NLPCC-ADL '17
Dalian, China
Nov. 9, 2017

# Deep Learning for Semantic Matching in Search

Jun Xu
junxu@ict.ac.cn

ICT, CAS

# Tutorial Talk @ ADL 52 & NLPCC 2014

ADL & NLPCC 2014 Tutorial
December 6, 2014
ShenZhen China

## Semantic Matching in Search


Jun Xu
junxu@ict.ac.cn

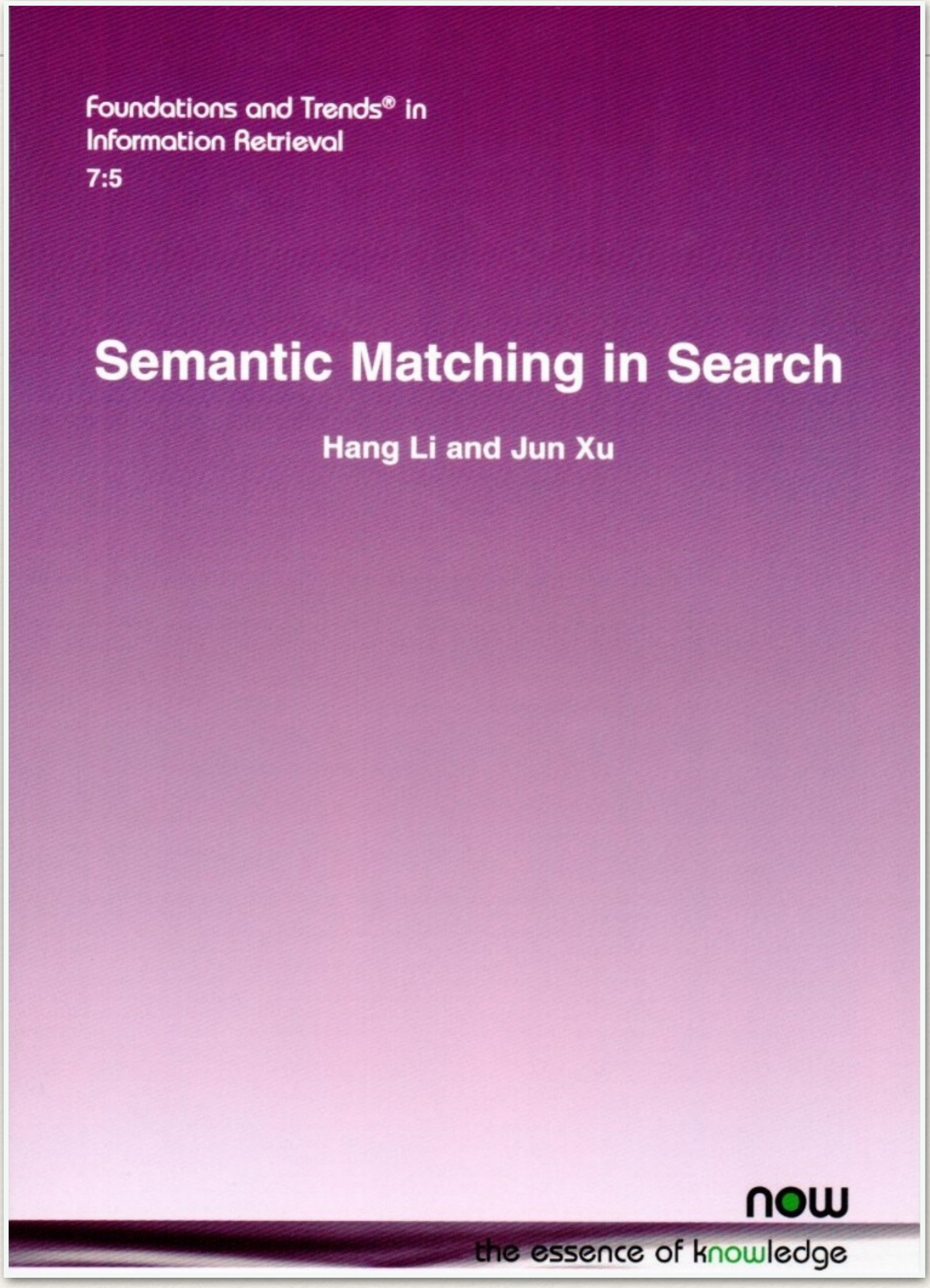
Institute of Computing Technology
Chinese Academy of Sciences

# Discussed Traditional Approaches to Semantic Matching in Search

## Outline of Tutorial

- Semantic Matching between Query and Document
- Approaches to Semantic Matching
  1. Matching by Query Reformulation
  2. Matching with Term Dependency Model
  3. Matching with Translation Model
  4. Matching with Topic Model
  5. Matching with Latent Space Model
- Summary

# Details Introduced in a Monograph




http://www.nowpublishers.com/articles/foundations-and-trends-in-information-retrieval/INR-035
http://www.hangli-hl.com/uploads/3/1/6/8/3168008/ml_for_match-step2.pdf

# Growing Interest in "Deep" IR in the Past Three Years

- Success of deep learning in other fields
  - Speech recognition, computer vision, and NLP
- Growing presence of deep learning in IR research
  - SIGIR 2016 keynote, Tutorial, and Neu-IR workshop
- Adopted by industry
  - ACM News: Google Turning its Lucrative Web Search Over to AI Machines (Oct. 26, 2015)
  - WIRED: AI is Transforming Google Search. The Rest of the Web is Next (April 2, 2016)
- Chris Manning (Stanford)'s SIGIR keynote:
  "I'm certain that *deep learning will come to dominate SIGIR over the next couple of years* … just like speech, vision, and NLP before it."

# “Deep” Semantic Matching also Gain a Lot of Attention

- Before 2014, a few studies, e.g.,
  - Paraphrase detection [Socher et al., 2011]
  - Ad-hoc retrieval (DSSM)[Huang et al., 2013]
- 2014 ~ 2017, a lot of studies (as summarized in the tutorial)
  - Paraphrase identification
  - Ad-hoc retrieval
  - Question answering
  - Dialog
  - Result diversification

  ……

This tutorial:

Update the survey with newly developed deep matching methods

# Outline

- Semantic matching in search
- Word-level matching: bridging the semantic gap
- Sentence-level matching: capturing the proximity
- Summary and discussion

# A Good Web Search Engine

- Must be good at
    - Relevance
    - Coverage
    - Diversity
    - Freshness
    - Response time
    - User interface……
- Relevance is particularly important

# Query-Document Mismatch Challenge

**Table 1.1:** Examples of query document mismatch.

| query | document | term match | semantic match |
|---|---|---|---|
| seattle best hotel | seattle best hotels | partial | yes |
| pool schedule | swimming pool schedule | partial | yes |
| natural logarithm transform | logarithm transform | partial | yes |
| china kong | china hong kong | partial | no |
| why are windows so expensive | why are macs so expensive | partial | no |

# Why Query-Document Mismatch Happens?

- Search is still mainly based on term-level matching signals
- Some search intent can be represented by different queries (representations)
- Query document mismatch occurs, when searcher and author use different terms (representations) to describe the same concept

# Same Search Intent, Different Query Representations

**Table 1.2:** Queries about "distance between sun and earth".

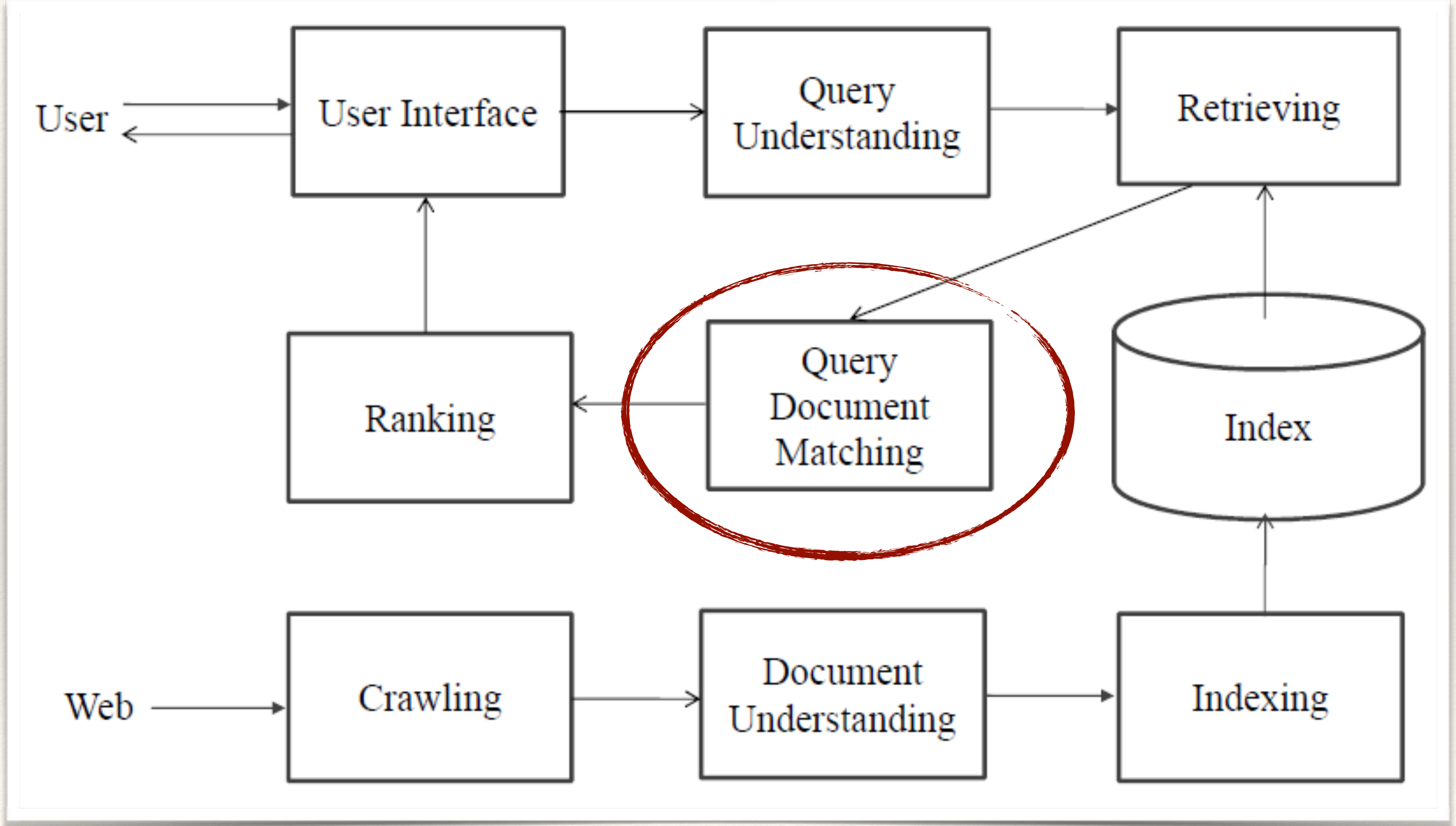
| | |
|---|---|
| "how far" earth sun | average distance from the earth to the sun |
| "how far" sun | how far away is the sun from earth |
| average distance earth sun | average distance from earth to sun |
| how far from earth to sun | distance from earth to the sun |
| distance from sun to earth | distance between earth and the sun |
| distance between earth & sun | distance between earth and sun |
| how far earth is from the sun | distance from the earth to the sun |
| distance between earth sun | distance from the sun to the earth |
| distance of earth from sun | distance from the sun to earth |
| "how far" sun earth | how far away is the sun from the earth |
| how far earth from sun | distance between sun and earth |
| how far from earth is the sun | how far from the earth to the sun |
| distance from sun to the earth | |

# Same Search Intent, Different Query Representations

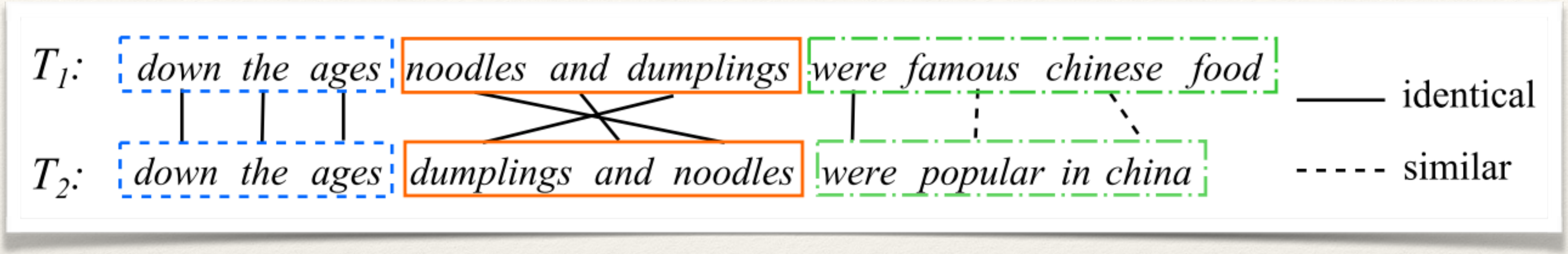
**Table 1.3:** Queries about "Youtube".

| | | |
|---|---|---|
| yutube | yuotube | yuo tube |
| ytube | youtubr | yu tube |
| youtubo | youtuber | youtubecom |
| youtube om | youtube music videos | youtube videos |
| youtube | youtube com | youtube co |
| youtub com | you tube music videos | yout tube |
| youtub | you tube com yourtube | your tube |
| you tube | you tub | you tube video clips |
| you tube videos | www you tube com | wwww youtube com |
| www youtube | www youtube com | www youtube co |
| yotube | www you tube | www utube com |
| ww youtube com | www utube | www u tube |
| utube videos | utube com | utube |
| u tube com | utub | u tube videos |
| u tube | my tube | toutube |
| outube | our tube | toutube |

# Semantic Matching in Search



# Challenges

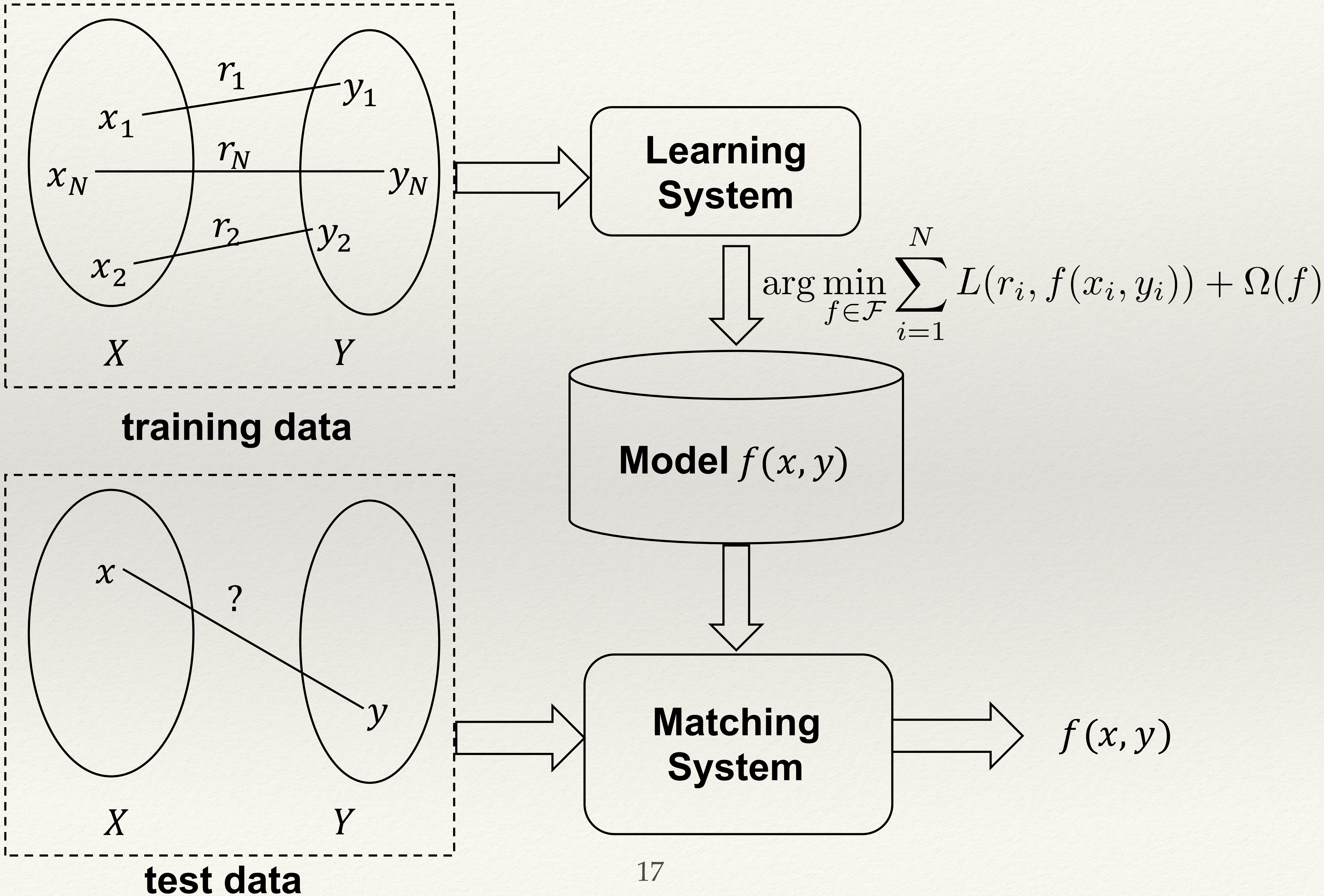



- Word-level matching: semantic gap between words
  - Two words has similar meanings
  - "popular" ~ "famous"; "china" ~ "chinese"
- Sentence-level: proximity matching between sentences
  - The matching positions do matter
  - "noodles and dumplings" ~ "dumplings and noodles"
- Need to consider them simultaneously

Ideally: Understanding the Natural Language

Current Approaches: Avoid Understanding and Conduct Matching

# Learning to Match

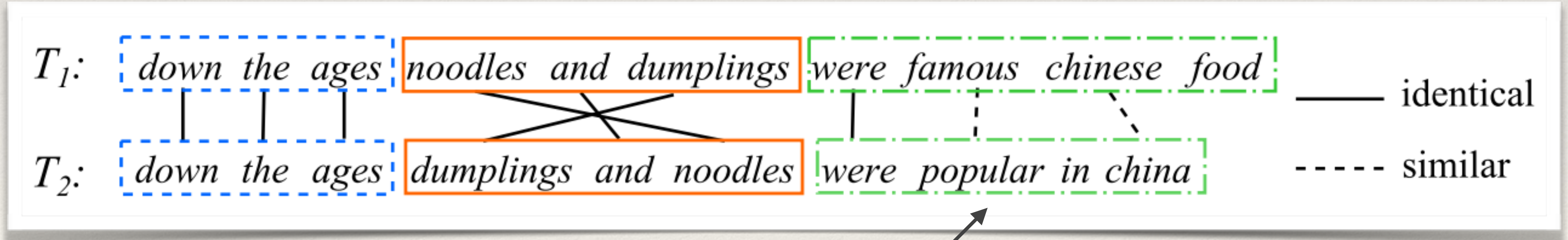


# Why Deep?

- Representation
  - Word: one hot  distributed
  - Sentence: bag-of-words —> distributed representation
  - Better representation ability, better generalization ability
- Matching function
  - Inputs (features): handcrafted —> automatically learned
  - Function: simple functions (e.g., cosine, dot product) —> nonlinear neural networks
  - Involving richer matching signals
  - Considering soft matching patterns

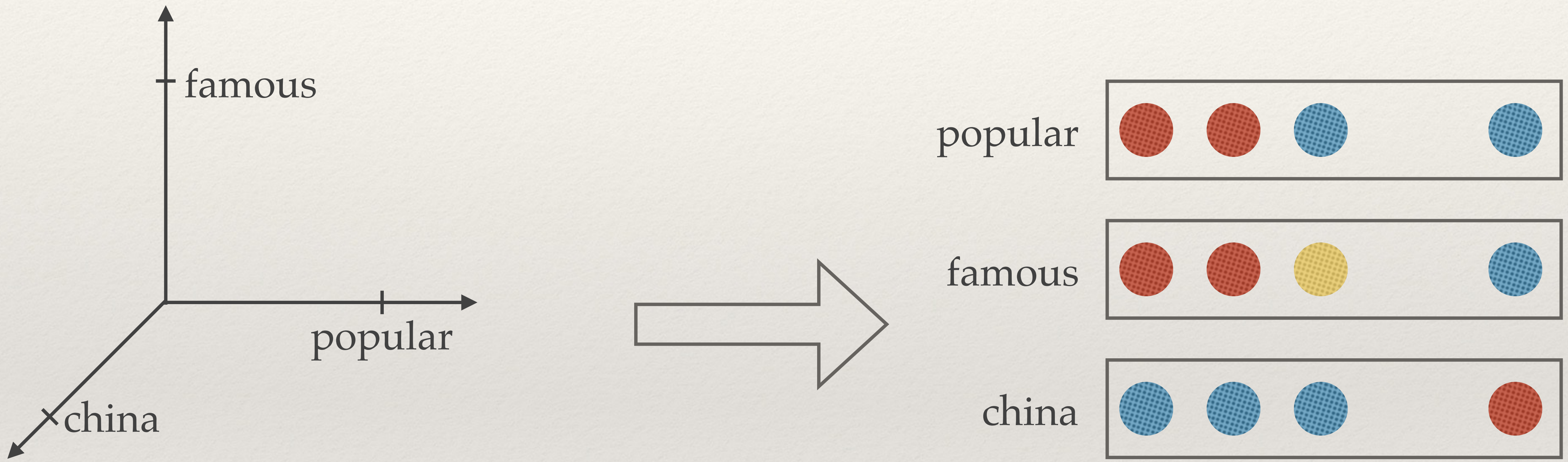
# Outline

- Semantic matching in search
- Word-level matching: bridging the semantic gap
- Sentence-level matching: capturing the proximity
- Summary and discussion




Measure the similarity between "famous" and "popular"

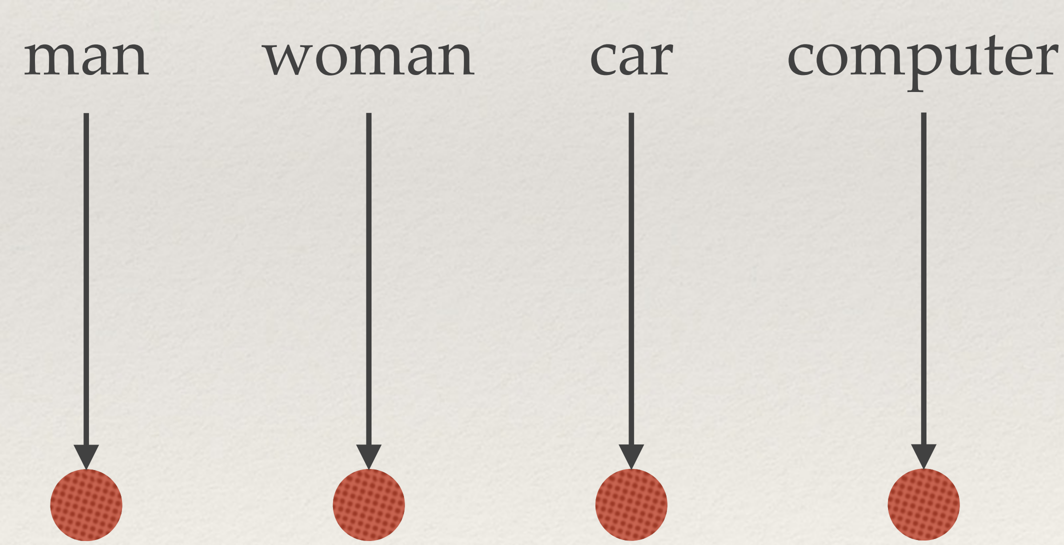
# From Local to Distributed Representations



# Local Representation of Words

- Words are building blocks of queries/documents
- Conventional IR models considers words as atomic symbols, also known as "one-hot" or local representations

| Local(One-Hot) Representation | |
|---|---|
| man | [1,0,...,0,0,...,0,0] |
| woman | [0,1,...,0,0,...,0,0] |
| car | [0,0,...,1,0,...,0,0] |
| automobile | [0,0,...,0,0,...,1,0] |




Each word is locally represented by a distinct node

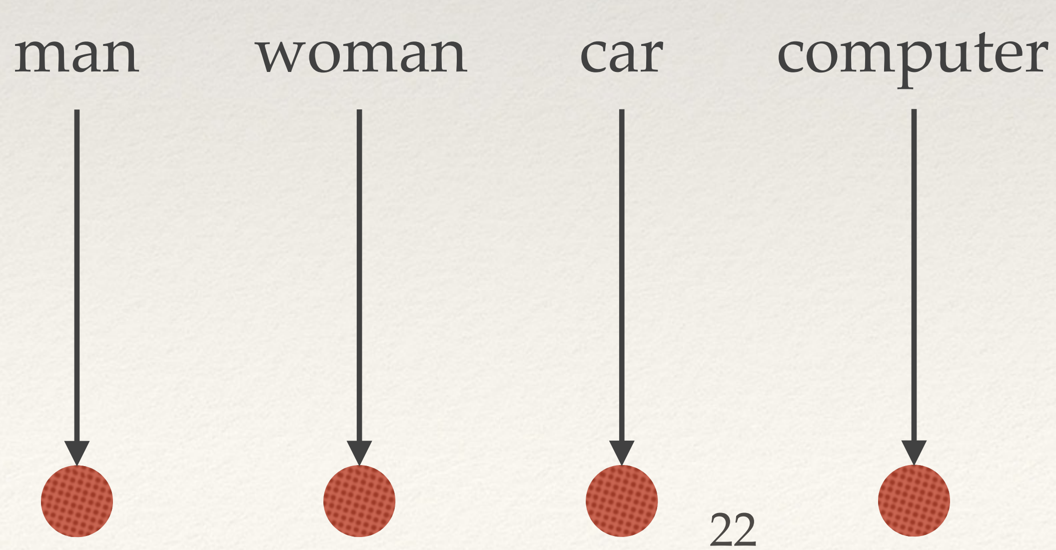
# Limitation of Local Representations

- Independent assumption

| Local (one-hot) representations | |
|---|---|
| man | [1,0,...,0,0,...,0,0] |
| woman | [0,1,...,0,0,...,0,0] |
| car | [0,0,...,1,0,...,0,0] |
| automobile | [0,0,...,0,0,...,1,0] |

cos(man, woman) = 0
cos(man, automobile) = 0

- Inefficient: N dimensions for N words



# Limitation of Local Representations (cont')

- Poor generalization ability
- Using language modeling as an example
  - Cannot generalize to unseen bigram "three groups"

Doc1: There are three teams left for the qualification

Doc2: Four teams have passed the first round

Doc3: Four groups are playing in the field

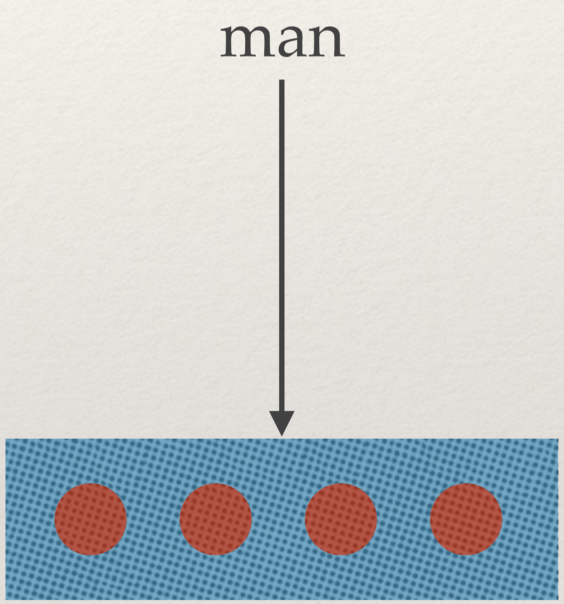
$$
\begin{aligned}
P(\text{teams}|\text{three}) &> 0 \\
P(\text{teams}|\text{four}) &> 0 \\
P(\text{groups}|\text{four}) &> 0
\end{aligned}
$$

$$
P(\text{groups}|\text{three}) = 0
$$

# Distributional Representation of Words

- Each word is represented by a low-dimensional dense vector

| Distributed representations | |
|---|---|
| man | [0.326172, . . . , 0.00524902, . . . , 0.0209961] |
| woman | [0.243164, . . . , −0.205078, . . . , −0.0294189] |
| car | [0.0512695, . . . , −0.306641, . . . , 0.222656] |
| automobile | [0.107422, . . . , −0.0375977, . . . , −0.0620117] |




Hinton, G. E., et al. Distributed representations. In Rumelhart, D. E., McClelland, J. L., and PDP Research Group, C., editors, Parallel Distributed Processing: Explorations in the Microstructur of Cognition, Vol. 1,1986, pages 77–109. MIT Press, Cambridge, MA, USA.

# Advantages of Distributed Representations

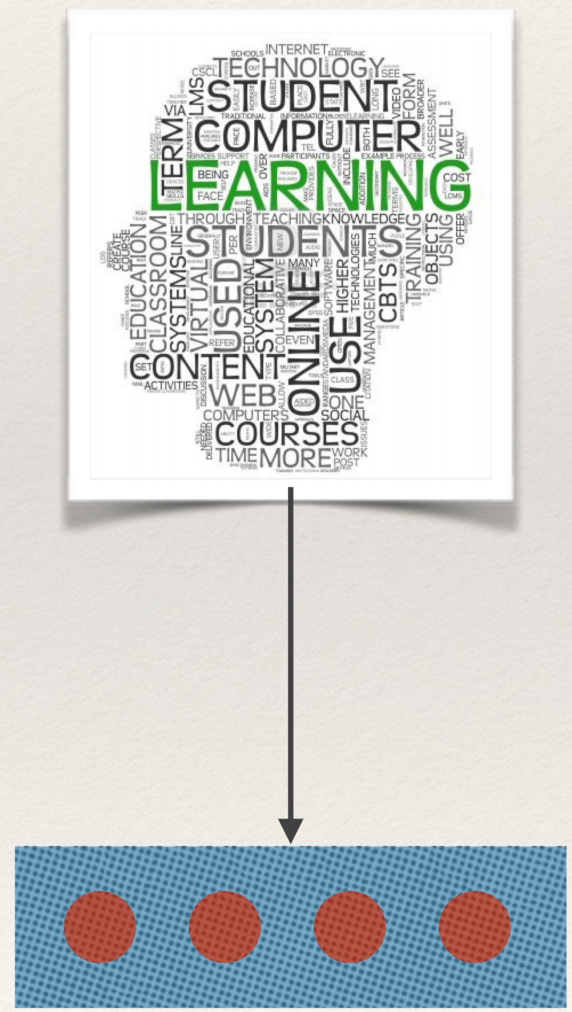
- Beyond the independent assumption

| Distributed representations | |
|---|---|
| man | [0.326172, . . . , 0.00524902, . . . , 0.0209961] |
| woman | [0.243164, . . . , −0.205078, . . . , −0.0294189] |
| car | [0.0512695, . . . , −0.306641, . . . , 0.222656] |
| automobile | [0.107422, . . . , −0.0375977, . . . , −0.0620117] |

$$\begin{aligned} \cos(\text{man}, \text{woman}) &= 0.77 \\ \cos(\text{man}, \text{automobile}) &= 0.25 \end{aligned}$$

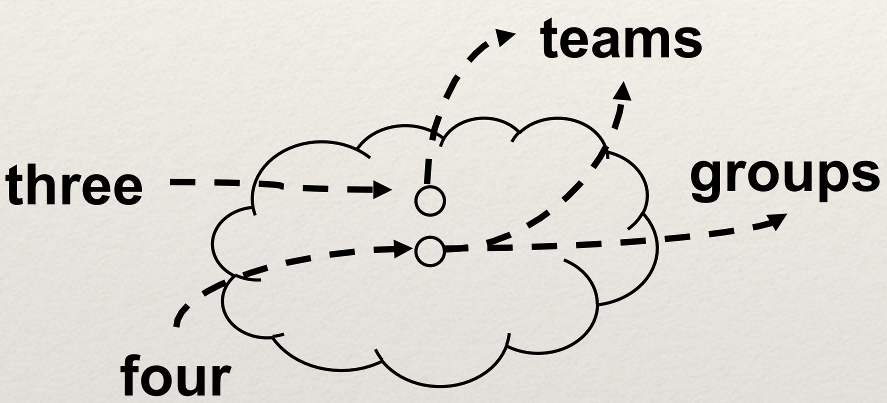
# Advantages of Distributed Representations (cont')

- Efficient word representations: N dimensions can represent $2^N$ words



# Advantages of Distributed Representations (cont')

- Better generalization ability: semantically similar words are mapped to nearby points




Doc1: There are three teams left for the qualification

Doc2: Four teams have passed the first round

Doc3: Four groups are playing in the field

- Assigning probability to unseen bigram "three groups" $P(\text{groups} \mid \text{three}) > 0$

Language modeling with distributed word representations can assign probabilities to unseen bigrams **according to their semantics**

"You shall know a word by the company it keeps!"

*–J. R. Firth (1957)*

"Words that occur in the same context tends to have similar meanings."

*–Zelling Harris (1954)*

# What is the Meaning of "badiwac"?

He handed her a glass of bardiwac.

Beef dishes are made to complement the bardiwacs.

Nigel staggered to his feet, face flushed from too much bardiwac.

