Microsoft

What Will Search Engines be Changed by NLP Advancements

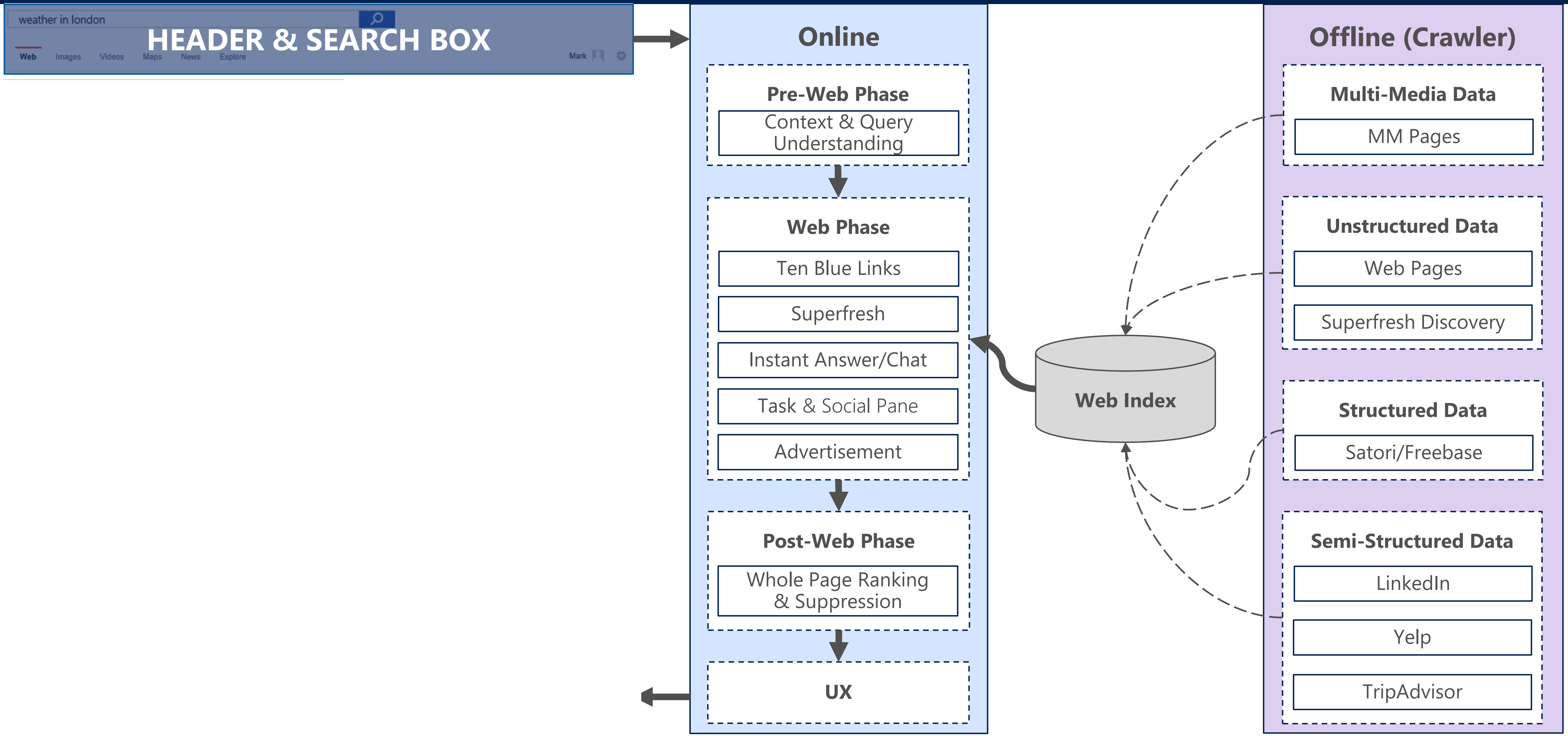
Dr. Ming Zhou
Microsoft Research Asia

ICTIR-2018, Tianjin

Introduction

- Intelligence and naturalness are two driving forces of a search engine
- Intelligence
  Precisely understanding user's intent and finding out accurate answers
- Naturalness
  Expressing user's intent and presenting the search results in a natural way
- NLP has been vital (though at shallow level) to both intelligence and naturalness
- With new advancements of NLP, what new changes will be brought to search?

Search engine architecture(Bing)



NLP in search engine

Search Result Page

Post-Web Phase

Whole Page Ranking	Suppression	Coherence	Ranking Model	Query Suggestion

Web Phase

Ten Bleu Links

- Bag-of-word model
- Translation model
- Click-through
- Summarization/snippet
- …

Instant Answer/Chat

- Knowledge-based QA
- Visual/Video QA
- MRC
- Task-oriented bot
- …

Task & Social Pane

- Knowledge graph
- Entity linking
- Social networks
- User profiling
- …

Super Fresh

- News recommendation
- Personalized news feed
- Event detection
- News classification
- …

Advertisement

- User intent detection
- Slot tagging
- Query-Ads matching
- Ads keyword generation
- …

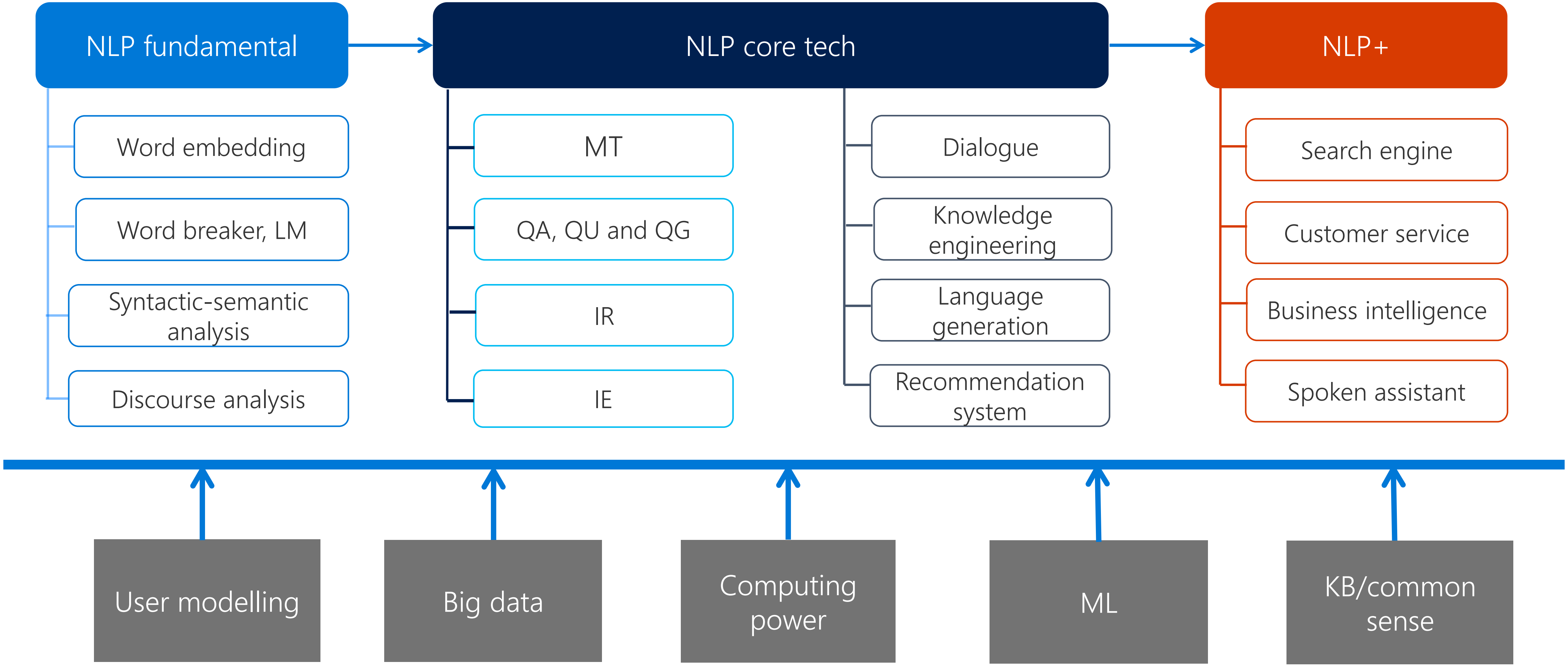
Pre-Web Phase

Word Breaking	Spelling	Synonym	Dependency Parsing	Entity Linking
Intent Classification	Semantic Parsing	Word/Sentence Embedding	User Profiling	Query Rewriting

In the following, I try to answer

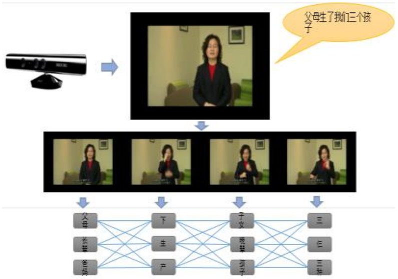
- What are new advancements of NLP?
- What new changes will be brought to search?

NLP technologies

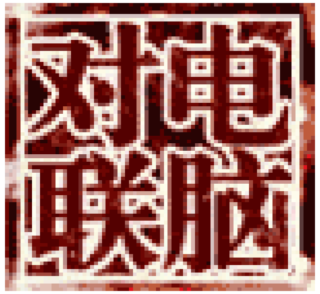


NLP innovations in MS

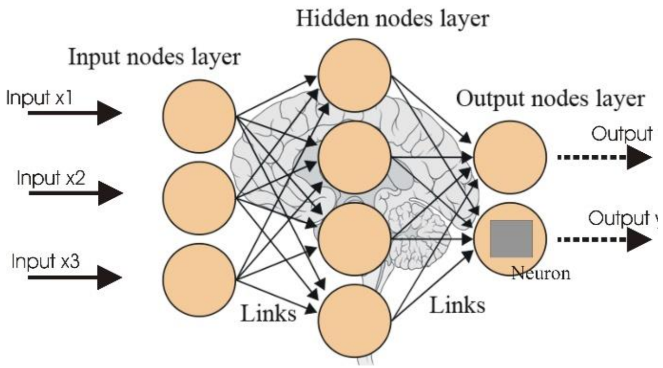
必应 输入法



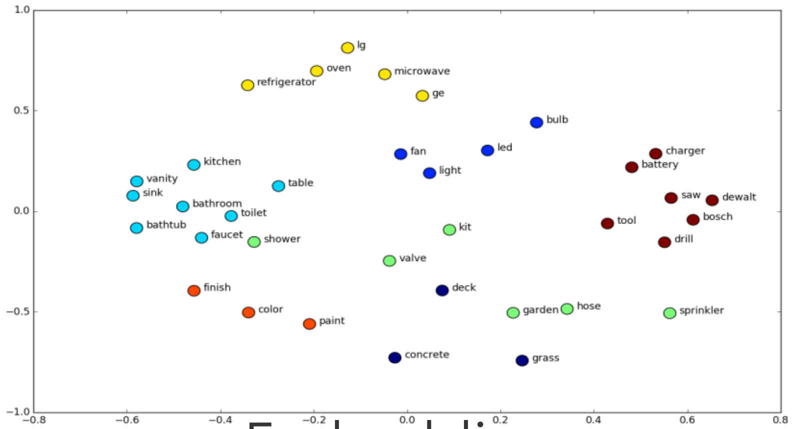
Bing



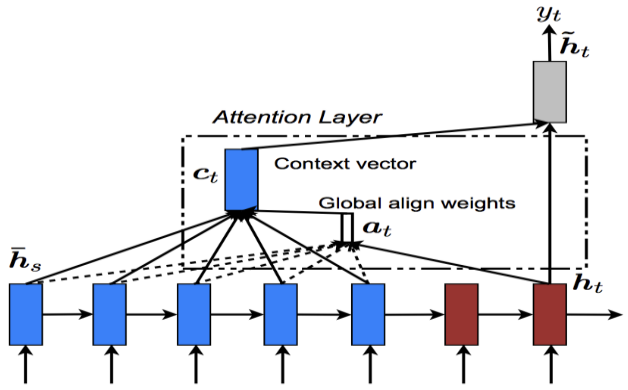
DNN techs that changed NLP



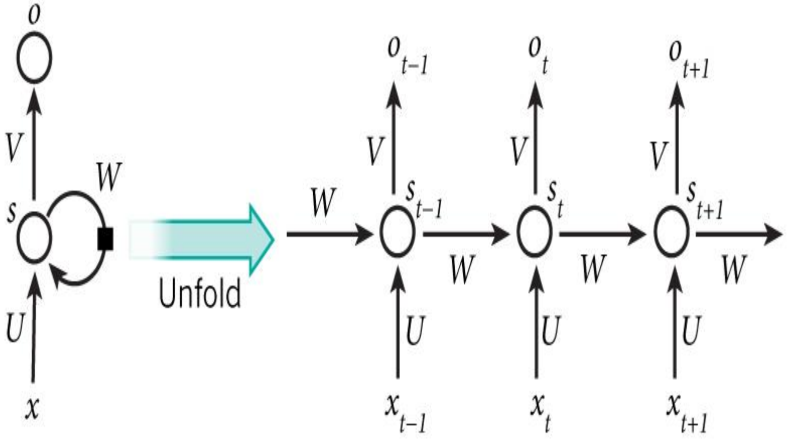
DNN



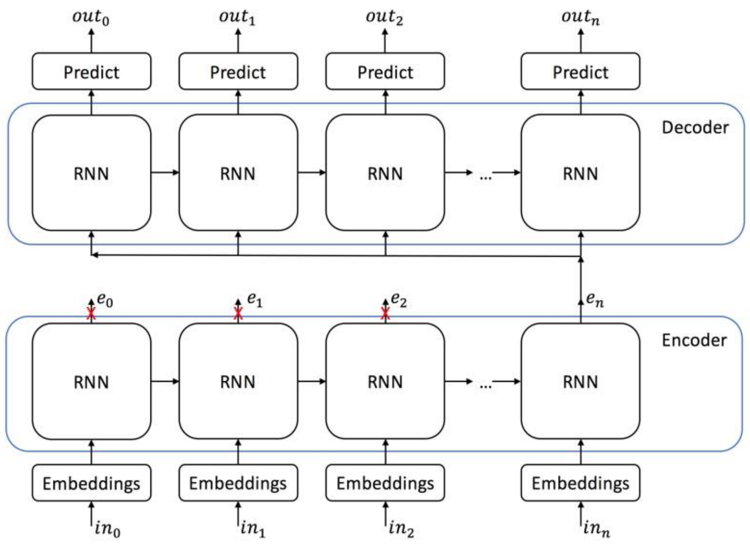
Embedding



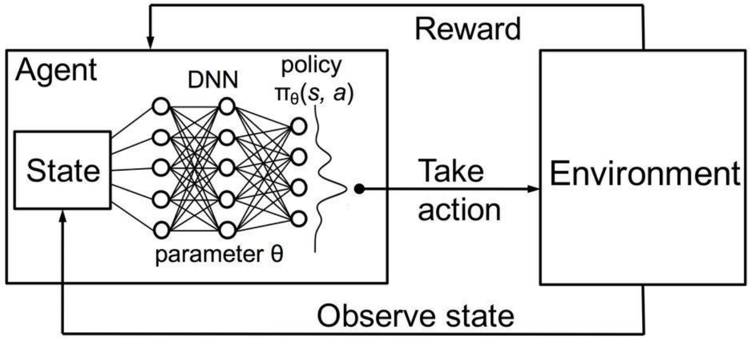
Attention



RNN/LSTM/GRU



Encoder-Decoder



Reinforcement learning

Recent breakthroughs in MSRA

2017	2018	2018	2018
23 CPS Chatbot	82.65% Machine Reading	69% Machine Translation	61.34% 76.88% 62.42% Grammar check
Xiaoice	SQuAD	WMT-2017	CoNLL-2014 CoNLL-10 JFLEG

