The Web Conference 2018, Lyon, France

DKN: Deep Knowledge-Aware Network for News Recommendation

Hongwei Wang[1,2], Fuzheng Zhang[2], Xing Xie[2], Minyi Guo[1]

[1] Shanghai Jiao Tong University
[2] Microsoft Research Asia

April 25, 2018

People read / listen to / watch news everyday…

Ancient Chinese newspaper (1803)


The New York Times.
HEIR TO AUSTRIA'S THRONE IS SLAIN WITH HIS WIFE BY A BOSNIAN YOUTH TO AVENGE SEIZURE OF HIS COUNTRY

New York Times (1914)

Radio (1920s)

TV (1960s)

The Era of Internet …

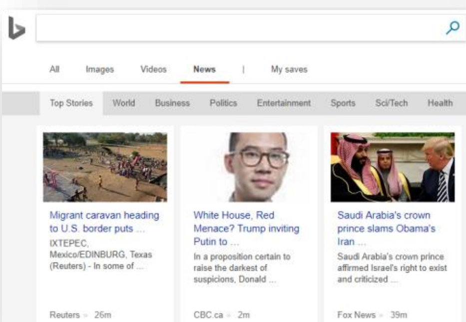
Web portals

CNN

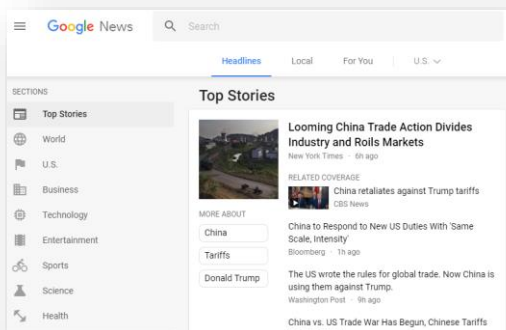
BBC

FOX

News platforms



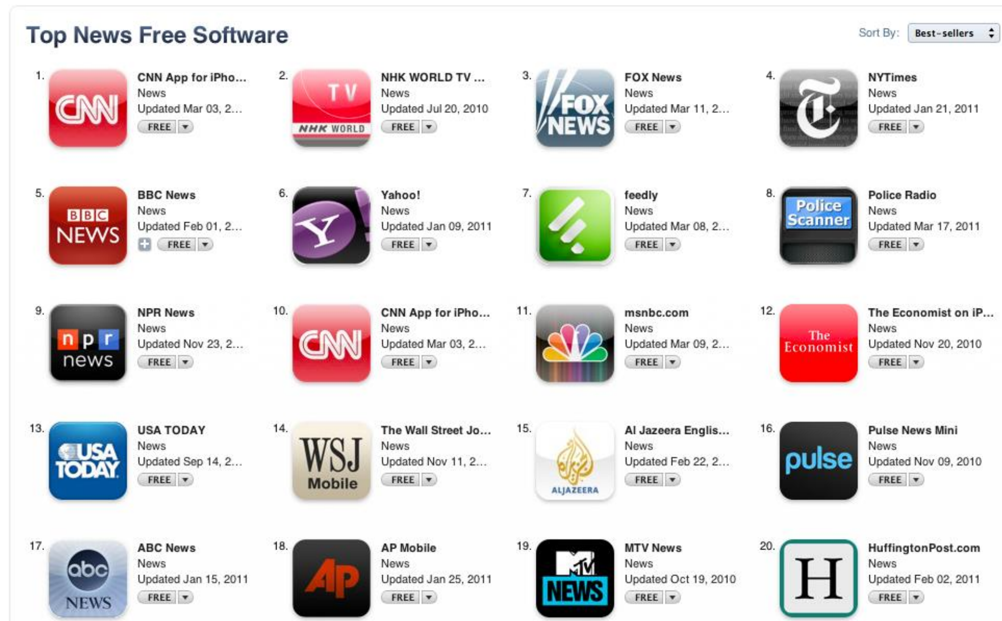
Bing News



Google News

Facebook Newsroom

The Era of Mobile Internet …

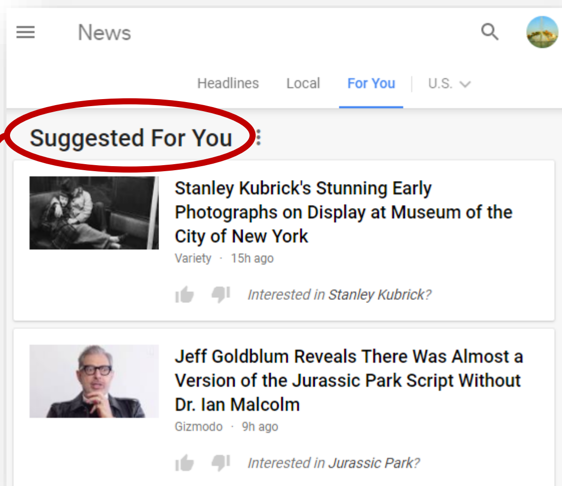


News Recommendation

The volume of articles can be overwhelming to users …

News Recommendation

It's critical to help users target their interests and make personalized recommendations …