Malbec, one of the lesser-known bardiwac grapes, responds well to Australia's sunshine.

I dined off bread and cheese and this excellent bardiwac.

The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.

- A red alcoholic beverage made from grapes

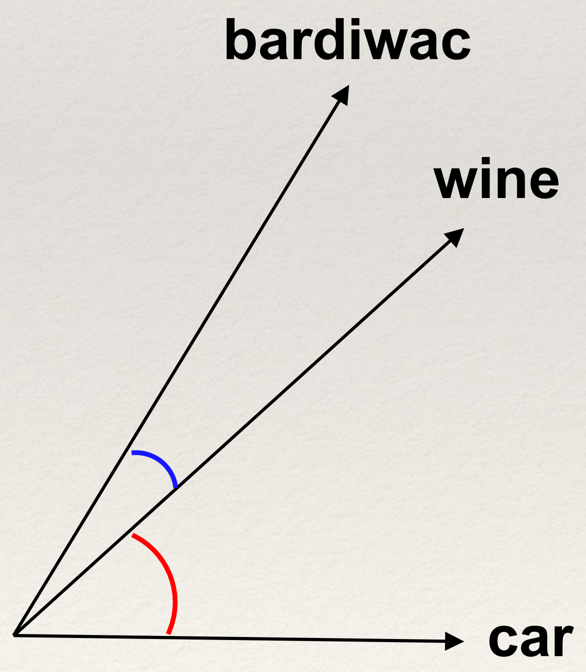
From Stefan Evert. Distributional Semantic Models . NAACL-HLT 2010 Tutorial

# Surrounding Words

| | | |
|---|---|---|
| Just checking on the | bardiwac | he boomed as he come back |
| I hope you'll take to a good French | bardiwac | chimed in Arthur Iverson jovially |
| our host did slip out to attend to the | bardiwac | that was before the shrimp |
| Iverson did when he went through to see to | bardiwac | before dinner. Henry rubbed his hands |
| and drinking red win from France -- sour | bardiwac | , which bad proved hard to sell. |
| eyes were alight and he was drinking the | bardiwac | down like water. It is like Hallow-fair |
| quizzically at him and offering him some more | bardiwac | . He shook his head. 'I will sleep |
| drinks (as Queen Victoria reputedly did with | bardiwac | and malt whisky), but still the result |
| do we really 'wash down' a good meal with | bardiwac | ? Port is immediately suggested by Stilton |
| completely different: cheap and cheerful | bardiwac | . Two good examples from Victoria Wine are |
| examples from Victoria Wine are its house | bardiwac | , juicy and touch almondy, a good buy |
| opened a bottle of rather rust-coloured | bardiwac | . I ate too much and drank nearly three-quaters |
| elections, it was apparent the SDP of ' | bardiwac | and chips' mould-breaking fame at the time |
| the black hills. Not a night of vintage | bardiwac | . burnley: Pearce, Measham, McGrory |
| SONS Old School – the Marlborian navy, | bardiwac | and slim-white stripe. Heavy woven silk |
| white-hot passion, We are like a good bottle of | bardiwac | ; we both have sediment in our shoes |
| few minutes later he was uncorking a fine | bardiwac | in Masha's room, saying he had something |
| the phone, Surkov silently offered me more | bardiwac | but I indicated a bottle of Perrier |

# Word-Word Co-occurrence

| | glass | drink | grape | rex | meal |
|---|---|---|---|---|---|
| bardiwac | 10 | 22 | 43 | 16 | 29 |
| wine | 14 | 10 | 4 | 15 | 45 |
| car | 5 | 0 | 0 | 10 | 0 |



# Two Interpretations of Distributed Hypothesis

syntagmatic

Albert Einstein was a physicist.

Paradigmatic

Richard Feynman was a physicist.

syntagmatic

- **Syntagmatic**: words co-occur in the same text region
- **Paradigmatic**: words occur in the same context, may not at the same time

Sahlgren, M. (2008). The distributional hypothesis. Italian Journal of Linguistics, 20(1):33–54.
Fei Sun et al. Learning Word Representations by Jointly Modeling Syntagmatic and Paradigmatic Relations. In Proceedings of ACL. 2015, 136–145

# Modeling the Syntagmatic Relation

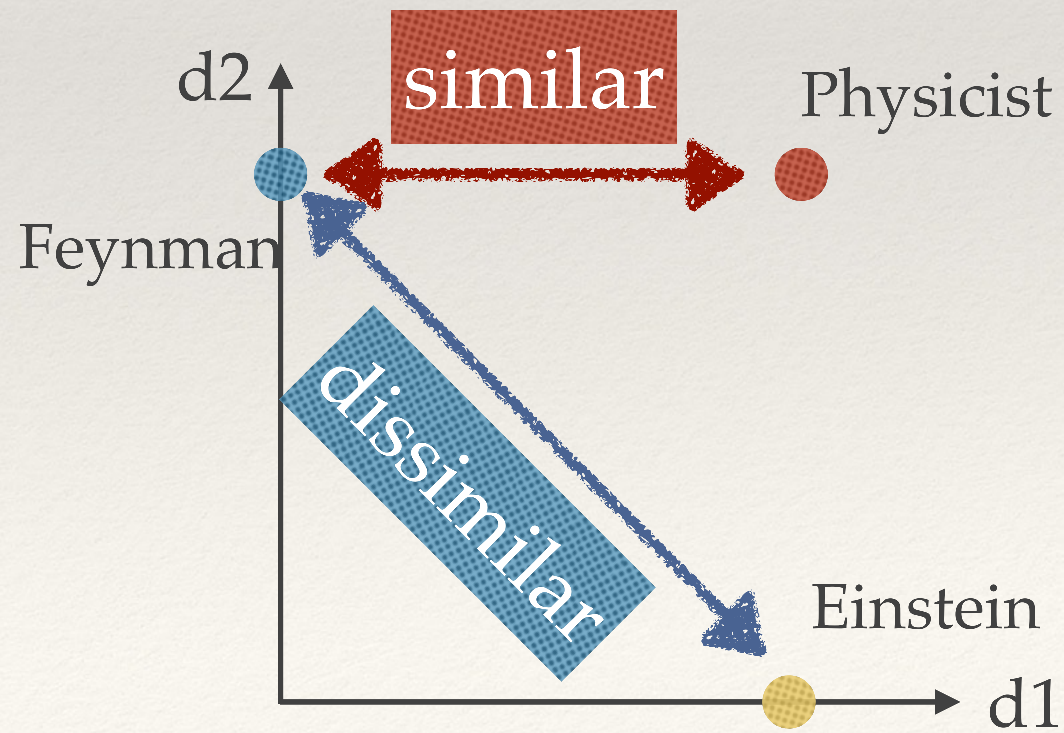
syntagmatic

Albert Einstein was a physicist.

Richard Feynman was a physicist.

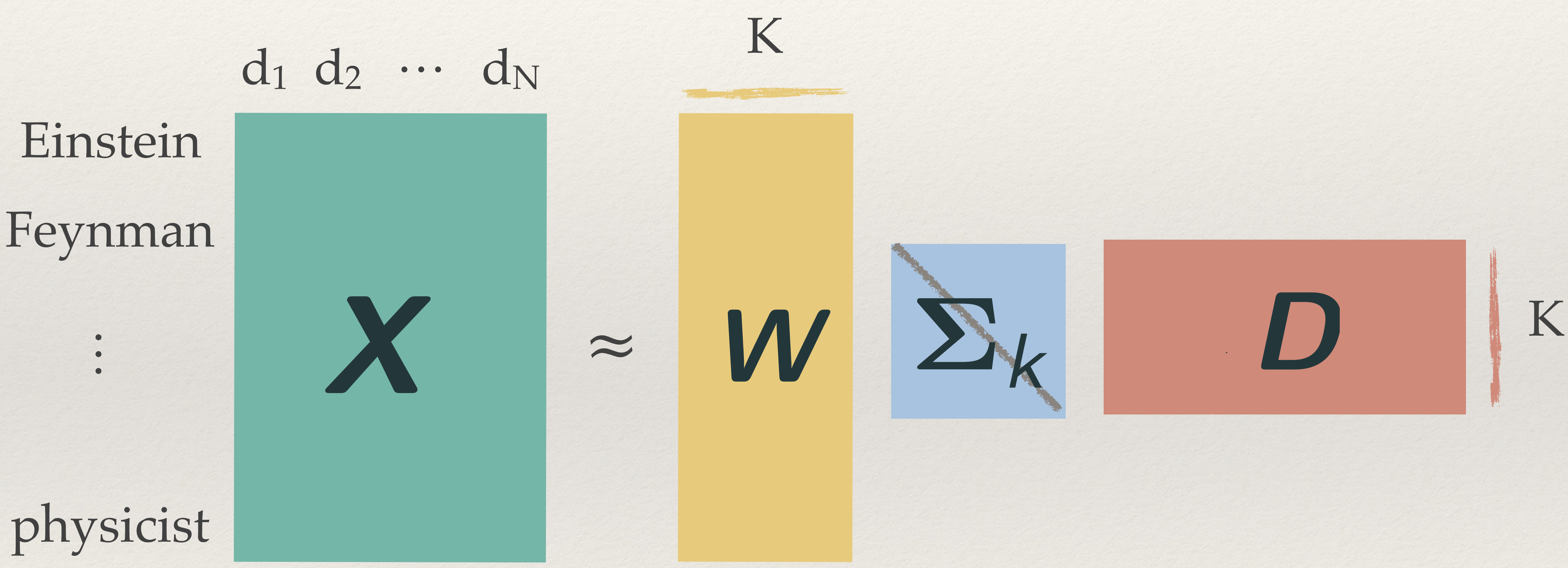
syntagmatic

| Word-document co-occurrence matrix (words represented by documents) | | |
|---|---|---|
| | d1 | d2 |
| Einstein | 1 | 0 |
| Feynman | 0 | 1 |
| Physicist | 1 | 1 |



# Modeling Syntagmatic Relation – LSI

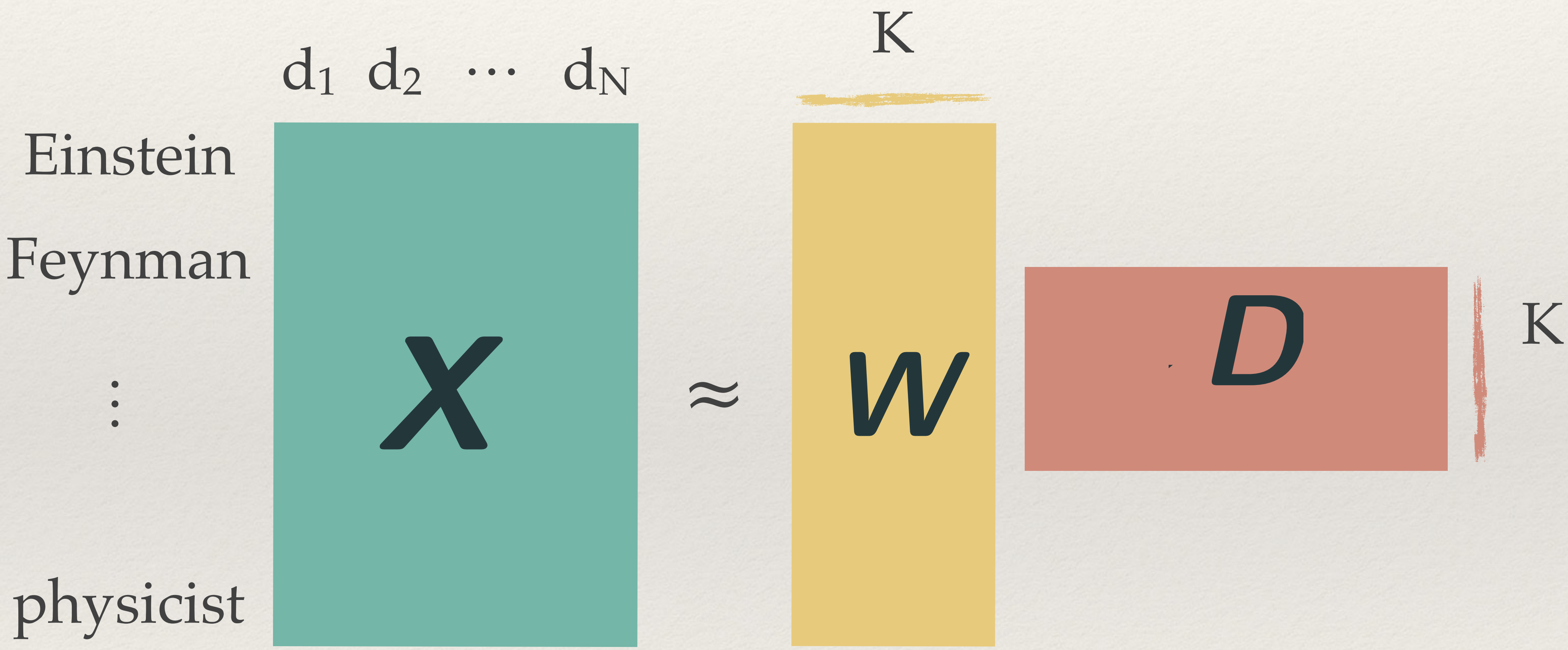
- Rank-reduced SVD of document-word co-occurrence matrix




$$\mathbf{X} \approx W \Sigma_k D^T$$

Deerwester et al. Indexing by latent semantic analysis. Journal of the American Society for Information Science. 1990, 41(6): 391–407.

# Modeling Syntagmatic Relation – NMF

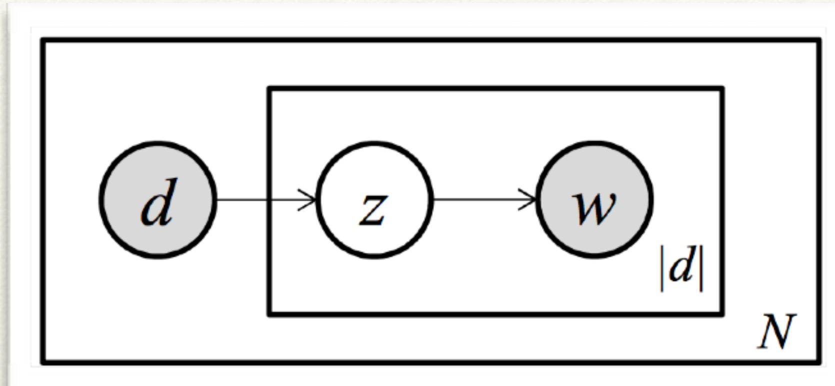



$$\mathbf{X} \approx W D^T$$

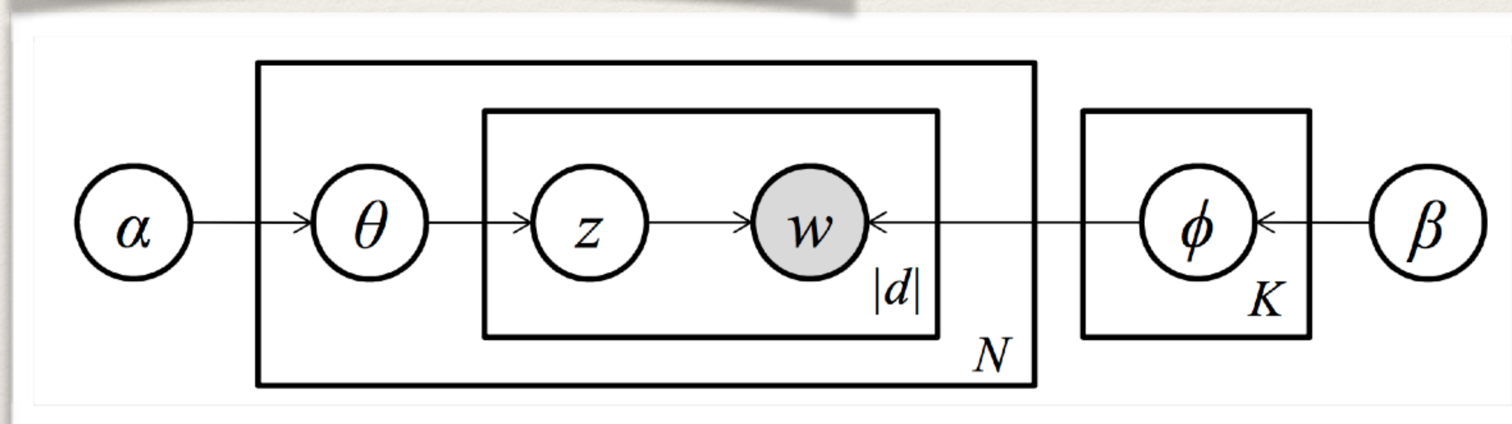
Lee, Daniel D., and H. Sebastian Seung. "Algorithms for non-negative matrix factorization." *Advances in neural information processing systems*. 2001

# Modeling Syntagmatic Relation – PLSA and LDA

PLSA



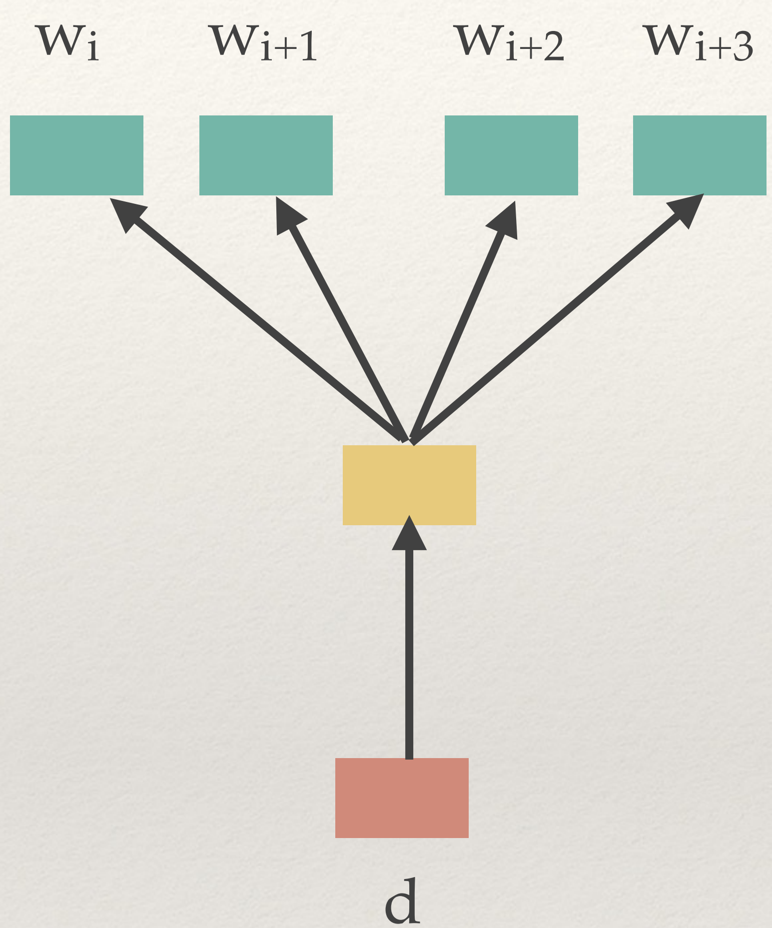

LDA




- Maximum likelihood solution of PLSA is NMF with KL divergence

Eric Gaussier, and Cyril Goutte. Relation between PLSA and NMF and implications. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005

# Modeling Syntagmatic Relation – Distributed Bag of Words Version of Paragraph Vector (PV-DBOW)




Predict word vector using document vector.

Quoc Leand Tomas Mikolov. Distributed Representations of Sentences and Documents. ICML 2014, 1188-1196.

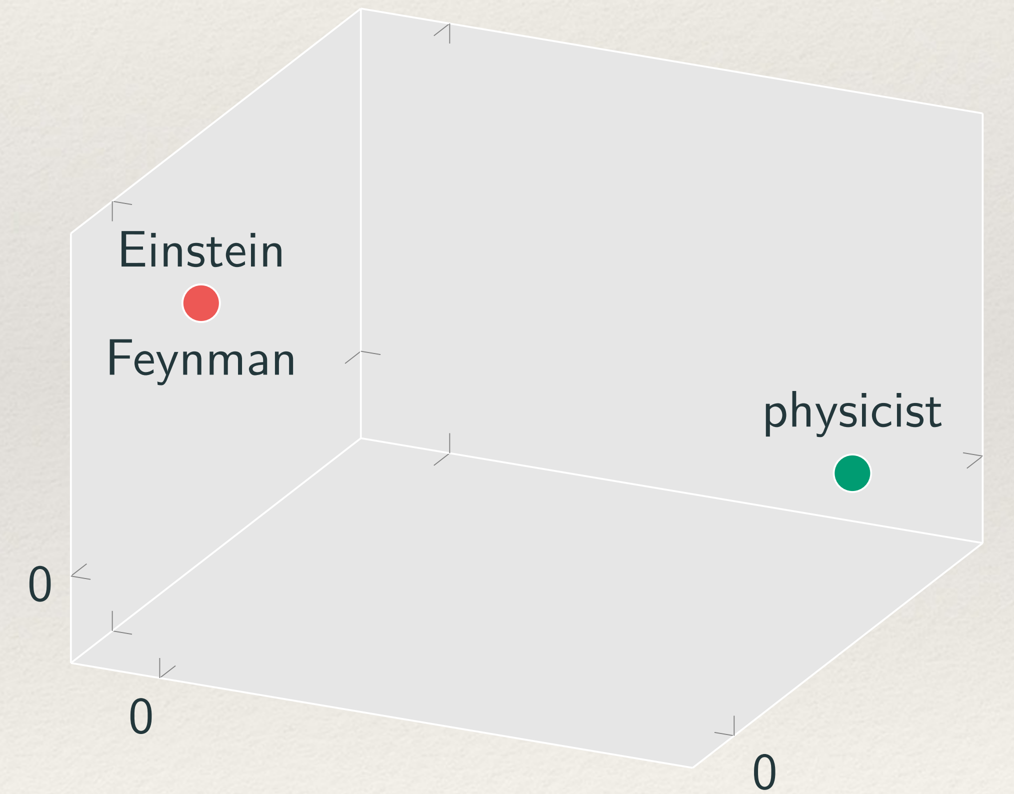
# Modeling the Paradigmatic Relation

Albert Einstein was a physicist.

Paradigmatic

Richard Feynman was a physicist.

| Word-word co-occurrence matrix (words represented by other words) | | | |
|---|---|---|---|
| | Einstein | Feynman | Physicist |
| Einstein | 0 | 0 | 1 |
| Feynman | 0 | 0 | 1 |
| Physicist | 1 | 1 | 0 |

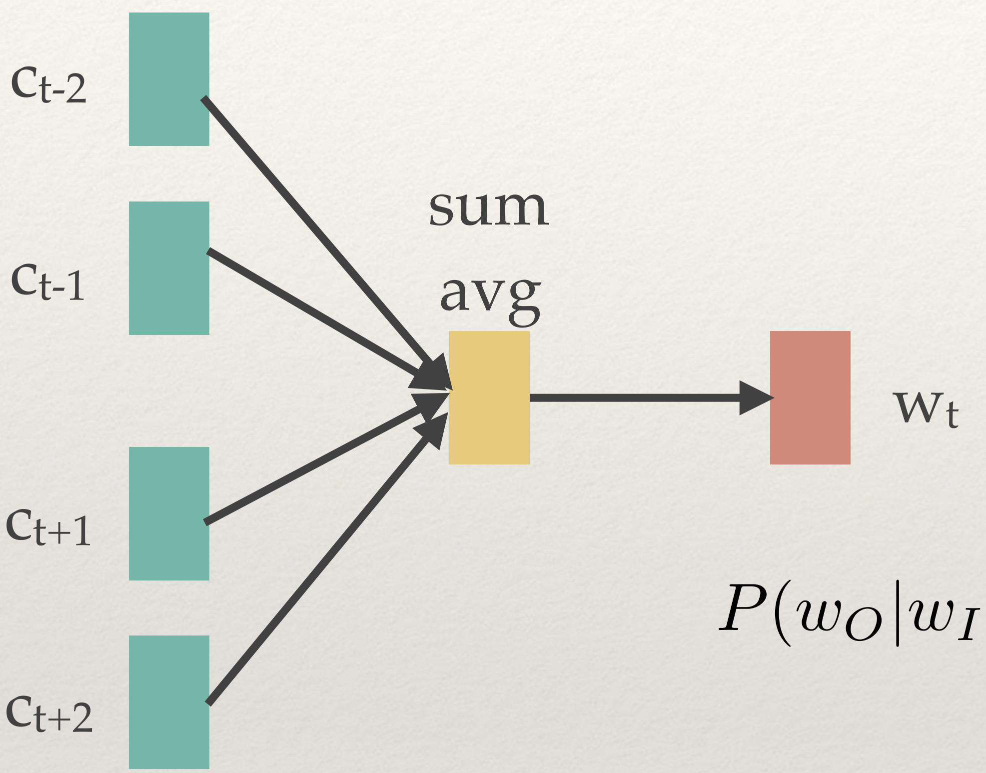


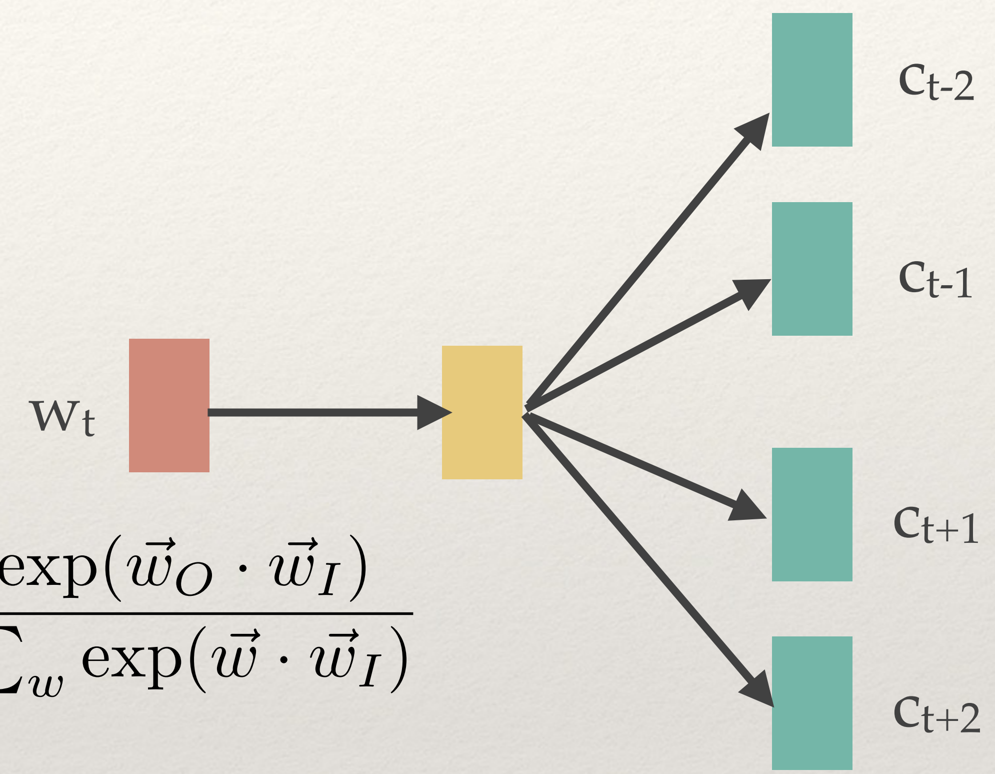

More suitable for learning the embeddings from short documents.



d1

# Modeling the Paradigmatic Relation – Word2Vec



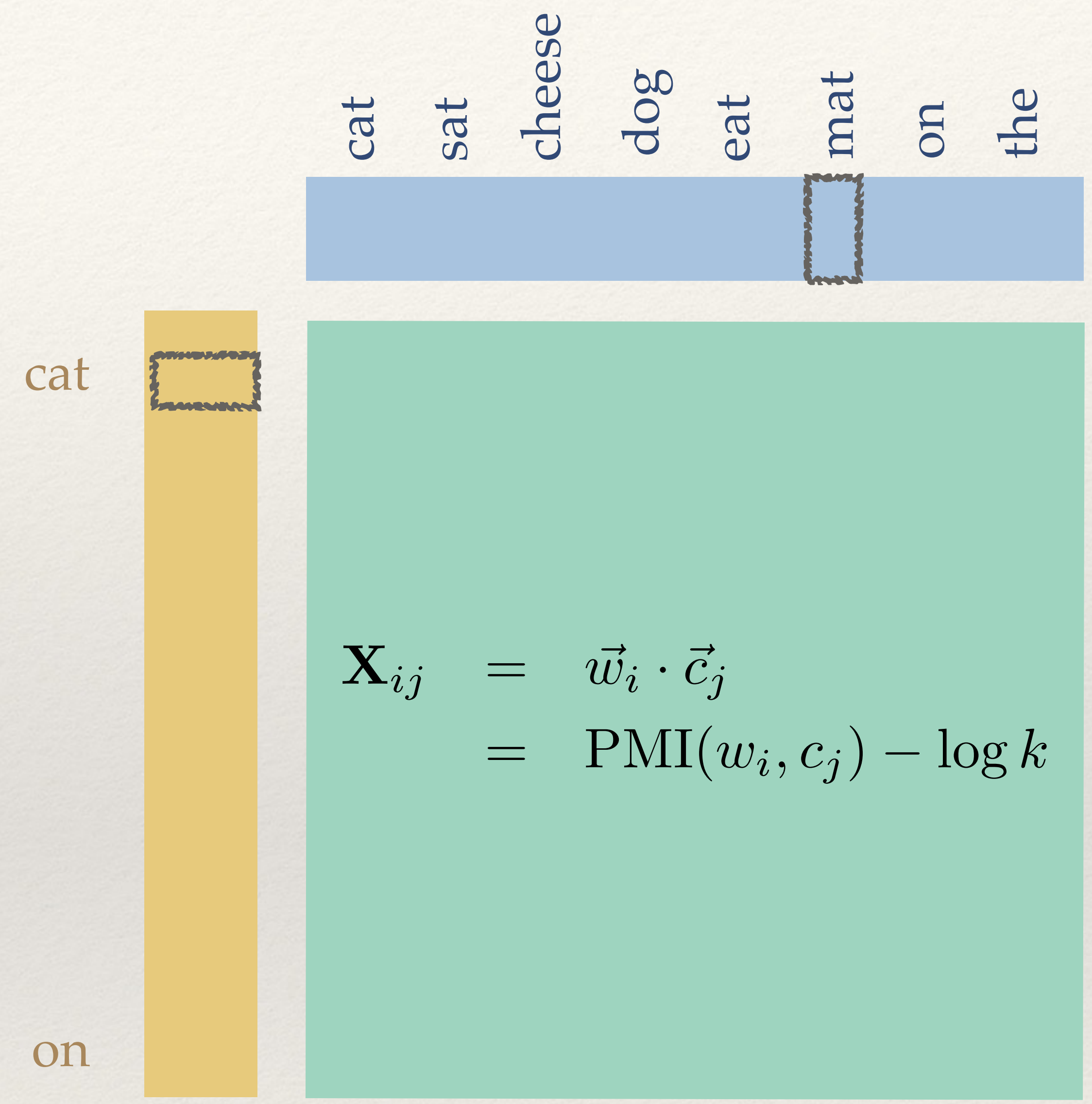




$$P(w_O|w_I) = \frac{\exp(\vec{w}_O \cdot \vec{w}_I)}{\sum_w \exp(\vec{w} \cdot \vec{w}_I)}$$

- Usually optimized with negative sampling

# Word Embedding as Matrix Factorization

- Skip-gram negative sampling (with sampling values k > 1) is factorizing the shifted point wise mutual information (PMI) matrix




$$\begin{aligned}\mathbf{X}_{ij} &= \vec{w}_i \cdot \vec{c}_j \\ &= \mathrm{PMI}(w_i, c_j) - \log k\end{aligned}$$

the **cat** sat on the **mat**

Levy, M. and Goldberg, Y. Neural word embedding as implicit matrix factorization. In NIPS. 2014, 2177–2185.

# Syntagmatic v.s. Paradigmatic

Syntagmatic

Albert Einstein was a physicist.

Paradigmatic

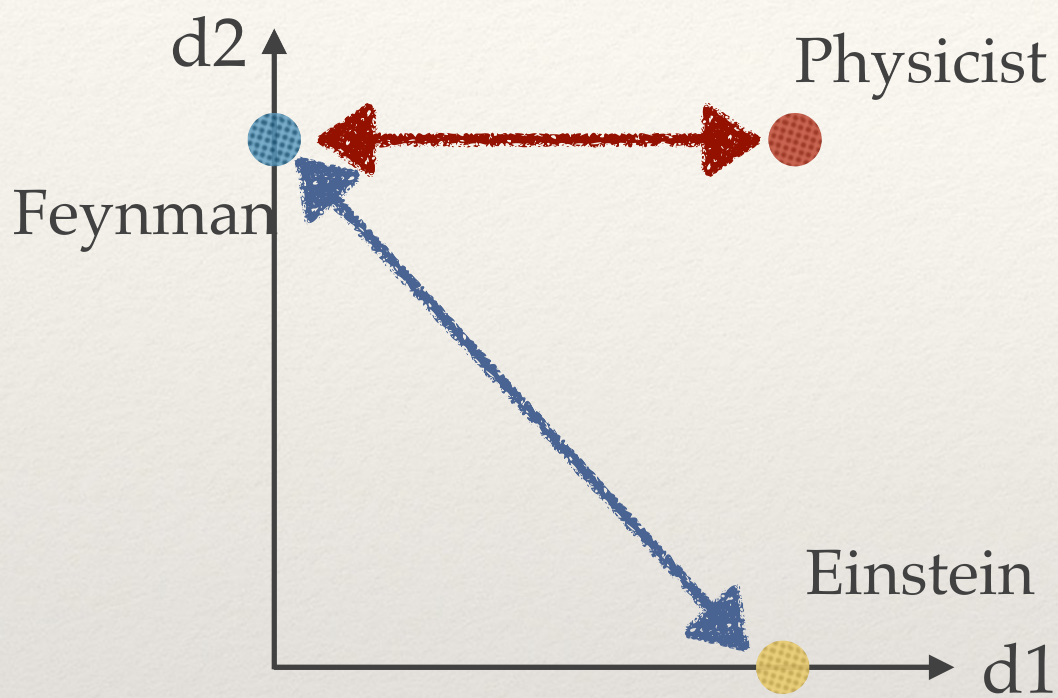
Richard Feynman was a physicist.