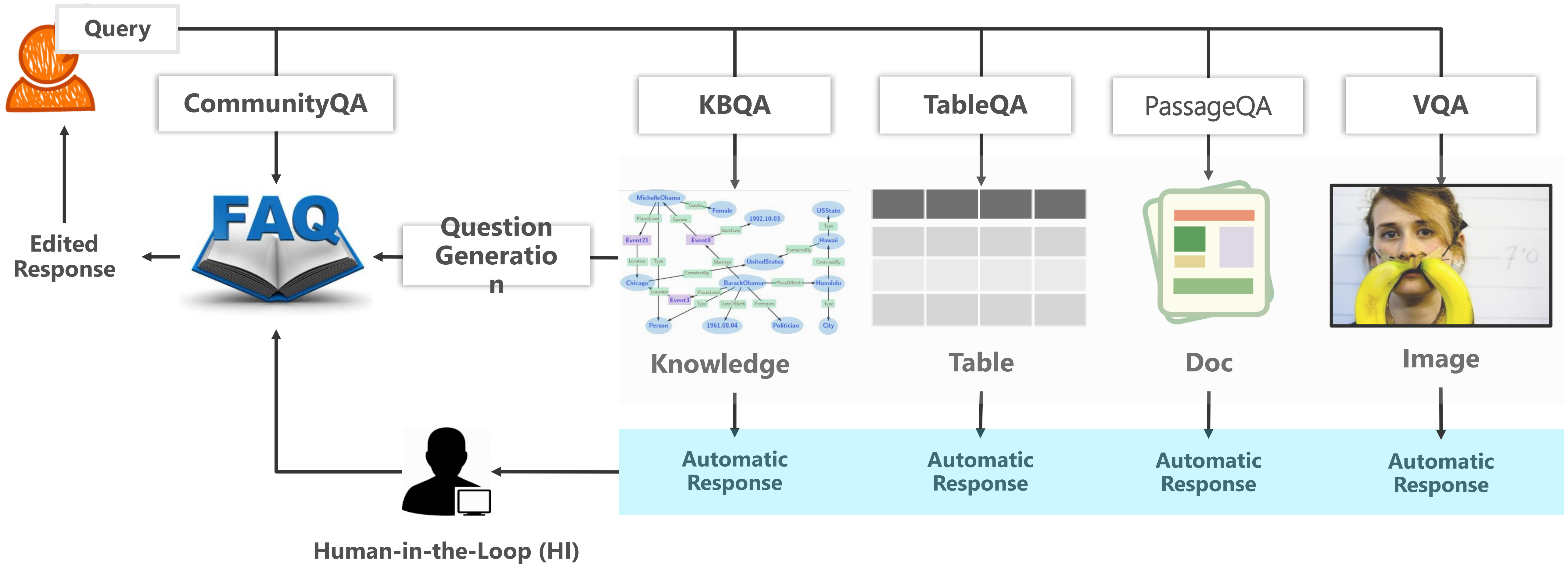
Important trends

- QA
- Multi-lingual
- Multi-modal
- MRC
- Personalized recommendation

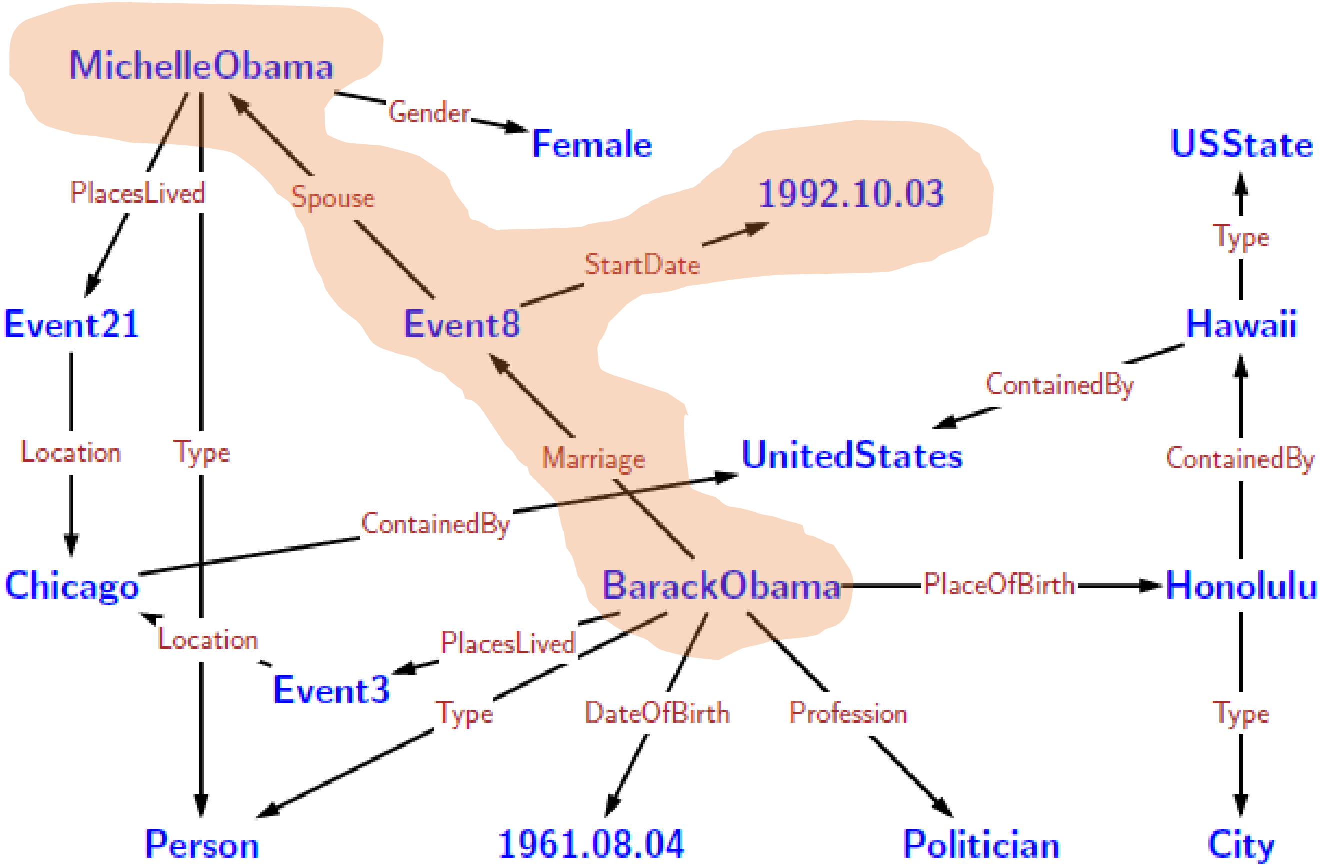
Important trends

- QA
- Multi-lingual
- Multi-modal
- MRC
- Personalized recommendation

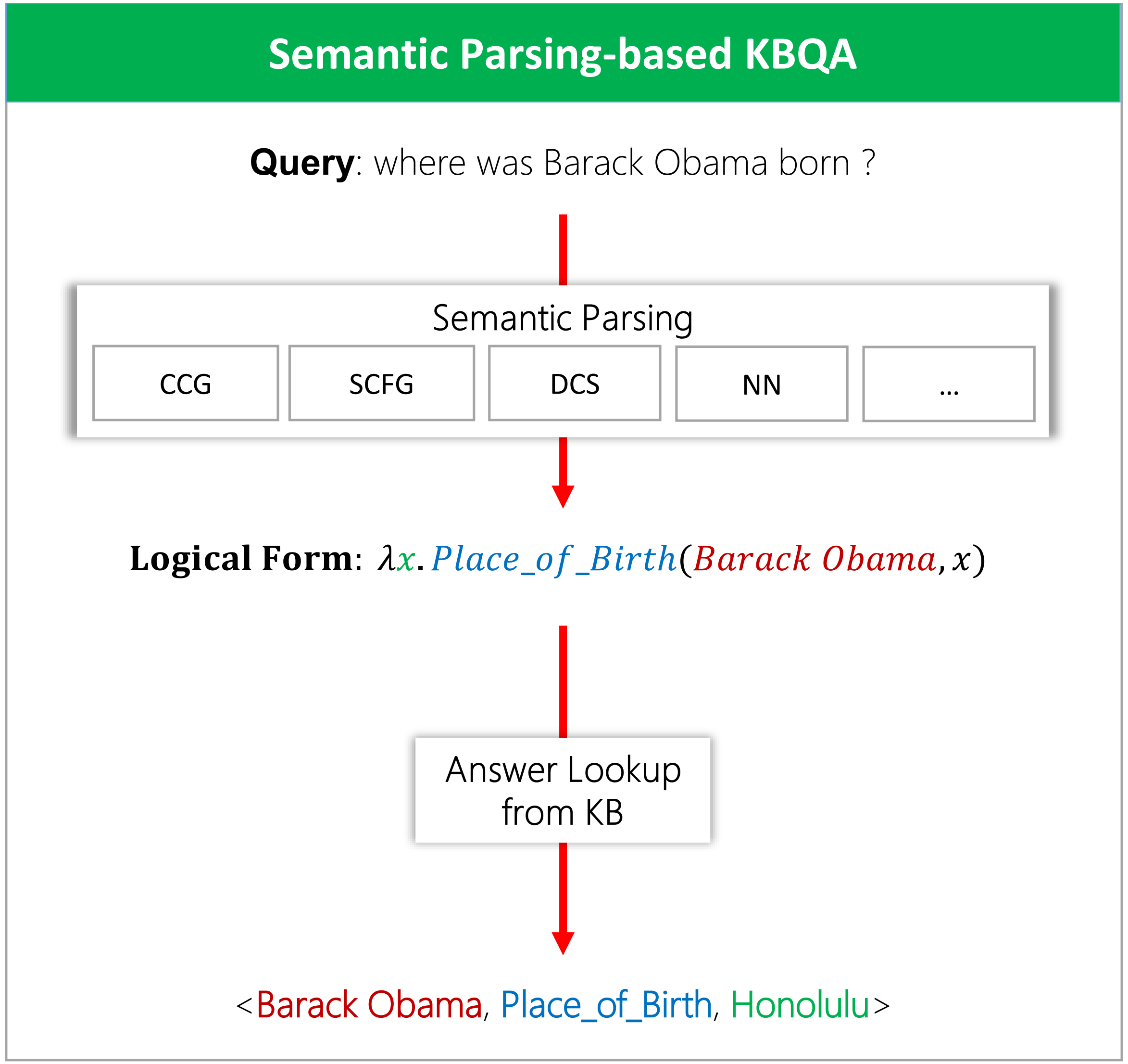
Question-Answering

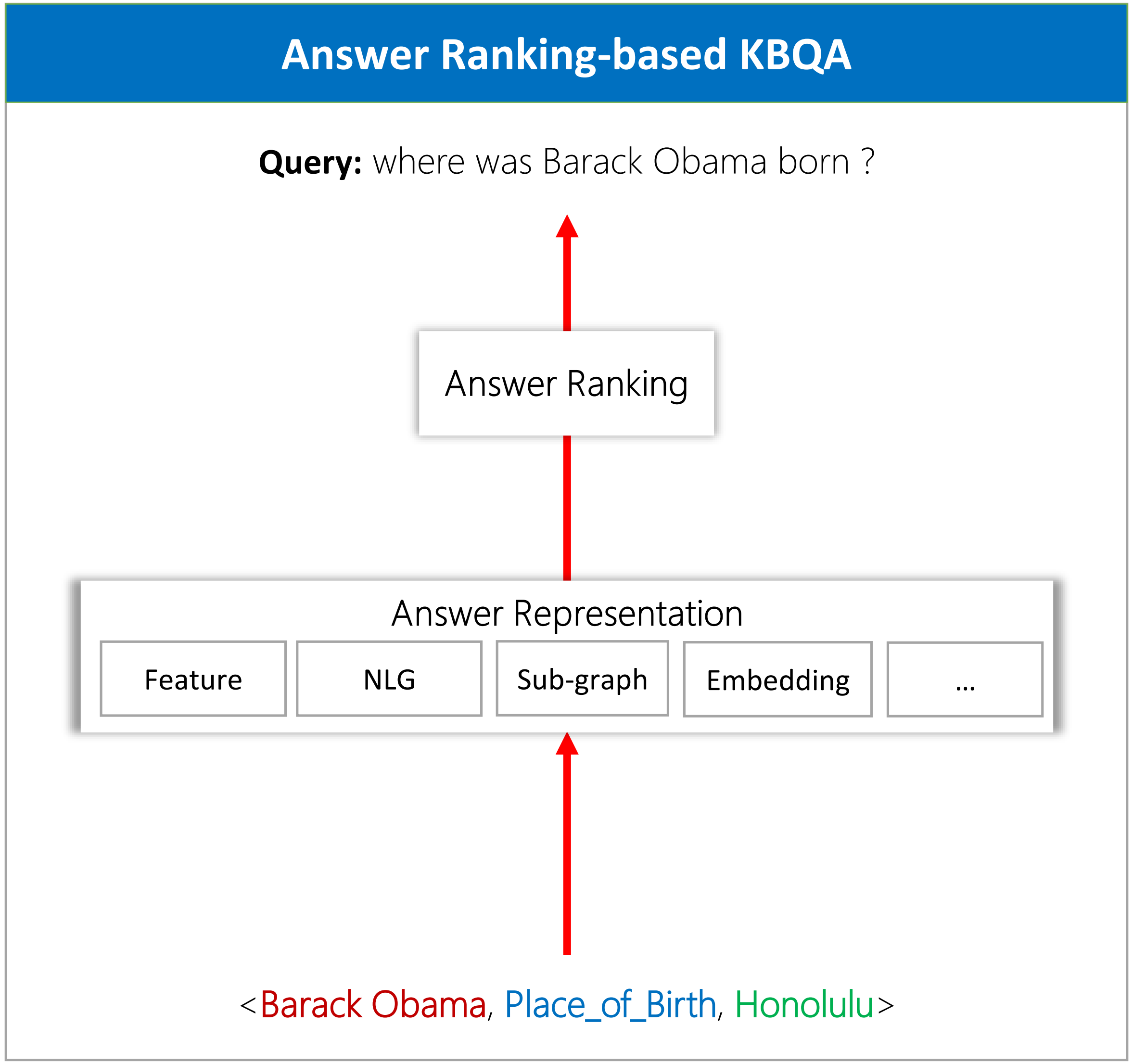


Knowledge-Base

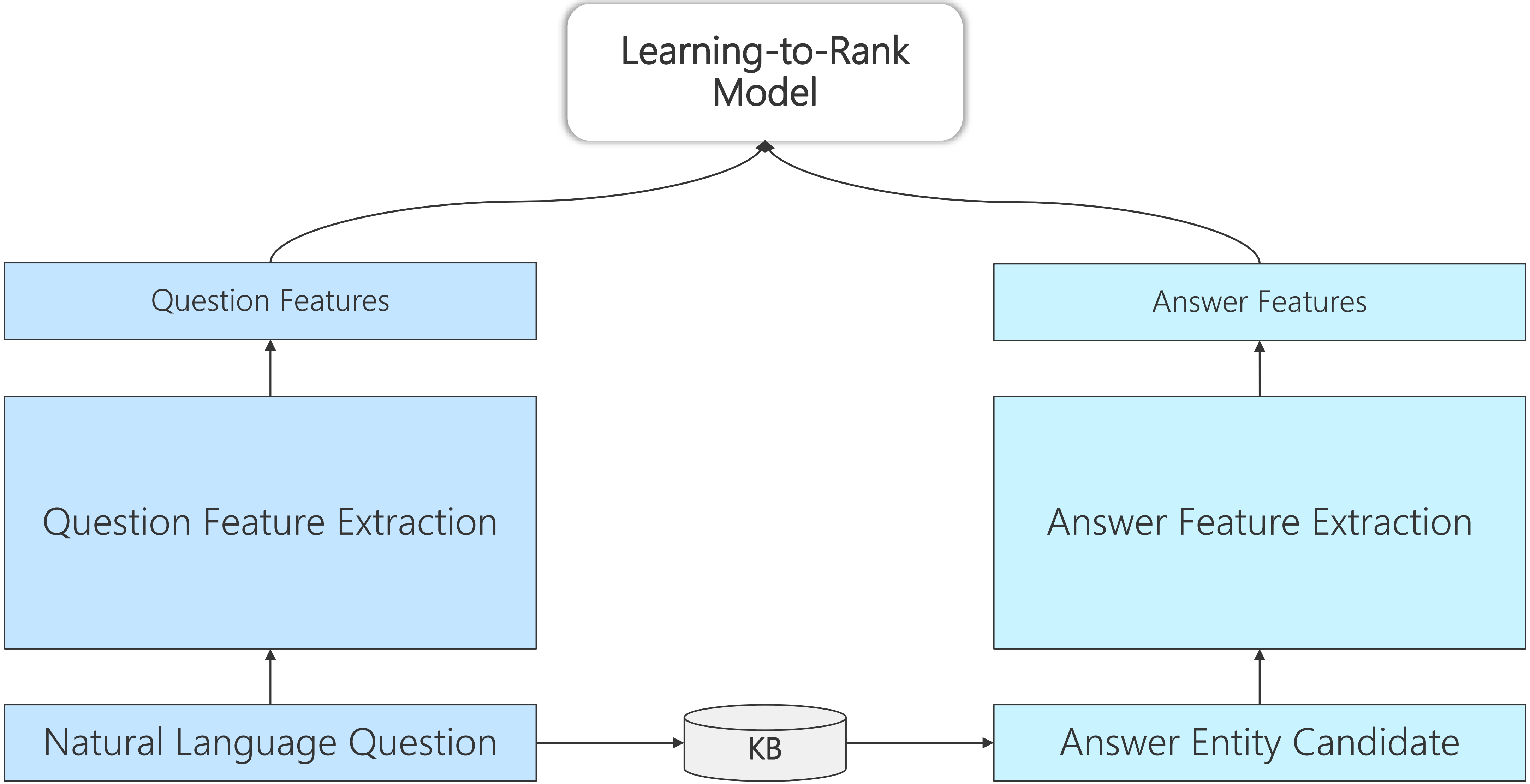


KBQA approaches



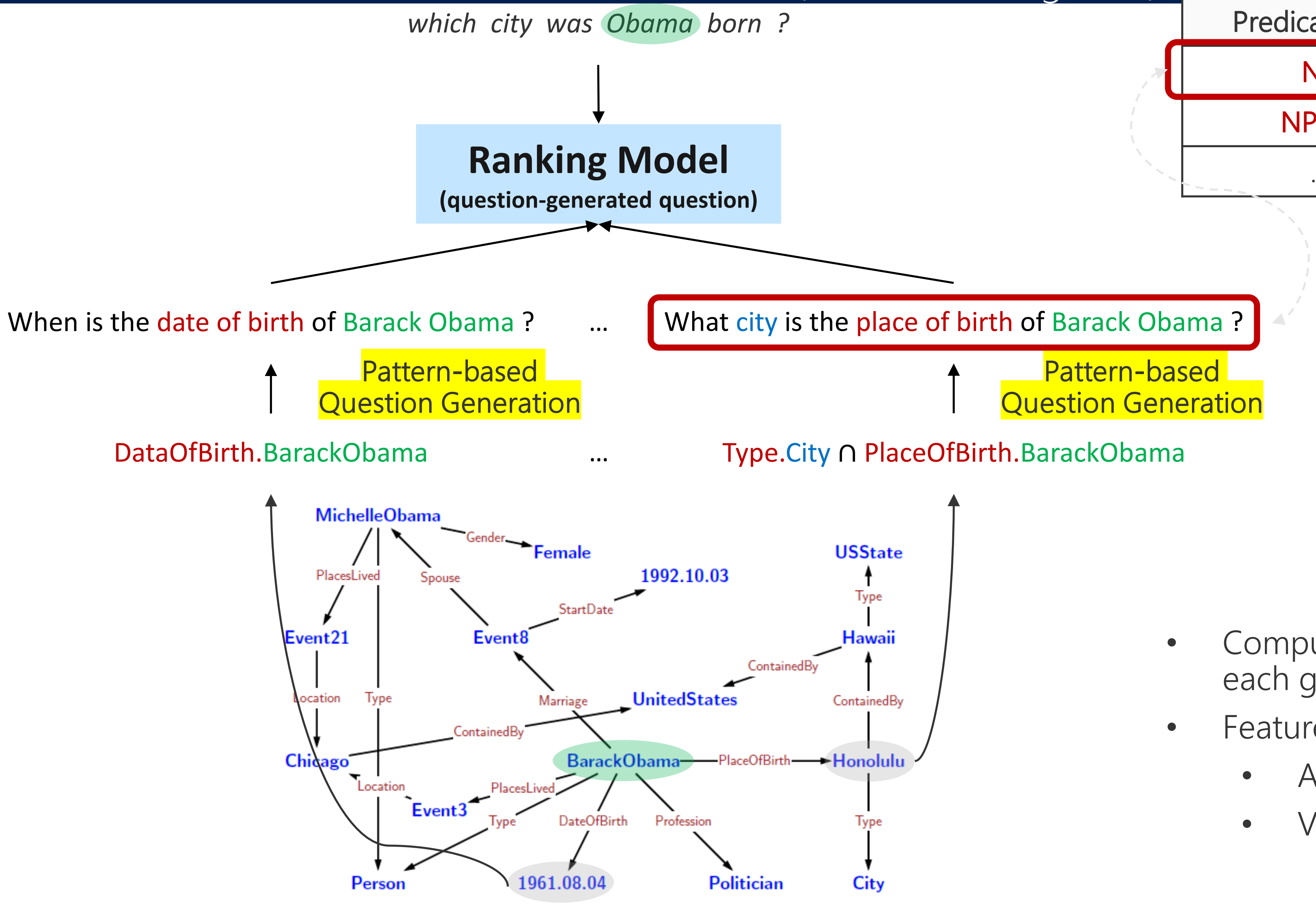


Answer Ranking-based KBQA



Ranking with NL description of the candidate answers

(Berant and Liang, 2014)



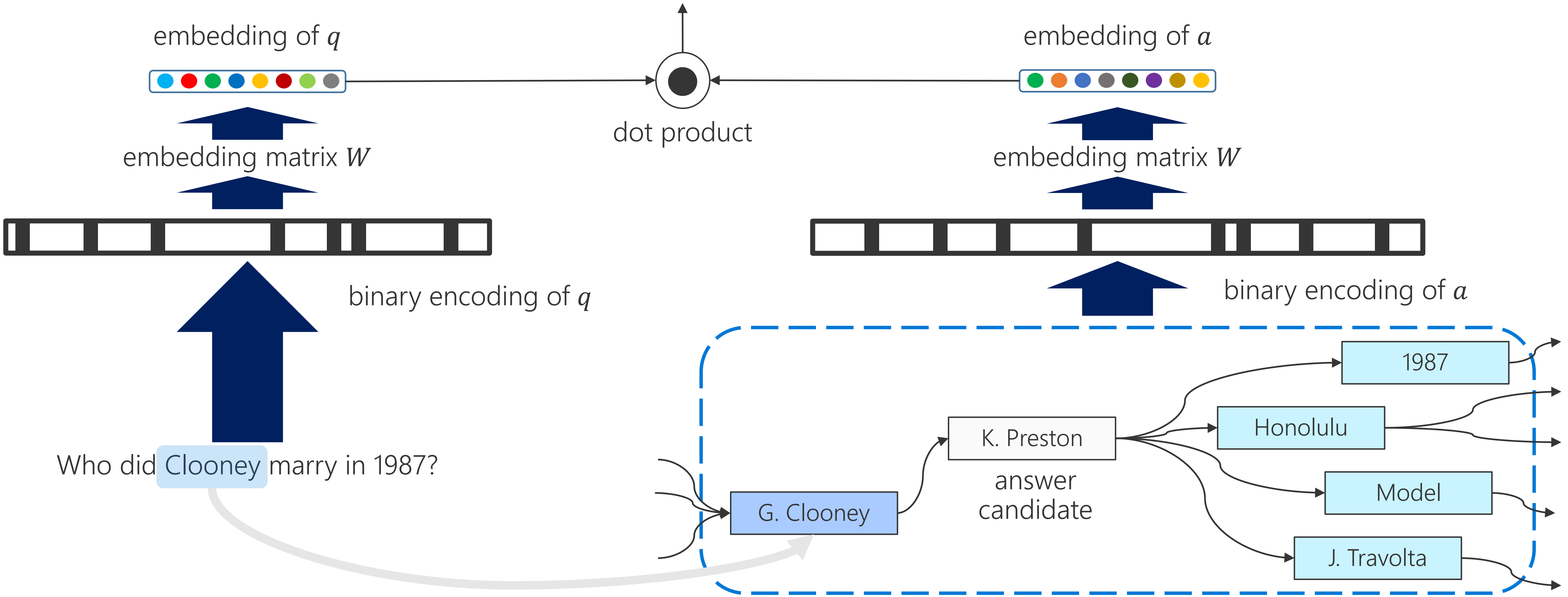
Predicate POS	Question Generation Pattern
NP	What TYPE is the NP of ENTITY ?
NP VP	What NP is VP by ENTITY ?
…	…

Ranking Model

- Compute similarity between input question and each generated question
- Features
  - Association Model (paraphrasing)
  - Vector Space Model (word embedding)

Ranking with embedding of candidate answers

(Bordes et al., 2014)