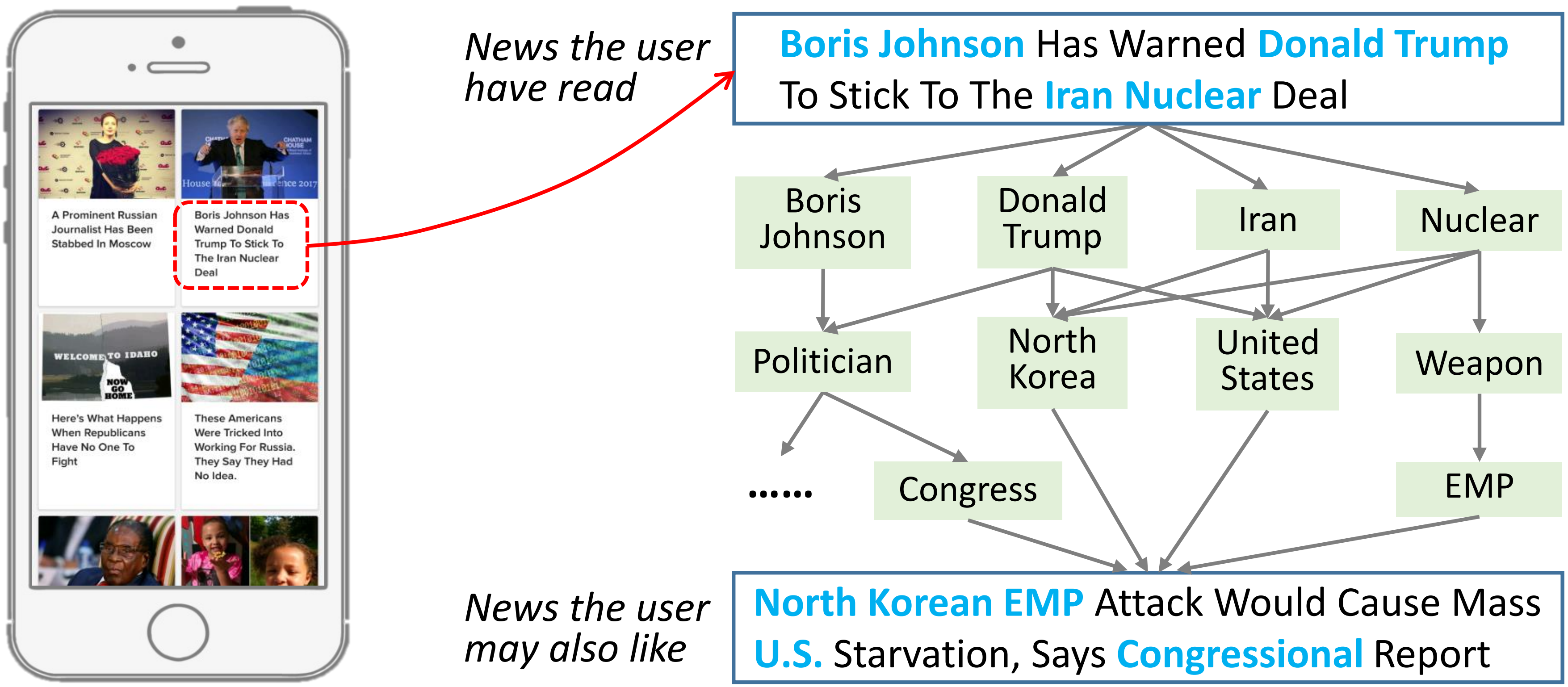


Suggested For You !

Challenges

- News articles are highly **time-sensitive**
  - News expires quickly
  - Out-of-date news are replaced by newer ones frequently
- Readers are **topic-sensitive**
  - They are usually interested in specific news categories
- News language is highly **condensed**, containing a large amount of **knowledge entities**
  - Topic models or semantic models can hardly find their latent knowledge-level connection

Challenges


A Prominent Russian Journalist Has Been Stabbed In Moscow
Boris Johnson Has Warned Donald Trump To Stick To The Iran Nuclear Deal
Here's What Happens When Republicans Have No One To Fight
These Americans Were Tricked Into Working For Russia. They Say They Had No Idea.
News the user have read
Boris Johnson Has Warned Donald Trump To Stick To The Iran Nuclear Deal
Boris Johnson
Donald Trump
Iran
Nuclear
Politician
North Korea
United States
Weapon
......
Congress
EMP
News the user may also like
North Korean EMP Attack Would Cause Mass U.S. Starvation, Says Congressional Report

Our Task

Click history

Elon Musk *offers* ***Tesla Model 3*** *sneak peek …*

Google *fumbles while* ***Tesla*** *sprints toward a driverless future …*

Trump *pledges aid to* ***Silicon Valley*** *during tech meeting …*

Knowledge graph

General Motors *is ramping up its self-driving car:* ***Ford*** *should be nervous …*