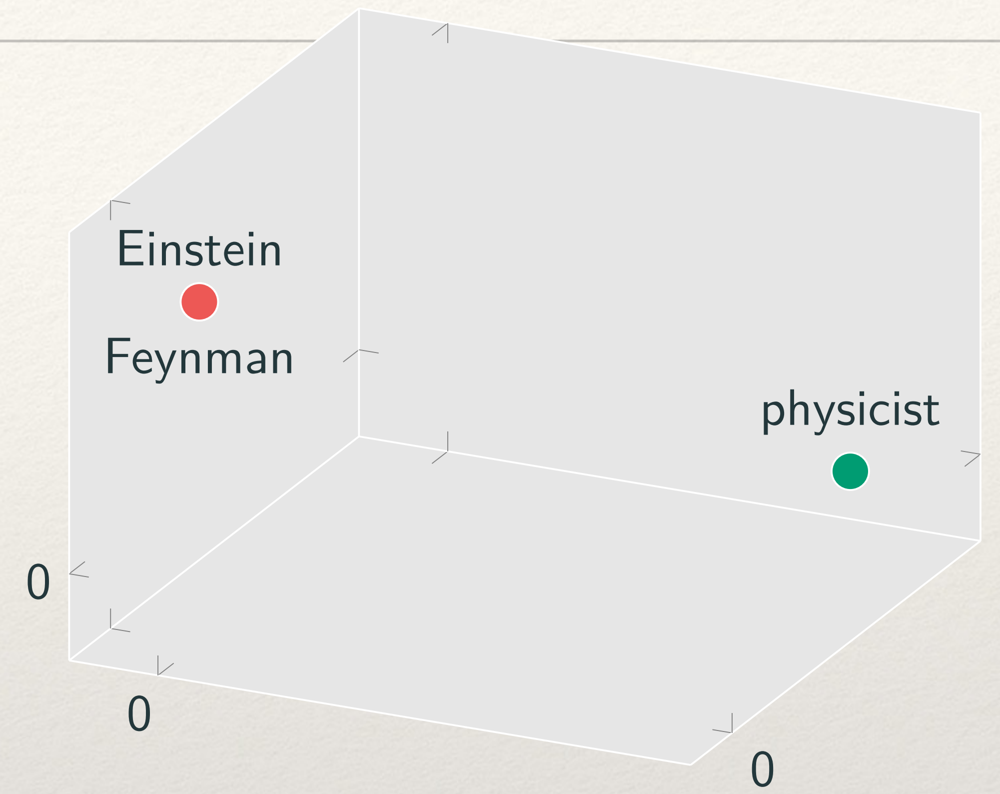
Syntagmatic

| Word-document co-occurrence matrix (words represented by documents) | | |
|---|---|---|
| | d1 | d2 |
| Einstein | 1 | 0 |
| Feynman | 0 | 1 |
| Physicist | 1 | 1 |

| Word-word co-occurrence matrix (words represented by other words) | | | |
|---|---|---|---|
| | Einstein | Feynman | Physicist |
| Einstein | 0 | 0 | 1 |
| Feynman | 0 | 0 | 1 |
| Physicist | 1 | 1 | 0 |

# Syntagmatic v.s. Paradigmatic (cont')

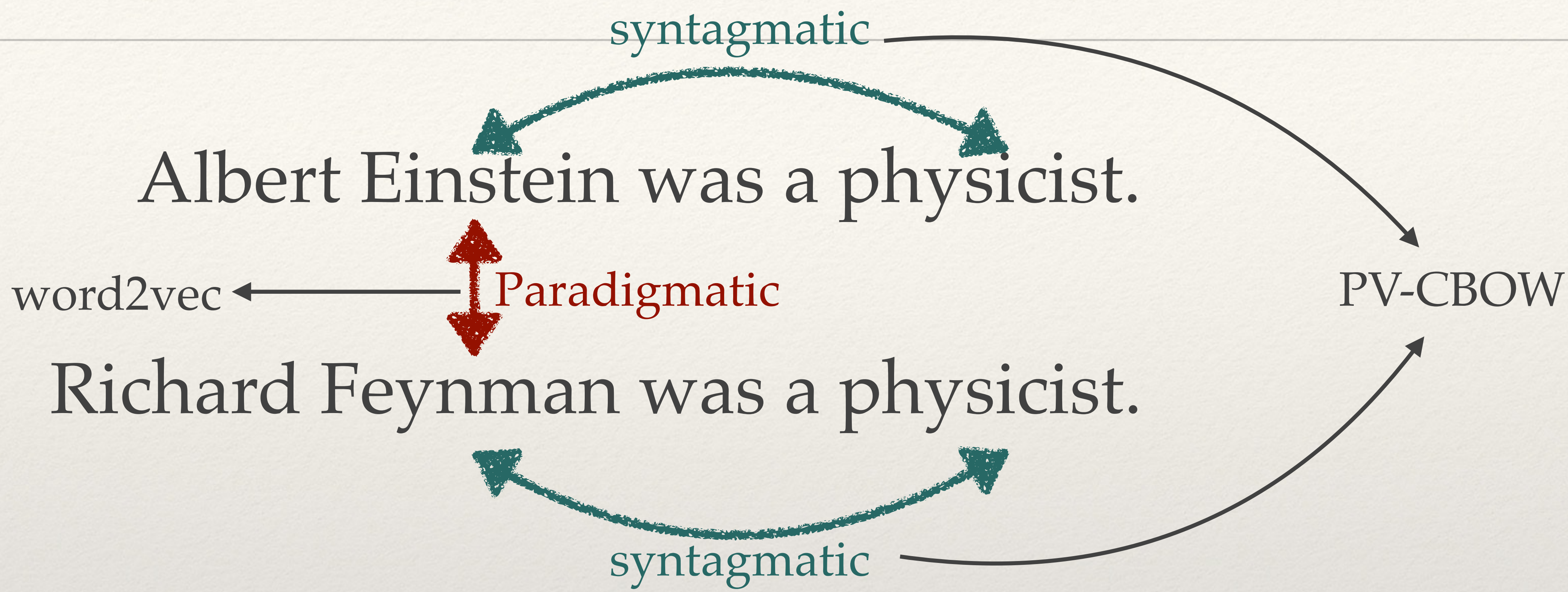



Similar words to "Feynman"

| Syntagmatic | Paradigmatic |
|---|---|
| quantum | Einstein |
| physicist | Schwinger |
| electrodynamics | Bethe |
| relativity | Bohm |

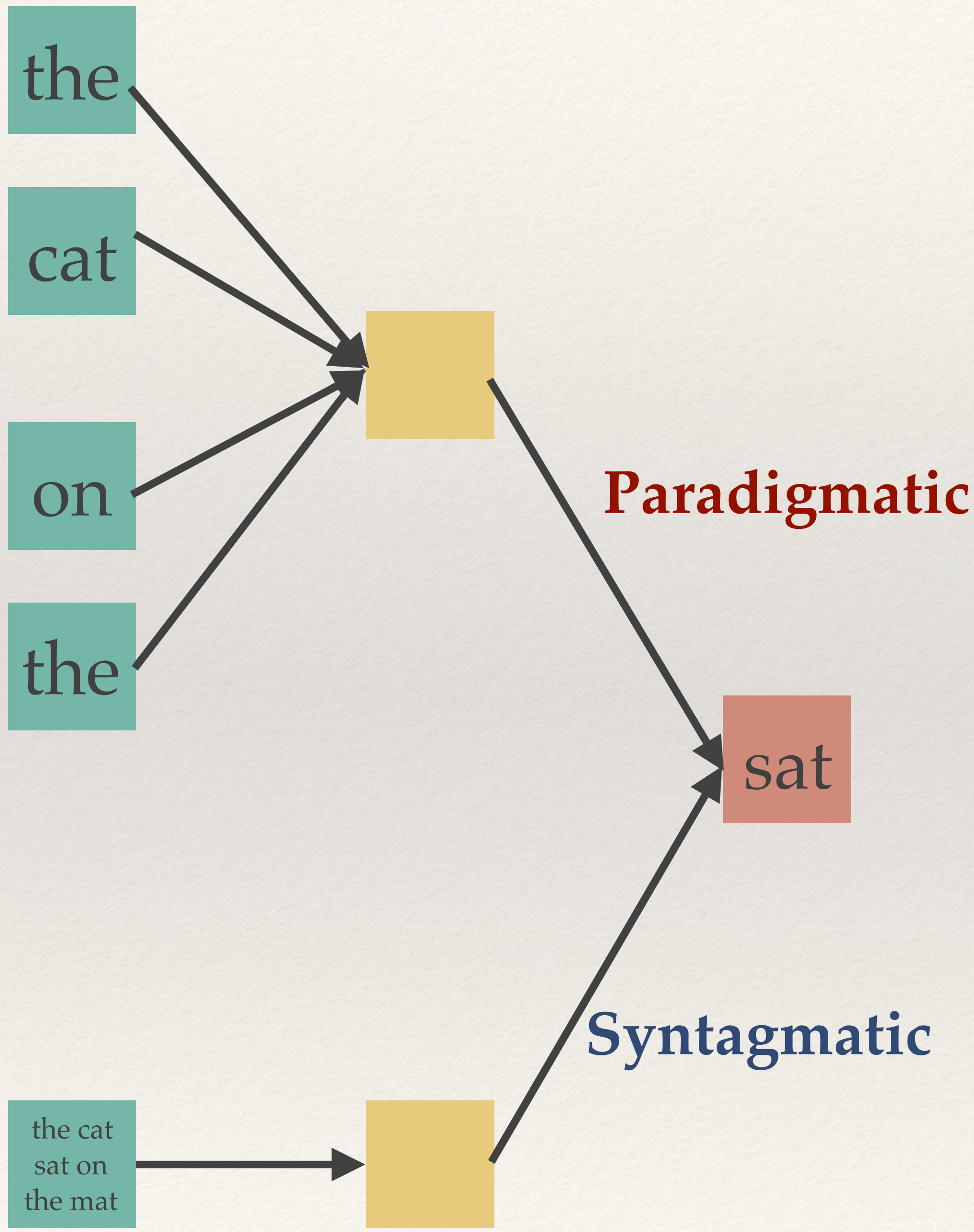
# A Natural Extension: Modeling them Jointly




- Construct the model under word2vec framework
  - Paradigmatic: modeling with word2vec
  - Syntagmatic: modeling with PV-CBOW

Sun et al., Learning Word Representations by Jointly Modeling Syntagmatic and Paradigmatic Relations. In Proc. ACL 2015.

# Parallel Document Content (PDC) Model




$$\ell = \sum_{n=1}^{N} \sum_{w_i^n \in d_n} \left( \log p(w_i^n | h_i^n) + \log p(w_i^n | d_n) \right)$$

Negative Sampling

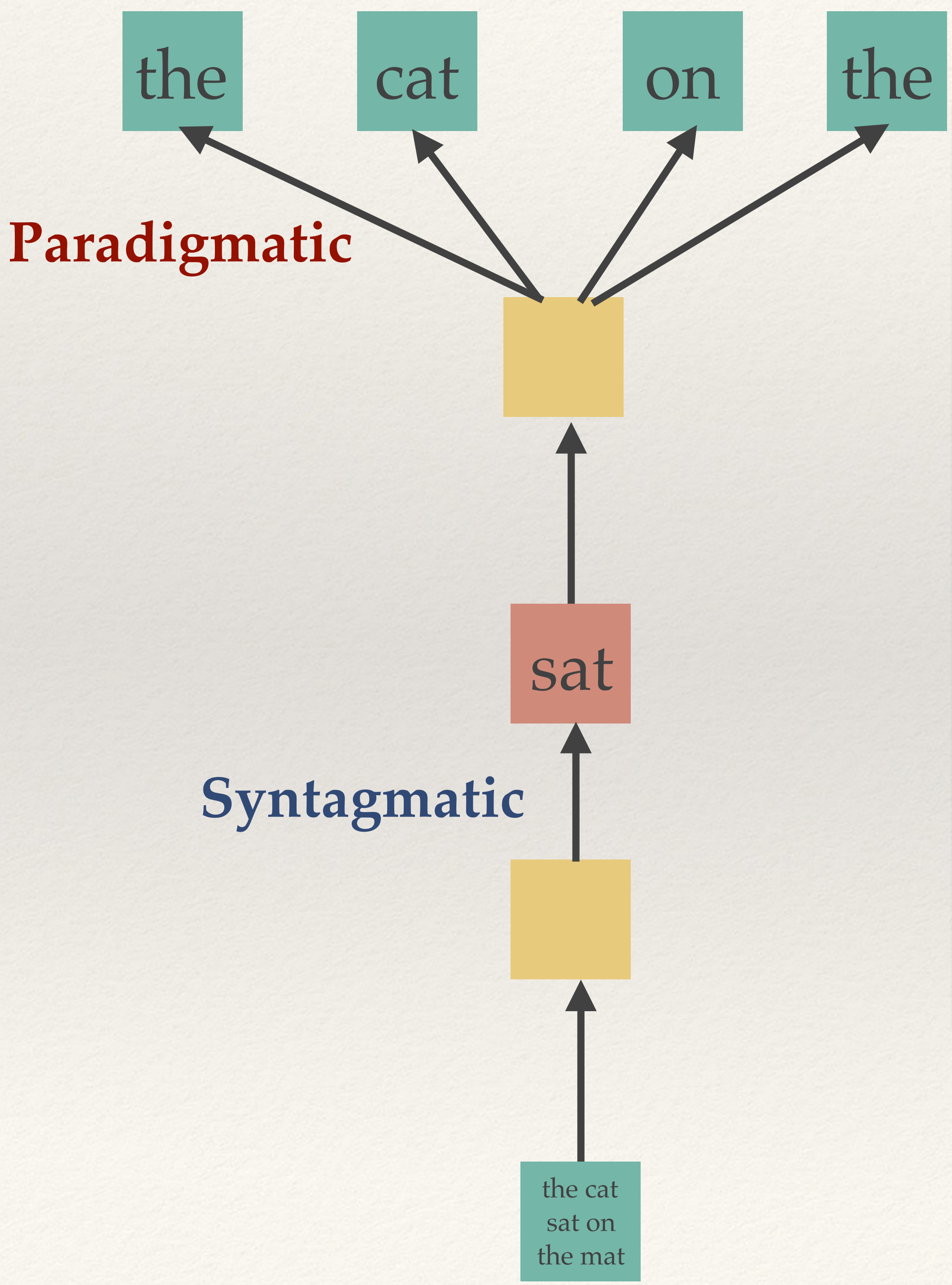
$$\ell = \sum_{n=1}^{N} \sum_{w_i^n \in d_n} \Big( \log \sigma(\vec{w}_i^n \cdot \vec{h}_i^n) + \log \sigma(\vec{w}_i^n \cdot \vec{d}_n)$$

$$+ k \cdot \mathbb{E}_{w' \sim P_{\mathrm{nw}}} \log \sigma(-\vec{w}' \cdot \vec{h}_i^n)$$

$$+ k \cdot \mathbb{E}_{w' \sim P_{\mathrm{nw}}} \log \sigma(-\vec{w}' \cdot \vec{d}_n) \Big)$$

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

# Hierarchical Document Context Model (HDC)



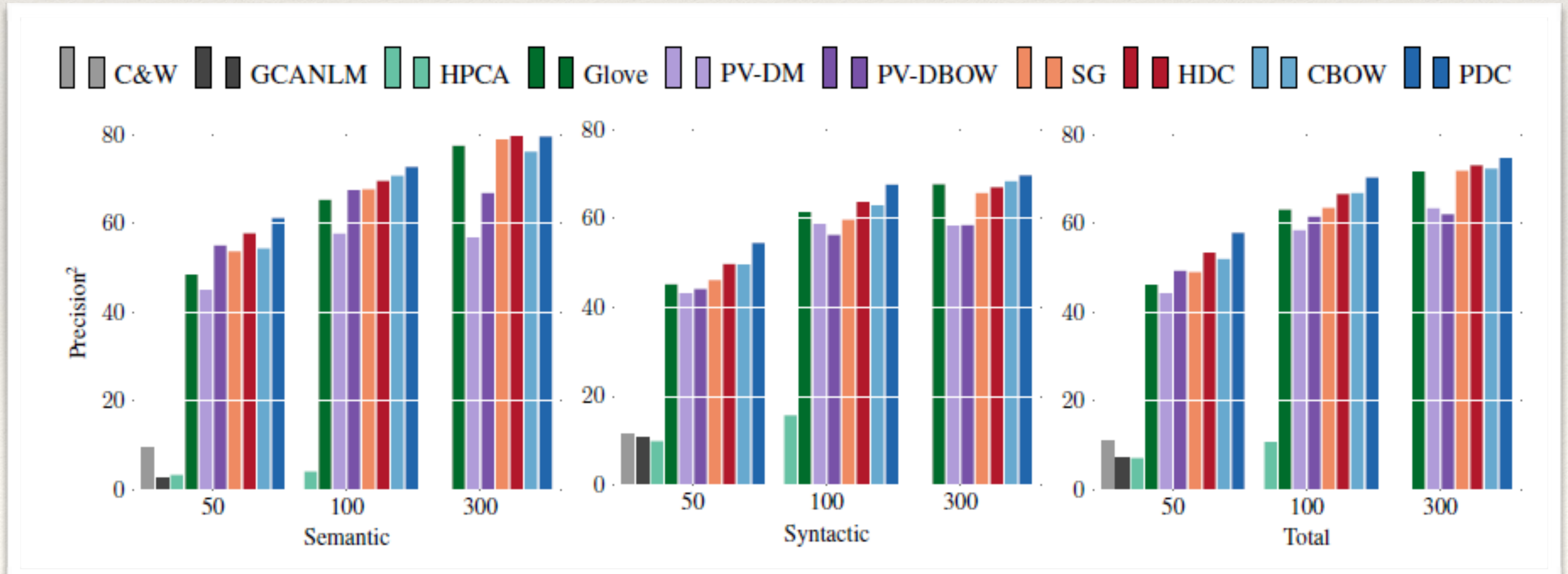

$$\ell=\sum_{n=1}^{N}\sum_{w_i^n\in d_n}\left(\log p(w_i^n|d_n)+\sum_{\substack{j=i-L\\j\neq i}}^{i+L}\log p(c_j^n|w_i^n)\right)$$

Negative Sampling

$$\ell=\sum_{n=1}^{N}\sum_{w_i^n\in d_n}\Bigg(\sum_{\substack{j=i-L\\j\neq i}}^{i+L}\Big(\log\sigma(\vec{c_j^n}\cdot\vec{w_i^n})+k\cdot\mathbb{E}_{c'\sim P_{\mathrm{nc}}}\log\sigma(-\vec{c'}\cdot\vec{w_i^n})\Big)+\log\sigma(\vec{w_i^n}\cdot\vec{d_n})+k\cdot\mathbb{E}_{w'\sim P_{\mathrm{nw}}}\log\sigma(-\vec{w'}\cdot\vec{d_n})\Bigg)$$

# Empirical Evaluation of PDC and HDC

- Word analogy based on Google test set
  - Semantic: "Beijing is to China as Paris is to ___"
  - Syntactic: "big is to bigger as deep is to ___"



# More Diverse Results

Paradigmatic → Syntagmatic

| Top five similar words to "Feynman" | | | | |
|---|---|---|---|---|
| **CBOW** | **SG** | **PDC** | **HDC** | **PV-DBOW** |
| Einstein | Schwinger | geometrodynamics | Schwinger | physicists |
| Schwinger | quantum | Bethe | electrodynamics | spacetime |
| Bohm | Bethe | semiclassical | Bethe | geometrodynamics |
| Bethe | Einstein | Schwinger | semiclassical | tachyons |
| relativity | semiclassical | peturbative | quantum | Einstein |

Paradigmatic    Syntagmatic

# Re-examine the Distributed Hypothesis

syntagmatic

Albert Einstein was a physicist.

Paradigmatic

Richard Feynman was a physicist.

syntagmatic

- Syntagmatic: words co-occur in the same text region
- Paradigmatic: words occur in the same context, may not at the same time
- Distributed hypothesis considers words as IDs
  - However, words are constructed by more fine-grained elements, e.g., breakable —> break, able

# Beyond Distributed Hypothesis

- Distributed hypothesis: discovering semantics of words from external information
- Beyond distributed hypothesis: discovering semantics of words from both external and internal information
  - External: distributed hypothesis
  - Internal: words are built from morphemes, e.g., breakable —> break, able

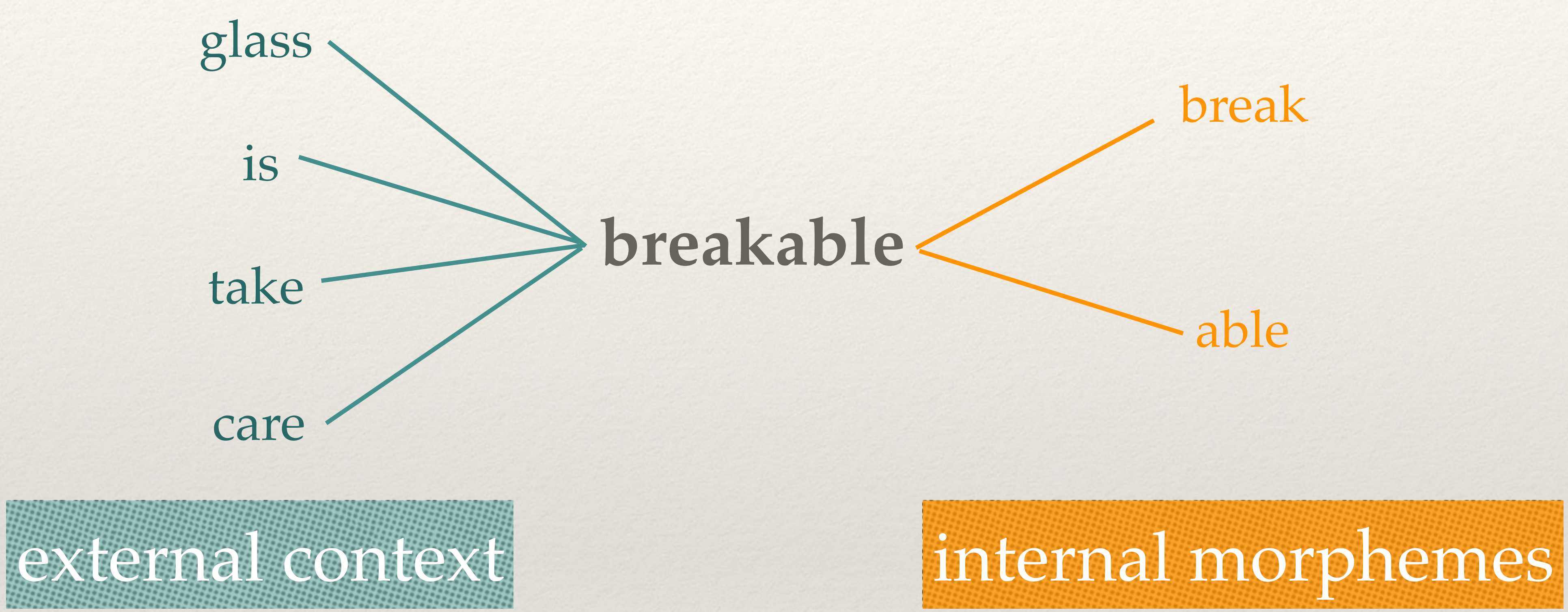
glass is **breakable**, take care

He **breaks** the glass

Similar embeddings for "breakable" and "break"

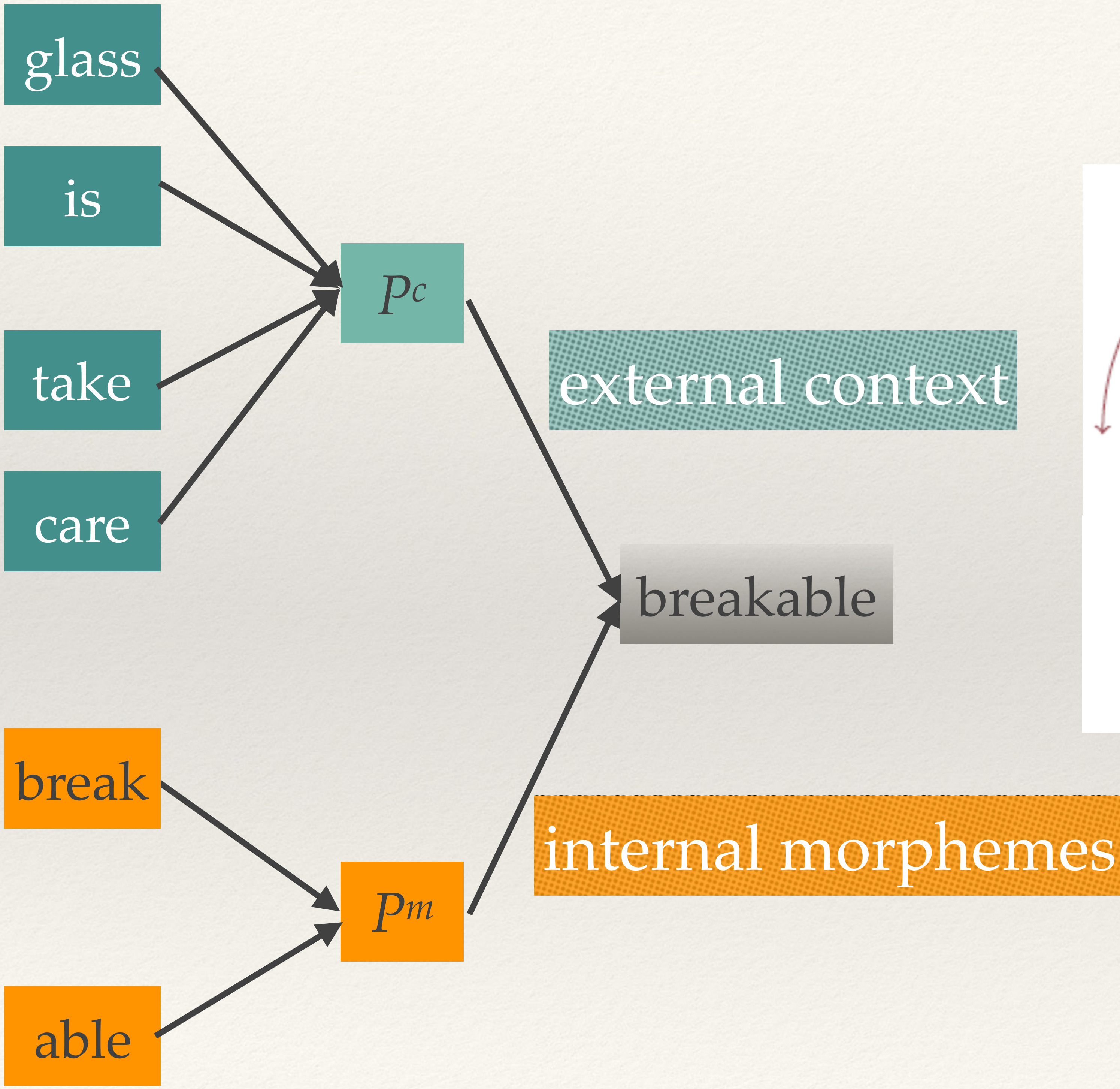
# Word Embedding with Morphemes

"… glass is **breakable**, take care …"




Sun et al., Inside Out: Two Jointly Predictive Models for Word Representations and Phrase Representations. In Proc. AAAI 2016.