Combinatorial Categorial Grammar (CCG)

- CCG captures syntactic and semantic information jointly

A CCG Rule Example

$$\underbrace{\mathrm{border}}_{\text{natural language}} \coloneqq \underbrace{(S\backslash NP)/NP}_{\text{syntax}} : \underbrace{\lambda x \lambda y.\,\mathrm{Border}(x,y)}_{\text{semantics}}$$

- Match natural language input

- Syntactic symbols: S, N, NP, ADJ and PP
- Syntactic combinator: / and \ which specify combination orders and directions

- λ-Calculus expression

CCG-based semantic parsing for KBQA

(Zettlemoyer and Collins, 2007; Kwiatkowski et al., 2011)

State	borders	New Mexico
NP	$(\mathrm{S}\backslash\mathrm{NP})/\mathrm{NP}$	NP
$\lambda x.\,\mathrm{Type}(State, x)$	$\lambda x \lambda y.\,\mathrm{Border}(x, y)$	$\mathrm{New\ Mexico}$

$>$

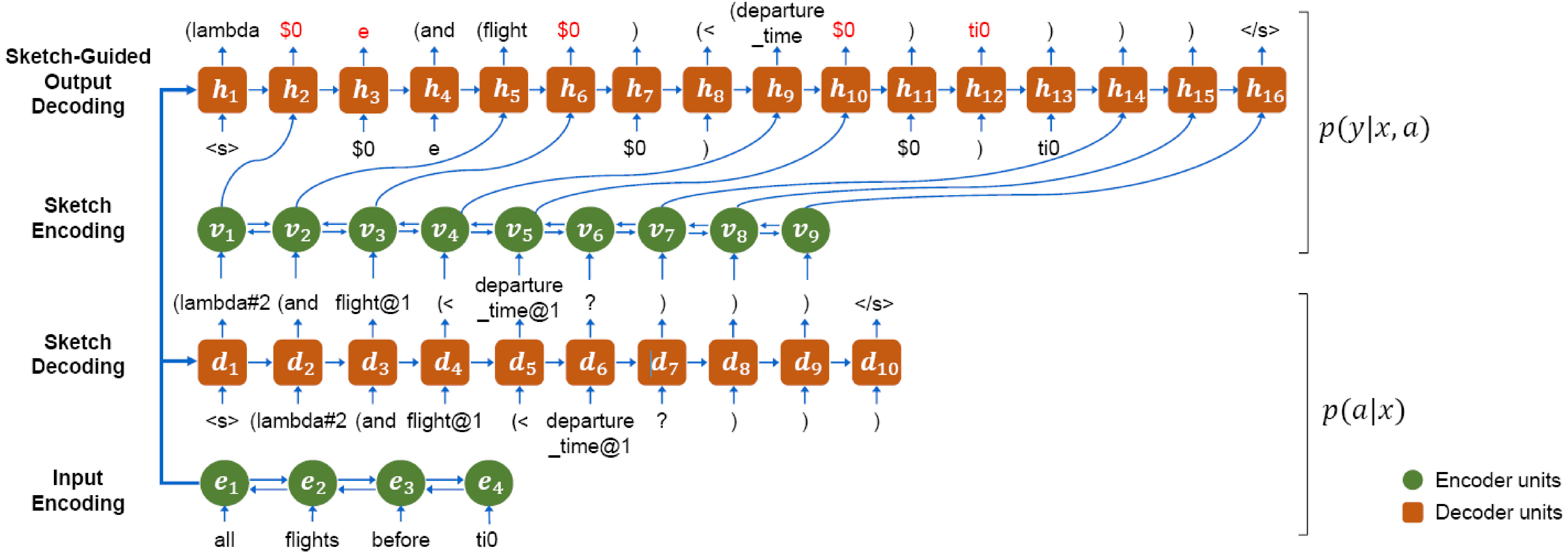
$$\mathrm{S}\backslash\mathrm{NP}$$
$$\lambda x.\,\mathrm{Border}(x, \mathrm{New\ Mexico})$$

$<$

$$\mathrm{S}$$
$$\lambda x.\,\mathrm{Type}(\mathrm{State}, x) \wedge \mathrm{Border}(x, \mathrm{New\ Mexico})$$