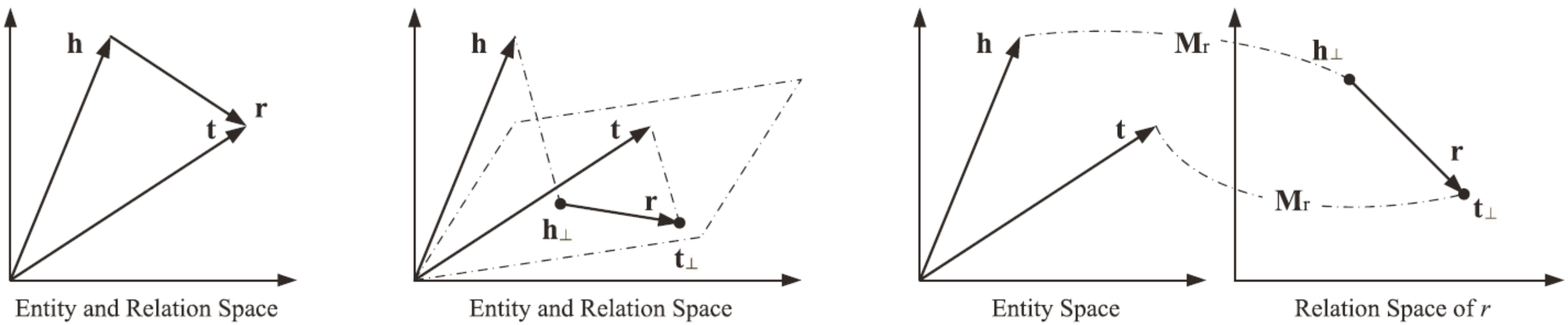
Candidate news

Will the user click it?

Knowledge Graph Embedding

- A knowledge graph consists of millions of triples (head, relation, tail)
- KGE aims to learn a low-dimensional representation vector for each entity and relation
- Translational distance models (TransX)
  - TransE: $f_r(h,t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_2^2$
  - TransH: $f_r(h,t) = \|\mathbf{h}_\perp + \mathbf{r} - \mathbf{t}_\perp\|_2^2$, where $\mathbf{h}_\perp = \mathbf{h} - \mathbf{w}_r^{\mathrm{T}}\mathbf{h}\mathbf{w}_r$ and $\mathbf{t}_\perp = \mathbf{t} - \mathbf{w}_r^{\mathrm{T}}\mathbf{t}\mathbf{w}_r$
  - TransR: $f_r(h,t) = \|\mathbf{h}_r + \mathbf{r} - \mathbf{t}_r\|_2^2$, where $\mathbf{h}_r = \mathbf{h} - \mathbf{h}\mathbf{M}_r$ and $\mathbf{t}_r = \mathbf{h} - \mathbf{t}\mathbf{M}_r$
  - TransD: $f_r(h,t) = \|\mathbf{h}_\perp + \mathbf{r} - \mathbf{t}_\perp\|_2^2$, where $\mathbf{h}_\perp = \left(\mathbf{r}_p\mathbf{h}_p^{\mathrm{T}} + \mathbf{I}\right)\mathbf{h}$ and $\mathbf{t}_\perp = \left(\mathbf{r}_p\mathbf{t}_p^{\mathrm{T}} + \mathbf{I}\right)\mathbf{t}$

Knowledge Graph Embedding



(a) TransE. (b) TransH. (c) TransR.

$$\mathcal{L} = \sum_{(h,r,t)\in\Delta} \sum_{(h',r,t')\in\Delta'} \max\left(0, f_r(h,t) + \gamma - f_r(h',t')\right)$$

Correct triples Incorrect triples Margin

Knowledge Distillation

Trump praises **Las Vegas** medical team

Apple CEO Tim Cook: **iPhone 8** and **Apple Watch Series 3** are sold out in some places

EU Spain: **Juncker** does not want **Catalonian** independence

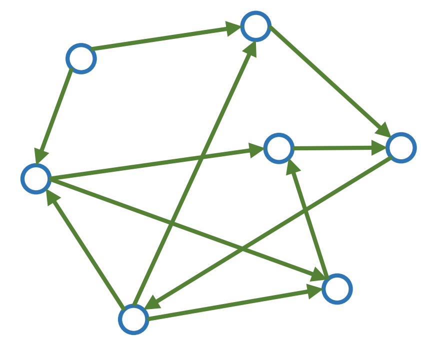
……

Entity linking

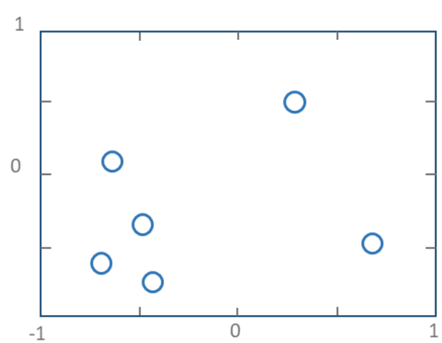
Donald Trump: *Donald Trump is the 45th president …*
Las Vegas: *Las Vegas is the 28th-most populated city …*
Apple Inc.: *Apple Inc. is an American multinational …*
CEO: *A chief executive officer is the position of the …*
Tim Cook: *Timothy Cook is an American business …*
iPhone 8: *iPhone 8 is smartphone designed, …*