# Continuous Bag of External and Internal Gram (BEING)




$$\mathcal{L} = \sum_{i=1}^{N} \left( \log p(w_i|\mathcal{P}_i^c) + \log p(w_i|\mathcal{P}_i^m) \right)$$

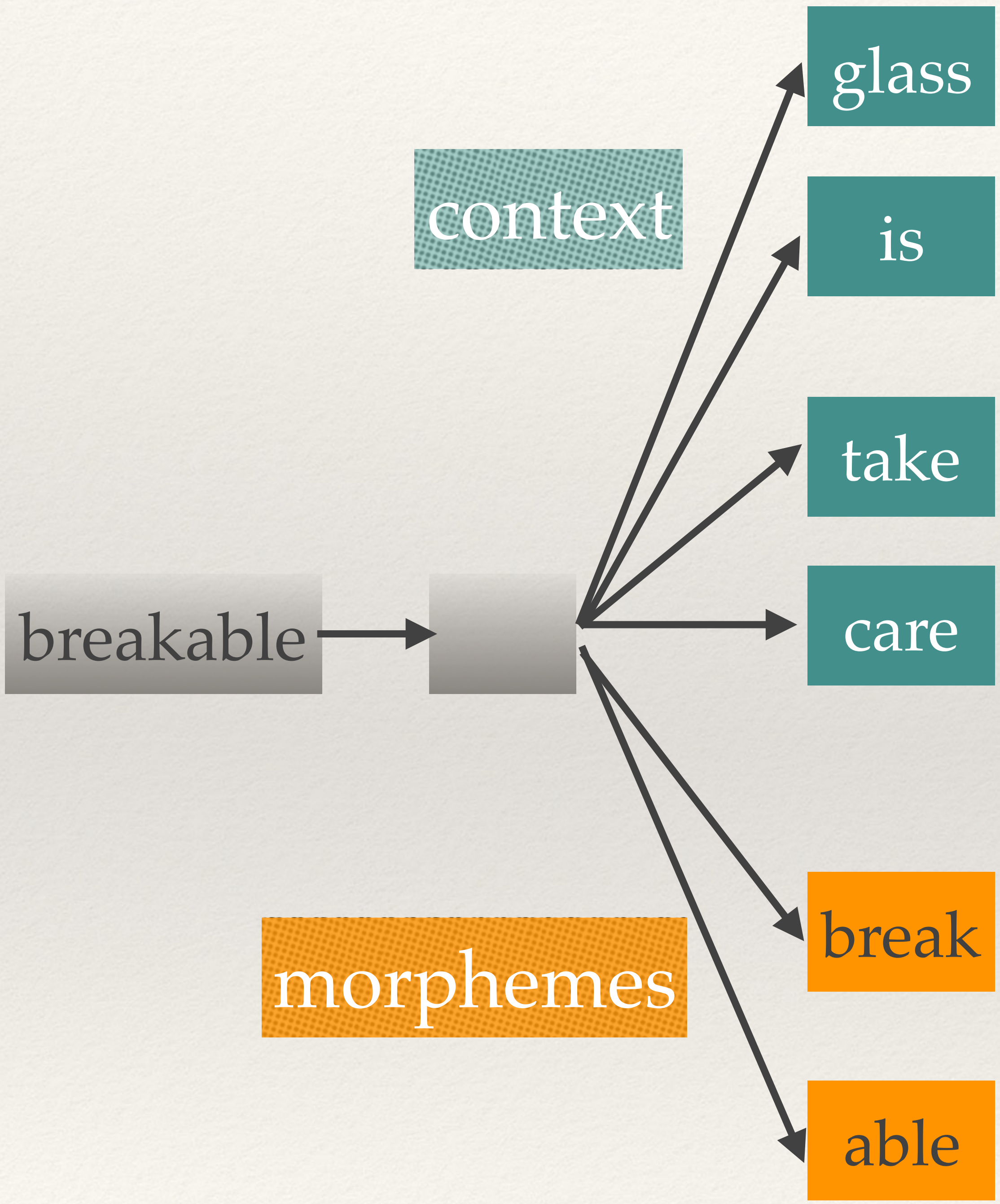
Negative Sampling

$$\mathcal{L} = \sum_{i=1}^{N} \Big( \log \sigma(\vec{w_i} \cdot \overrightarrow{\mathcal{P}_i^c}) + k \cdot \mathbf{E}_{\tilde{w} \sim P_W} \log \sigma(-\vec{\tilde{w}} \cdot \overrightarrow{\mathcal{P}_i^c})$$

$$+ \log \sigma(\vec{w_i} \cdot \overrightarrow{\mathcal{P}_i^m}) + k \cdot \mathbf{E}_{\tilde{w} \sim P_W} \log \sigma(-\vec{\tilde{w}} \cdot \overrightarrow{\mathcal{P}_i^m}) \Big)$$

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

# Continuous Skip External and Internal Gram (SEING)




$$\mathcal{L} = \sum_{i=1}^{N} \left( \sum_{\substack{j=i-l \\ j \neq i}}^{i+l} \log p(c_j | w_i) + \sum_{z=1}^{s(w_i)} \log p(m_i^{(z)} | w_i) \right)$$

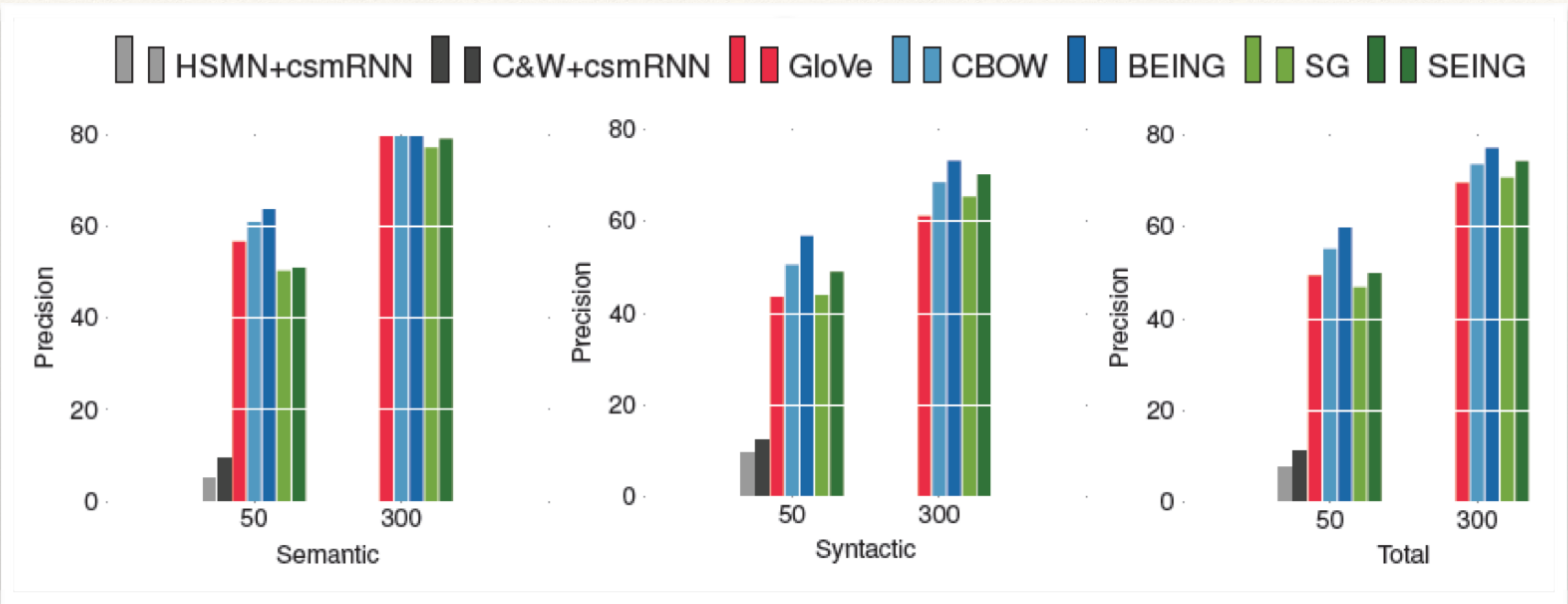
Negative Sampling

$$\mathcal{L} = \sum_{i=1}^{N} \left( \sum_{\substack{j=i-l \\ j \neq i}}^{i+l} \left( \log \sigma(\vec{c_j} \cdot \vec{w_i}) + k \cdot \mathbf{E}_{\tilde{c} \sim P_{\tilde{C}}} \log \sigma(-\vec{\tilde{c}} \cdot \vec{w_i}) \right) \right.$$

$$\left. + \sum_{z=1}^{s(w_i)} \left( \log \sigma(\overrightarrow{m_i^{(z)}} \cdot \vec{w_i}) + k \cdot \mathbf{E}_{\tilde{m} \sim P_{\tilde{M}}} \log \sigma(-\vec{\tilde{m}} \cdot \vec{w_i}) \right) \right)$$

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

# Empirical Evaluation of BEING and SEING




- BEING and SEING outperformed CBOW and SG, respectively
- Significant improvements achieved on syntactic task

# Direct Matching with Word Embeddings

# Query-Document Matching based on Local Representations

**VSM score**

cos

[0 0 0 1.6 0 1.4 0 …. 0 4.8 0 1.2 0 0]

tf-idf

[5.2 0 2.5 0 0 0 …. 0 1.3 1.2 0 0 0 1.9]

| One-hot embedding | idf |
|---|---|
| [0 0 0 0 0 0 0 0 …. … 0 1 0 0 0 0 0] | 4.8 |
| [0 0 0 0 0 0 0 0 …. … 0 0 0 1 0 0 0] | 1.2 |
| [0 0 0 0 0 0 1 0 …. … 0 0 0 0 0 0 0] | 1.4 |
| [0 0 0 1 0 0 0 0 …. … 0 0 0 0 0 0 0] | 1.6 |

peace process in the Middle East

| One-hot embedding | idf |
|---|---|
| *[0 0 0 0 0 0 0 0 …. … 0 0 0 0 0 1 0]* | *1.2* |
| *[0 0 0 0 0 0 0 0 …. … 0 0 2 0 0 0 0]* | 1.3 |
| *[0 0 3 0 0 0 0 0 …. … 0 0 0 0 0 0 0]* | *2.5* |
| *[1 0 0 0 0 0 0 0 …. … 0 0 0 0 0 0 0]* | 5.2 |
| *[0 0 0 0 0 0 0 0 …. … 0 2 0 0 0 0 0]* | *1.3* |
| … | … |

As for the Arabian and Palestinean voices that are against the current negotiations and the so-called peace process, they are not against peace per se, but rather for their well-founded predictions that Israel would NOT give an inch of the West bank (and most probably the same for Golan Heights) back to the Arabs. An 18 months of "negotiations" in Madrid, and Washington proved these predictions. Now many will jump on me saying why are you blaming israelis for no-result negotiations. I would say why would the Arabs stall the negotiations, what do they have to loose ?



Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings, I. Vulic et. al. 2015 SIGIR.

# When Embedding Comes …

WE-VS score

cos

[0.4 1.2 1.6 0.2 1.4 …. 0.7 1.2 4.8 1.5 ]

[2.6 2.5 5.2 0.5 1.4 3.1 …. 0.9 0.72 1.2]

weighted sum

| embedding | idf |
|---|---|
| [0.1 0.3 0.03 0 0.4 …. 0.05 0.12 0.02] | 4.8 |
| [0.2 0.13 0.03 0 0 …. 0.07 0.09 0.01] | 1.2 |
| [0.13 0.3 0.3 0 0.2 …. 0.08 0.87 0.02] | 1.4 |
| [0.3 0.4 0.09 0 0.3 …. 0.05 0.34 0.14] | 1.6 |

| embedding | idf |
|---|---|
| [0.4 0.3 0.45 0 0.3 …. 0.09 0.24 0.7 0.01] | 1.2 |
| [0.2 0.1 0.03 0 0.1 …. 0.18 0.91 0.2 0.02] | 1.3 |
| [0.1 0.3 0.13 0 0.1 …. 0.16 0.25 0.8 0.03] | 2.5 |
| [0.3 0.5 0.11 0 0.2 …. 0.03 0.170.1 0.15] | 5.2 |
| [0.7 0.9 0.01 0 0.6 …. 0.15 0.35 0.4 0.26] | 1.3 |
| … | … |

peace process in the Middle East

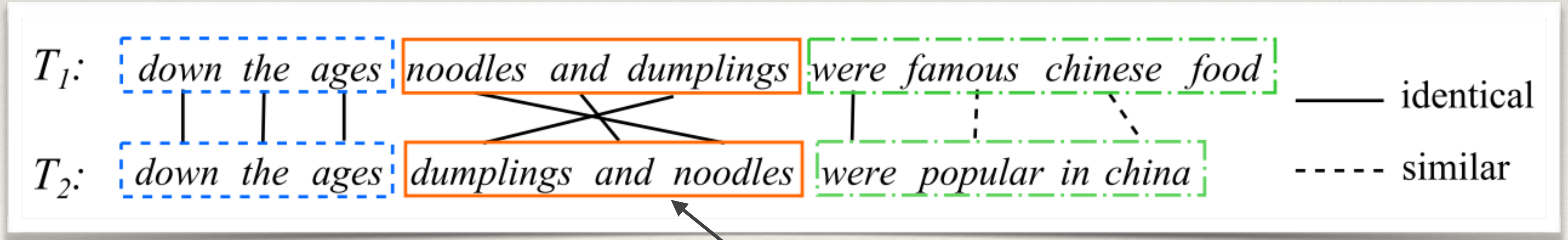
As for the Arabian and Palestinean voices that are against the current negotiations and the so-called peace process, they are not against peace per se, but rather for their well-founded predictions that Israel would NOT give an inch of the West bank (and most probably the same for Golan Heights) back to the Arabs. An 18 months of "negotiations" in Madrid, and Washington proved these predictions. Now many will jump on me saying why are you blaming israelis for no-result negotiations. I would say why would the Arabs stall the negotiations, what do they have to loose ?



Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings, I. Vulic et. al. 2015 SIGIR.

# Outline

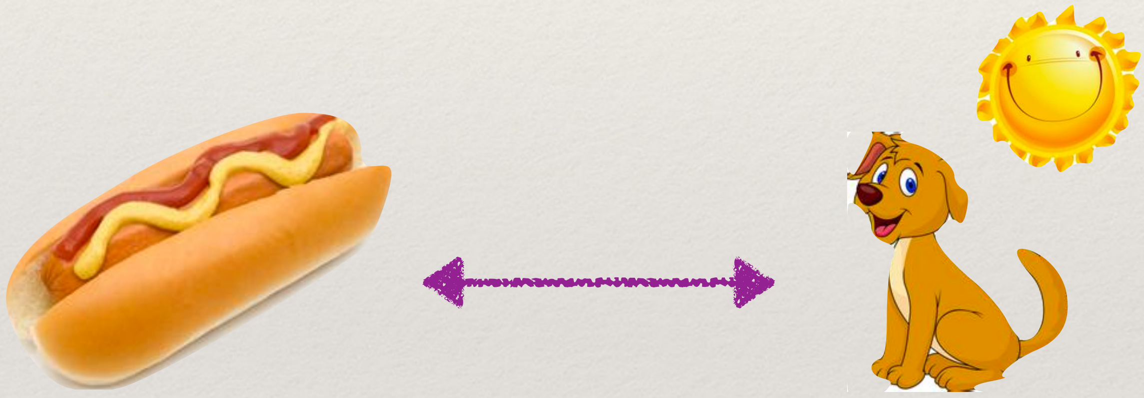
- Semantic matching in search
- Word-level matching: bridging the semantic gap
- Sentence-level matching: capturing the proximity
- Summary and discussion




Similarity between "noodles and dumplings" and "dumplings and noodles"

# Problems with Direct Matching

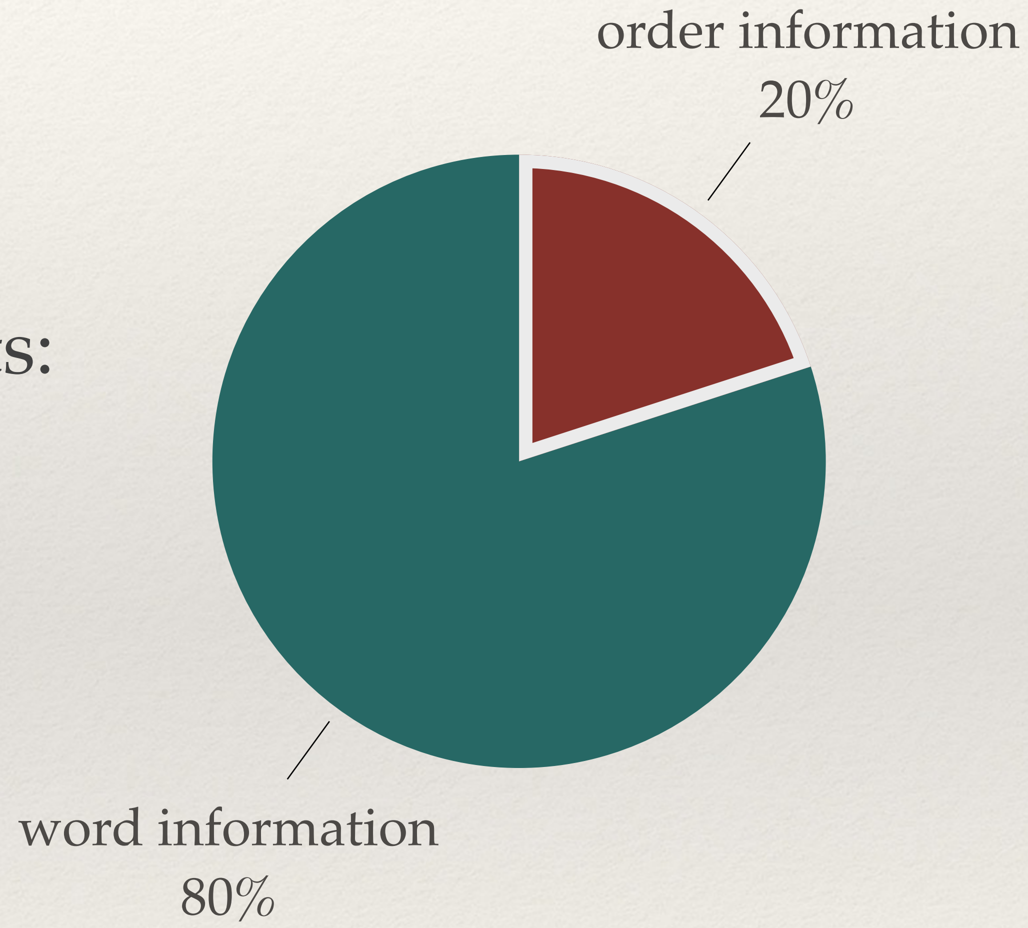
- ***Problem 1: information on the words order is missing***



  - Bag of words: Dog Hot = Hot Dog
  - In real world: Dog Hot ≠ Hot Dog

# The Importance of Words Order

- Assume:
  - size of vocabulary = 100,000
  - average sentences length = 20
- Rough contributions of information in bits:
  - From the selection of words: $\log_2(100000^{20})$
  - From the order of words: $\log_2(20!)$
- Conclusion: over 80% of the potential information in language being in the choice of words without regard to the order in which they appear



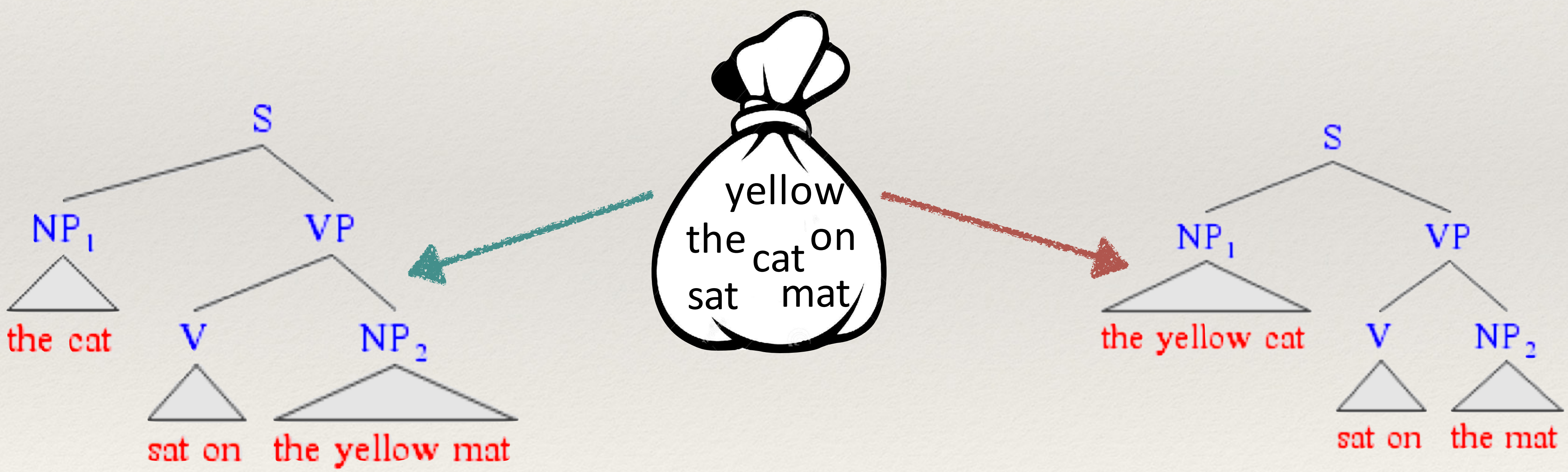

Landauer T K. On the computational basis of learning and cognition: Arguments from LSA[J]. Psychology of learning and motivation, 2002, 41: 43-84.

# Problem with Direct Methods

- ***Problem 2: simple sentence representation***

With bag-of-words assumption:

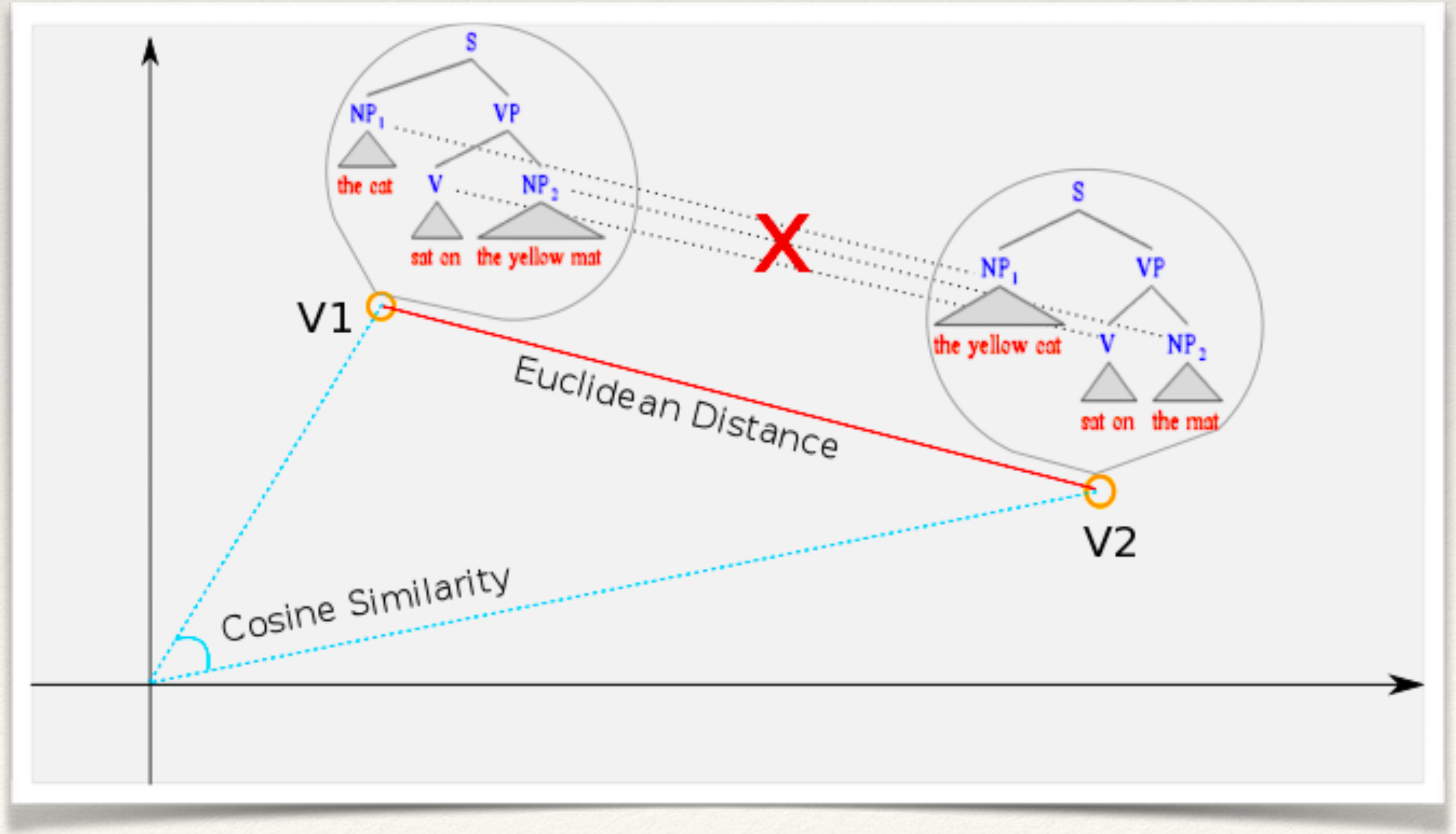
"the **yellow** cat sat on the mat" = "the mat sat on the **yellow** mat"



# Problem with Direct Methods

- ***Problem 3: Heuristic matching function***
  - A vector for representing the whole sentence
  - Based on distance measures between two vectors, e.g., Cosine, dot product, Euclidean distance

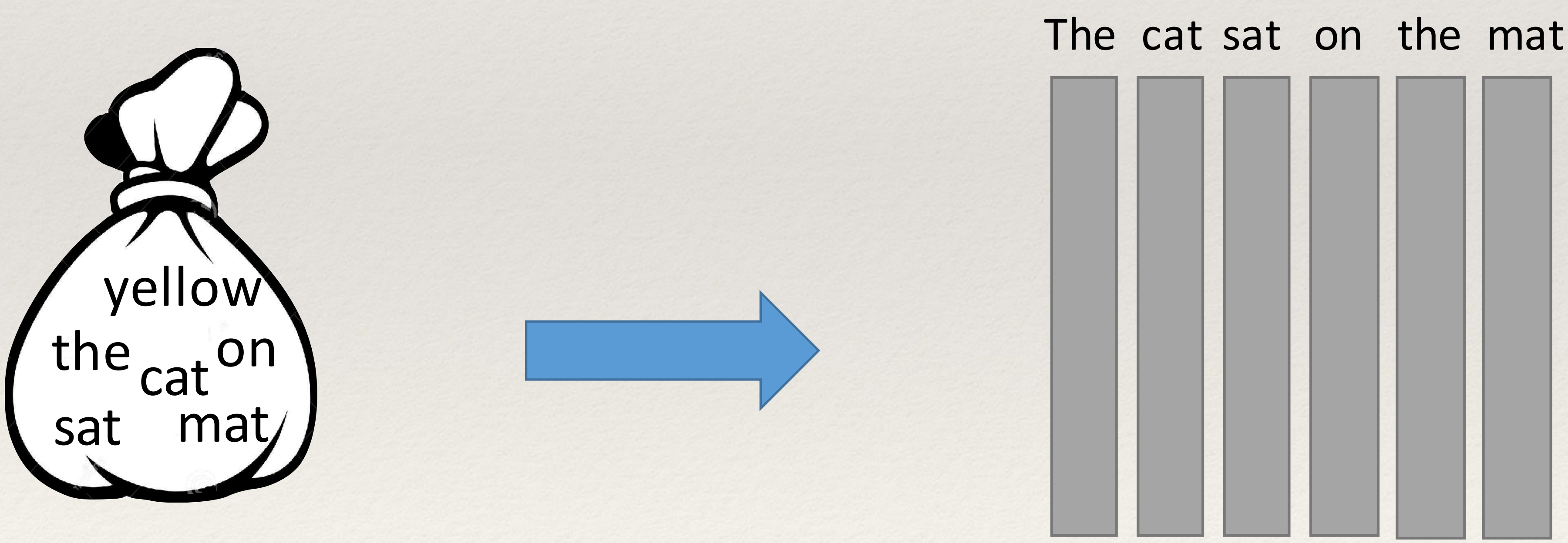
Limited information is taken into consideration



# How to design a deep model for semantic text matching?

# Keeping Order Information

- Sequence of word embeddings as the inputs
  - Convert words to embeddings (e.g. word2vec)
  - Concatenate embeddings to a sequence

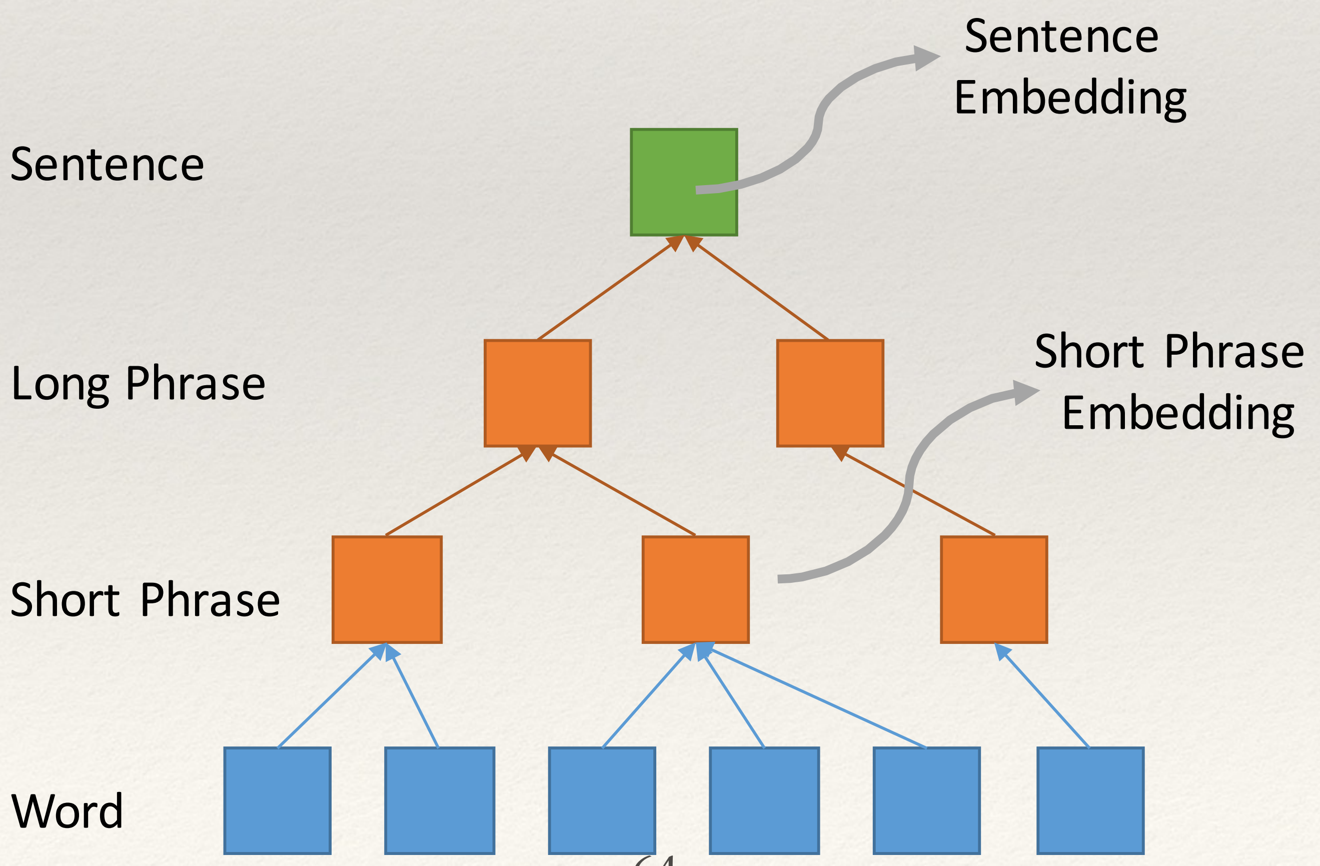



**Bag of Word Embeddings**

**Sequence of Word Embeddings**

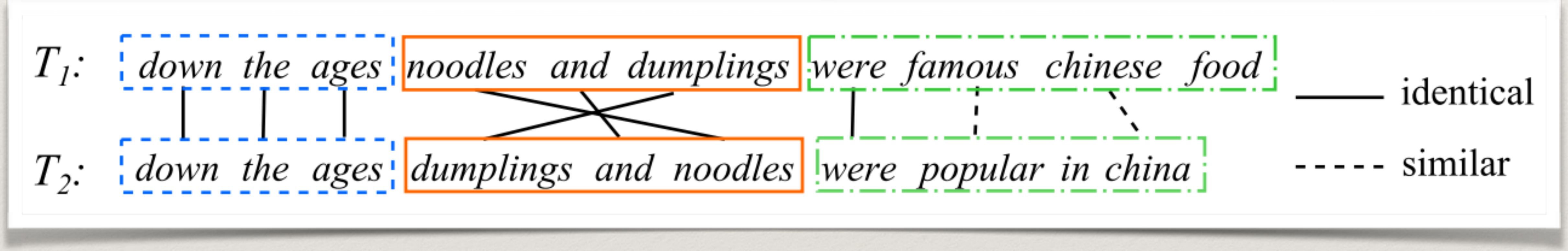
# Rich Sentence Representation

- Hierarchical structure of sentence representation
  - Different levels of embeddings
  - Involving sentence structure



# Powerful Matching Function

- Considering different levels/types of matching signals




N-gram

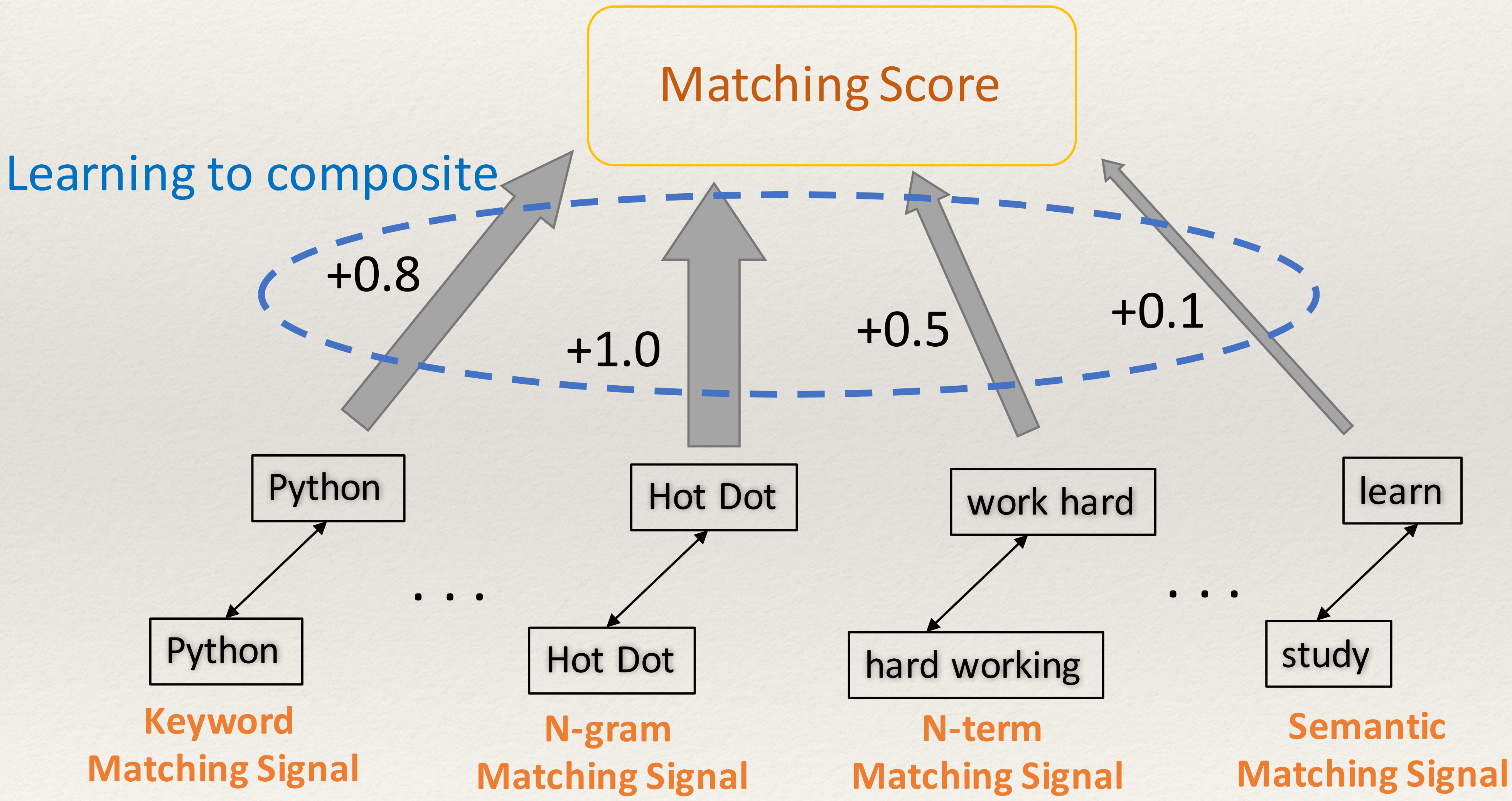
N-term

Proximal N-term


Pang L, Lan Y, Guo J, et al. Text matching as image recognition//Proceedings of the 30th AAAI Conference onArtificial Intelligence. Phoenix, USA, 2016: 2793-2799.

