parsing for Single-Relation Question Answering, in ACL.
- Zeiler, M. and Fergus, R. 2013. Visualizing and understanding convolutional networks, arXiv:1311.2901, pp. 1-11.