……

Knowledge subgraph construction



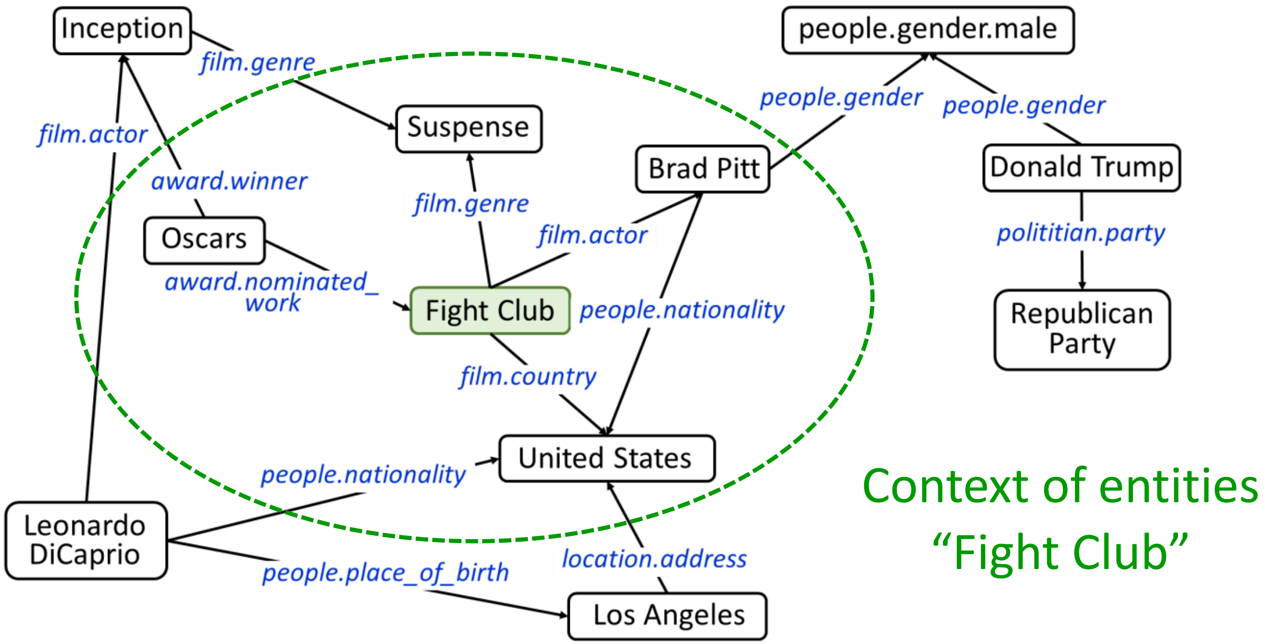
Knowledge graph embedding



Entity embedding

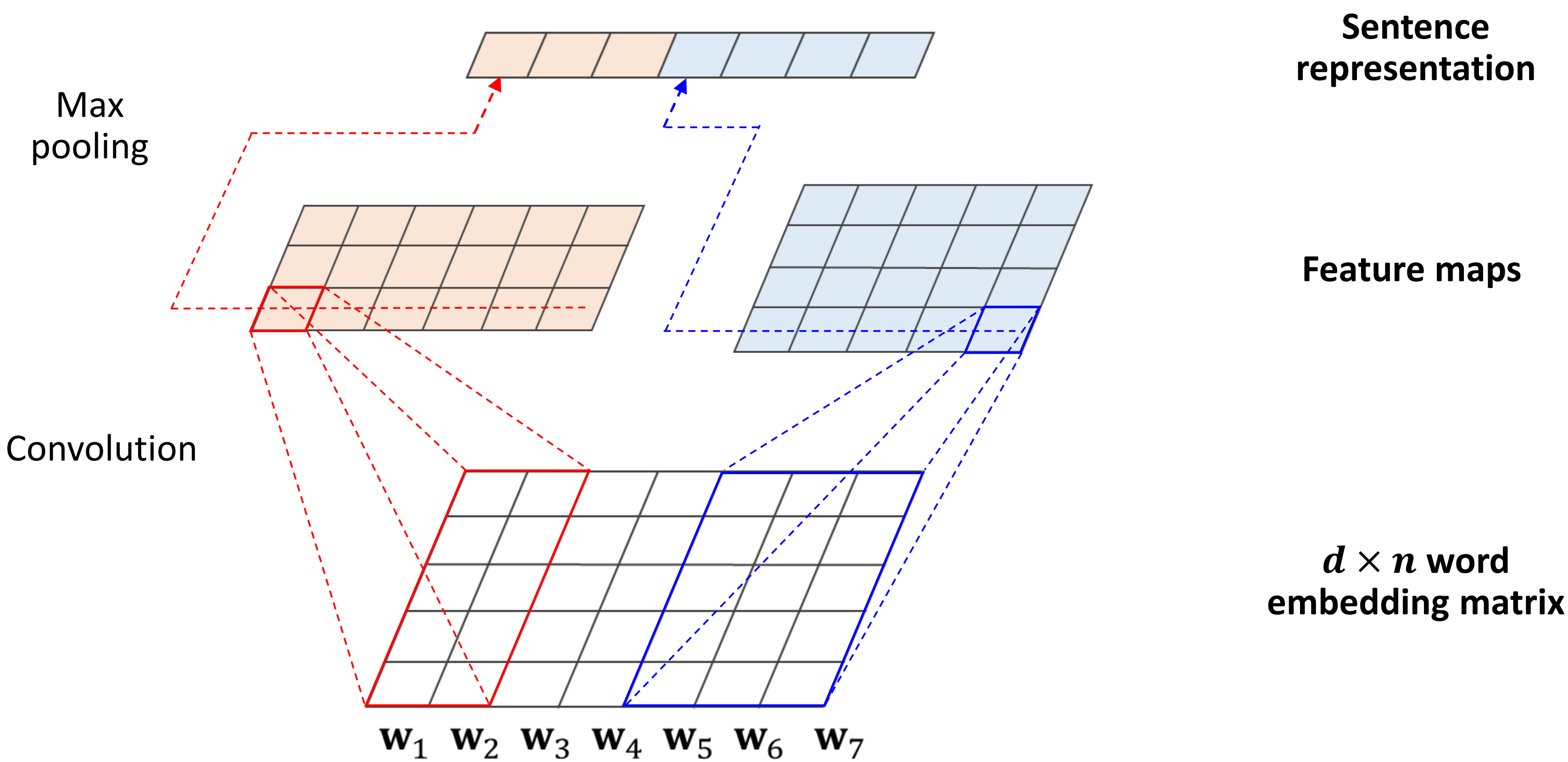
Donald Trump: (0.32, 0.48)
Las Vegas: (0.71, -0.49)
Apple Inc.: (-0.48, -0.41)
CEO: (-0.57, 0.06)
Tim Cook: (-0.61, -0.59)
iPhone 8: (-0.46, -0.75)