# Learning the Matching Function

- Data-driven approaches for determining the parameters



# Existing Deep Matching Models for Semantic Text Matching

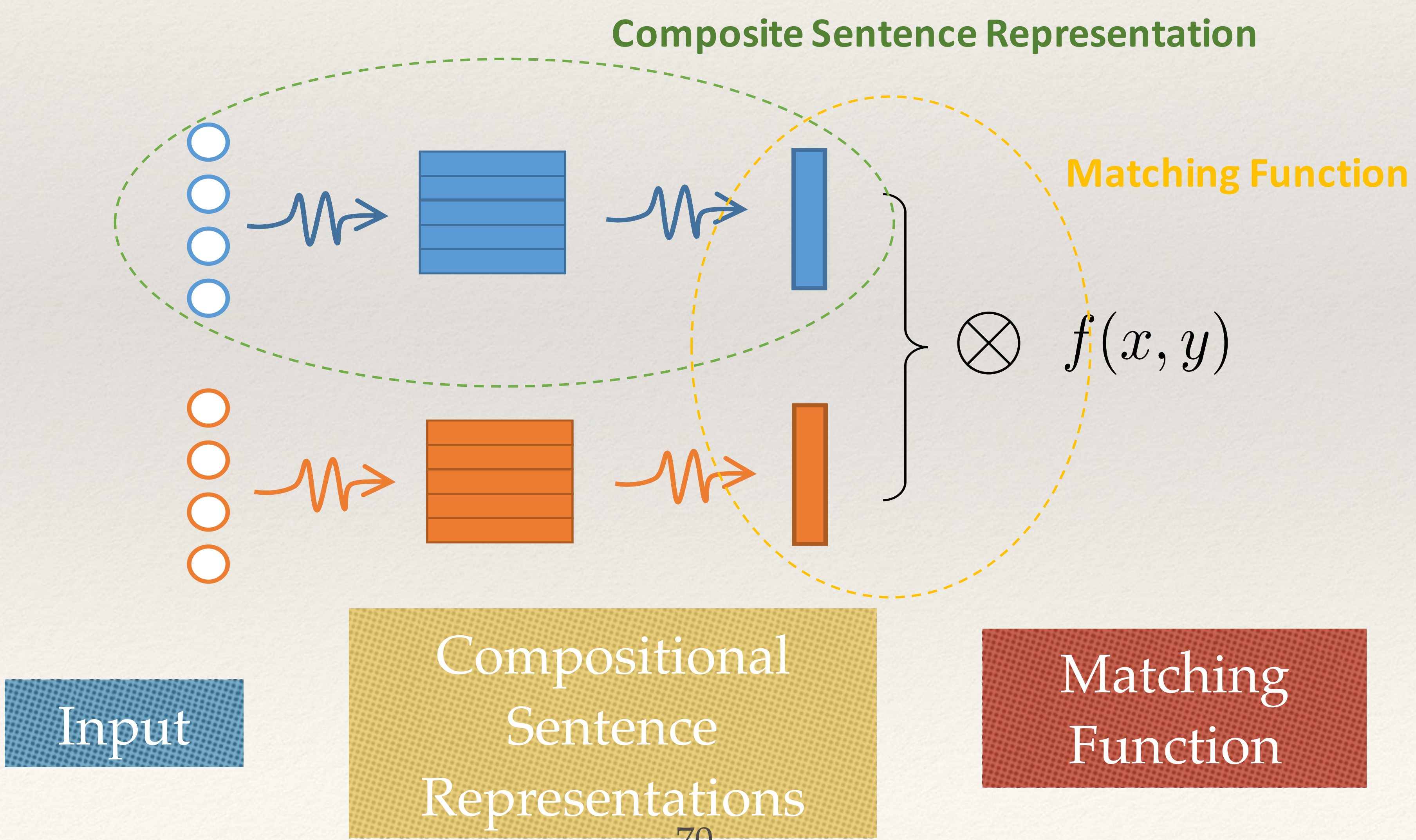
# Existing Deep Text Matching Models

- Composition Focused Methods [Problem 1] [Problem 2]
  - Composite each sentence into one embedding
  - Measure the similarity between the two embeddings
- Interaction Focused Methods [Problem 1] [Problem 3]
  - Two sentences meet before their own high-level representations mature
  - Capture complex matching patterns

# Composition Focused Methods

# Composition Focused Methods

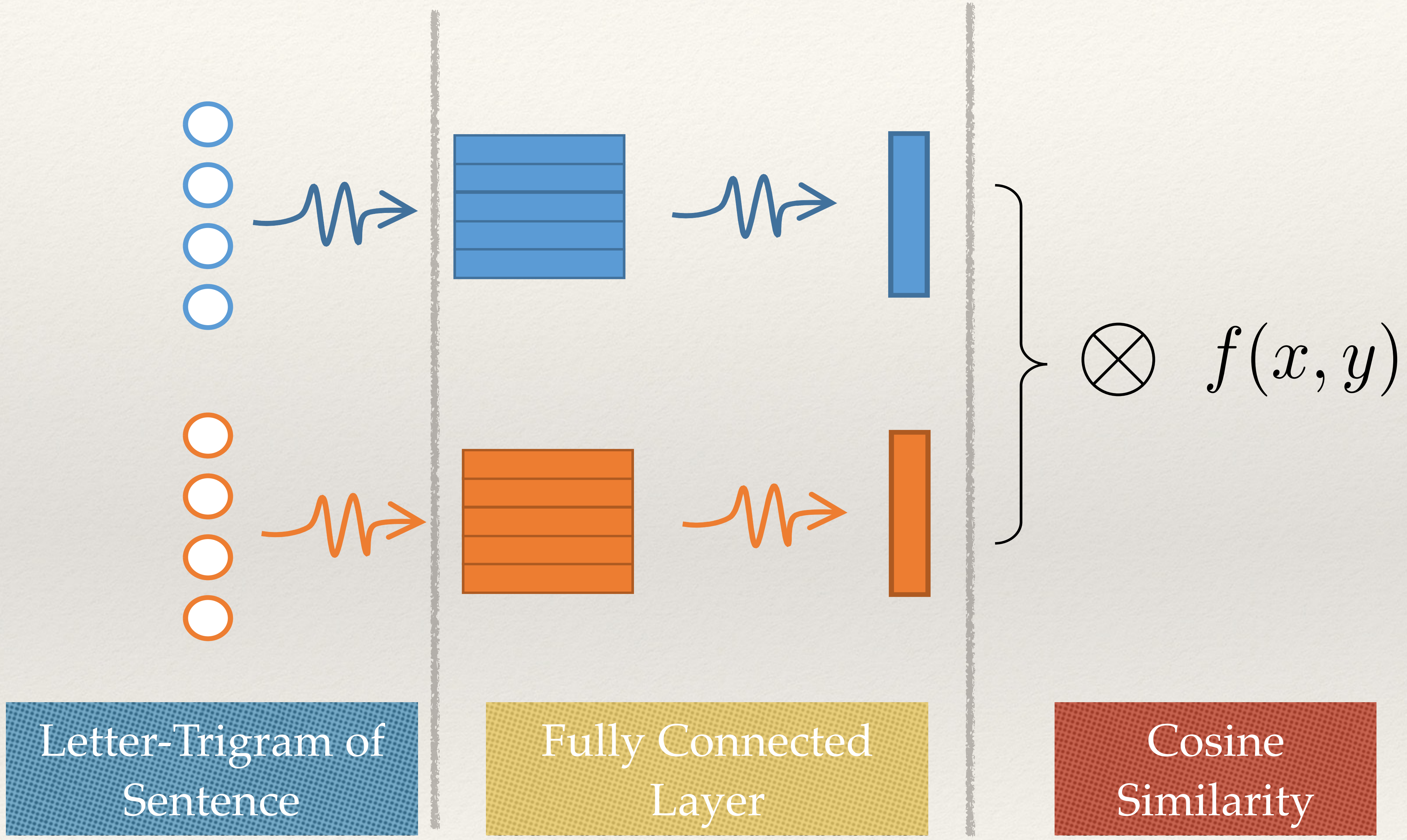
- Step 1: Compositional sentence representation $\phi(x)$
- Step 2: Matching function $F(\phi(x), \phi(y))$



# Typical Composition Focused Deep Matching Models

- Based on DNN
  - DSSM: Learning Deep Structured Semantic Models for Web Search using Click-through Data (Huang et al., CIKM'13)
- Based on CNN
  - CDSSM: A latent semantic model with convolutional-pooling structure for information retrieval (Shen et al. CIKM'14)
  - ARC I: Convolutional Neural Network Architectures for Matching Natural Language Sentences (Hu et al., NIPS'14)
  - CNTN: Convolutional Neural Tensor Network Architecture for Community-Based Question Answering (Qiu and Huang., IJCAI'15)
- Based on RNN
  - LSTM-RNN: Deep Sentence Embedding Using the Long Short Term Memory Network: Analysis and Application to Information Retrieval (Palangi et al., TASLP'2016)

# Deep Structured Semantic Model (DSSM)




Huang P-S, He X, Gao J, et al. Learning deep structured semantic models for web search using clickthrough data//Proceedings of the 22nd ACM international conference on CIKM. Amazon, India, 2013: 2333-2338

# DSSM Inputs: Letter-trigram

- Bag of words representation
  - "candy store": [ 0 0 0 1 0 0 0 1 0 0 0 … ]
- Letter-trigram representation
  - "#candy# #store#" ⟹ #ca | can | and | ndy | dy# | #st | sto | tor | ore | re#
  - [ 0 0 1 0 0 … 0 1 0 1 … 0 0 …]
- Advantages:
  - Compact representation: # words: 500K ⟹ # letter-trigram: 30K
  - Generalize to unseen words
  - Robust to misspelling, inflection, etc.

# DSSM Sentence Representation: DNN

- Model: DNN (auto-encoder) to capture the compositional sentence representations

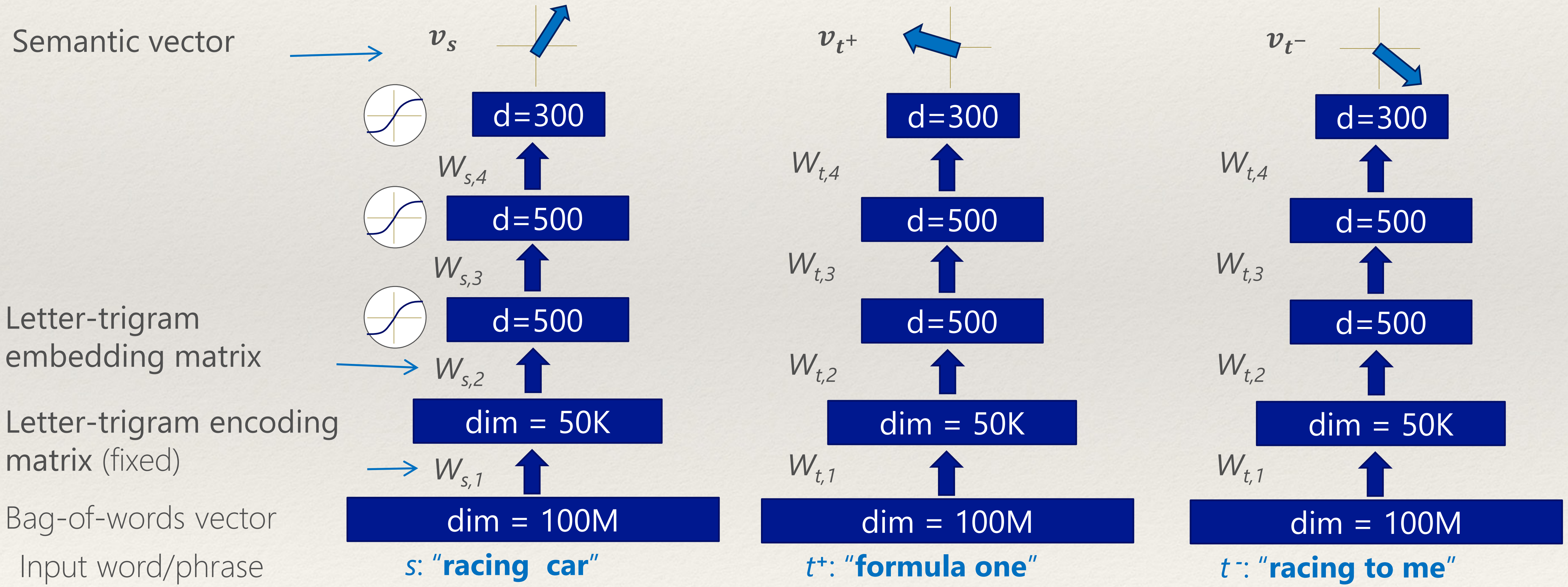



Figure from He et al., CIKM '14 tutorial

# DSSM Matching Function

- Cosine similarity between semantic vectors

$$S = \frac{x^T \cdot y}{|x| \cdot |y|}$$

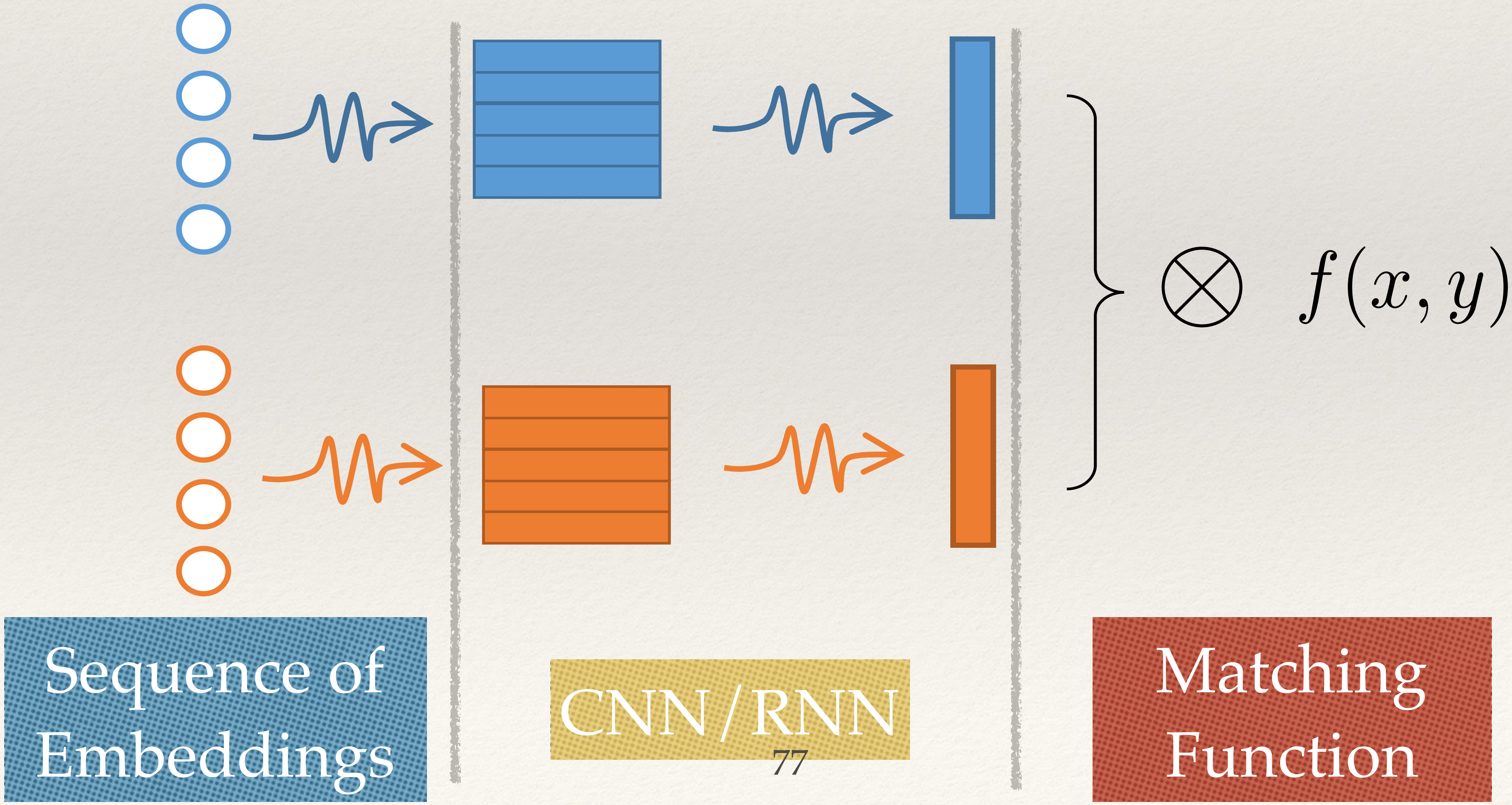
- Training
  - A query q and a list of docs $D = \{d^+, d_1^-, \cdots, d_k^-\}$
  - $d^+$ positive doc, $d_1^-, \cdots, d_k^-$ negative docs to query
  - Objective: $P(d^+|q) = \frac{\exp(\gamma \cos(q, d^+))}{\sum_{d \in D} \exp(\gamma \cos(q, d))}$
  - Optimizing with SGD

# DSSM: Brief Summary

- Inputs: sub-word units (i.e. letter-trigram) as input for scalability and generalizability
- Representations: mapping sentences to vectors (i.e. DNN): semantically similar sentences close to each other
- Matching: cosine similarity as the matching function
- Problem: bag of letter-trigrams, **the order information of words is missing**

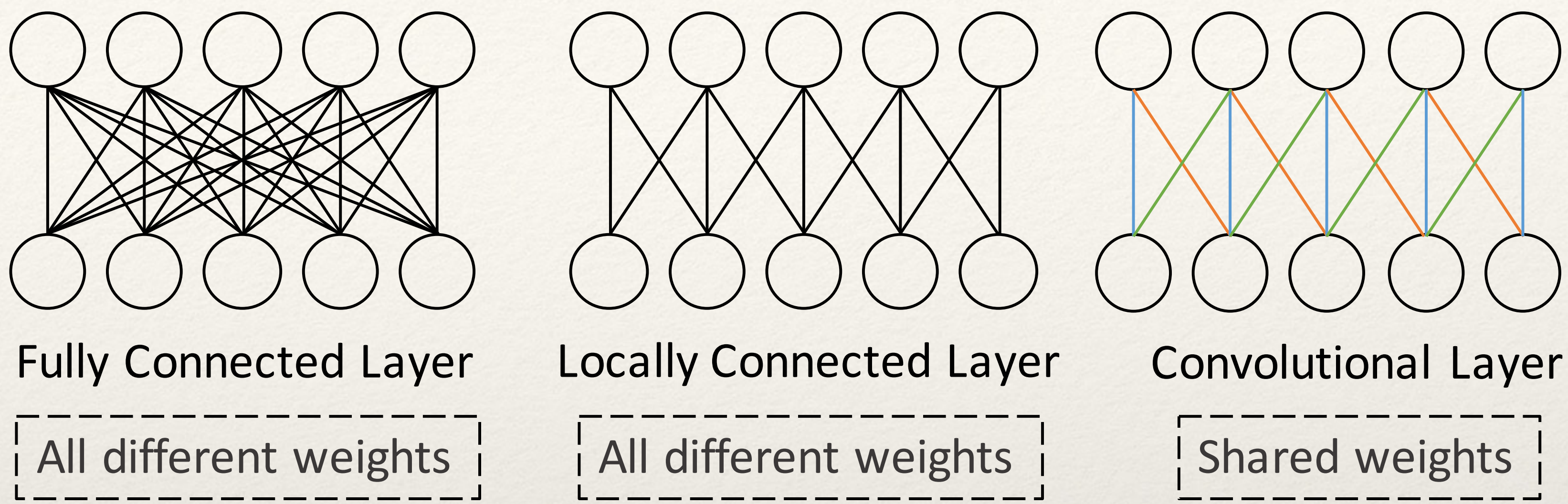
# Capturing Order Information?

- Input: word sequence instead of bag of letter-trigrams
- Model:
  - Convolutional based methods can keep locally order
  - Recurrent based methods can keep long dependence relations

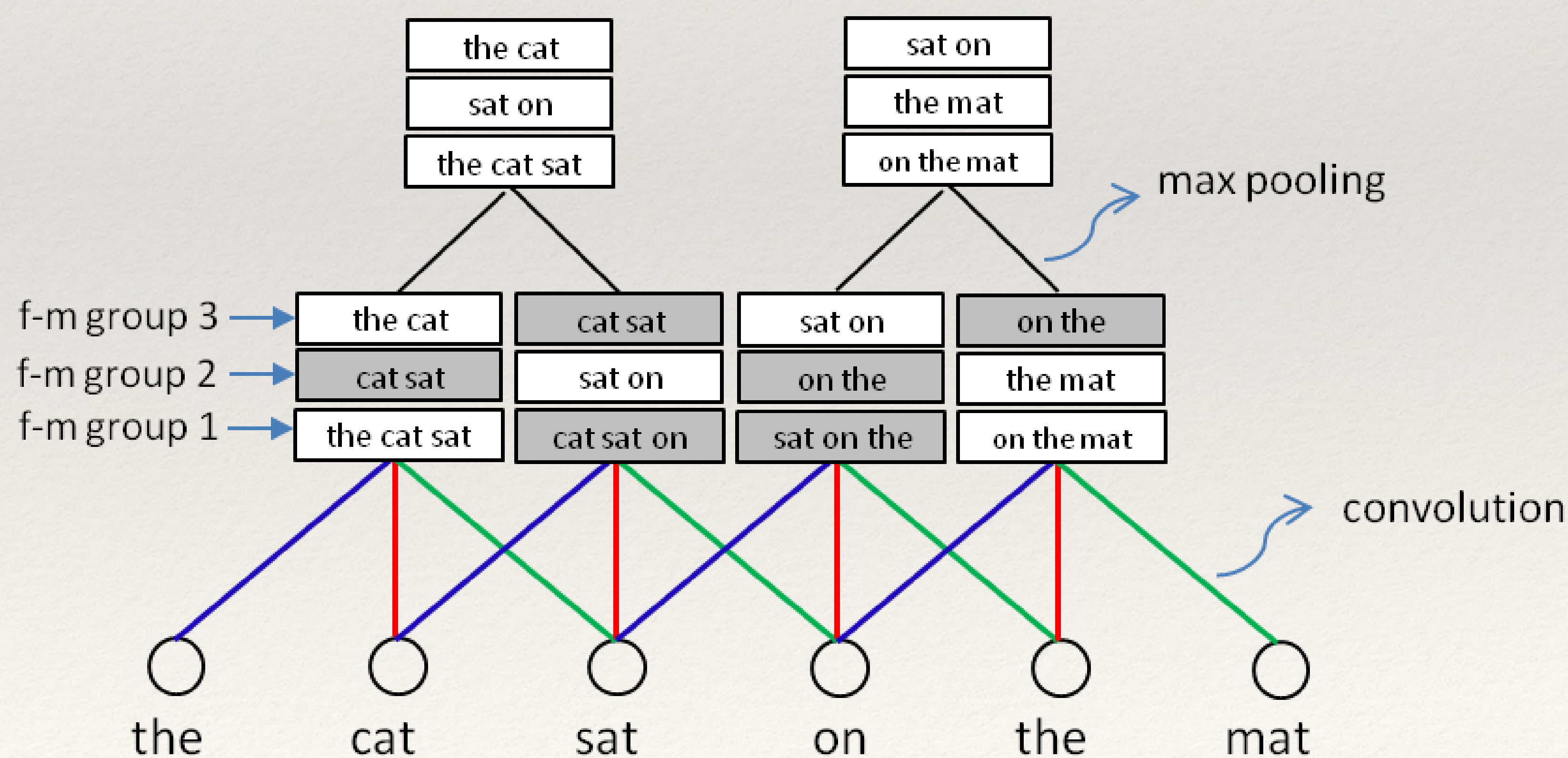


# CNN can Model the Order Information

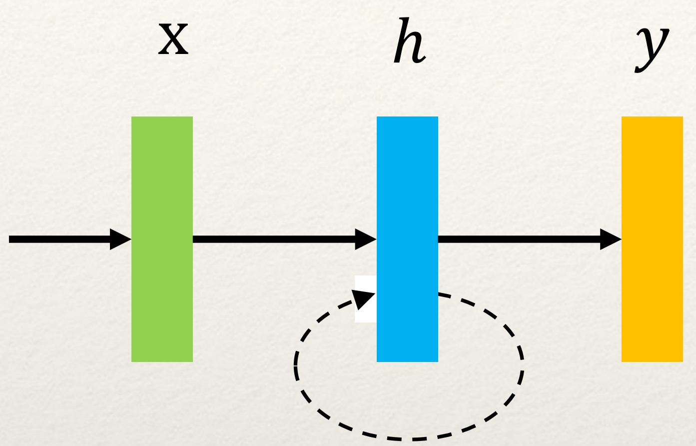
- Inspired by the cat's visual cortex [Hubel '68]



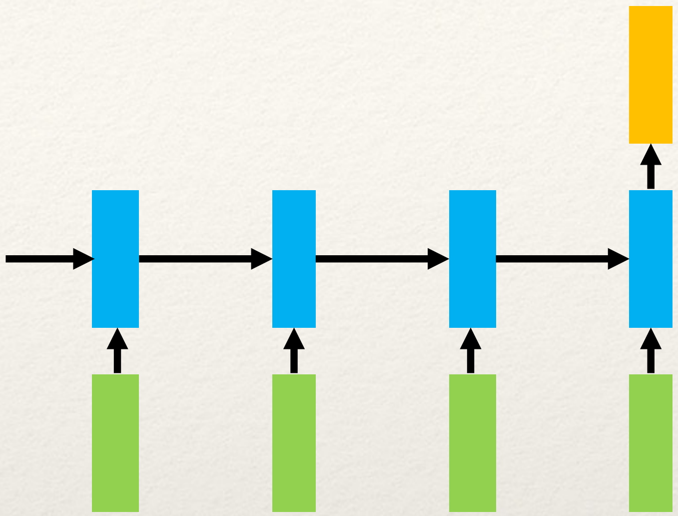

- Convolution & max pooling operations on text



# RNN can Model the Order Information




RNN – Self Recurrent Link



Expand RNN

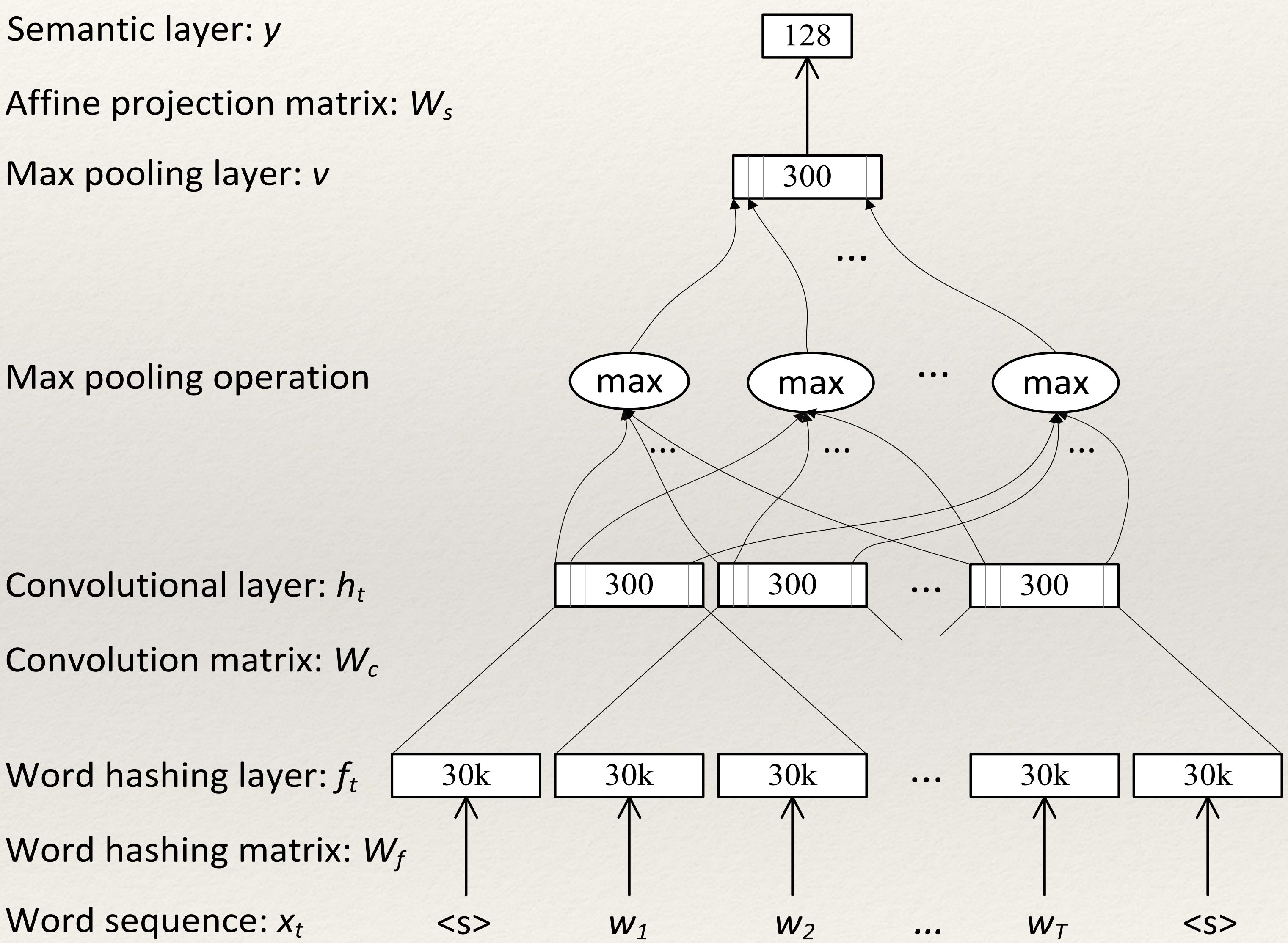
- RNNs implement dynamical systems
- RNNs can approximate arbitrary dynamical systems with arbitrary precision
- Training: Back Propagation Through Time

$$s(t) = f(\mathbf{U}w(t) + \mathbf{W}s(t-1) + b)$$

- Two popular variations: long-short term memory (LSTM) and gated recurrent unit (GRU)

# Using CNN: CDSSM

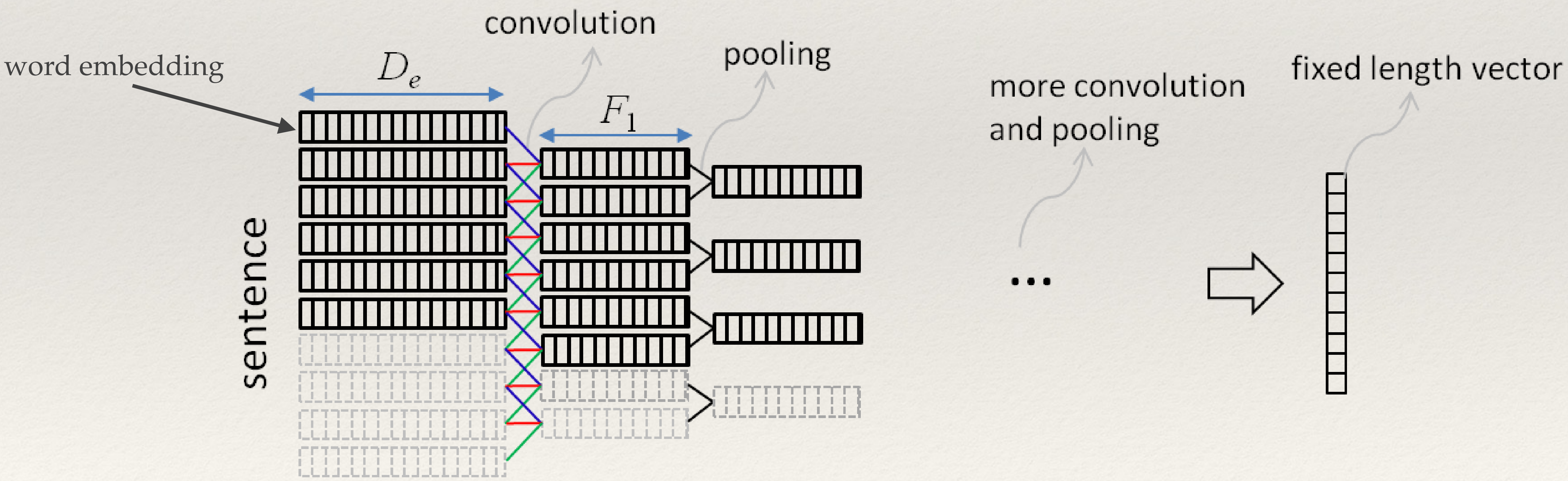
- Input: encode each word as bag of letter-trigram
- Model: the convolutional operation in CNN compacts each sequence of k words




Shen Y, He X, Gao J, et al. A latent semantic model with convolutional-pooling structure for information retrieval//Proceedings of the 23rd ACM international conference on CIKM. New York, USA, 2014: 101-110.