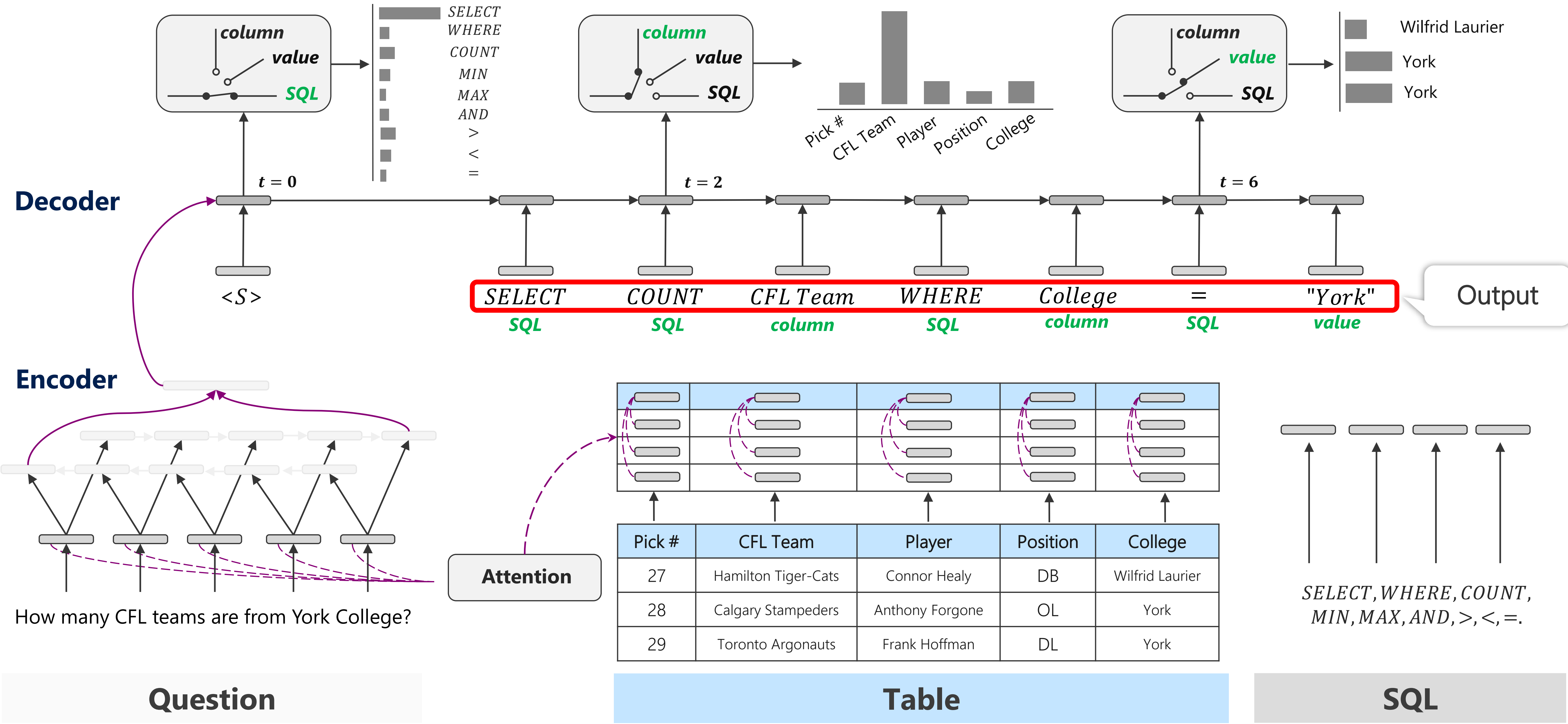
Semantic parsing with encoder-decoder

(Dong and Lapata, 2018)



Semantic parsing with multi-gates and syntax constraints

MSRA-NLC @ ACL 2018



Pick #	CFL Team	Player	Position	College
27	Hamilton Tiger-Cats	Connor Healy	DB	Wilfrid Laurier
28	Calgary Stampeders	Anthony Forgone	OL	York
29	Toronto Argonauts	Frank Hoffman	DL	York

Conversational QA

Entity Coreference

- Who is the president of the United States?
- what is its population?

- Donald Trump

Subsequence Coreference

- Where did president of the United States born?
- Where did he graduate from?

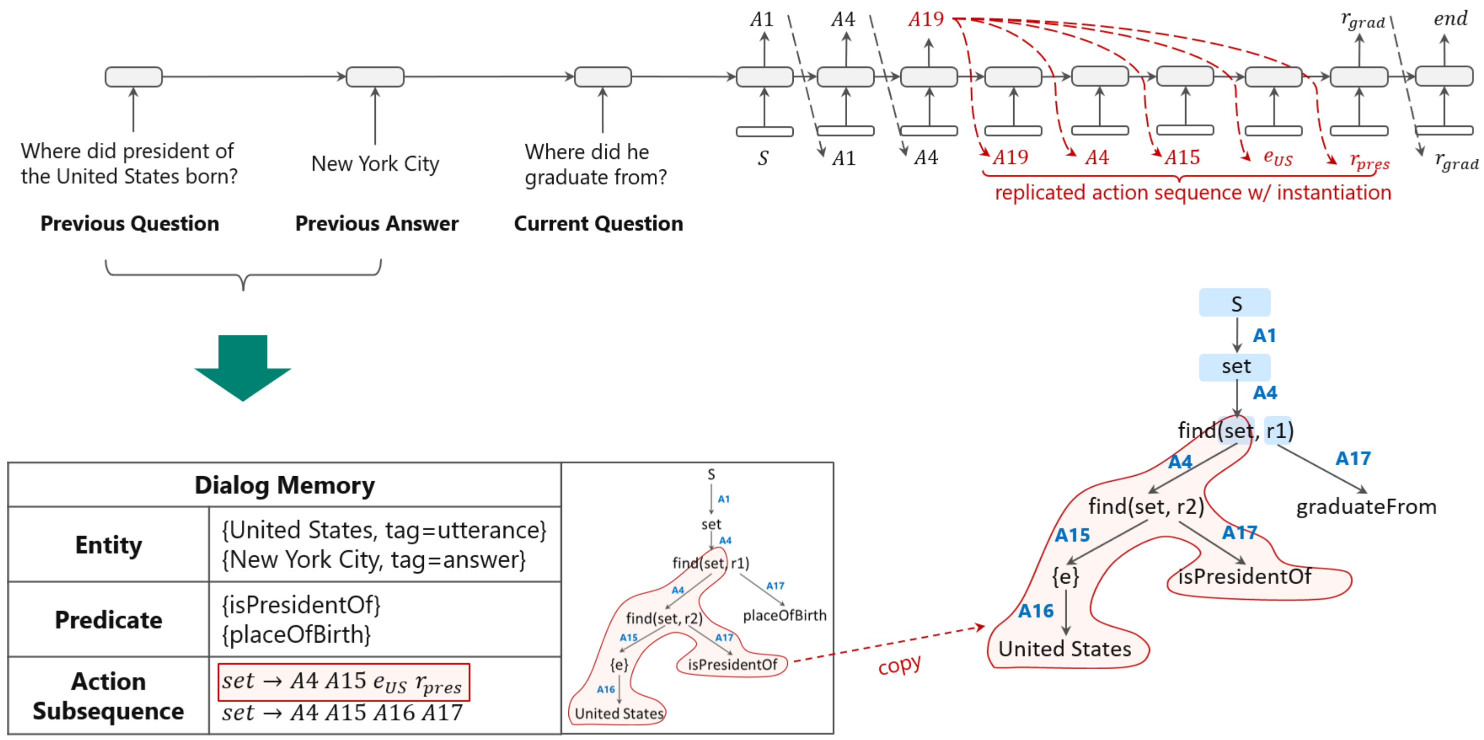
- New York City

Relation Ellipsis

- Who is the president of the United States?
- How about China

- New York City

Context-aware semantic parser



Action	Operation	Note
A1-A3	$start \rightarrow set\|num\|bool$	
A4	$set \rightarrow find(set, r)$	set of entities with a r edge to e
A5	$num \rightarrow count(set)$	total number of set
A6	$bool \rightarrow in(e, set)$	whether e is in set
A7	$set \rightarrow union(set_1, set_2)$	union of set_1 and set_2
A8	$set \rightarrow inter(set_1, set_2)$	intersection of set_1 and set_2
A9	$set \rightarrow diff(set_1, set_2)$	instances included in set_1 but not included in set_2
A10	$set \rightarrow larger(set, r, num)$	subset of set linking to more than num entities with relation r
A11	$set \rightarrow less(set, r, num)$	subset of set linking to less than num with relation r
A12	$set \rightarrow equal(set, r, num)$	subset of set linking to num entities with relation r
A13	$set \rightarrow argmax(set, r)$	subset of set linking to most entities with relation r
A14	$set \rightarrow argmin(set, r)$	subset of set linking to least entities with relation r
A15	$set \rightarrow \{e\}$	
A16-A18	$e\|r\|num \rightarrow constant$	instantiation for entity e, predicate r or number num
A19-A21	$set\|num\|bool \rightarrow action_{i-1}$	replicate previous action sequence (w/o or w/ instantiation)

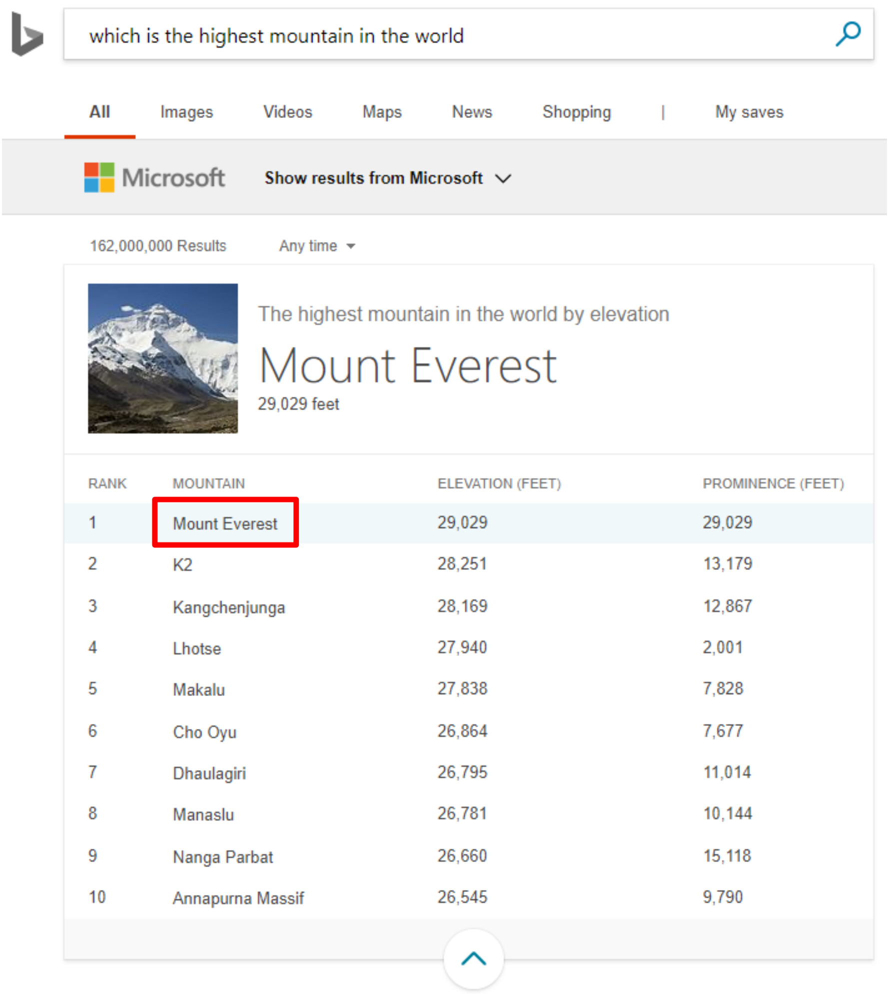
Table 2: Performance of different approaches on the CSQA dataset.

Methods	HRED+KVmem		ContxIndp-SP		D2A (our approach)	
Question Type	Recall	Precision	Recall	Precision	Recall	Precision
Overall	18.40%	6.30%	42.18%	40.88%	64.04%	61.76%
Simple Question (Direct)	33.30%	8.58%	94.04%	88.32%	93.67%	89.26%
Simple Question (Co-referenced)	12.67%	5.09%	40.29%	38.55%	71.31%	68.41%
Simple Question (Ellipsis)	17.30%	6.98%	14.09%	13.28%	86.58%	77.85%
Logical Reasoning (All)	15.11%	5.75%	36.23%	35.91%	42.49%	44.82%
Quantitative Reasoning (All)	0.91%	1.01%	43.75%	49.91%	48.59%	52.03%
Comparative Reasoning (All)	2.11%	4.97%	41.49%	38.91%	44.73%	43.69%
Clarification	25.09%	12.13%	0.01%	0.01%	19.36%	17.36%
Question Type	Accuracy		Accuracy		Accuracy	
Verification (Boolean)	21.04%		20.38%		45.05%	
Quantitative Reasoning (Count)	12.13%		30.60%		40.94%	
Comparative Reasoning (Count)	8.67%		15.54%		17.78%	

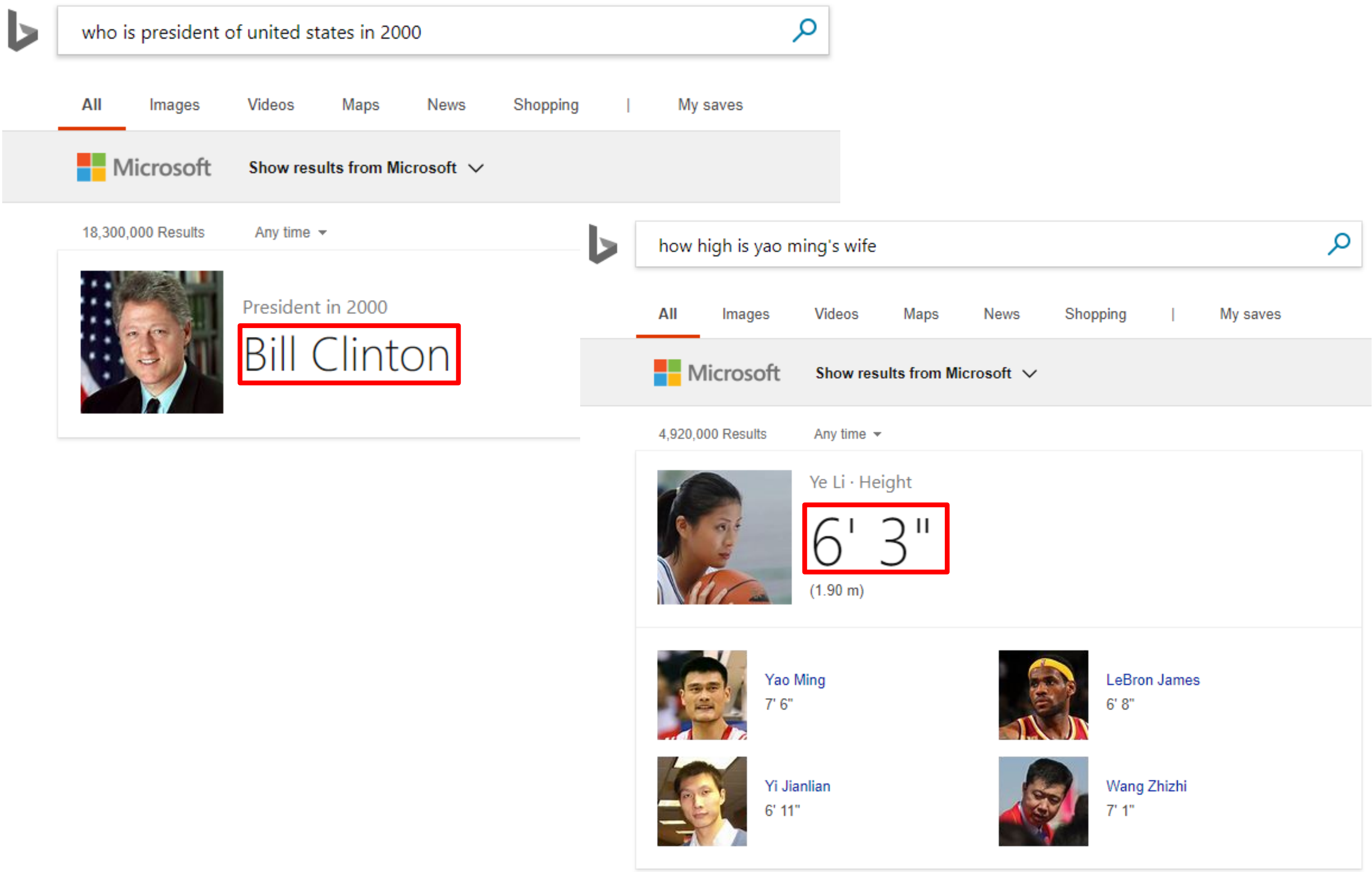
Neural Network-based Semantic Parsing for Conversational KBQA