Context Embedding



$$\bar{\mathbf{e}} = \frac{1}{|context(e)|} \sum_{e_i \in context(e)} \mathbf{e}_i$$

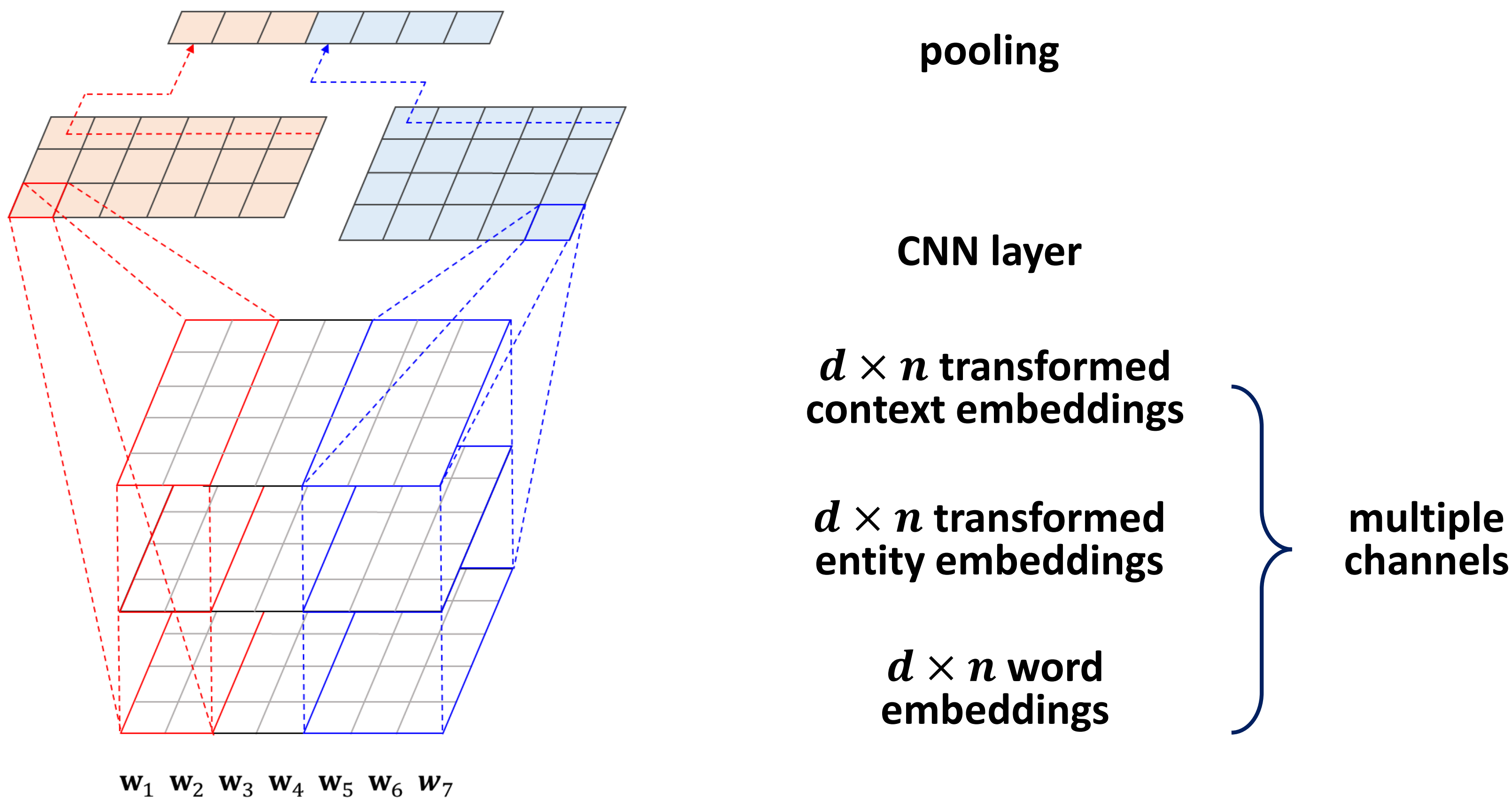
Kim's CNN



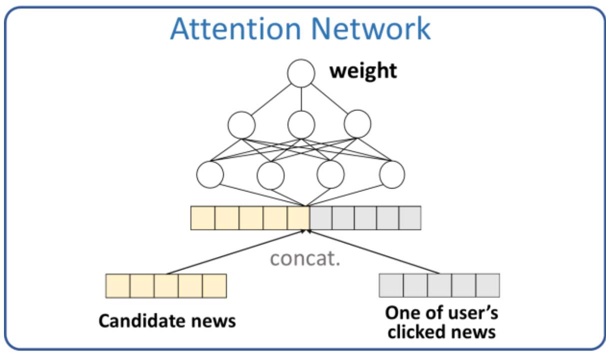
$w_{1:n}$ = [*Donald Trump praises Las Vegas medical team*]

Sentence

Knowledge-aware CNN (KCNN)



Attention-based User Interest Extraction