# Using CNN: ARC-I / CNTN

- Input: sequence of word embeddings
  - Word embeddings from word2vec model train on large dataset
- Model: the convolutional operation in CNN compacts each sequence of k words

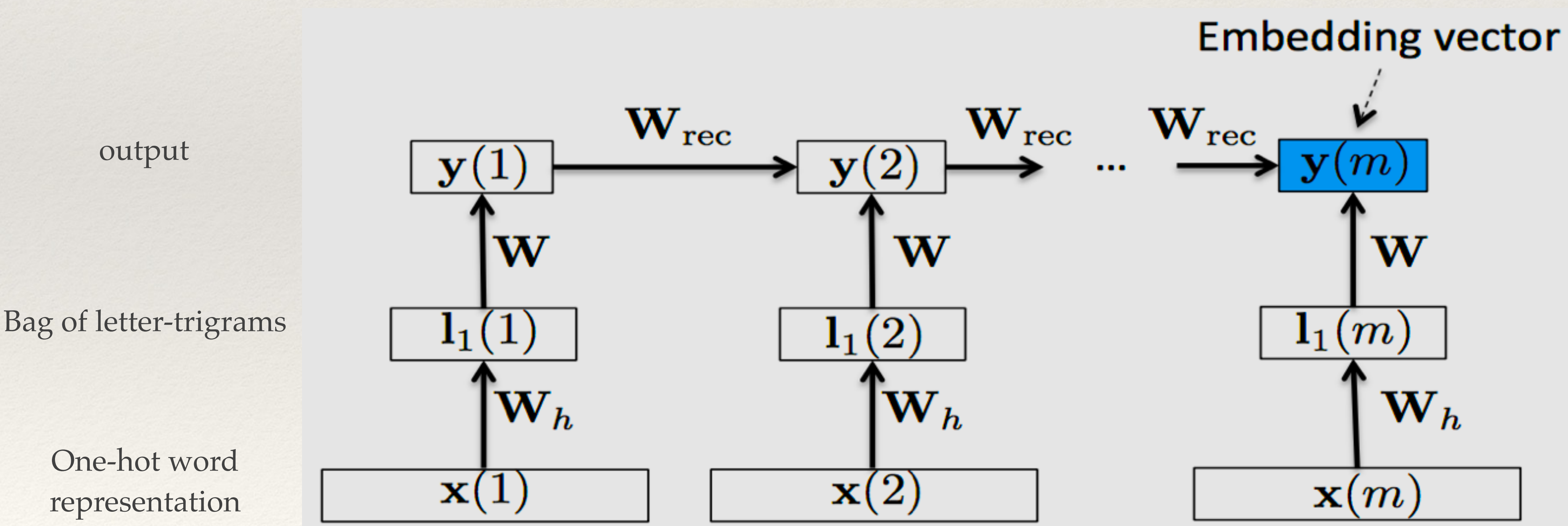



Baotian Hu, Zhengdong Lu, Hang Li, Qingcai Chen. Convolutional Neural Network Architectures for Matching Natural Language Sentences. Proceedings of Advances in Neural Information Processing Systems 27 (NIPS'14), 2042-2050, 2014.

Qiu X, Huang X. Convolutional neural tensor network architecture for community-based question answering//Proceedings of the 24th (IJCAI), Buenos Aires, Argentina, 2015: 1305-1311.

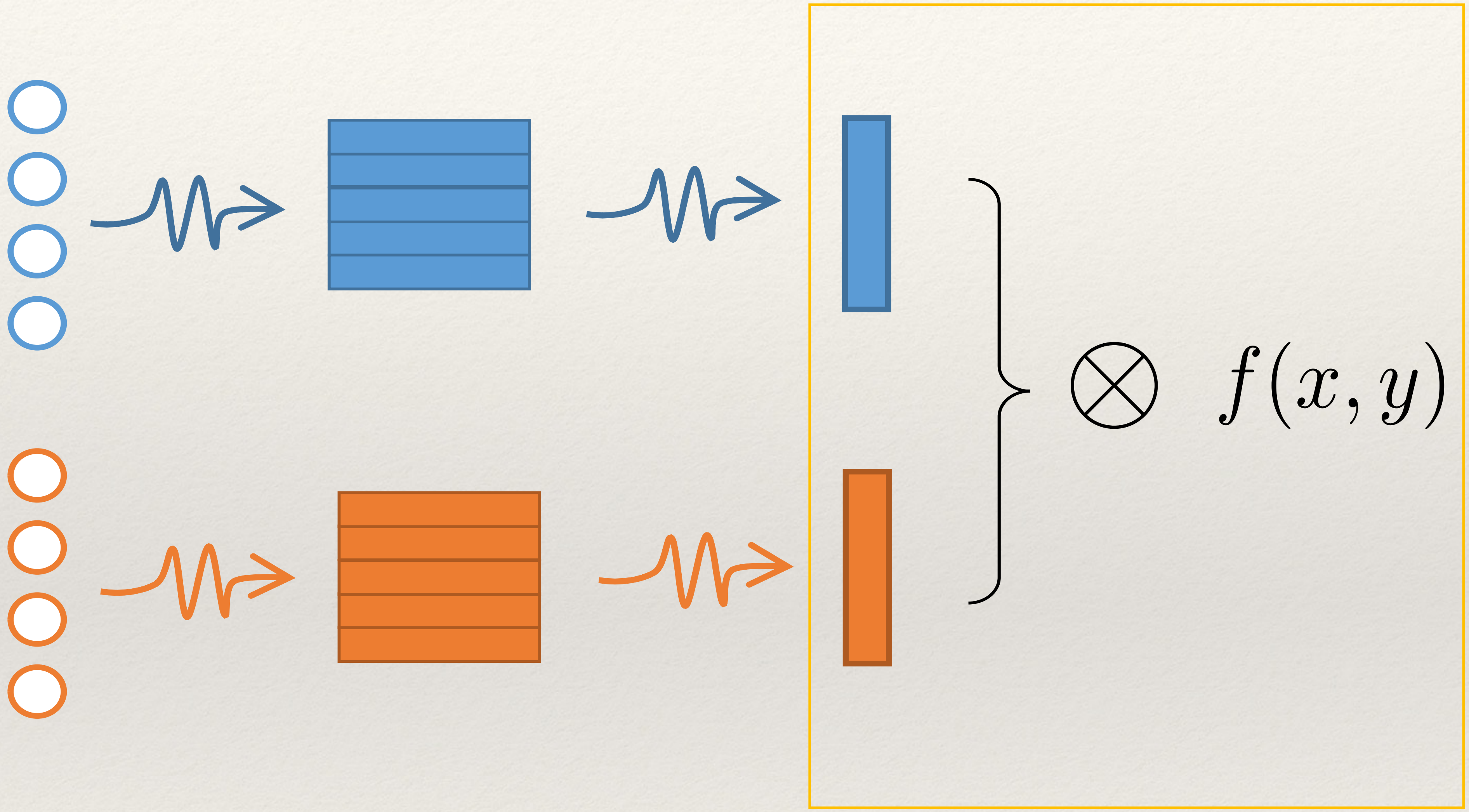
# Using RNN: LSTM-RNN

- Input: sequence letter trigrams
- Model: Long-short term memory (LSTM)
  - The last output as the sentence representation




Palangi H, Deng L, Shen Y, et al. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(4): 694-707.

# Matching Function




Heuristic: cosine, dot product
Learning: MLP, Neural tensor networks

# Matching Functions (cont')

- Given the representations of two sentences: *x* and *y*.
- Similarity between these two embeddings:
  - Cosine Similarity (DSSM, CDSSM, RNN-LSTM)

$$S = \frac{x^T \cdot y}{|x| \cdot |y|}$$

  - Dot Product

$$S = x^T \cdot y$$

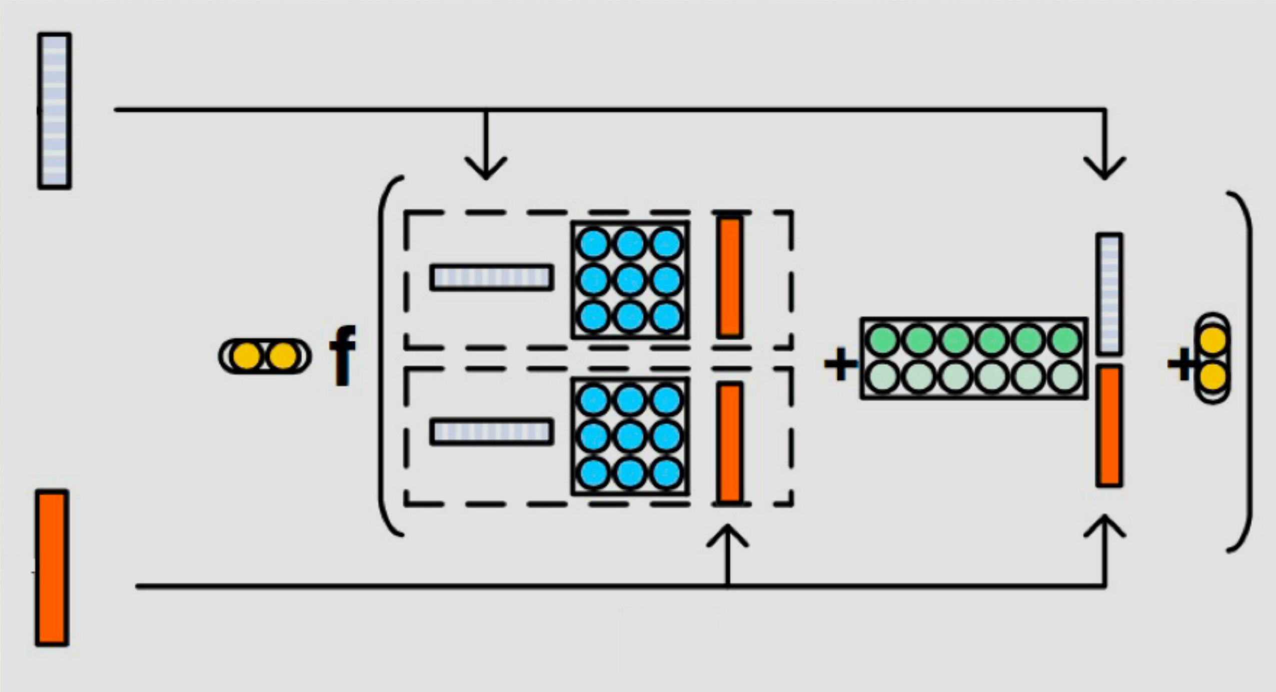
  - Multi-Layer Perception (ARC-I)

$$S = W_2 \cdot \left( W_1 \cdot \begin{bmatrix} x \\ y \end{bmatrix} + b_1 \right) + b_2$$

# Matching Functions (cont')

- Neural Tensor Networks (CNTN)

$$S = u^T f\left(x^T \mathbf{M}^{[1:r]} y + V \begin{bmatrix} x \\ y \end{bmatrix} + b\right)$$




Qiu X, Huang X. Convolutional neural tensor network architecture for community-based question answering//Proceedings of the 24th (IJCAI), Buenos Aires, Argentina, 2015: 1305-1311.

# Performance Evaluation on QA Task

- Dataset: Yahoo! Answers
  - 60,564 (question, answer) pairs
- Example:
  - *Q: How to get rid of memory stick error of my sony cyber shot?*
  - *A: You might want to try to format the memory stick but what is the error message you are receiving.*

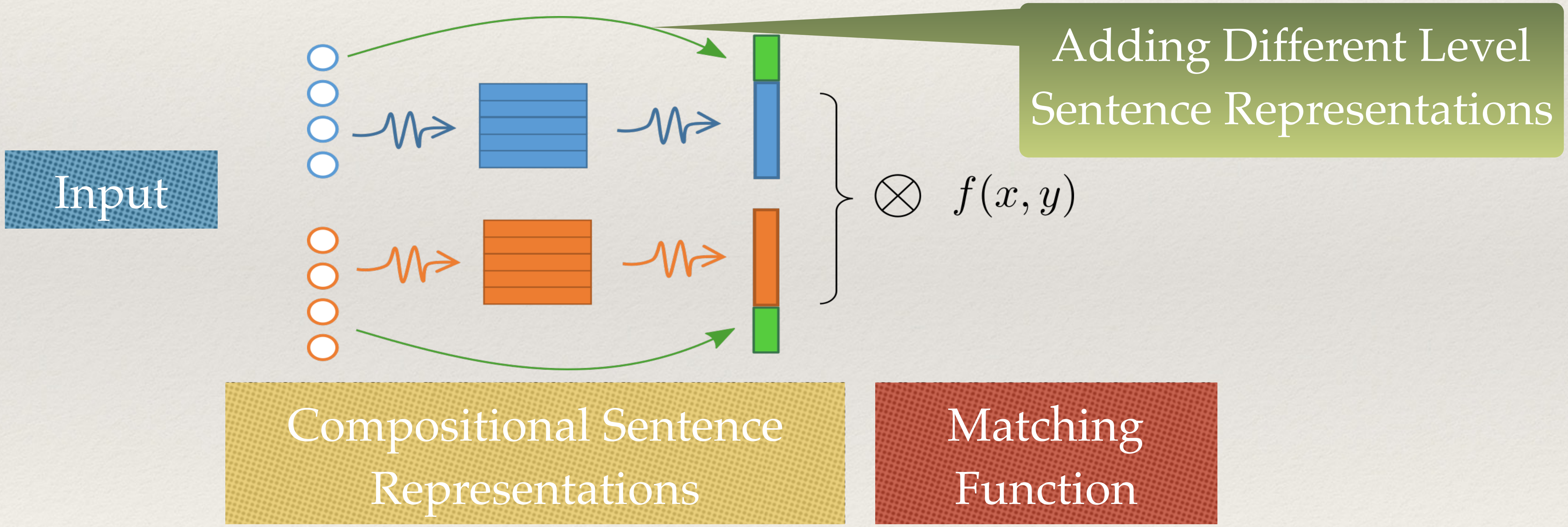
# Experimental Results

| | Model | P@1 | MRR |
|---|---|---|---|
| Random | Random | 0.200 | 0.457 |
| Traditional | BM25 | 0.579 | 0.726 |
| Comosition Focused | ARC-I | 0.581 | 0.756 |
| | CNTN | 0.626 | 0.781 |
| | LSTM-RNN | 0.690 | 0.822 |

- Composition focused methods outperformed the baselines
  - Semantic representation is important
- LSTM-RNN is the best performed method
  - Modeling the order information does help

# Extensions to Composition Focused Methods

- Problem: sentence representations are too coarse to conduct text match
  - Experience in IR: combining topic level and word level matching signals usually achieve better performancesAdding more fine-grained matching signals
- Solution: add fine-grained signals




- MultiGranCNN: An Architecture for General Matching of Text Chunks on Multiple Levels of Granularity. (Yin W, Schütze T, Hinrich. ACL2015)
- U-RAE: Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection, (Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, Christopher D. Manning, NIPS2011)
- MV-LSTM: A Deep Arhitecture for Semantic Matching with Multiple Positional Sentence Representations. (Shengxian Wan, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. AAAI 2016)

# Performance Evaluations on QA Task

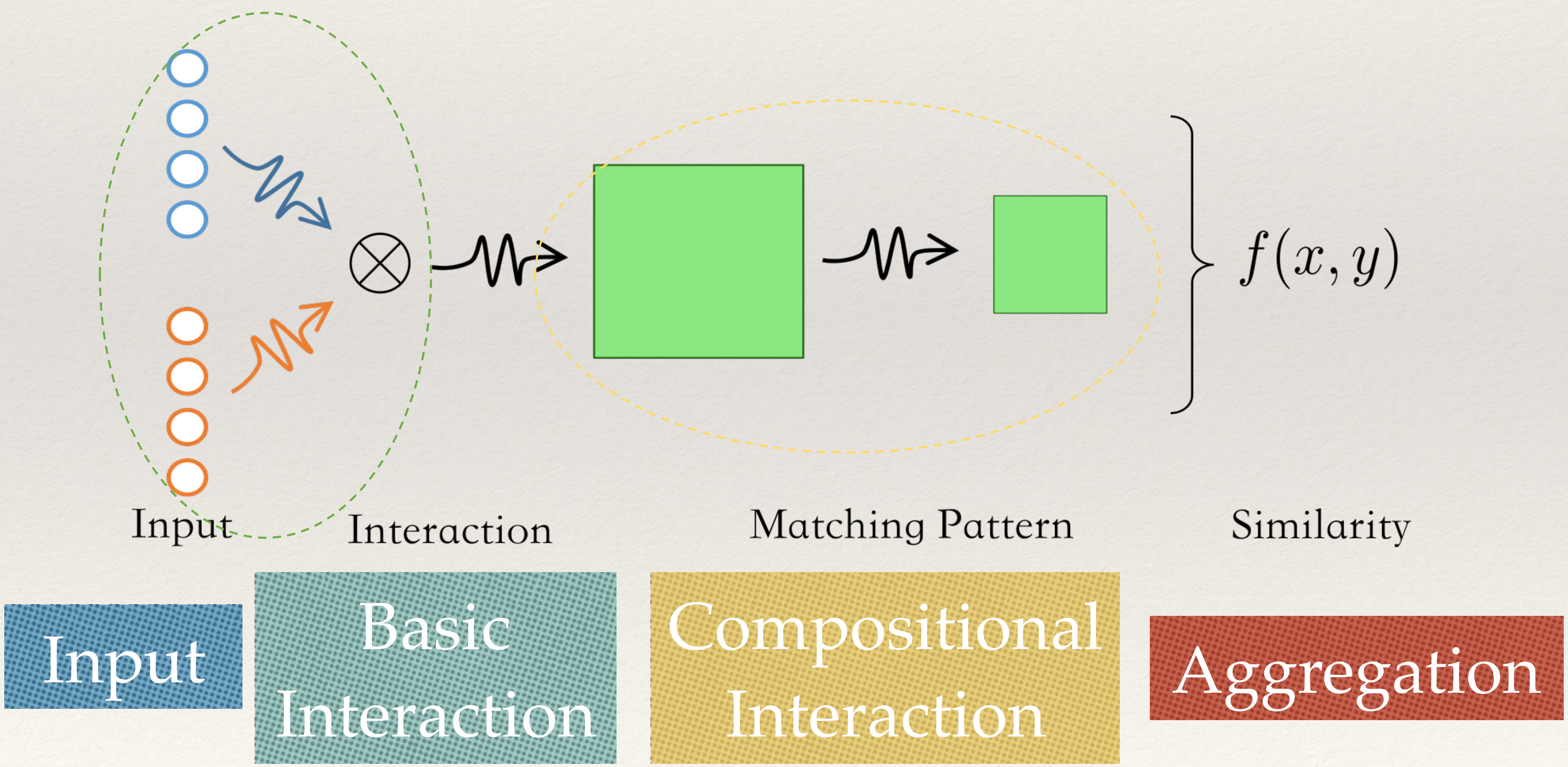
| | Model | P@1 | MRR |
|---|---|---|---|
| Statistic Traditional | Random | 0.200 | 0.457 |
| | BM25 | 0.579 | 0.726 |
| Comosition Focused | ARC-I | 0.581 | 0.756 |
| | CNTN | 0.626 | 0.781 |
| | LSTM-RNN | 0.690 | 0.822 |
| | uRAE | 0.398 | 0.652 |
| | MultiGranCNN | 0.725 | 0.840 |
| | MV-LSTM | 0.766 | 0.869 |

- MultiGranCNN and MV-LSTM achieved the best performances
    - Fine-grained matching signals are useful

# Interaction Focused Methods

# Interaction Focused Methods

- Step 1: Construct basic low-level interaction signals
- Step 2: Aggregate matching patterns

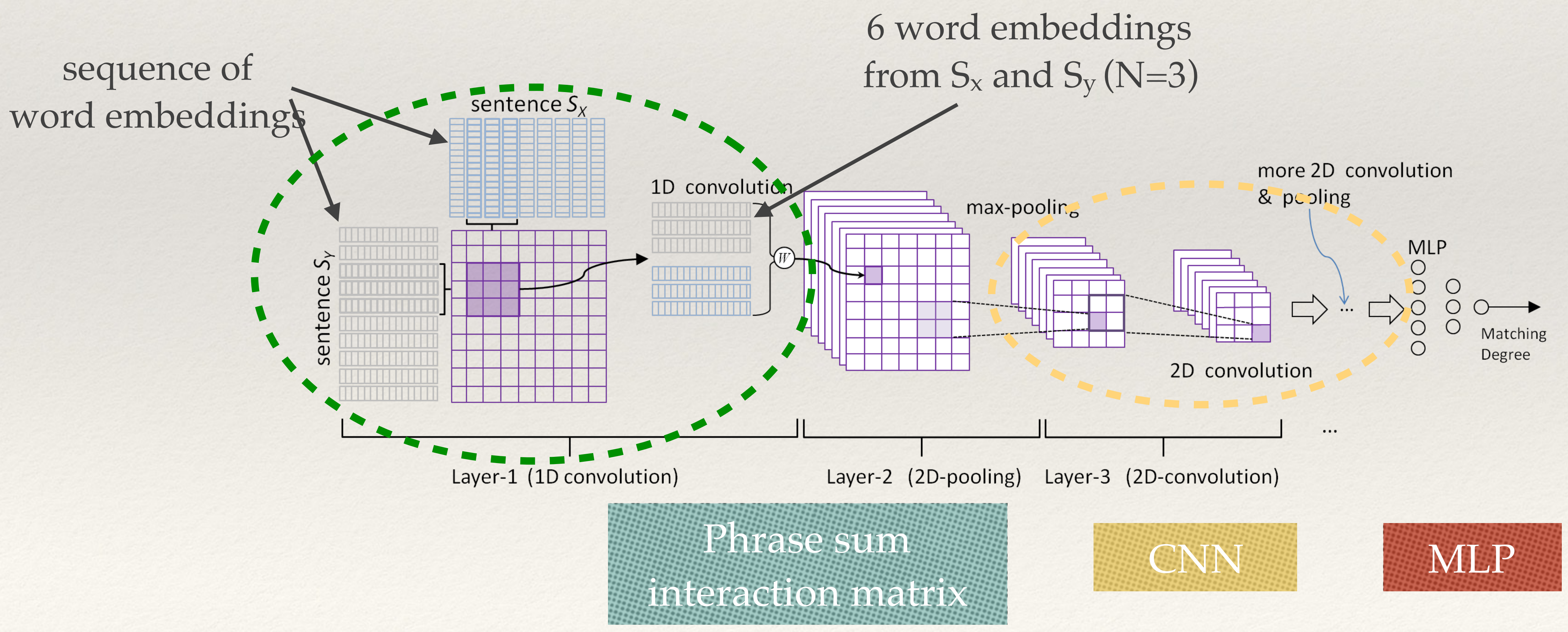


# Typical Interaction Focused Methods

- ARC II: Convolutional Neural Network Architectures for Matching Natural Language Sentences (Hu et al., NIPS'14)
- MatchPyramid: Text Matching as Image Recognition. (Pang et al. AAAI'16)
- Match-SRNN: Modeling the Recursive Matching Structure with Spatial RNN. (Wan et al. IJCAI'16)

# ARC-II

- Let two sentences meet before their own high-level representations mature.
- Basic interaction: phrase sum interaction matrix
- Compositional interaction: CNN to capture the local interaction structure
- Aggregation Function: MLP




Hu B, Lu Z, Li H, et al. Convolutional neural network architectures for matching natural language sentences//Proceedings of the Advances in NIPS, Montreal, Canada, 2014: 2042-2050.

# ARC-II (cont')

- Order Preservation
  - Both the convolution and pooling have order preserving property

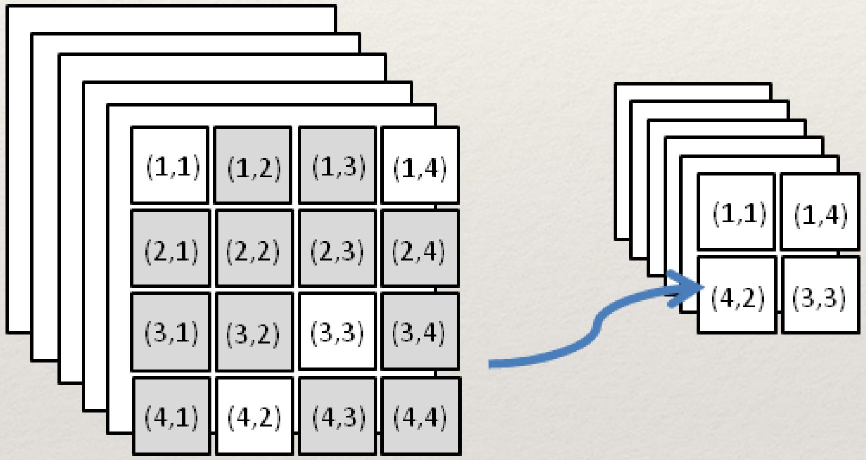
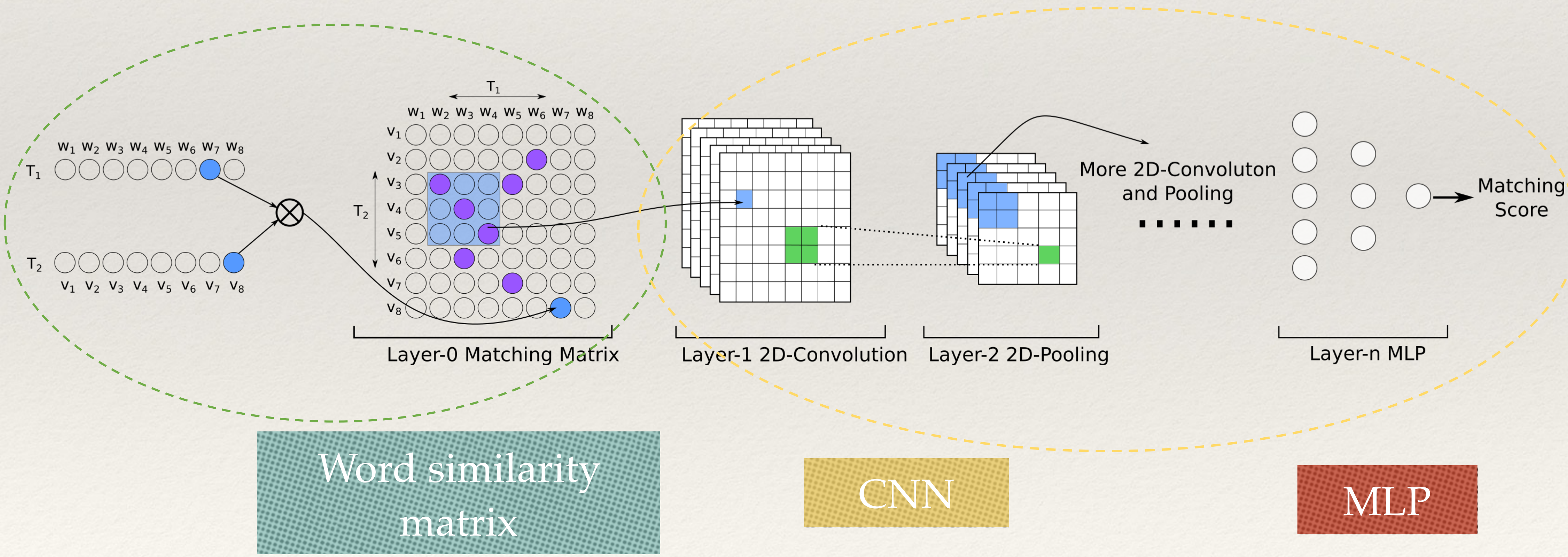



Figure 5: Order preserving in 2D-pooling.

- However, the word level matching signals are lost
  - 2-D matching matrix is construct based on the embedding of the words in two N-grams

# MatchPyramid

- Inspired by image recognition task
- Basic Interaction: word-level matching matrix
- Compositional interaction: hierarchical convolution
- Aggregation: MLP

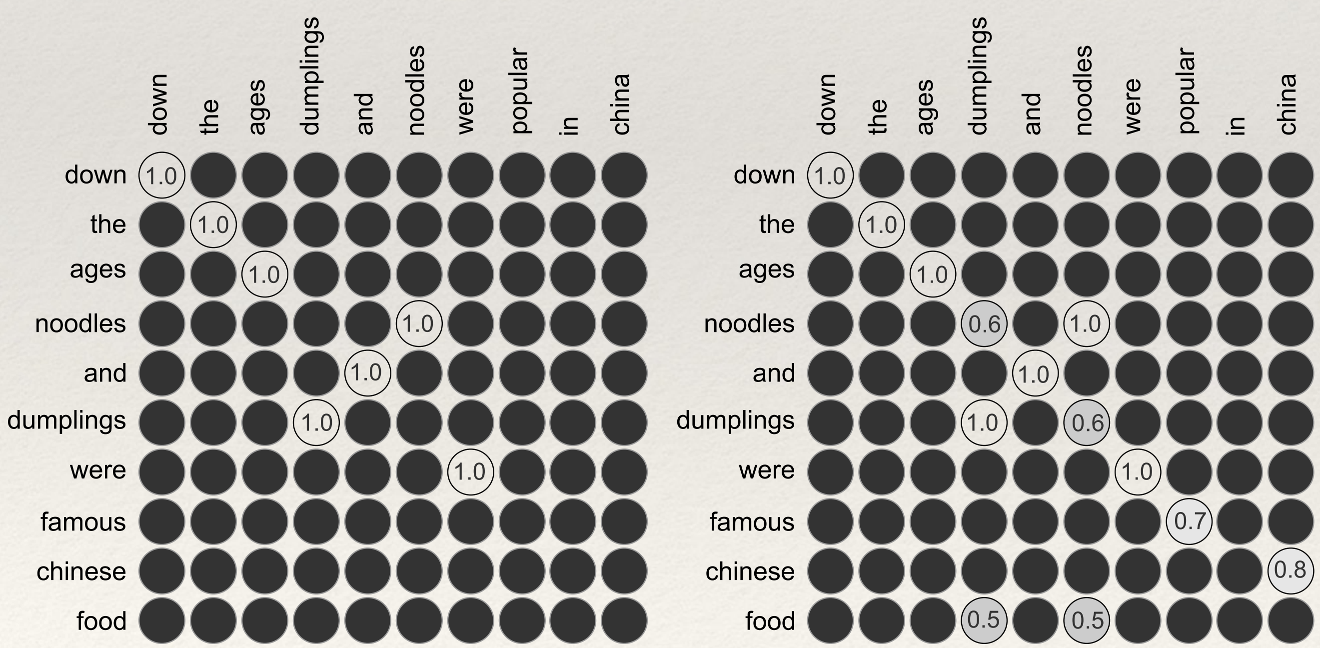



Pang L, Lan Y, Guo J, et al. Text matching as image recognition//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 2793-2799