(MSRA-NLC @ NIPS 2018)

Applications in Bing

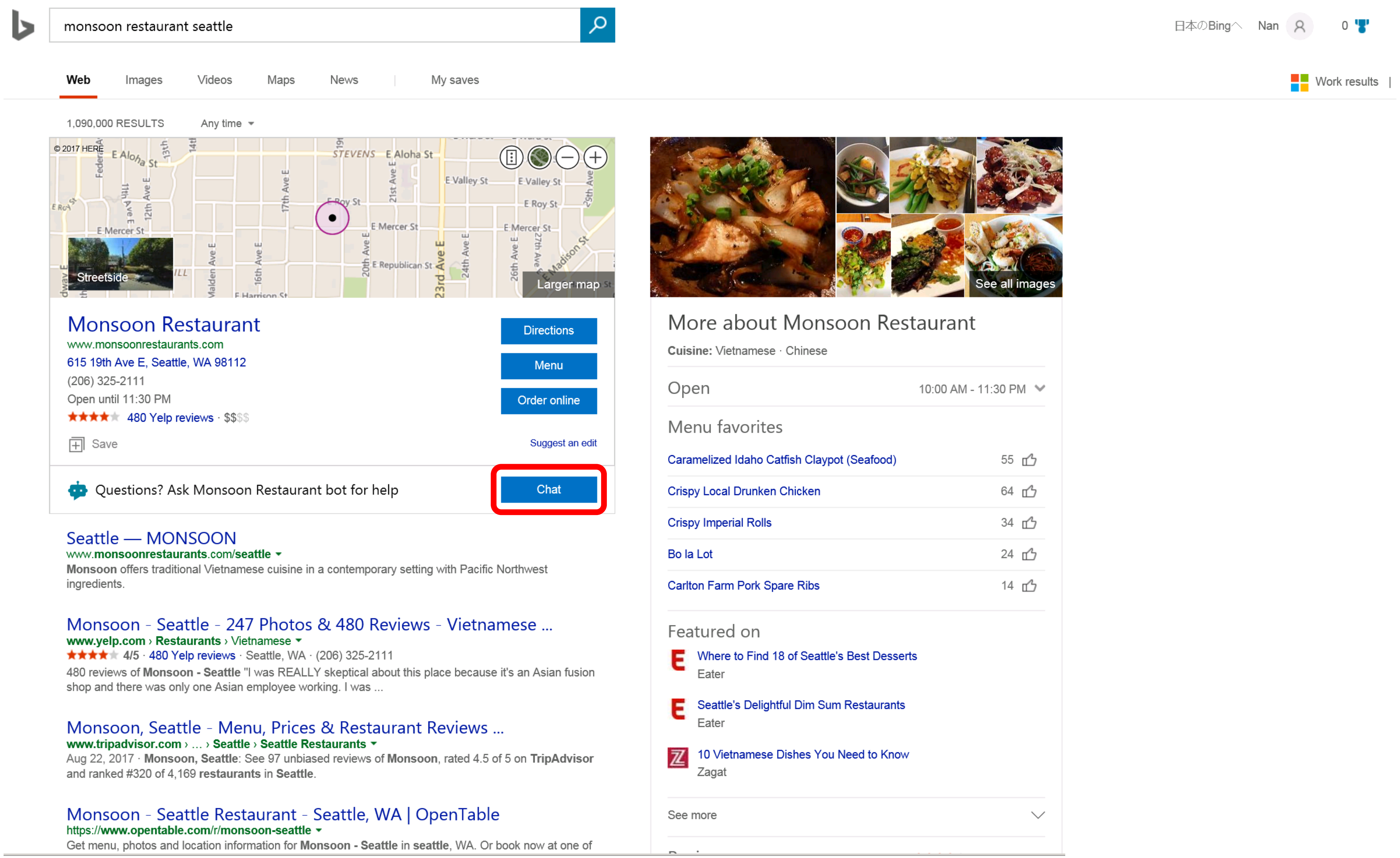


Bing Table QA



Bing Knowledge QA

Conversational search starts to appear



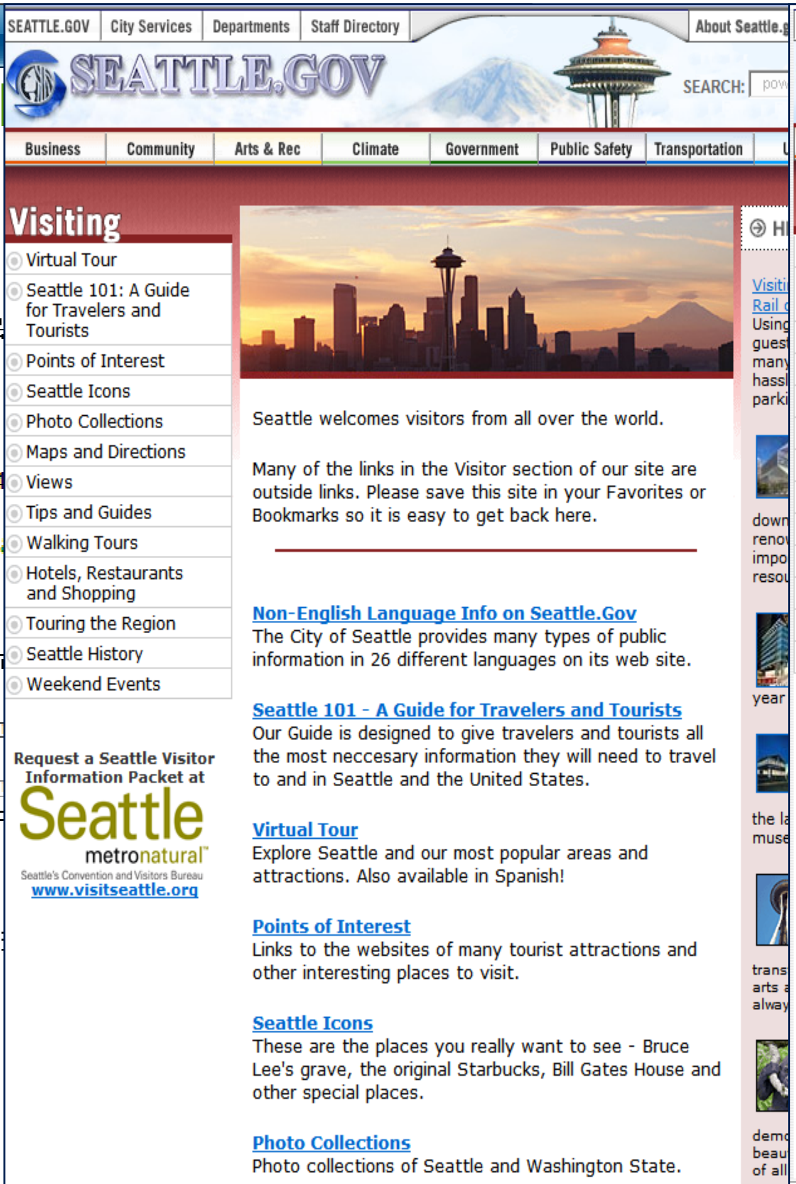
A summary about QA

1. Big progress was made in question understanding, knowledge graph and answer extraction. QA for simple question has been successfully used in search engine.
2. Traditionally, grammar-based semantic parser was applied, but now encoder-decoder approach becomes mainstreams. But I think that no conclusion can be made on which method is the best.
3. Context-aware semantic parser, as a vital technologies for conversational QA, has achieved promising result, but it needs more efforts in data annotation, modeling and testing.

New trends brought by NLP

- QA
- Multi-lingual
- Multi-modal
- MRC
- Personalized recommendation

Cross language search

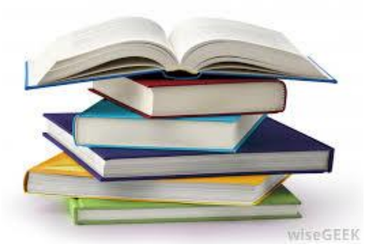


NMT research

Rich Resource MT
(Zh-En, Fr-En, De-En,...)

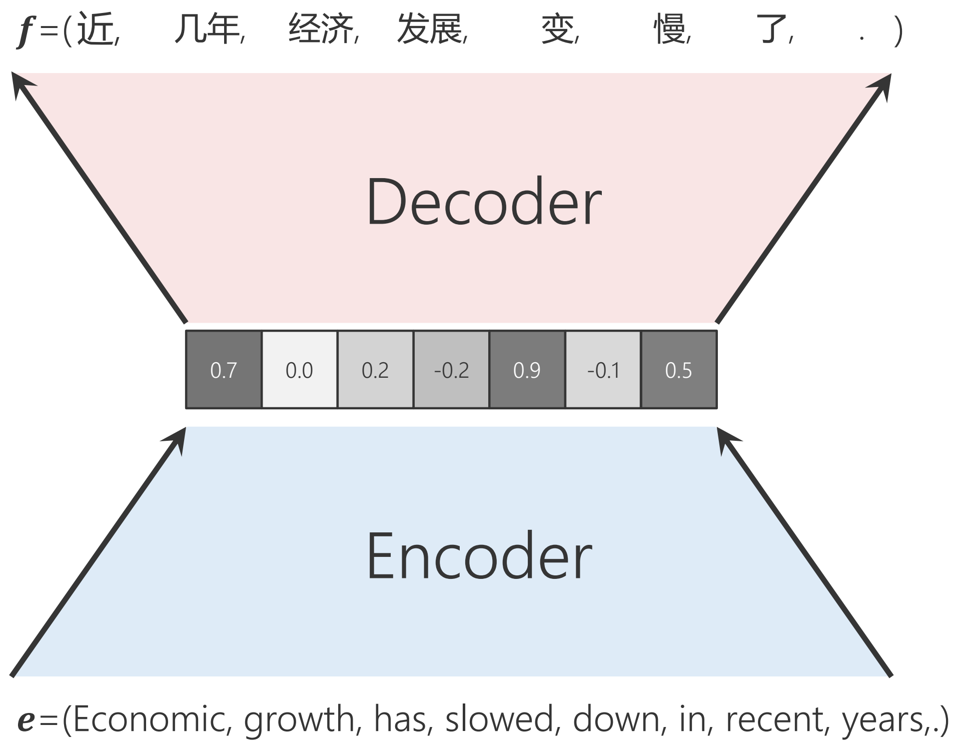
Large Bilingual Corpus

Low Resource MT
(En-He, Fr-Ro, He-Ro,...)