Attention net:

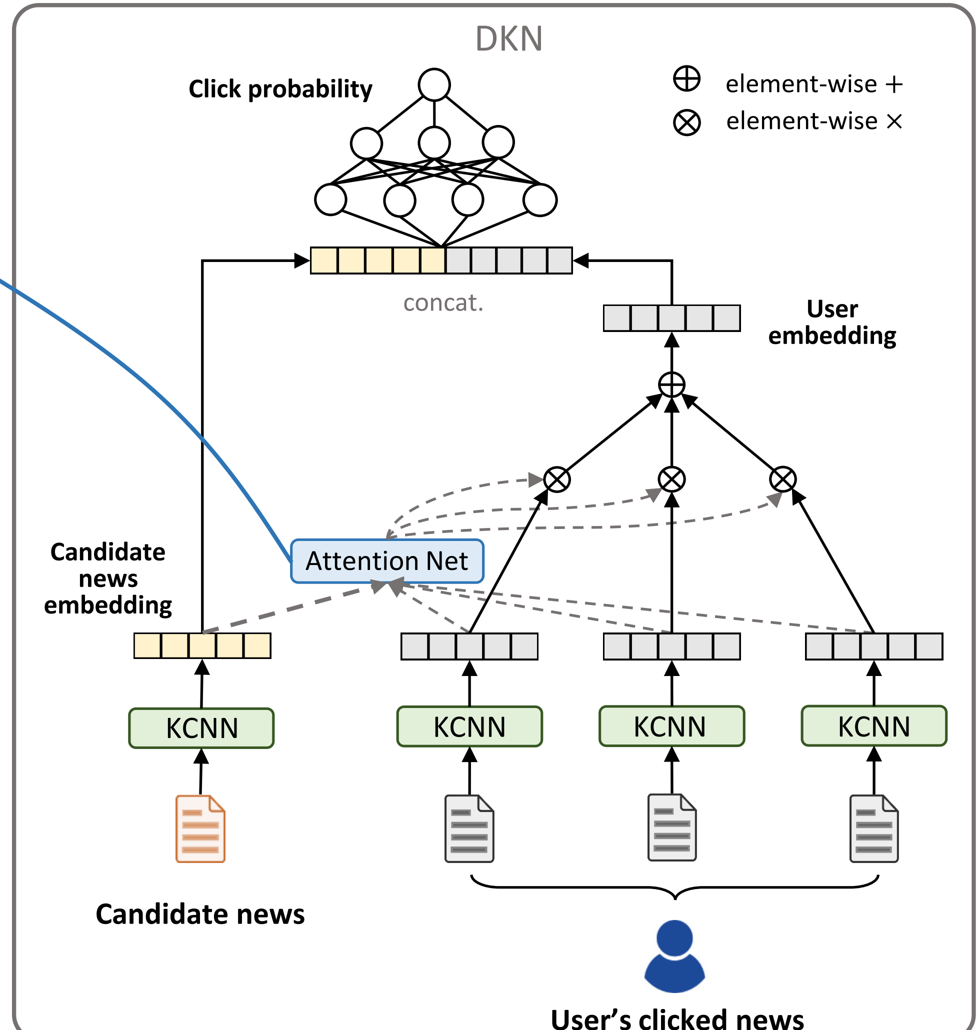
$$s_{t_k^i, t_j} = \mathrm{softmax}\left(\mathcal{H}\left(\mathrm{e}(t_k^i), \mathrm{e}(t_j)\right)\right) = \frac{\exp\left(\mathcal{H}\left(\mathrm{e}(t_k^i), \mathrm{e}(t_j)\right)\right)}{\sum_{k=1}^{N_i} \exp\left(\mathcal{H}\left(\mathrm{e}(t_k^i), \mathrm{e}(t_j)\right)\right)}$$