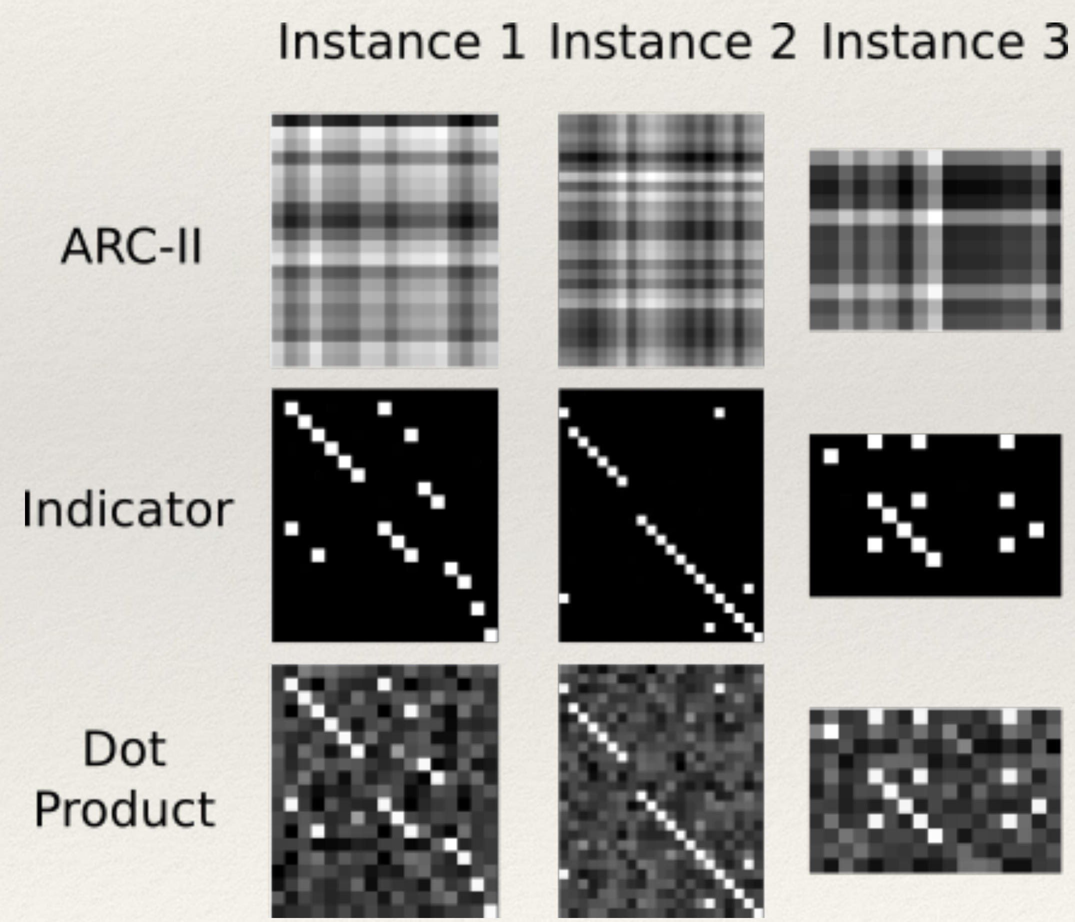
# MatchPyramid: Matching Matrix

- Basic Interaction: word similarity matrix
  - Strength of the word-level matching
  - Positions of the matching occurs

$$\mathbf{M}_{ij} = w_i \otimes v_j$$

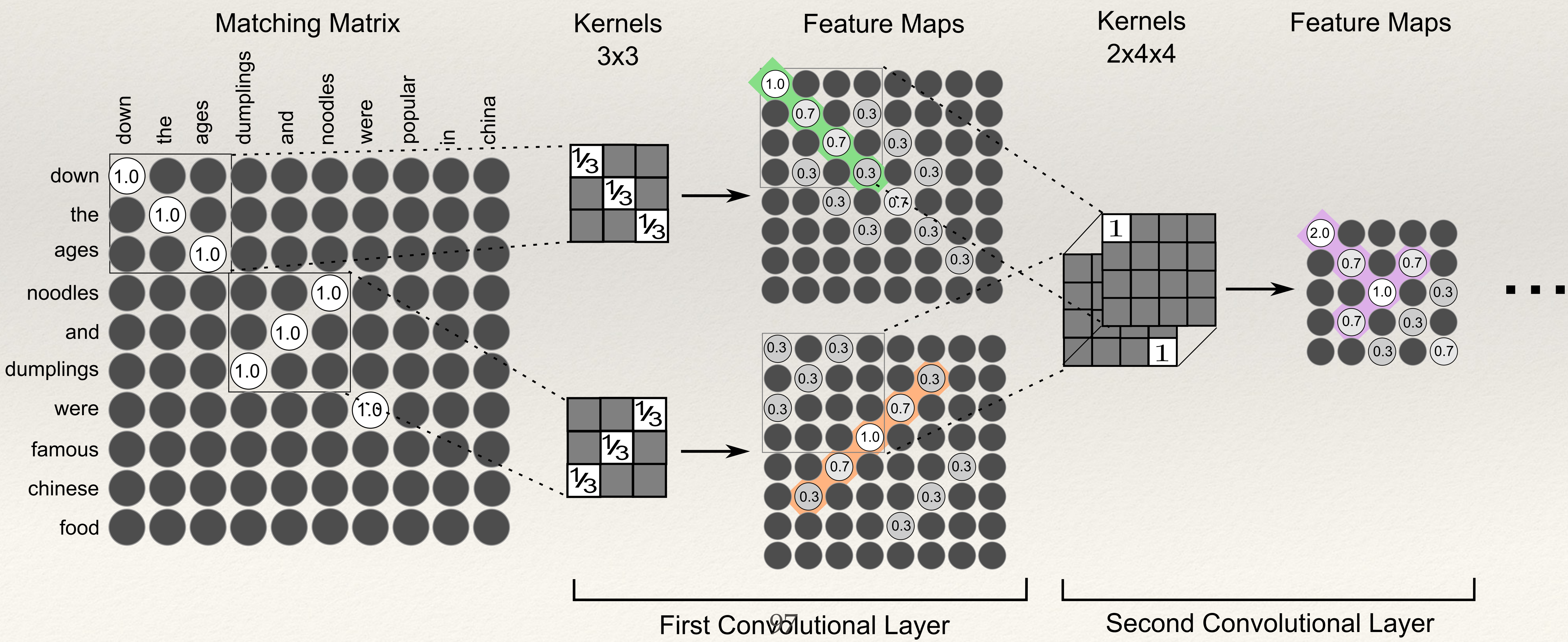



(a) Indicator (b) Cosine

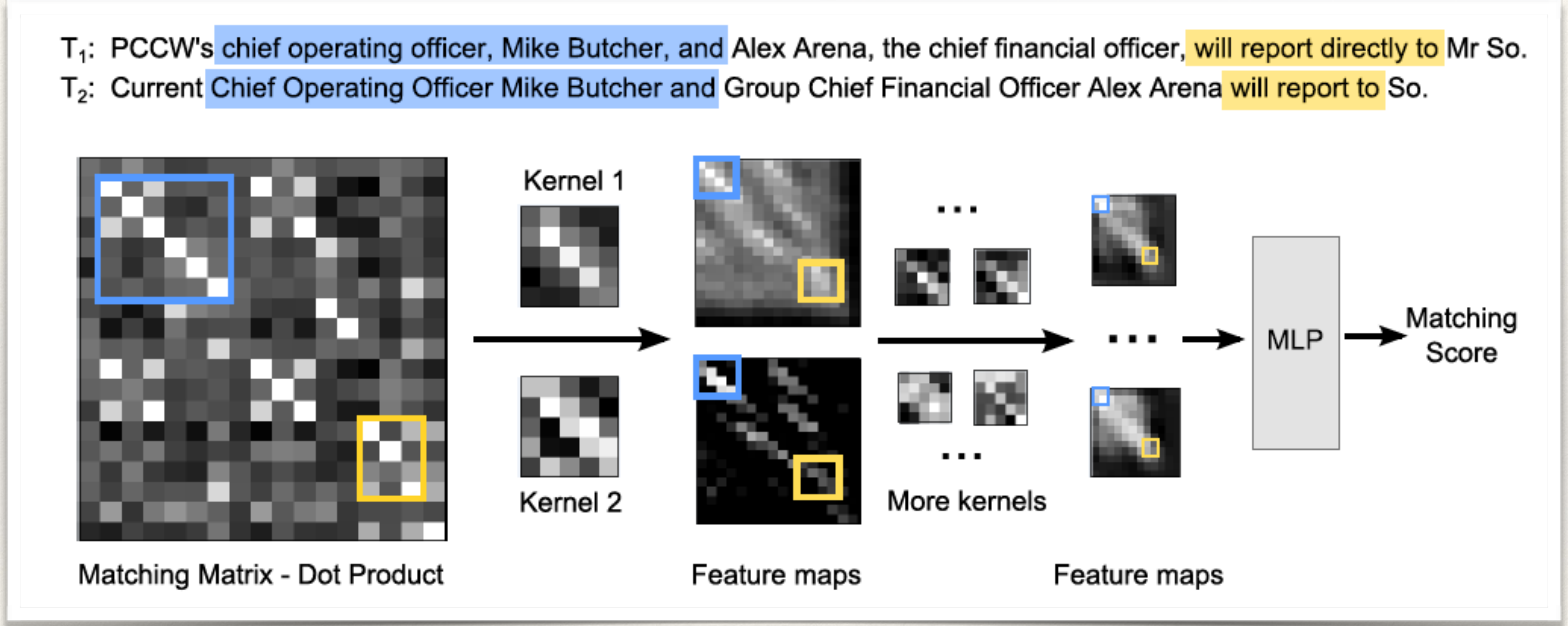


# MatchPyramid - Hierarchical Convolution

- Compositional interaction: CNN to capture different levels of matching patterns, based on word-level matching signals

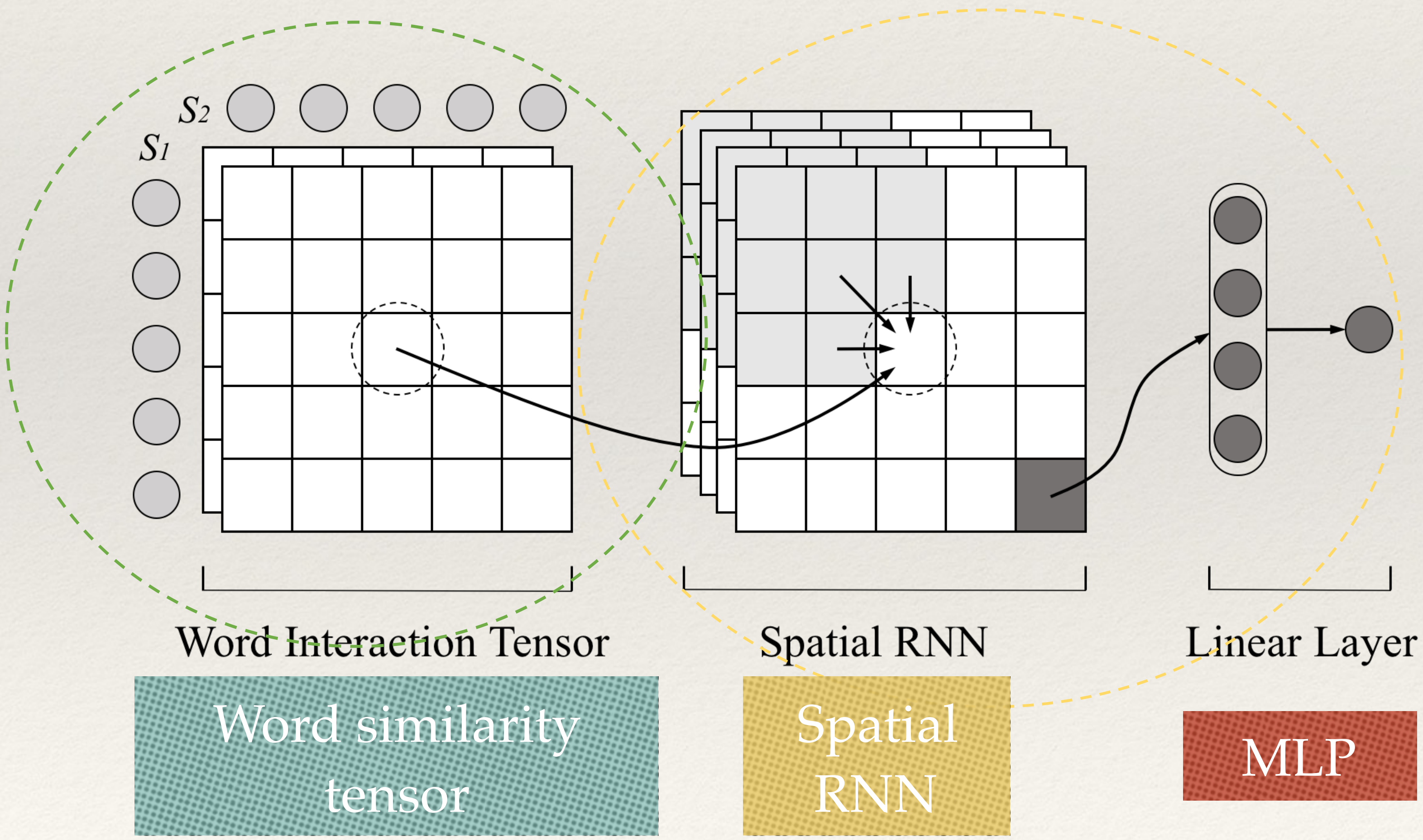


# Matching Patterns Discovered by MathPyramid



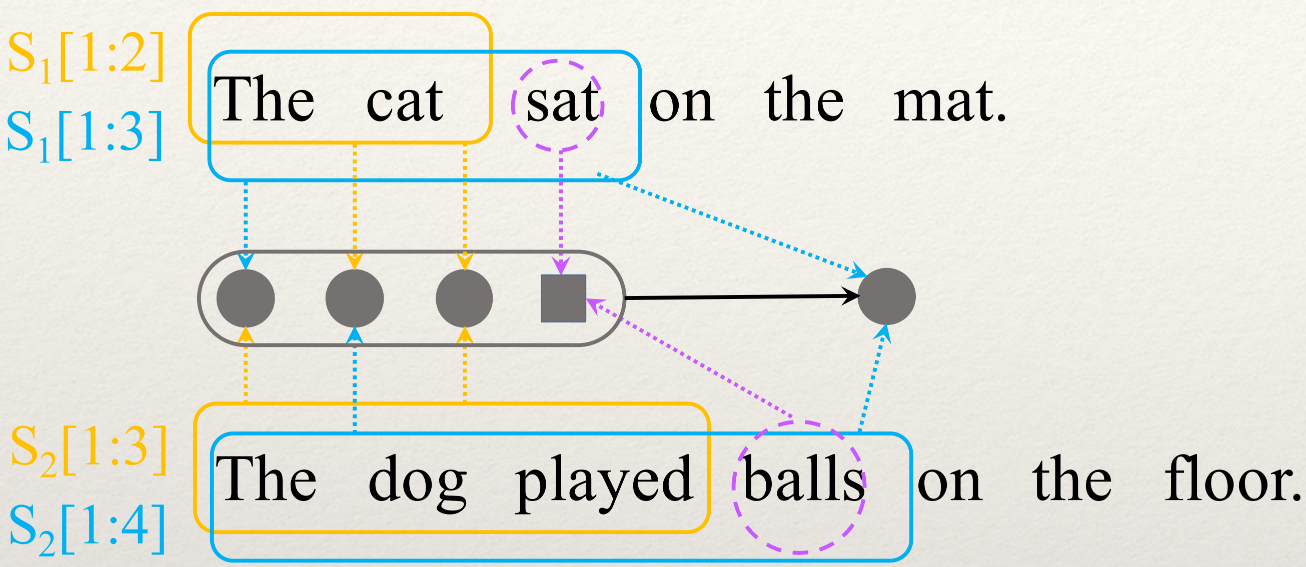
# Match-SRNN

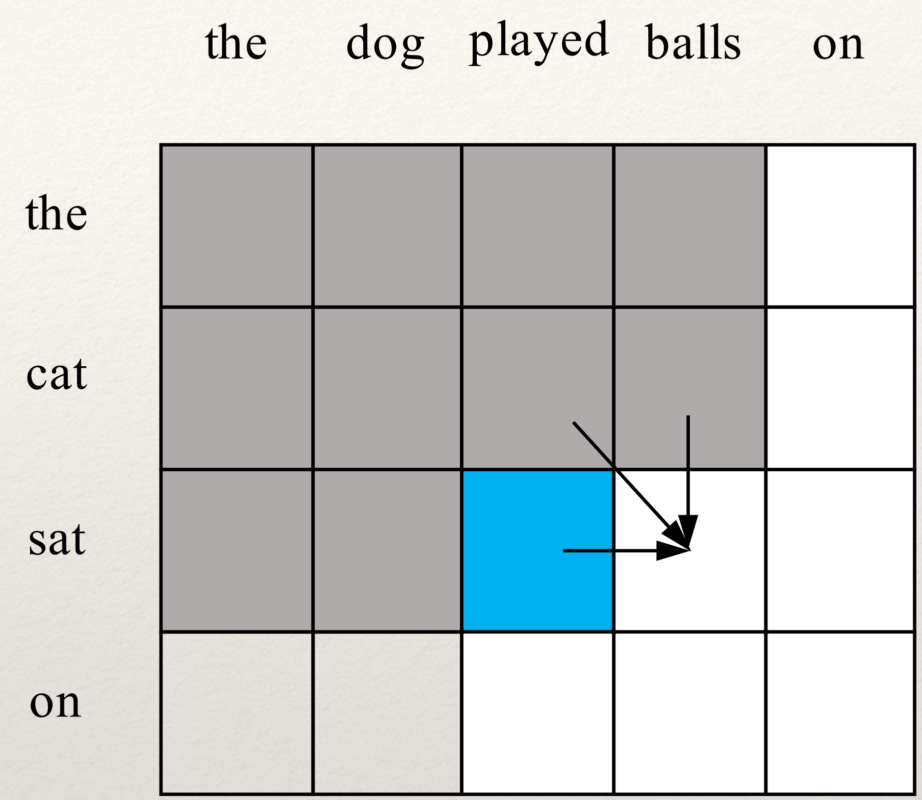
- Spatial recurrent neural network (SRNN) for text matching
- Basic interaction: word similarity tensor
- Compositional interaction: recursive matching
- Aggregation: MLP




Wan S, Lan Y, Guo J, et al. Match-SRNN: Modeling the recursive matching structure with spatial RNN//Proceedings of the 25th IJCAI, New York, US, 2016: 1022-1029.

# Match-SRNN: Recursive Matching Structure




- Matching scores are calculated recursively (from top left to bottom right)
- We can see all matching between sub sentences have been utilized
  - sat <—> balls
  - The cat <—> the dog played
  - The cat ←→ The dog played balls
  - The cat sat ←→ The dog played

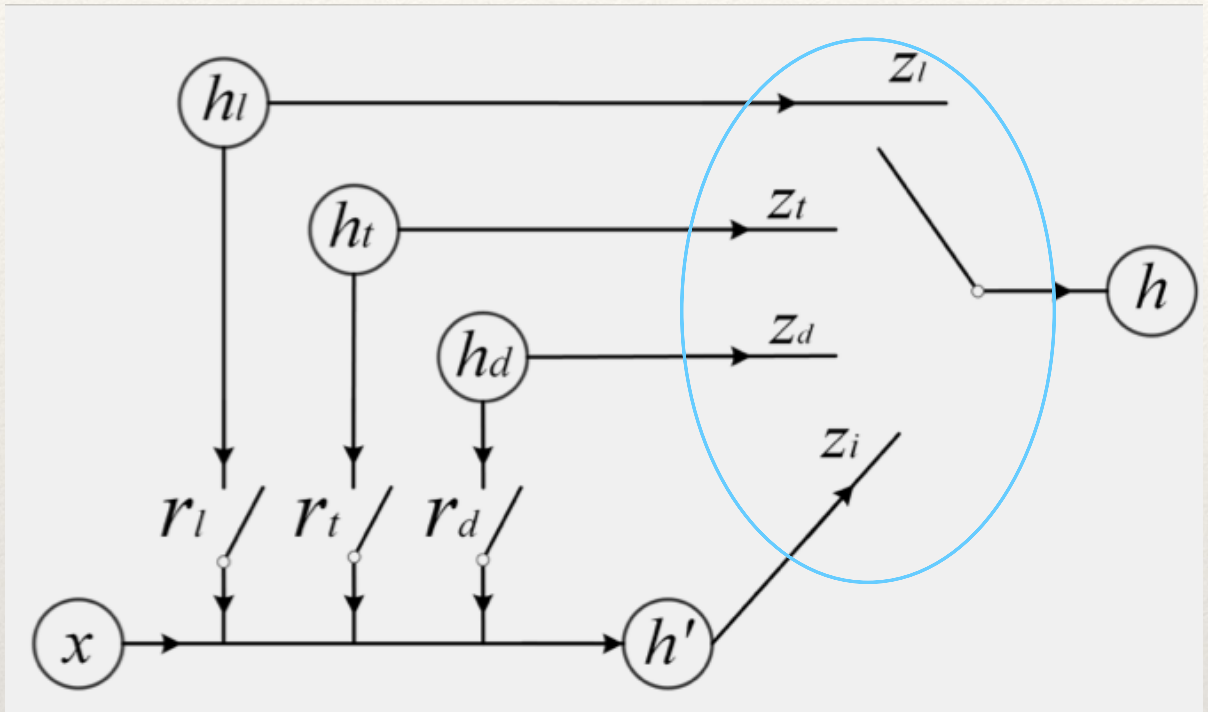
# Match-SRNN: Recursive Matching Structure (cont')

- Definition

  $S_1 = w_1, \cdots, w_m, S_2 = v_1, \cdots, v_n$

  - $S_1[1:i]$: prefix of of length i
  - $S_2[1:j]$: prefix of length j
  - $h_{ij}$: match representation between $S_1[1:i]$ and $S_2[1:j]$
- We have $h_{i,j} = f(h_{i-1,j} h_{i,j-1} h_{i-1,j-1} s(w_i, v_j))$

# Using Spatial GRU (two dimensions)




Softmax function is used to 'soft' choose connections among four choices.

$$
\begin{aligned}
q^T &= [h^T_{i-1,j}, h^T_{i,j-1}, h^T_{i-1,j-1}, s^T_{ij}]^T,\\
r_l &= \sigma(W^{(r_l)}q + b^{(r_l)}),\\
r_t &= \sigma(W^{(r_t)}q + b^{(r_t)}),\\
r_d &= \sigma(W^{(r_d)}q + b^{(r_d)}),\\
r^T &= [r_l^T, r_t^T, r_d^T]^T,\\
z_i^{'} &= W^{(z_i)}q + b^{(z_i)},\\
z_l^{'} &= W^{(z_l)}q + b^{(z_l)},\\
z_t^{'} &= W^{(z_t)}q + b^{(z_t)},\\
z_d^{'} &= W^{(z_d)}q + b^{(z_d)},\\
&\boxed{[z_i, z_l, z_t, z_d] = \mathrm{SoftmaxByRow}([\mathrm{z}_\mathrm{i}^{'}, \mathrm{z}_\mathrm{l}^{'}, \mathrm{z}_\mathrm{t}^{'}, \mathrm{z}_\mathrm{d}^{'}]),}\\
h_{i,j}^{'} &= \phi(Ws_{ij} + U(r \odot [h^T_{i,j-1}, h^T_{i-1,j}, h^T_{i-1,j-1}]^T) + b),\\
h_{i,j} &= z_l \odot h_{i,j-1} + z_t \odot h_{i-1,j} + z_d \odot h_{i-1,j-1} + z_i \odot h_{i,j}^{'}.
\end{aligned}
$$

# Connection with LCS

- Longest Common Sub-Sequence
  - S1: A B C D E
  - S2: F A C G D
  - LCS: A C D

| | A | B | C | D | E |
|---|---|---|---|---|---|
| F | 0 | 0 | 0 | 0 | 0 |
| A | 1 | 1 | 1 | 1 | 1 |
| C | 1 | 1 | 2 | 2 | 2 |
| G | 1 | 1 | 2 | 2 | 2 |
| D | 1 | 1 | 2 | 3 | 3 |

- Solving LCS with dynamic programming
- Step function: $c[i,j] = max(c[i,j-1], c[i-1,j], c[i-1,j-1] + \mathbb{I}_{x_i=y_j})$
  - Backtrace: depends on the selection of "max" operation

# Connection with LCS

- Matching-SRNN can be explained with LCS
- Simplify Match-SRNN
  - Using exact word level matching signals only
  - remove the reset gate r and set hidden dimensions to 1

$$h_{i,j} = z_l \cdot \boxed{h_{i,j-1}} + z_t \cdot \boxed{h_{i-1,j}} + \boxed{z_d \cdot h_{i-1,j-1} + z_i \cdot h'_{ij}}$$

- Simplified Match-SRNN simulates LCS

$$c[i,j] = max(\boxed{c[i,j-1]}, \boxed{c[i-1,j]}, \boxed{c[i-1,j-1] + \mathbb{I}_{x_i=y_j}})$$

  - *z* is obtained by SOFTMAX
  - Backtrace by the value of *z* in simplified Match-SRNN

# Simulation with Simplified Math-SRNN

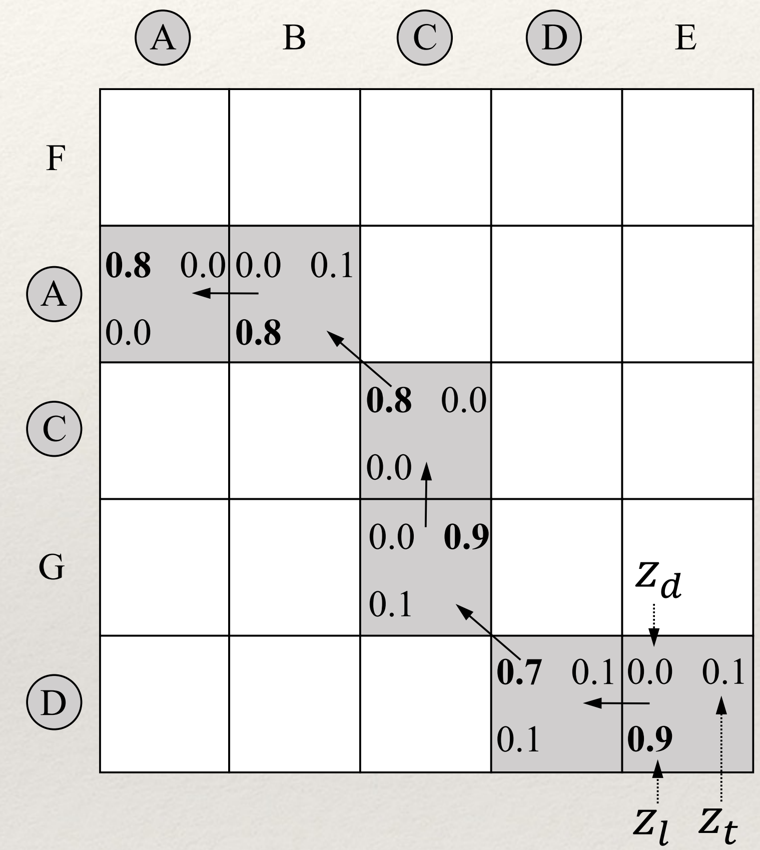
- Simulation data
  - random sampled sequence
  - ground truth obtained by DP
  - the label is the length of LCS

| | A | B | C | D | E |
|---|---|---|---|---|---|
| F | 0 | 0 | 0 | 0 | 0 |
| A | 1 | 1 | 1 | 1 | 1 |
| C | 1 | 1 | 2 | 2 | 2 |
| G | 1 | 1 | 2 | 2 | 2 |
| D | 1 | 1 | 2 | 3 | 3 |

(a)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| F | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| A | 1.0 | 1.0 | 1.0 | 1.0 | 0.9 |
| C | 1.0 | 1.0 | 2.1 | 2.1 | 2.0 |
| G | 1.0 | 1.0 | 2.1 | 2.0 | 2.0 |
| D | 1.0 | 1.0 | 2.0 | 3.1 | 3.1 |

(b)

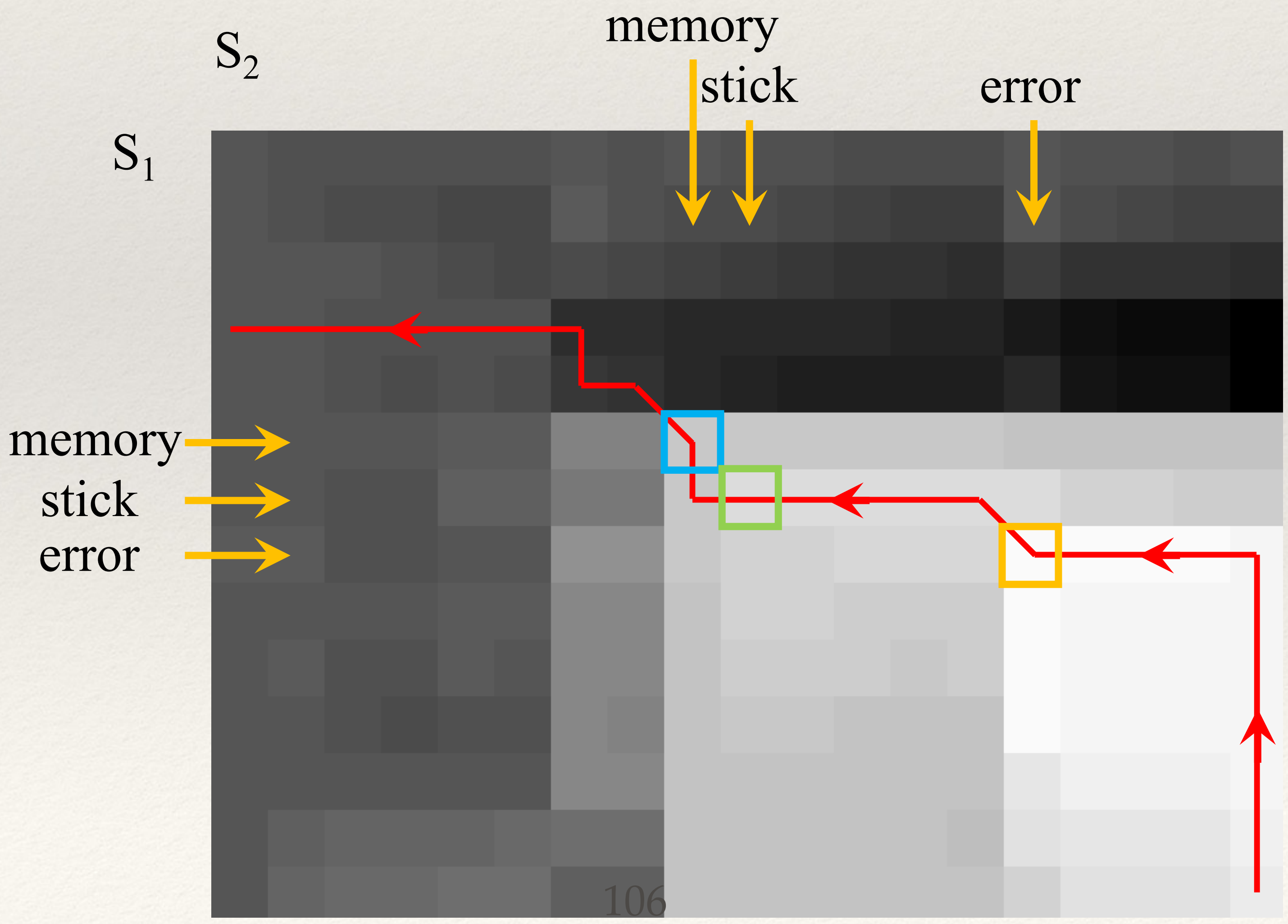



(c)

Match-SRNN simulates LCS!

# On Real Data

- *Question*: *"How to get rid of memory stick error of my sony cyber shot?"*
- *Answer*: *"You might want to try to format the memory stick but what is the error message you are receiving."*



# Performance Evaluations on QA Task

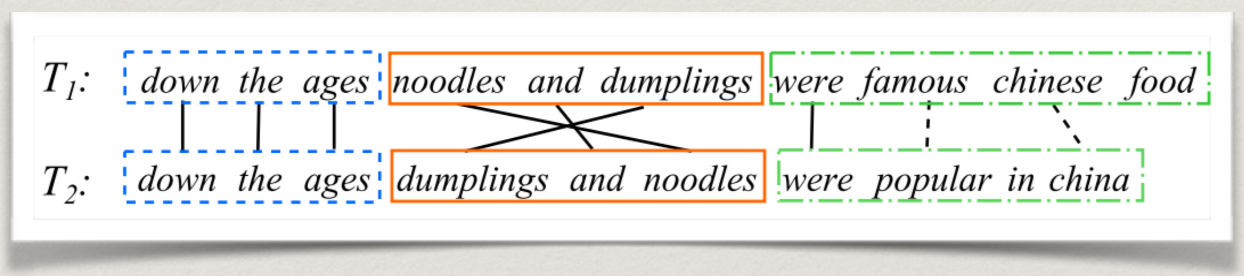
| | Model | P@1 | MRR |
|---|---|---|---|
| Statistic Traditional | Random | 0.200 | 0.457 |
| | BM25 | 0.579 | 0.726 |
| Comosition Focused | ARC-I | 0.581 | 0.756 |
| | CNTN | 0.626 | 0.781 |
| | LSTM-RNN | 0.690 | 0.822 |
| | uRAE | 0.398 | 0.652 |
| | MultiGranCNN | 0.725 | 0.840 |
| | MV-LSTM | 0.766 | 0.869 |
| Interaction Focused | DeepMatch | 0.452 | 0.679 |
| | ARC-II | 0.591 | 0.765 |
| | MatchPyramid | 0.764 | 0.867 |
| | Match-SRNN | 0.790 | 0.882 |

- Interaction focused methods outperformed the composition focused ones
    - Low level interaction (word level) signals are important
- Match-SRNN performs the best
    - Powerful recursive matching structure

# Application to Search —— Document Level Matching

# Document Level Matching: Aggregating Matching Signals

## Sentence-sentence matching

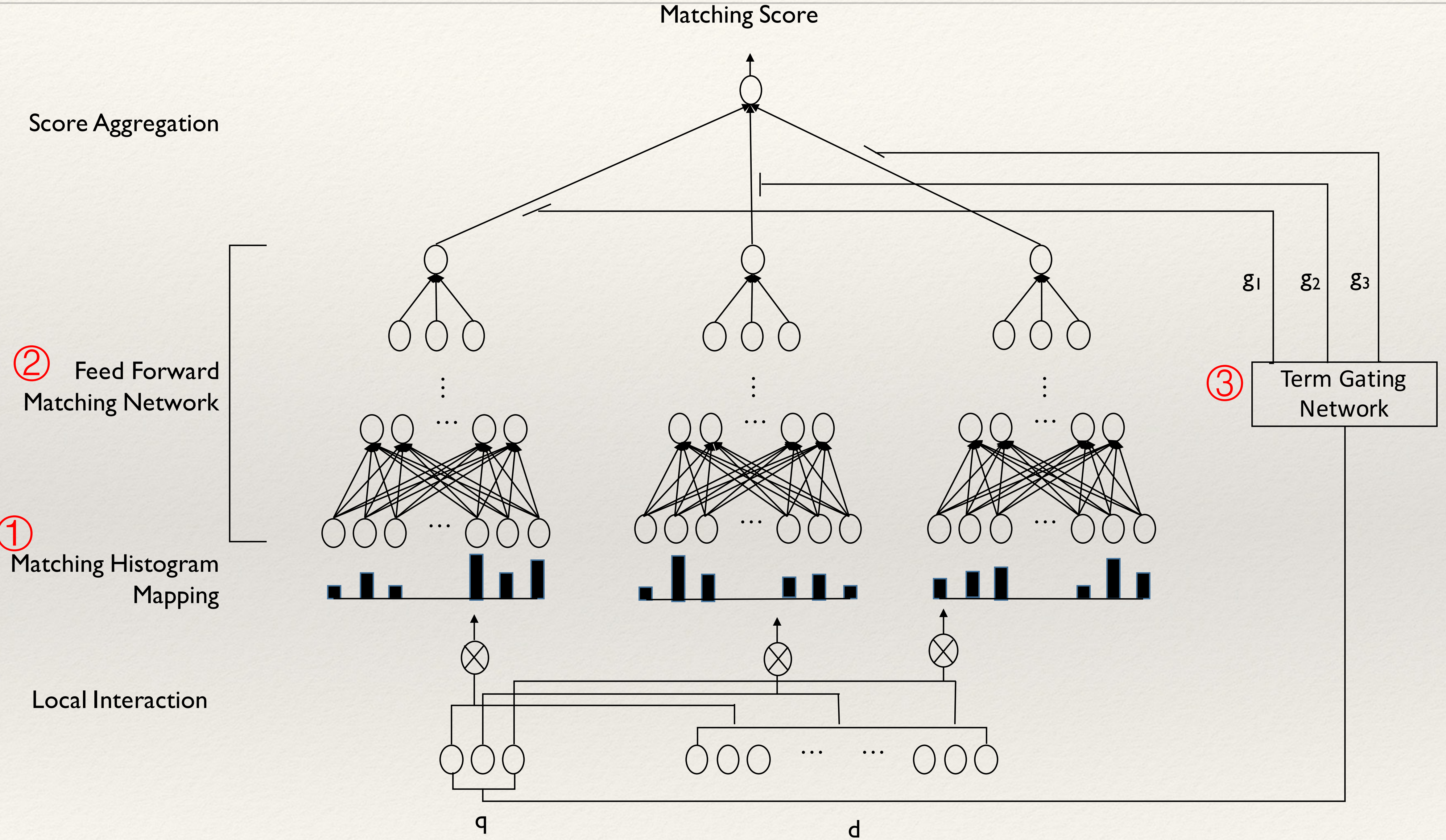



## Query-document matching

query


document

# Deep Relevance Matching Model (DRMM)




Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A Deep Relevance Matching Model for Ad-hoc Retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (CIKM '16). ACM, New York, NY, USA, 55-64.