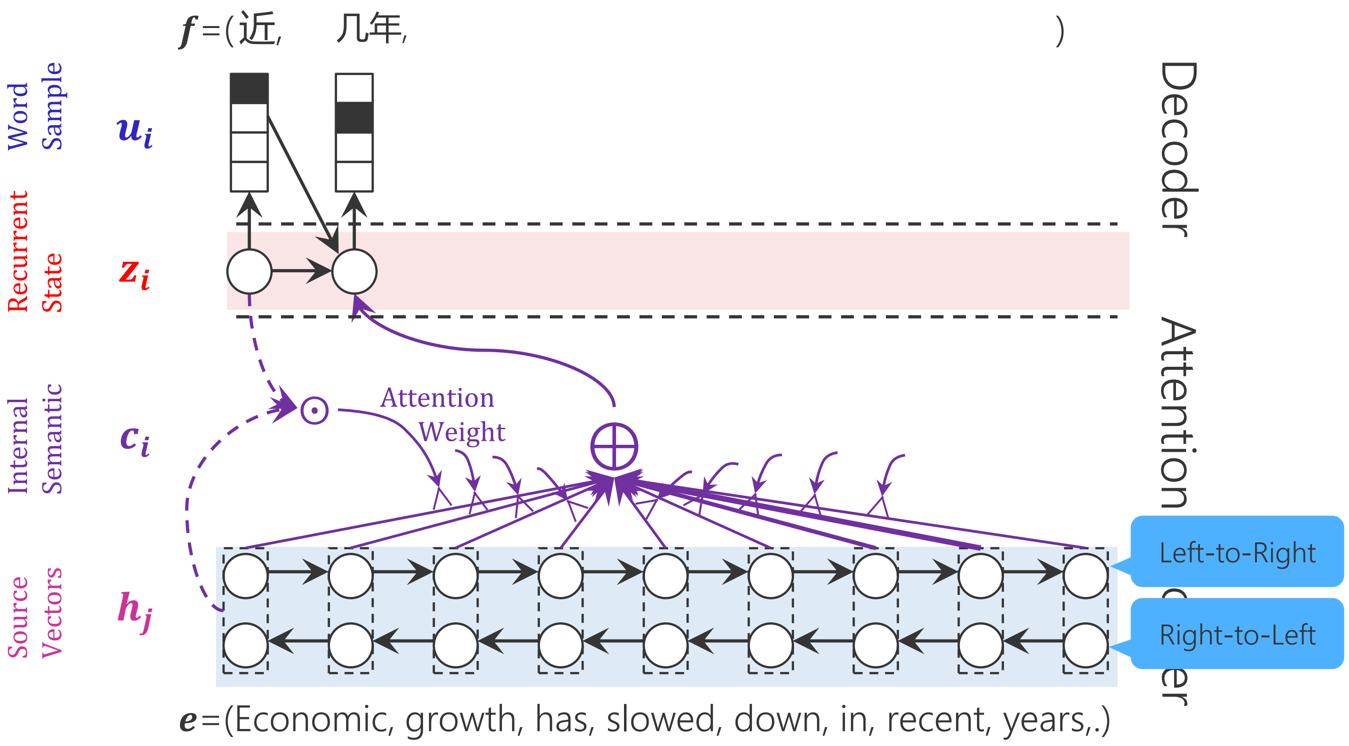


Small or No Bilingual Corpus

Neural machine translation

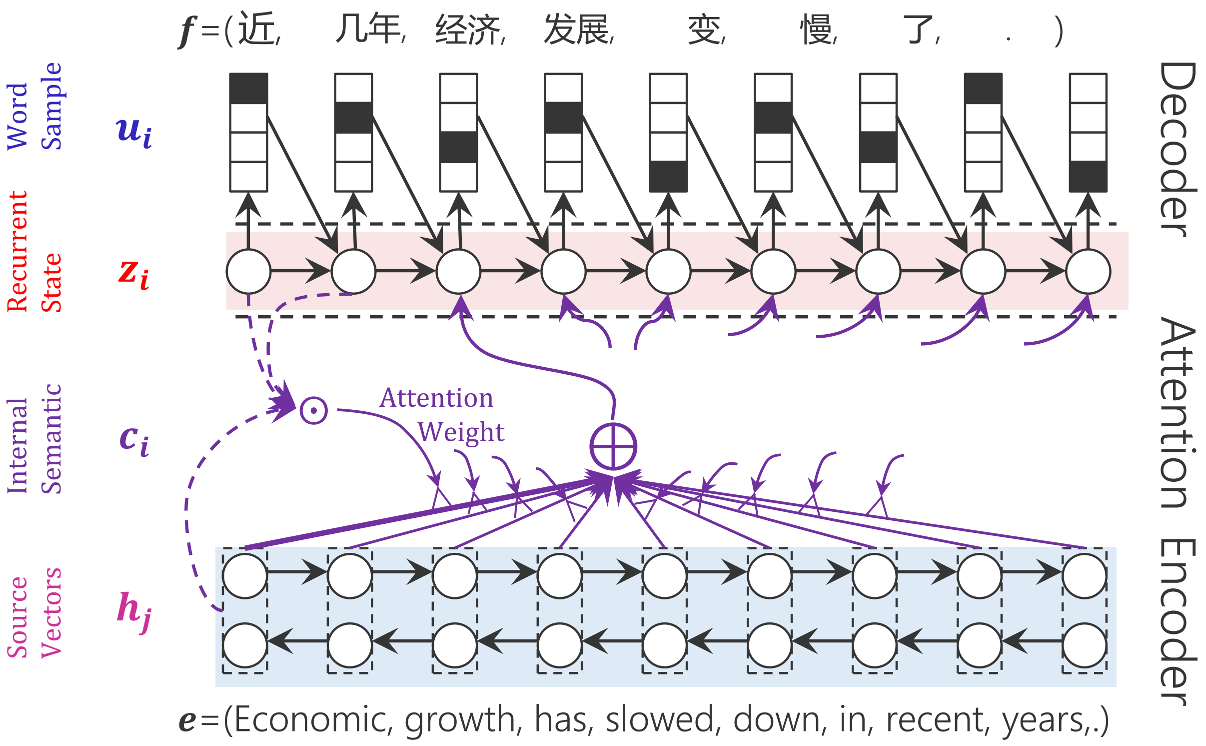


Neural machine translation



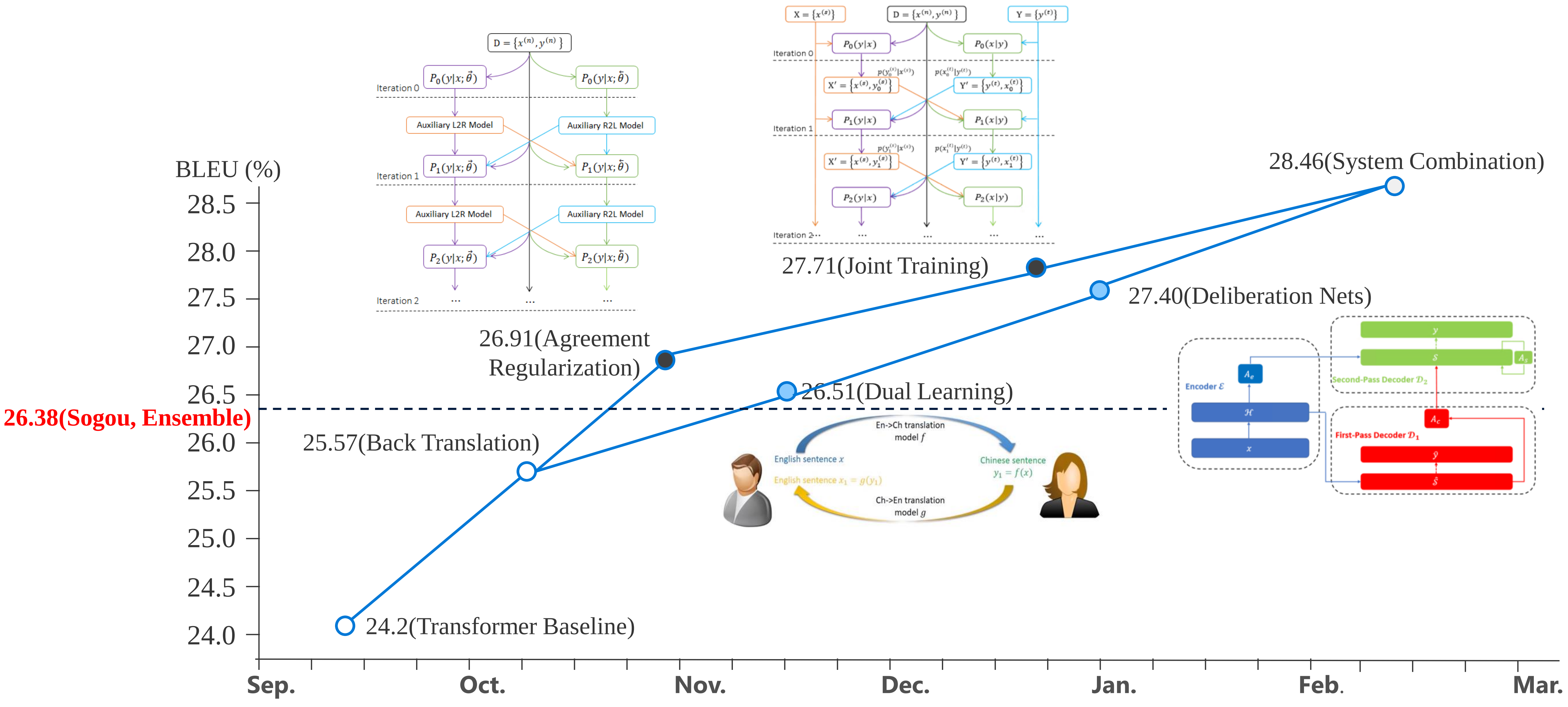
Bahdanau et al., ICLR, 2015

Neural machine translation



Bahdanau et al., ICLR, 2015

Human-Parity results on WMT 2017



News translation results

- Sampled from WMT2017 Chinese-English task

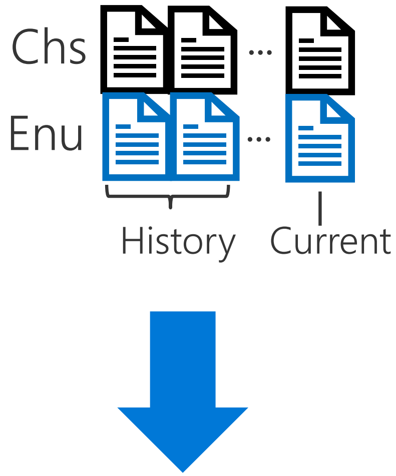
Source input	**有 线索 人士** 请 拨打 旧金山 警察局 举报 电话 4 15- 575 - 44 44 。
NMT output	For clues, call the San Francisco Police Department at 415-575 - 4444.
Human reference	Anyone with information is asked to call the SFPD Tip Line at 415-575-4444 .

Source input	他 的 职业 生涯 如 过 山 车 一般 。
NMT output	It has been a rollercoaster ride .
Human reference	His career is like a roller coaster.

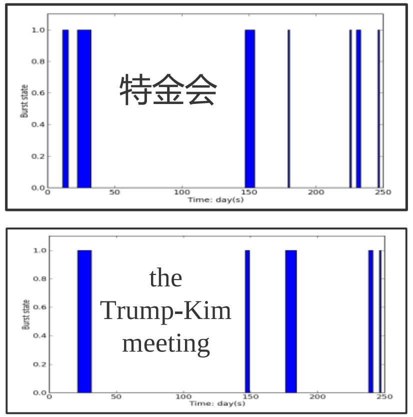
Source input	霍夫 施泰特尔 表示 : " **这 将 由 检察官 来 确定** " 。
NMT output	" That 's what the prosecutor must determine , " said Hofstetter .
Human reference	Mr Hoff Steitel said: "It will be up to the prosecutors to determine.

Term translation mining

News stream data

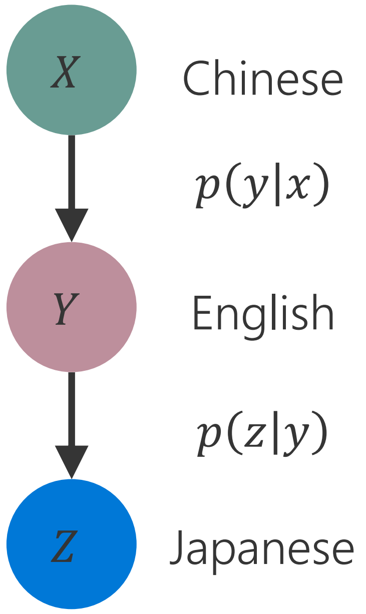


Burst distribution



Low-resource neural machine translation

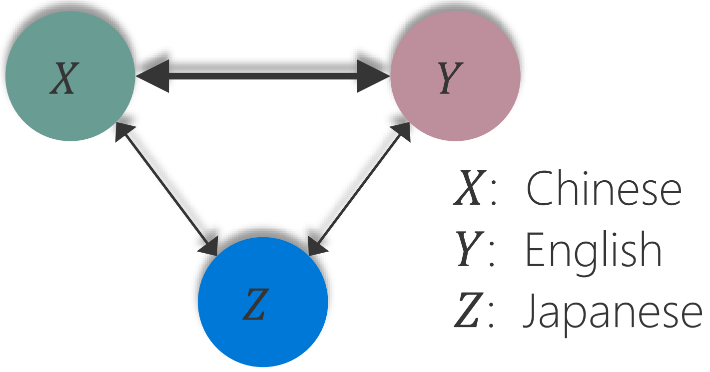
Pivot-based NMT



To translate a Chinese sentence x to a Japanese sentence z, we first translate x to an English sentence y, then translation y to Japanese z.

Chen,et al., 2017. A teacher-student framework for zero-resource neural machine translation.

Tri-Language NMT



Language Z is the hidden space between language X and language Y.

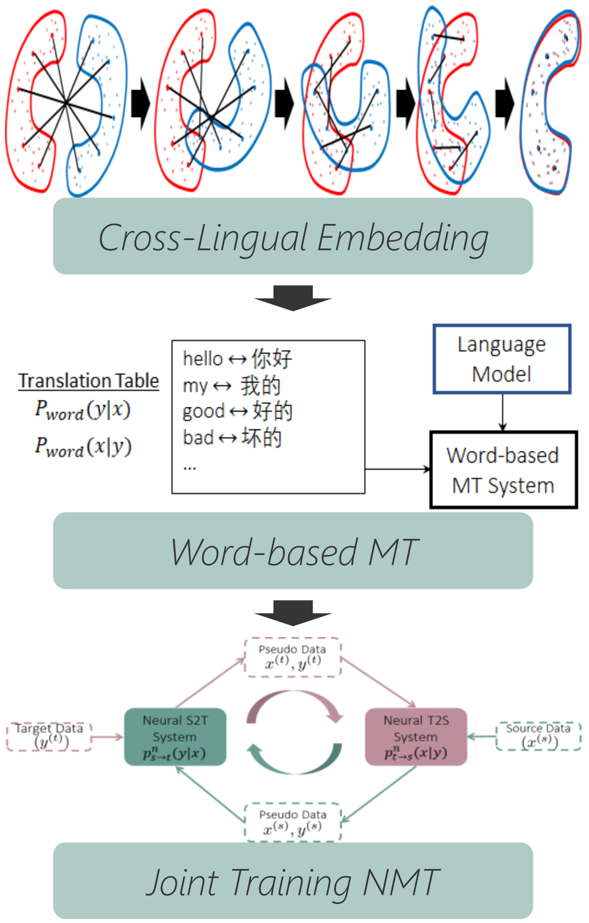
$$p(y|x) = \sum_{z \in Z} p(z|x)p(y|z)$$

$$p(x|y) = \sum_{z \in Z} p(z|y)p(x|z)$$

EM training is leveraged for fine tuning.

Ren, et al., 2018. Triangular architecture for rare language translation.

Unsupervised NMT



Lample, et al., 2018. Phrase-based & neural unsupervised machine translation.

Ren, et al., 2018. Unsupervised neural machine translation with SMT as posterior regularization. Submitted

A summary about multi-lingual search

1. Amazing progress in rich language translation at single-sentence level, such as news translation.
2. We have witnessed some progress in low-resource language translation, but it is still at very early stage. I think that it will be the most important topic in NMT.
3. We need to develop better models for term translation and query translation.

New trends brought by NLP

- QA
- Multi-lingual
- Multi-modal
- MRC
- Personalized recommendation

ImageNet

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) evaluates algorithms for object detection and image classification at large scale.

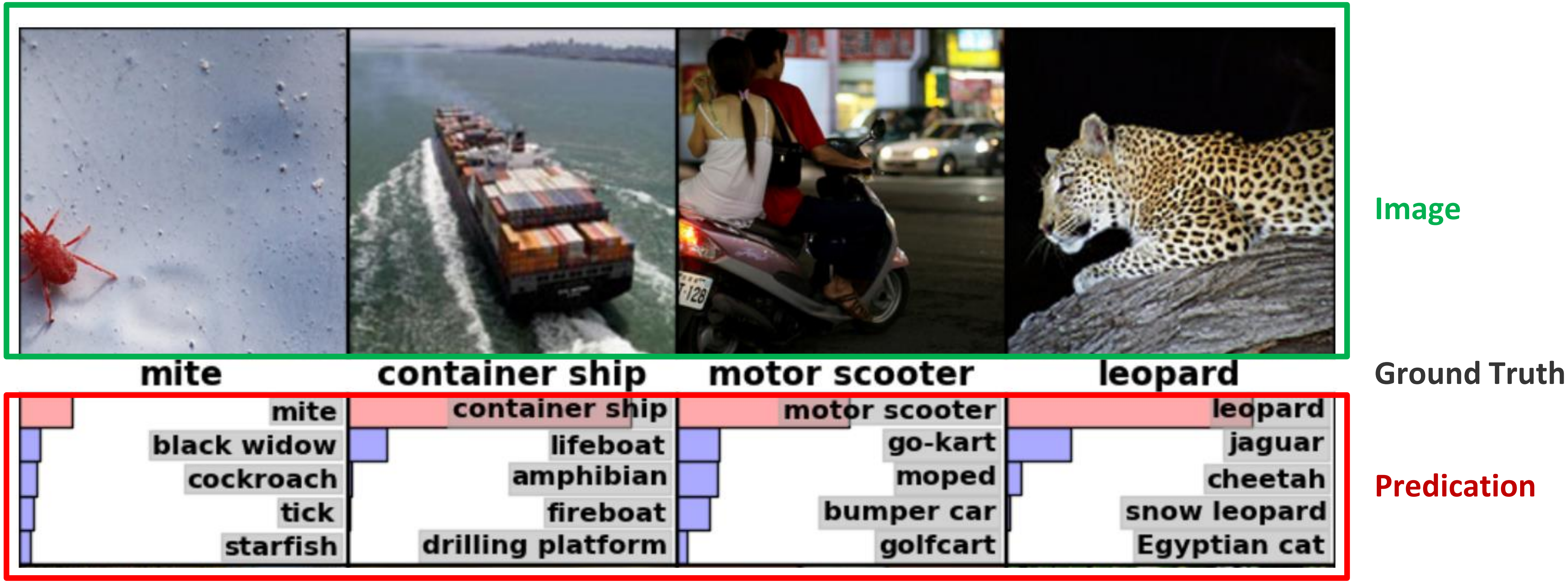


Image captioning

I think it's a man flying a kit on the beach.