User interest extraction:

$$\mathrm{e}(i) = \sum_{k=1}^{N_i} s_{t_k^i, t_j} \mathrm{e}(t_k^i).$$

CTR prediction:

$$p_{i, t_j} = \mathcal{G}\left(\mathrm{e}(i), \mathrm{e}(t_j)\right)$$



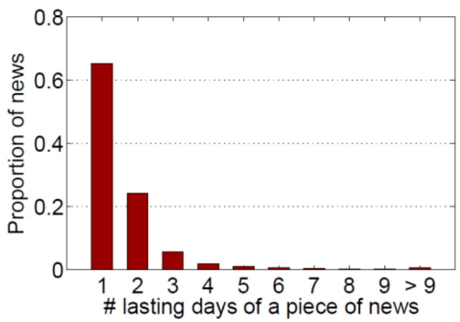
Dataset

- Dataset: Bing news
  - (timestamp, user_id, news_url, news_title, click_label)
  - Training set: October 16, 2016 ~ June 11, 2017
  - Test set: June 12, 2017 ~ August 11, 2017

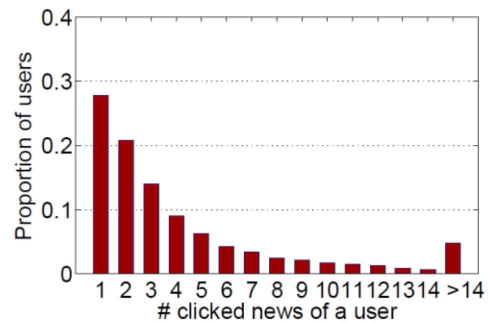
- Knowledge graph: Microsoft Satori

# users	141,487	# triples	7,145,776
# news	535,145	avg. # words per title	7.9
# logs	1,025,192	avg. # entities per title	3.7
# entities	336,350	avg. # contextual entities per entity	42.5
# relations	4,668		

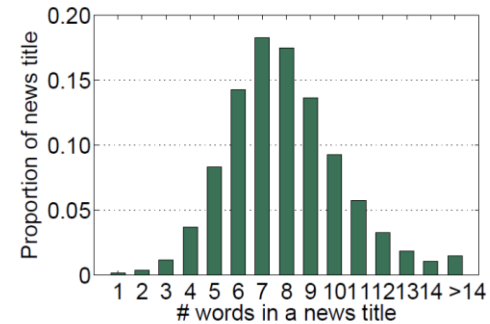
Statistics of the Dataset



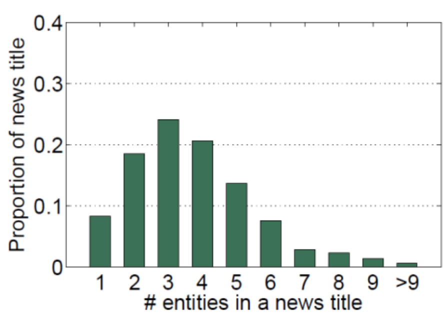
(a) Distribution of the length of news life cycle



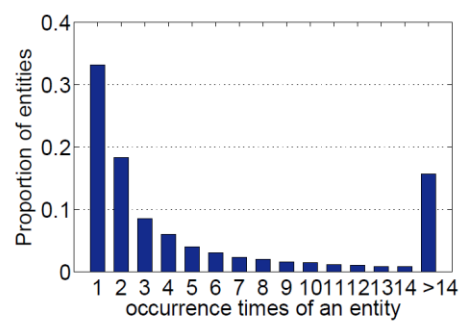
(b) Distribution of the number of clicked news of a user



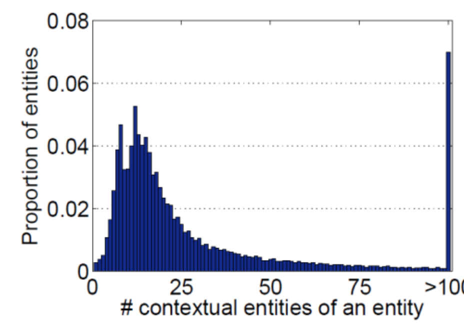
(c) Distribution of the number of words in a news title



(d) Distribution of the number of entities in a news title



(e) Distribution of the occurrence times of an entity in the news dataset



(f) Distribution of the number of contextual entities of an entity in the knowledge graph