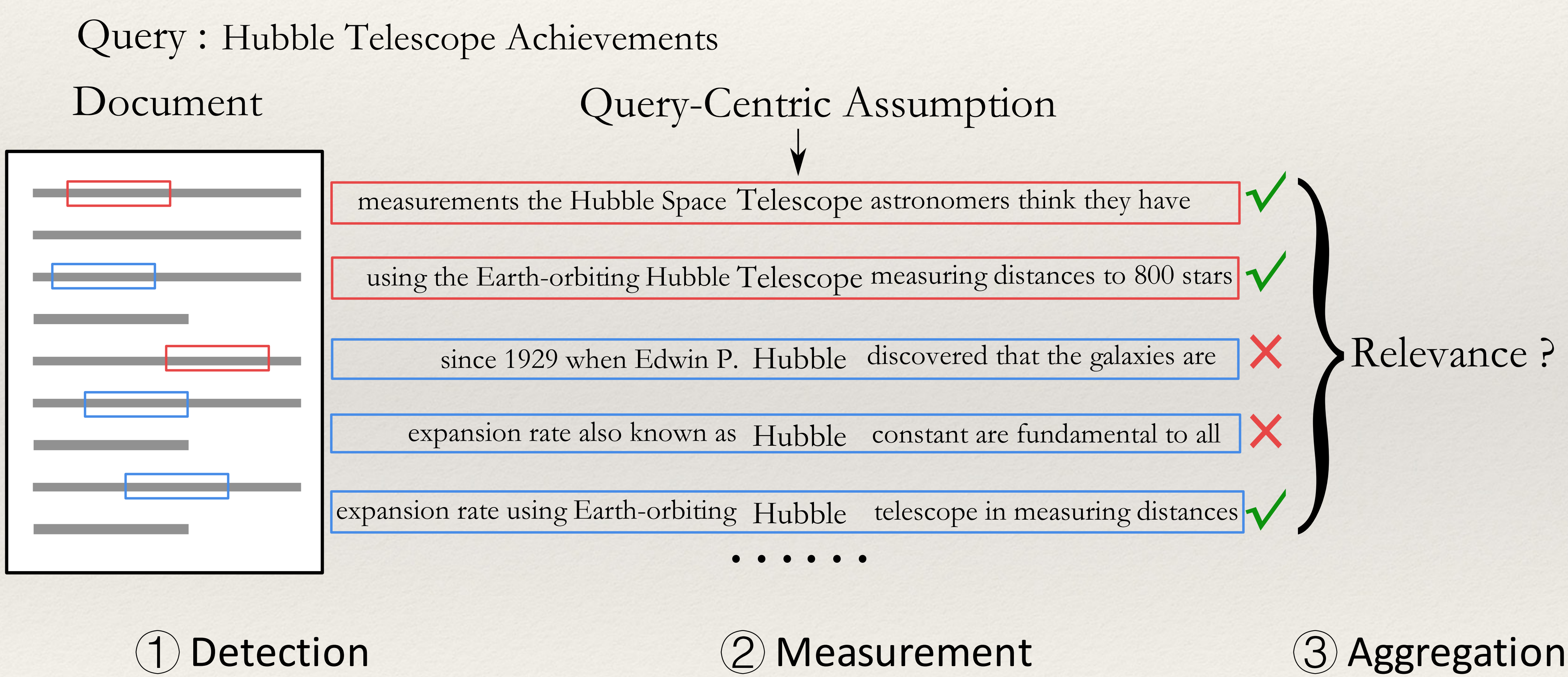
# Deep Relevance Matching Model (cont')

- Learning the parameters
  - Pairwise loss: $\ell(\mathbf{q}, \mathbf{d}^+, \mathbf{d}^-) = \max(0, 1 - F(\mathbf{q}, \mathbf{d}^+) + F(\mathbf{q}, \mathbf{d}^-))$
  - Optimization with stochastic gradient descent
- Experimental results (Robust-04 collection)

| | Using topic titles | | | Using topic descriptions | | |
|---|---|---|---|---|---|---|
| | MAP | nDCG@20 | P@20 | MAP | nDCG@20 | P@20 |
| DSSM | 0.095 | 0.201 | 0.171 | 0.078 | 0.169 | 0.145 |
| CDSSM | 0.067 | 0.146 | 0.125 | 0.050 | 0.113 | 0.093 |
| ARC-I | 0.041 | 0.066 | 0.065 | 0.030 | 0.047 | 0.045 |
| ARC-II | 0.067 | 0.147 | 0.128 | 0.042 | 0.086 | 0.074 |
| MP-IND | 0.169 | 0.319 | 0.281 | 0.067 | 0.142 | 0.118 |
| MP-COS | 0.189 | 0.330 | 0.290 | 0.094 | 0.190 | 0.162 |
| MP-DOT | 0.083 | 0.159 | 0.155 | 0.047 | 0.104 | 0.092 |
| DRMM | **0.279** | **0.431** | **0.382** | **0.275** | **0.437** | **0.371** |

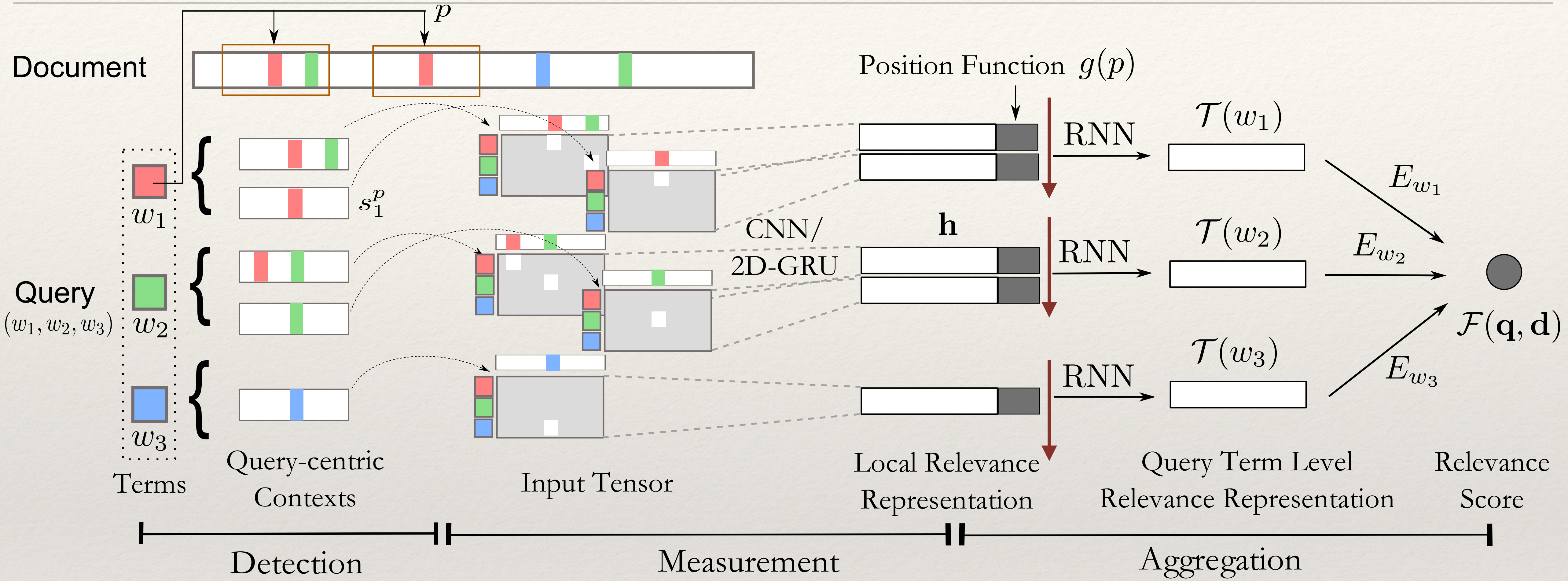
# DeepRank: Semantic Query-Document Matching

- Motivation: mimicking human-judgment of relevance




Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. DeepRank: a New Deep Architecture for Relevance Ranking in Information Retrieval. Proceedings of the 26th ACM International Conference on Information and Knowledge Management (CIKM '17)

# DeepRank




Focusing on the location of query terms when scanning the whole document

Determine local relevance – relevance between query and each query term -centric context, using MatchPyramid / MatchSRNN

Query term level aggregation

$$F(\mathbf{q}, \mathbf{d}) = \sum_{w \in \mathbf{q}} (E_w \mathbb{I})^T \cdot \mathcal{T}(w)$$

# Learning and Empirical Evaluation

- Learning the parameters
  - Pairwise loss: $\ell(\mathbf{q}, \mathbf{d}^+, \mathbf{d}^-) = \max(0, 1 - F(\mathbf{q}, \mathbf{d}^+) + F(\mathbf{q}, \mathbf{d}^-))$
  - Optimization: stochastic gradient decent
- Experimental results

MQ2007

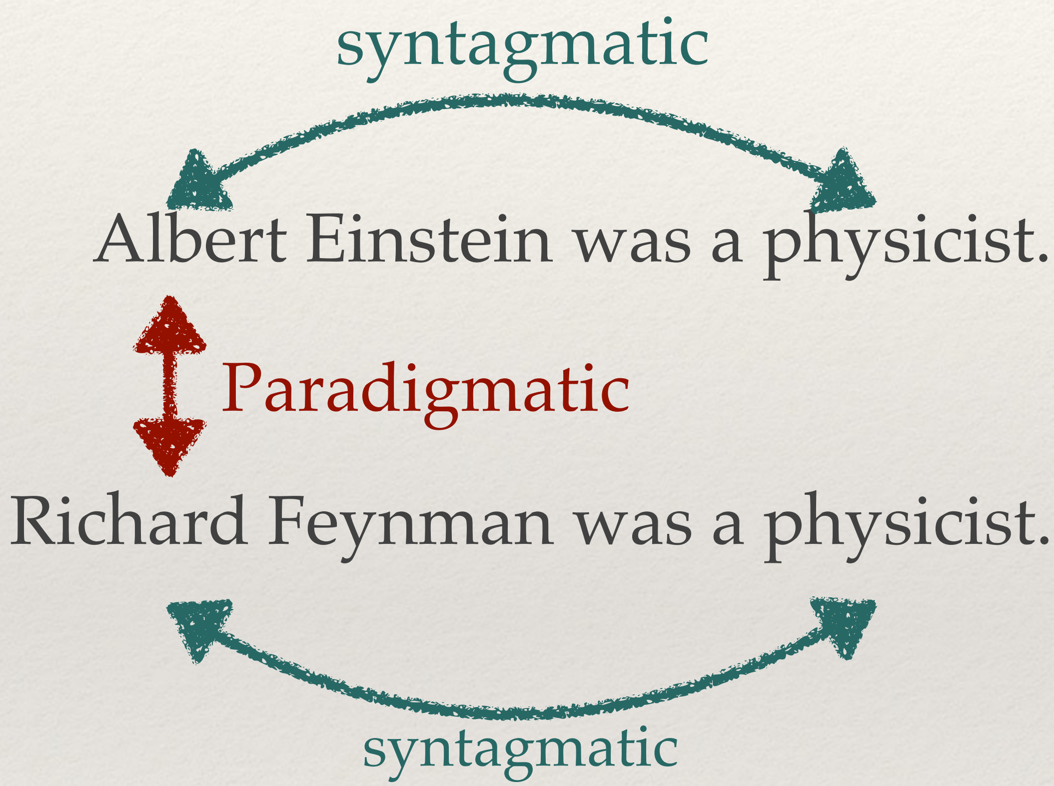
| Model | NDCG@1 | NDCG@3 | NDCG@5 | NDCG@10 | P@1 | P@3 | P@5 | P@10 | MAP |
|---|---|---|---|---|---|---|---|---|---|
| BM25-TITLE | 0.358$^-$ | 0.372$^-$ | 0.384$^-$ | 0.414$^-$ | 0.427$^-$ | 0.404$^-$ | 0.388$^-$ | 0.366$^-$ | 0.450$^-$ |
| RANKSVM | 0.408$^-$ | 0.405$^-$ | 0.414$^-$ | 0.442$^-$ | 0.472$^-$ | 0.432$^-$ | 0.413$^-$ | 0.381$^-$ | 0.464$^-$ |
| RANKBOOST | 0.401$^-$ | 0.404$^-$ | 0.410$^-$ | 0.436$^-$ | 0.462$^-$ | 0.428$^-$ | 0.405$^-$ | 0.374$^-$ | 0.457$^-$ |
| ADARANK | 0.400$^-$ | 0.410$^-$ | 0.415$^-$ | 0.439$^-$ | 0.461$^-$ | 0.431$^-$ | 0.408$^-$ | 0.373$^-$ | 0.460$^-$ |
| LAMBDAMART | 0.412$^-$ | 0.418$^-$ | 0.421$^-$ | 0.446$^-$ | 0.481$^-$ | 0.444$^-$ | 0.418$^-$ | 0.384$^-$ | 0.468$^-$ |
| DSSM | 0.290$^-$ | 0.319$^-$ | 0.335$^-$ | 0.371$^-$ | 0.345$^-$ | 0.359$^-$ | 0.359$^-$ | 0.352$^-$ | 0.409$^-$ |
| CDSSM | 0.288$^-$ | 0.288$^-$ | 0.297$^-$ | 0.325$^-$ | 0.333$^-$ | 0.309$^-$ | 0.301$^-$ | 0.291$^-$ | 0.364$^-$ |
| ARC-I | 0.310$^-$ | 0.334$^-$ | 0.348$^-$ | 0.386$^-$ | 0.376$^-$ | 0.377$^-$ | 0.370$^-$ | 0.364$^-$ | 0.417$^-$ |
| SQA-NOFEAT | 0.309$^-$ | 0.333$^-$ | 0.348$^-$ | 0.386$^-$ | 0.375$^-$ | 0.373$^-$ | 0.372$^-$ | 0.364$^-$ | 0.419$^-$ |
| DRMM | 0.380$^-$ | 0.396$^-$ | 0.408$^-$ | 0.440$^-$ | 0.450$^-$ | 0.430$^-$ | 0.417$^-$ | 0.388$^-$ | 0.467$^-$ |
| ARC-II | 0.317$^-$ | 0.338$^-$ | 0.354$^-$ | 0.390$^-$ | 0.379$^-$ | 0.378$^-$ | 0.377$^-$ | 0.366$^-$ | 0.421$^-$ |
| MATCHPYRAMID | 0.362$^-$ | 0.364$^-$ | 0.379$^-$ | 0.409$^-$ | 0.428$^-$ | 0.404$^-$ | 0.397$^-$ | 0.371$^-$ | 0.434$^-$ |
| MATCH-SRNN | 0.392$^-$ | 0.402$^-$ | 0.409$^-$ | 0.435$^-$ | 0.460$^-$ | 0.436$^-$ | 0.413$^-$ | 0.384$^-$ | 0.456$^-$ |
| DEEPRANK-2DGRU | 0.439 | 0.439 | 0.447 | 0.473 | **0.513** | 0.467 | 0.443 | 0.405 | 0.489 |
| DEEPRANK-CNN | **0.441** | **0.447** | **0.457** | **0.482** | 0.508 | **0.474** | **0.452** | **0.412** | **0.497** |

# Outline

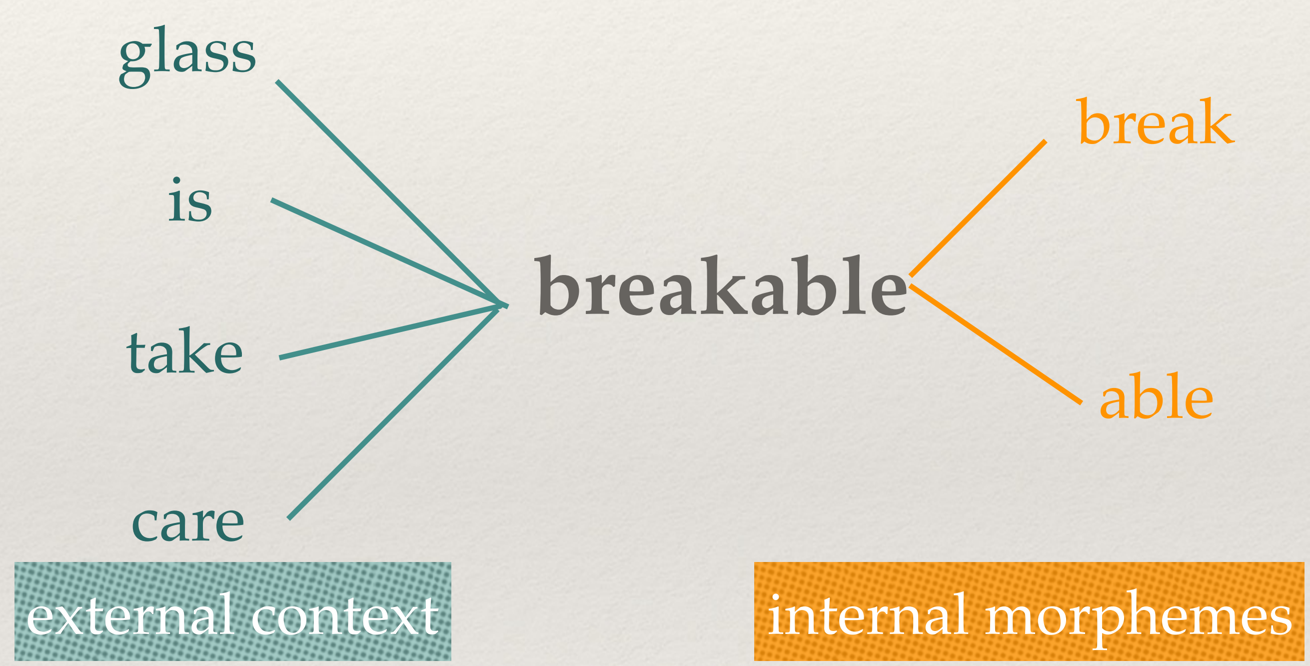
- Semantic matching in search
- Word-level matching: bridging the semantic gap
- Sentence-level matching: capturing the proximity
- Summary and discussion

# Summary

- Word level matching: bridging the semantic gap




"… glass is **breakable**, take care …"



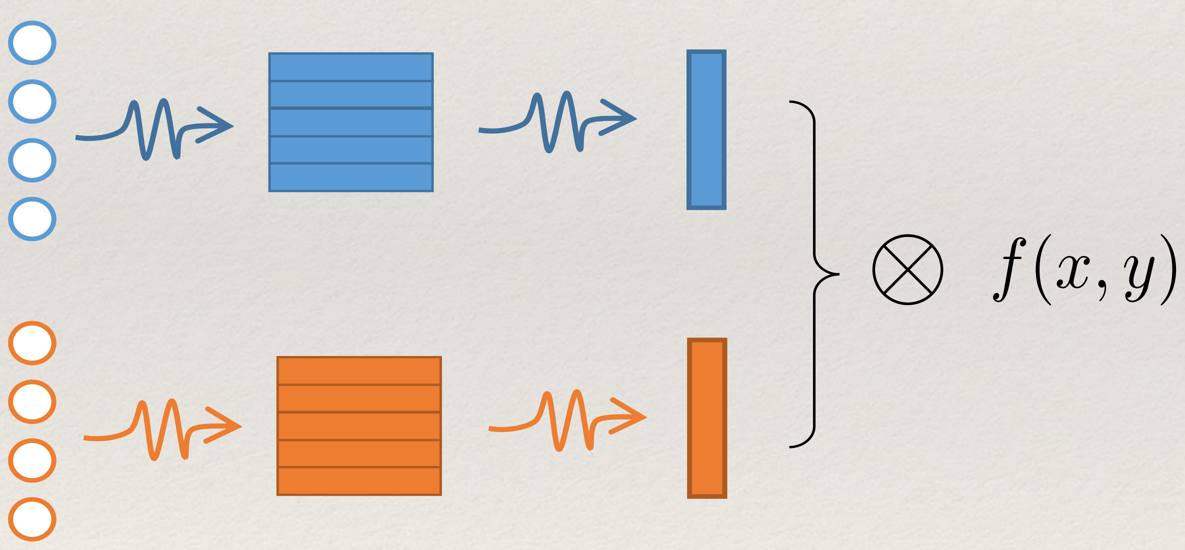

**Two interpretations of distributed hypothesis**

**Beyond distributed hypothesis**

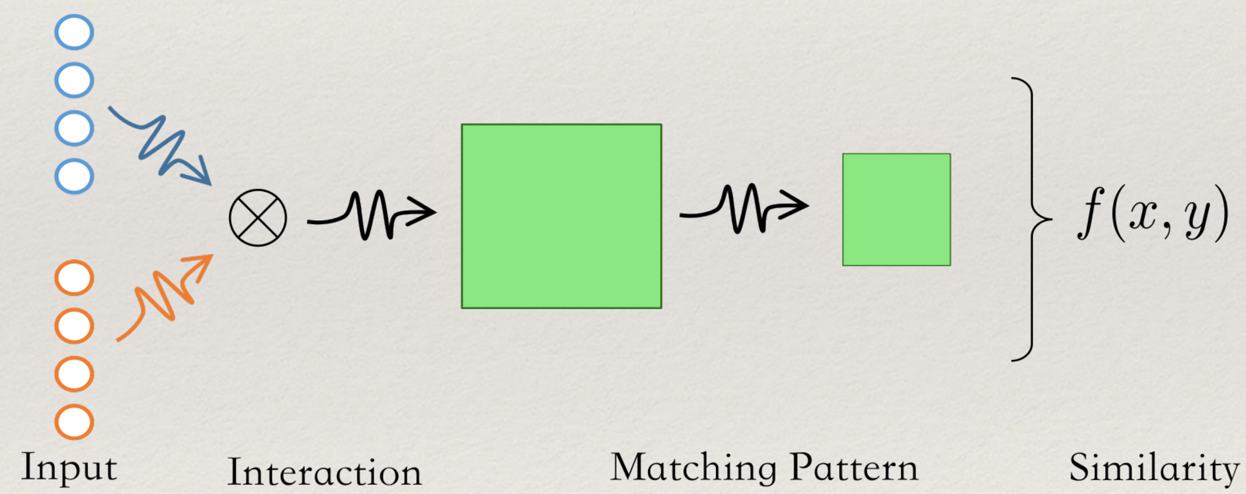
# Summary

- Sentence level matching: capturing the proximity

| Semantic representation of sentences | Aggregating fine-grained matching signals |
| --- | --- |



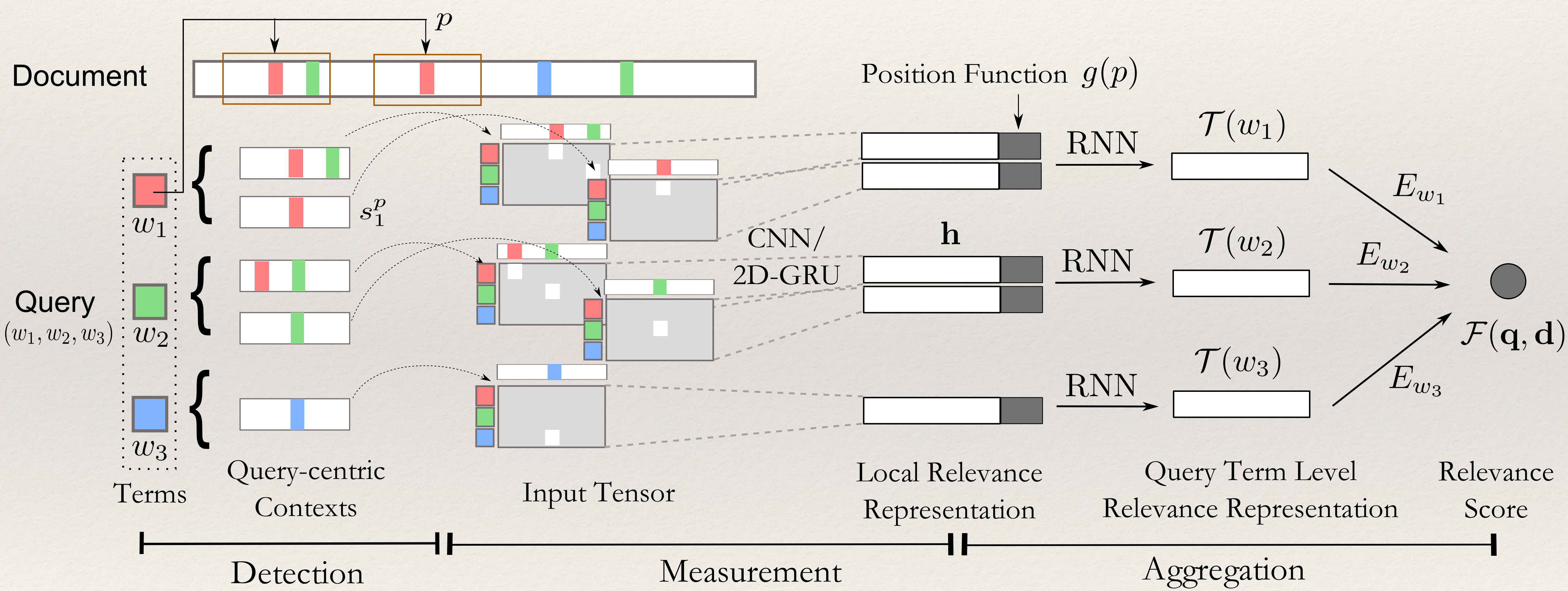
Compositional focused methods: representing queries and document in semantic space




Interaction focused methods: discovering the query-document matching patterns

# Summary

- Document level matching: aggregating matching signals



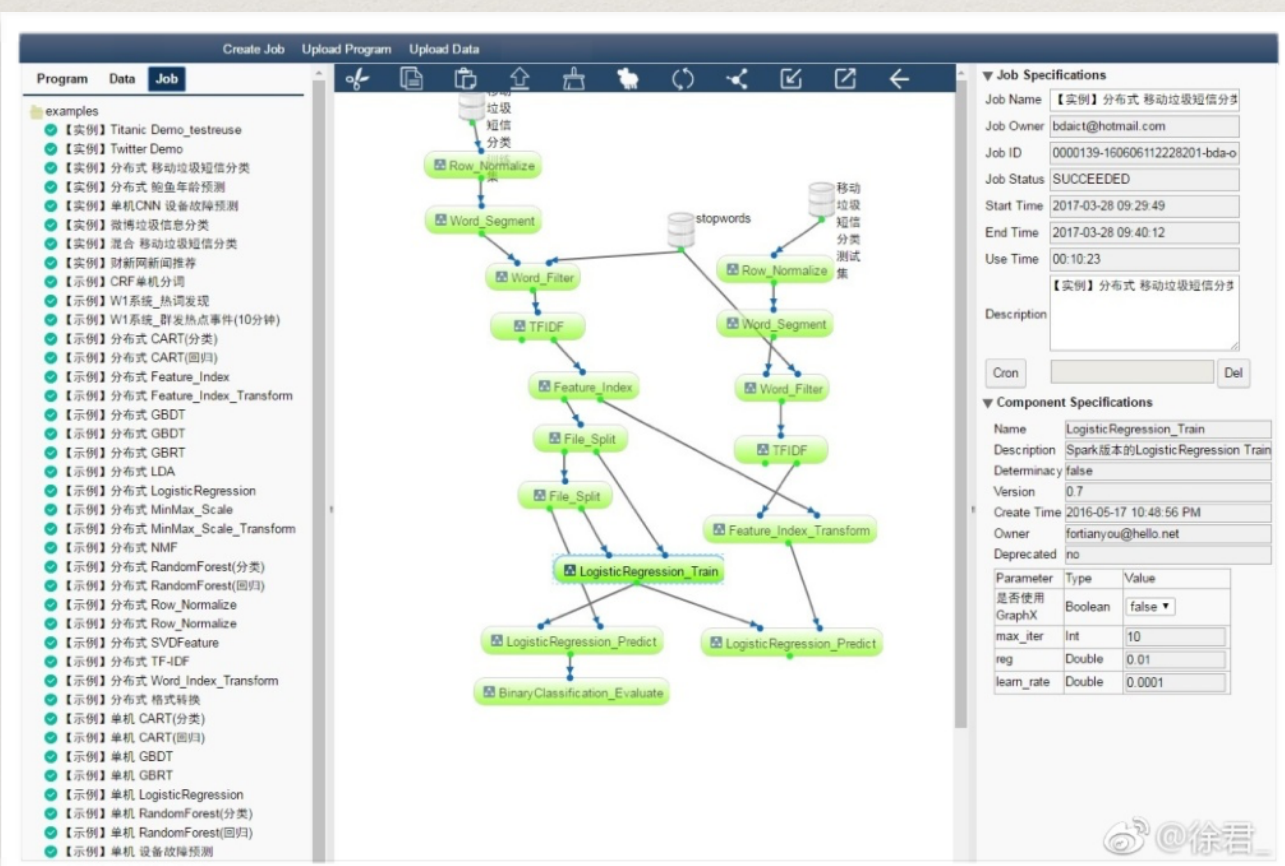
# Challenges

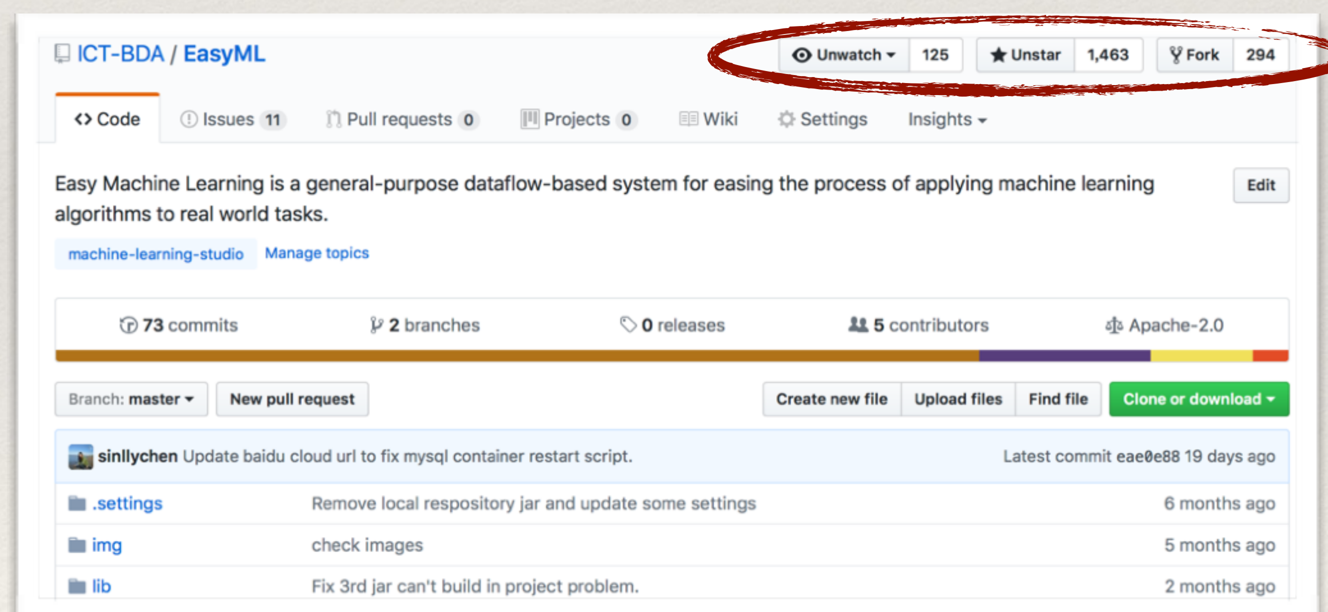
- Data: building benchmarks
  - Current: lack of large scale text matching data
  - Deep learning models has a lot of parameters
- Model: leveraging human knowledge
  - Current: most models are purely data-driven
  - Prior information (e.g., large scale knowledge base) should be helpful
- Application
  - Domain specific matching models: different application have different matching goal, e.g., in IR, relevance != similarity

# Easy Machine Learning Github Project

https://github.com/ICT-BDA/EasyML

- Purpose: ease the process of applying machine learning algorithms to real tasks
  - Machine learning tasks as data-flow DAG
  - Interactive GUI for creating, running, and managing scalable machine learning tasks
  - Deployed as web service http://159.226.40.104:18080/dev/





# Thanks!