I think it's a colorful bird perched on a tree branch.

I think it's a boat that is lit u at night in a city.

I think it's a little boy sitting in front of a birthday cake and he seems 😁.

Video captioning

A car is running

A man is cutting a piece of meat

A man is performing on a stage

A man is riding a bike

A man is singing

A panda is walking

A woman is riding a horse

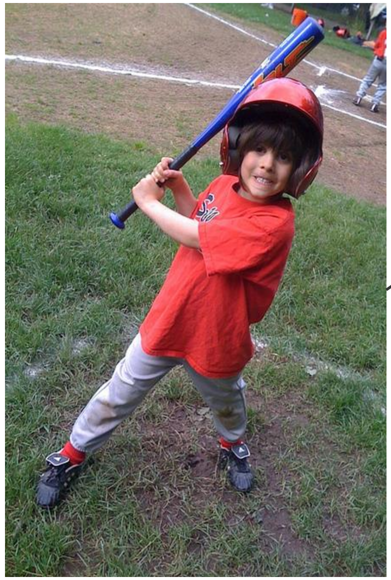
A man is flying in a field

Visual question answering (VQA)

Given an **image** and a related **question**, predict the most possible **answer**

question

What sport is the boy playing?



Image

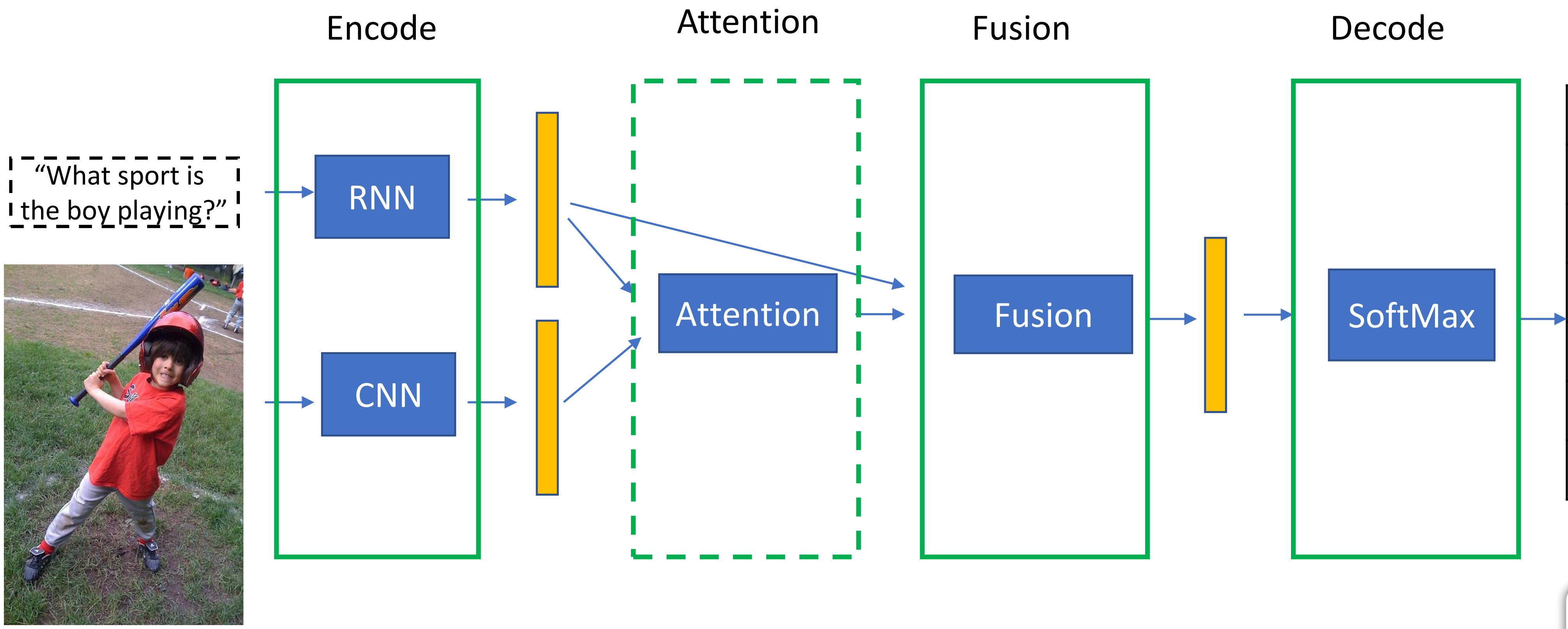
Visual QA System

answer

baseball

VQA dataset (Agrawal et al., 2016), which contains ~0.25M images, ~0.76M questions, and ~10M answers.

A typical VQA framework

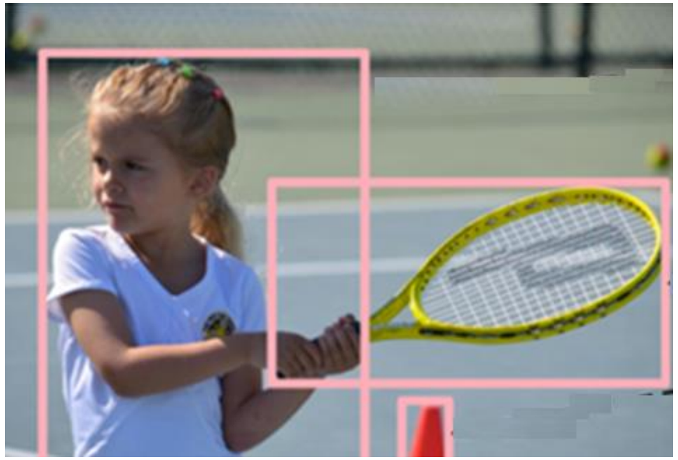


Answer	Label
"yes"	0
"baseball"	1
"brown"	0
"horse"	0
…	
"grass"	0

The size of answer candidates is ~3,000.

Motivating example of using knowledge

- **Image mis-understanding due to relation missing**

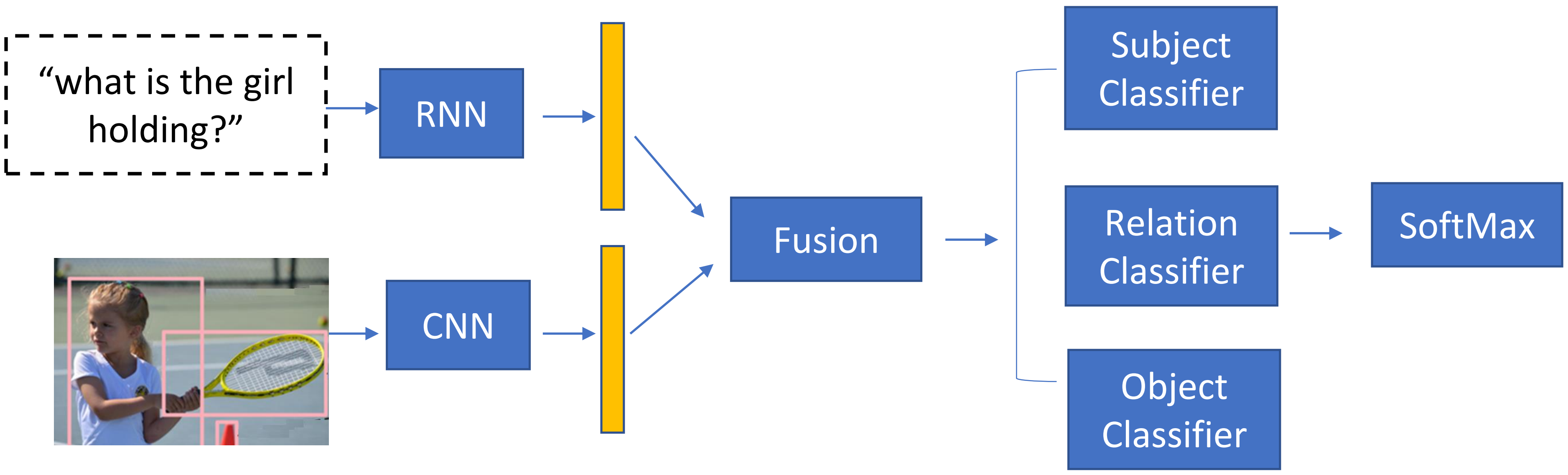


Question: what is the girl holding?
Ground Truth Answer: racket
Predicted Answer: [hand: 0.630] [racket: 0.21] [cone: 0.072] [mouse: 0.004]

If we know (girl, hold, racket) is valid, and (girl, hold, hand) is not valid, we probably can infer "racket" as a better answer.

Knowledge mining from image

Training data comes from **Visual Genome dataset** (Krishna et al., 2017)

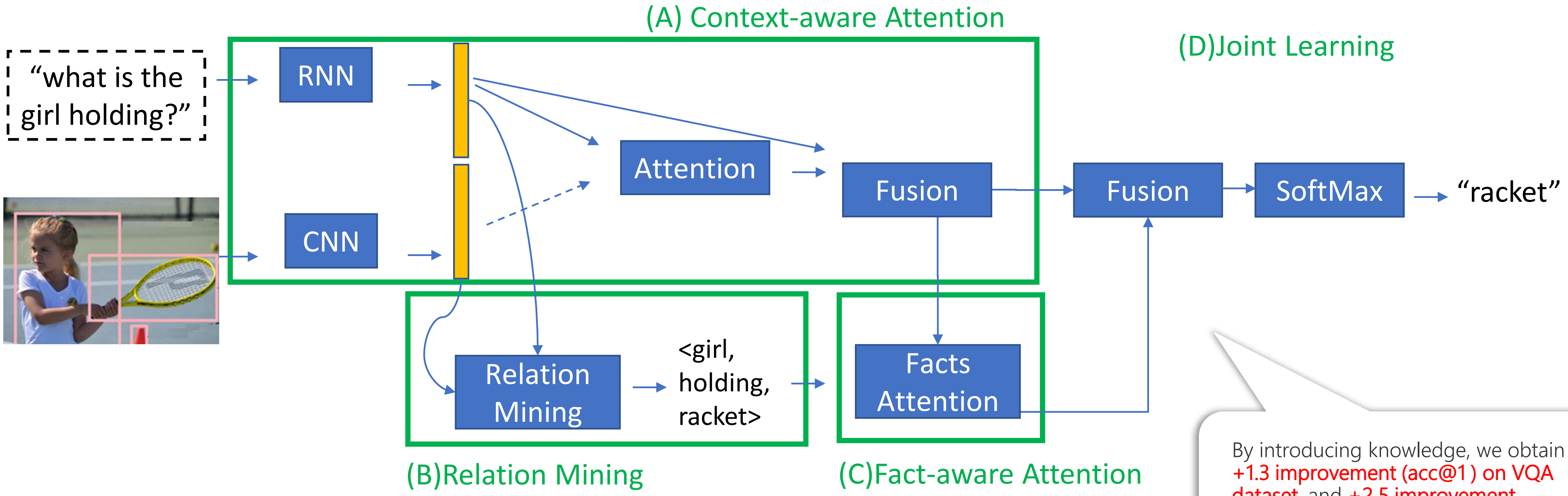


girl	by	cone
girl	holding	racket
girl	is	standing
racket	is	yellow
cone	is	orange

$$L_t = \lambda_s L(s, \hat{s}) + \lambda_r L(r, \hat{r}) + \lambda_o L(o, \hat{o}) + \lambda_w \|W\|_2$$

Lu et al. R-VQA: Learning Visual Relation Facts with Semantic Attention for Visual Question Answering. KDD, 2018.

VQA with knowledge



By introducing knowledge, we obtain +1.3 improvement (acc@1) on VQA dataset, and +2.5 improvement (acc@1) on COCO QA dataset.

[Lu et al. R-VQA: Learning Visual Relation Facts with Semantic Attention for Visual Question Answering. KDD, 2018.]