Comparison with Baselines

Models*	F1	AUC	p-value**
DKN	**68.9 ± 1.5**	**65.9 ± 1.2**	–
LibFM	61.8 ± 2.1 (-10.3%)	59.7 ± 1.8 (-9.4%)	$< 10^{-3}$
LibFM(-)	61.1 ± 1.9 (-11.3%)	58.9 ± 1.7 (-10.6%)	$< 10^{-3}$
KPCNN	67.0 ± 1.6 (-2.8%)	64.2 ± 1.4 (-2.6%)	0.098
KPCNN(-)	65.8 ± 1.4 (-4.5%)	63.1 ± 1.5 (-4.2%)	0.036
DSSM	66.7 ± 1.8 (-3.2%)	63.6 ± 2.0 (-3.5%)	0.063
DSSM(-)	66.1 ± 1.6 (-4.1%)	63.2 ± 1.8 (-4.1%)	0.045
DeepWide	66.0 ±1.2 (-4.2%)	63.3 ± 1.5 (-3.9%)	0.039
DeepWide(-)	63.7 ± 0.9 (-7.5%)	61.5 ± 1.1 (-6.7%)	0.004
DeepFM	63.8 ± 1.5 (-7.4%)	61.2 ± 2.3 (-7.1%)	0.014
DeepFM(-)	64.0 ± 1.9 (-7.1%)	61.1 ± 1.8 (-7.3%)	0.007
YouTubeNet	65.5 ± 1.2 (-4.9%)	63.0 ± 1.4 (-4.4%)	0.025
YouTubeNet(-)	65.1 ± 0.7 (-5.5%)	62.1 ± 1.3 (-5.8%)	0.011
DMF	57.2 ± 1.2 (-17.0%)	55.3 ± 1.0 (-16.1%)	$< 10^{-3}$

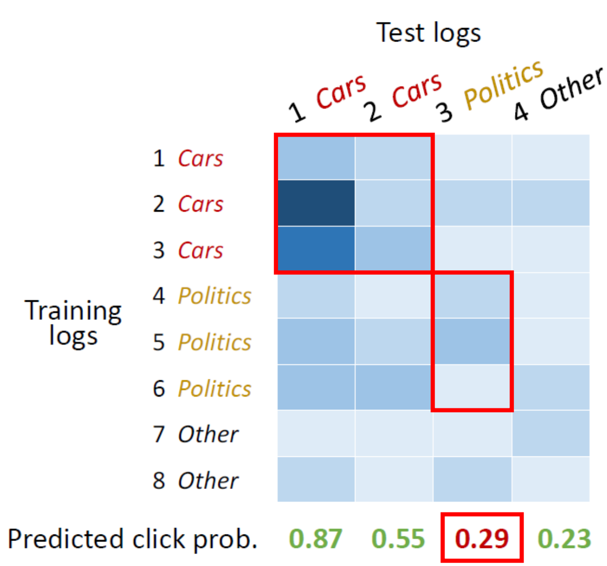
* "(-)" denotes "without input of entity embeddings".

** p-value is the probability of no significant difference with DKN on AUC by t-test.

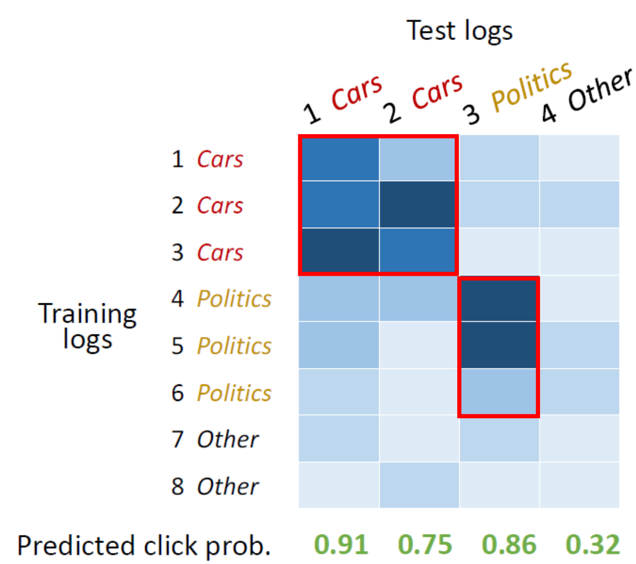
Comparison with Variants

Variants	F1	AUC
DKN with entity and context emd.	**68.8 ± 1.4**	**65.7 ± 1.1**
DKN with entity emd. only	67.2 ± 1.2	64.8 ± 1.0
DKN with context emd. only	66.5 ± 1.5	64.2 ± 1.3
DKN without entity nor context emd.	66.1 ±1.4	63.5 ± 1.1
DKN + TransE	67.6 ± 1.6	65.0 ± 1.3
DKN + TransH	67.3 ± 1.3	64.7 ± 1.2
DKN + TransR	67.9 ± 1.5	65.1 ± 1.5
DKN + TransD	**68.8 ± 1.3**	**65.8 ± 1.4**
DKN with non-linear mapping	**69.0 ± 1.7**	**66.1 ± 1.4**
DKN with linear mapping	67.1 ± 1.5	64.9 ± 1.3
DKN without mapping	66.7 ± 1.6	63.7 ± 1.6
DKN with attention	**68.7 ± 1.3**	**65.7 ± 1.2**
DKN without attention	67.0 ± 1.0	64.8 ± 0.8

Visualization of Attention



(a) without knowledge graph



(b) with knowledge graph

Q & A

Thanks!