VQA with knowledge mining examples

Q: Does this building have a clock ?
A: yes
R: *(clock, on, building)* 0.81
(window, on, building) 0.07
(building, has, clock) 0.03
(clock, near, tower) 0.01
(clock tower, near, building) 0.01

(a)

Q: What are the animals standing on ?
A: grass
R: *(animal, standing on, grass)* 0.27
(animal, standing on, field) 0.20
(zebra, standing on, grass) 0.19
(zebra, standing on, field) 0.14
(animal, standing on, gazing) 0.13

(b)

Q: What color is the ball ?
A: yellow
R: *(ball, is, yellow)* 0.43
(tennis ball, is, yellow) 0.36
(ball, is, *green)* 0.12
(there, is, *tennis ball)* 0.05
(ball, in, air) 0.02

(c)

Q: What kind of court is this ?
A: tennis court
R: *(there, is, tennis court)* 0.52
(white lines, in, tennis court) 0.21
(line, in, tennis court) 0.09
(grass, in, tennis court) 0.08
(court, is, orange) 0.03

(d)

Image retrieval

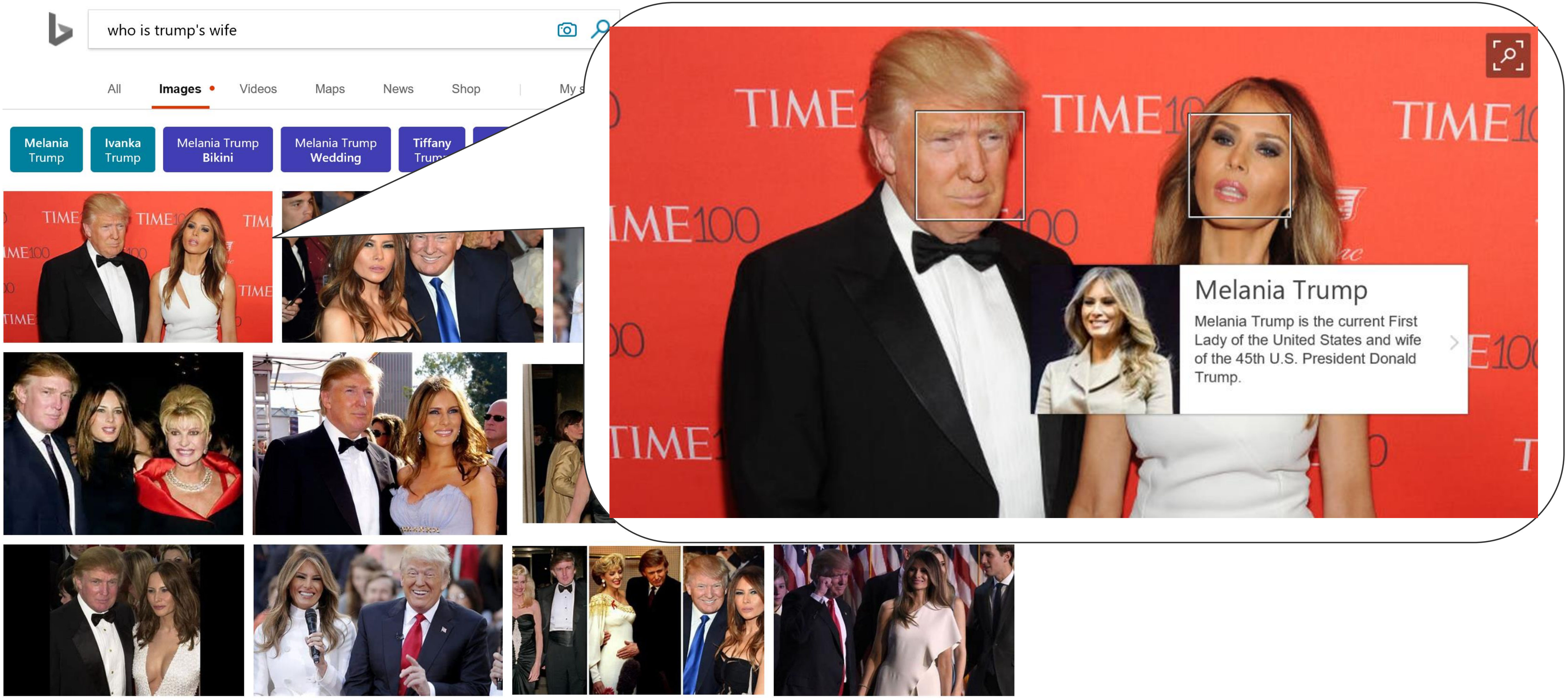
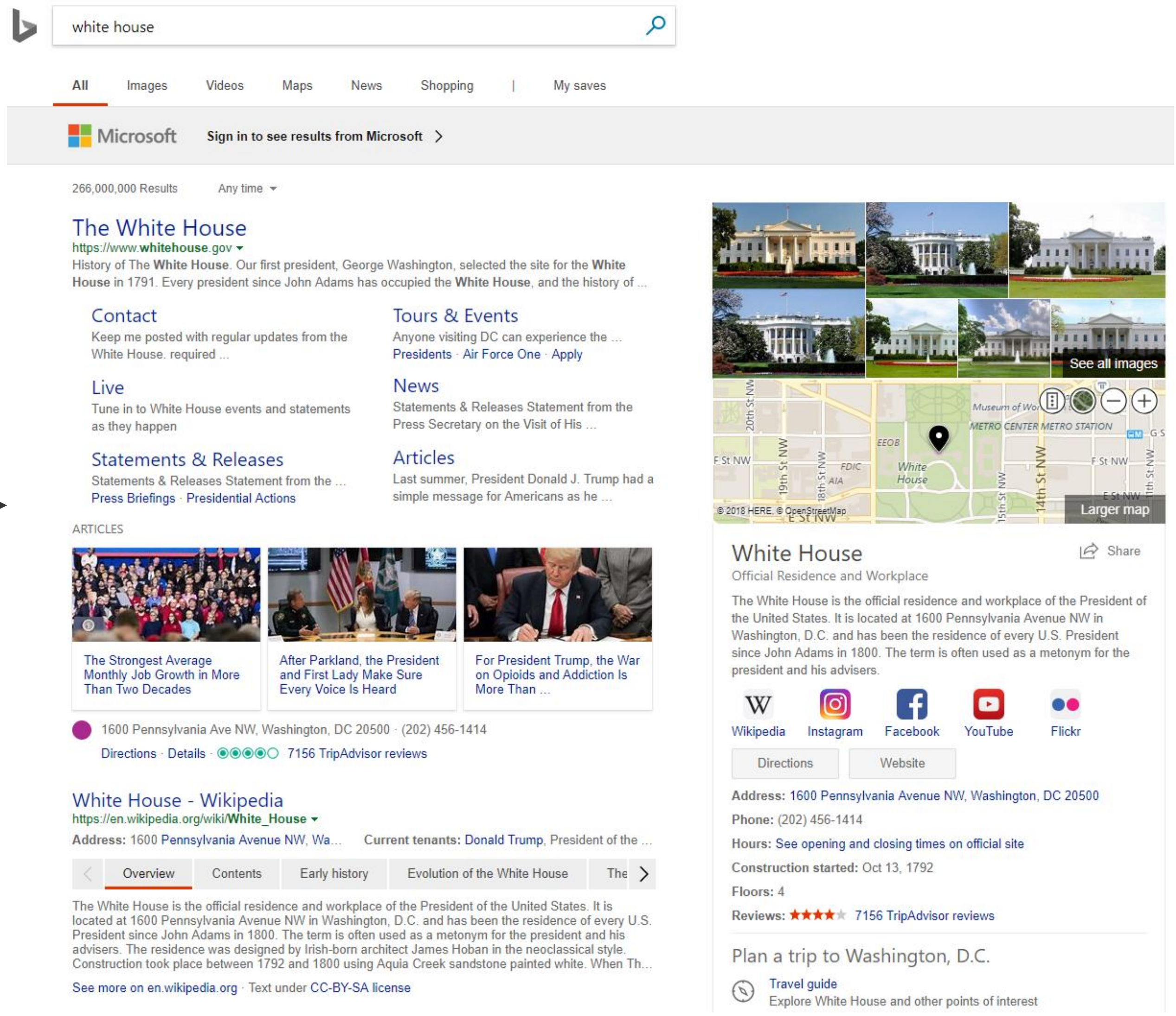


Image query

Image query:

Search Engine



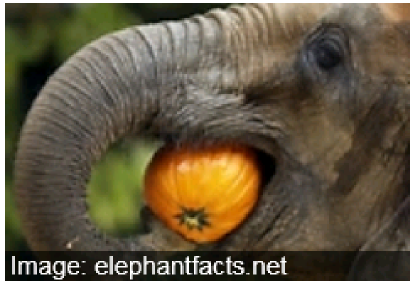
Enrich search with images

what do elephants eat

Web Images Videos Maps News | My saves

57,700,000 RESULTS Any time

What Elephants Eat. An elephants size means that, when it is fully grown, it is safe from all predators other than humans. To support its large size it needs huge amounts of food. An adult elephant eats about 150 kg (300 lb) of grass, **leaves**, twigs and **fruit** each day.



What Elephants Eat And Drink - Live Animals List
www.liveanimalslist.com/mammals/what-elephants-eat-and-drink.php

Improve this answer · Is this answer helpful?

A summary about multi-modal search

1. Recent progress of ImageNet and image captioning inspires researchers to conduct research on multi-modal understanding and search.
2. VQA is still at very early stage. There are many challenging topics such as use of knowledge and common sense.
3. There are a few applications in search engine but all are very simple. We expect that with the technologies continually advance, new scenarios will appear.

New trends brought by NLP

- QA
- Multi-lingual
- Multi-modal
- MRC
- Personalized recommendation

Machine reading comprehension

Read a document (passage) and then answer questions about it

Passage (P) + Question (Q) Answer (A)

P

Tesla later approached Morgan to ask for more funds to build a more powerful transmitter. **When asked where all the money had gone, Tesla responded by saying that he was affected by the Panic of 1901**, which he (Morgan) had caused. Morgan was shocked by the reminder of his part in the stock market crash and by Tesla's breach of contract by asking for more funds.

Q On what did Tesla blame for the loss of the initial money?

A Panic of 1901

SQuAD

The Stanford Question Answering Dataset

SQuAD: 100,000+ Questions for Machine Comprehension of Text

Best Resource Paper in EMNLP 2016

Pranav Rajpurkar and **Jian Zhang** and **Konstantin Lopyrev** and **Percy Liang**
{pranavsr,zjian,klopyrev,pliang}@cs.stanford.edu
Computer Science Department
Stanford University

passage → In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

query → What causes precipitation to fall?

answer → **gravity**

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

graupel

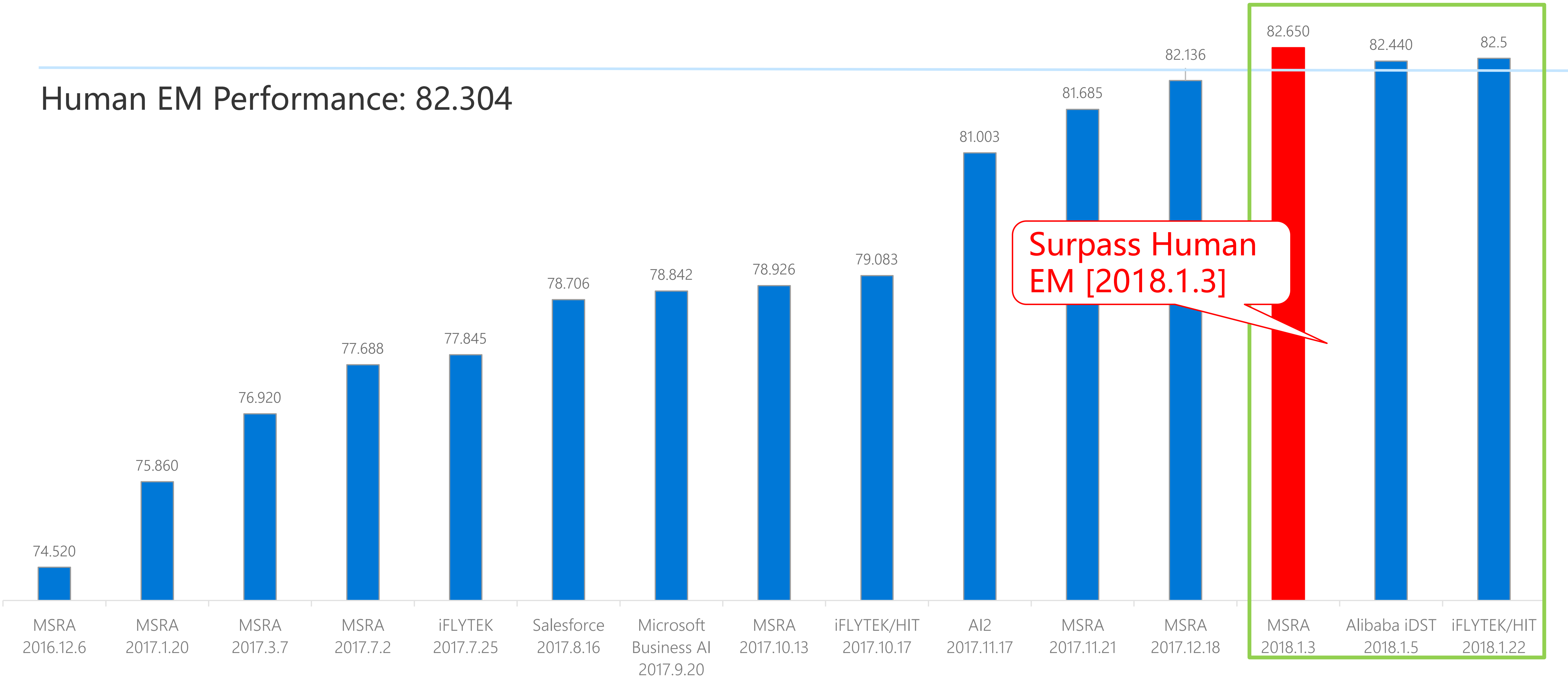
Where do water droplets collide with ice crystals to form precipitation?

within a cloud

Dataset	# of <question, passage> pairs
Training	87,599
Dev	10,570
Test (not available to participants)	> 10K

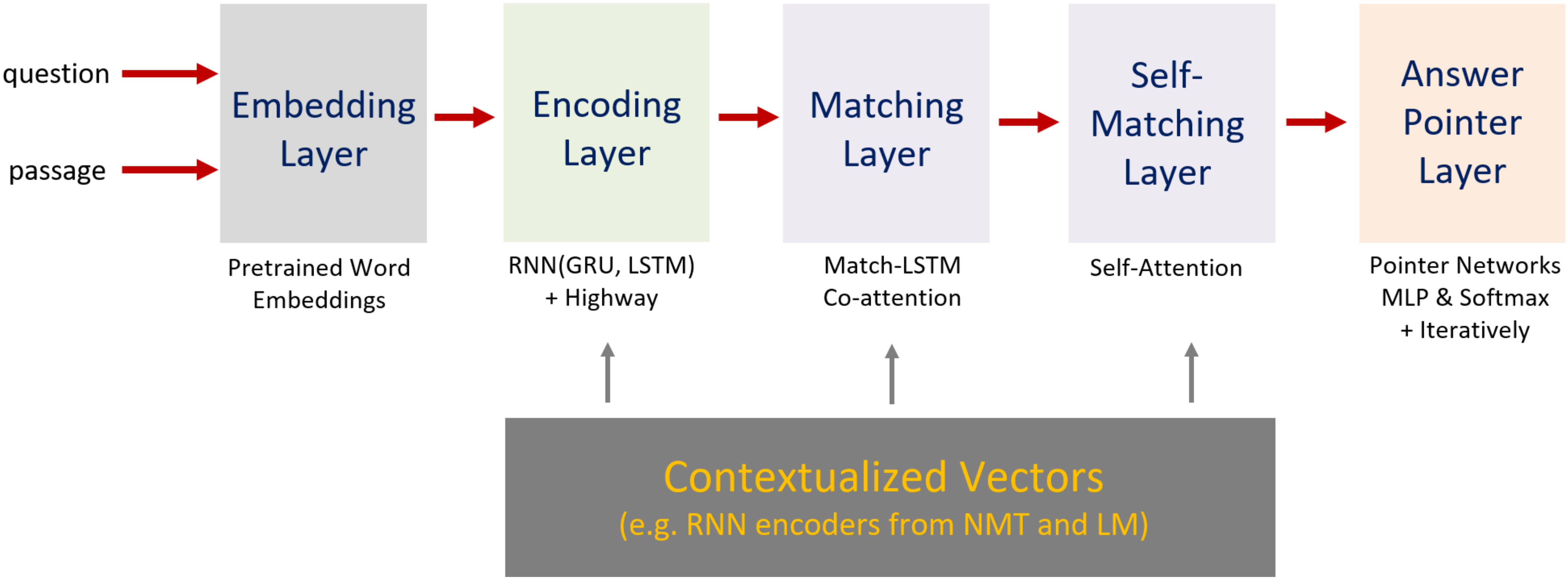
ImageNet style competition for machine reading comprehension

Progress of MRC experiments



Best System EM Scores on SQuAD Machine Reading Comprehension Dataset (Dec. 6, 2016-Jan. 26, 2018)

MRC architecture (r-net)



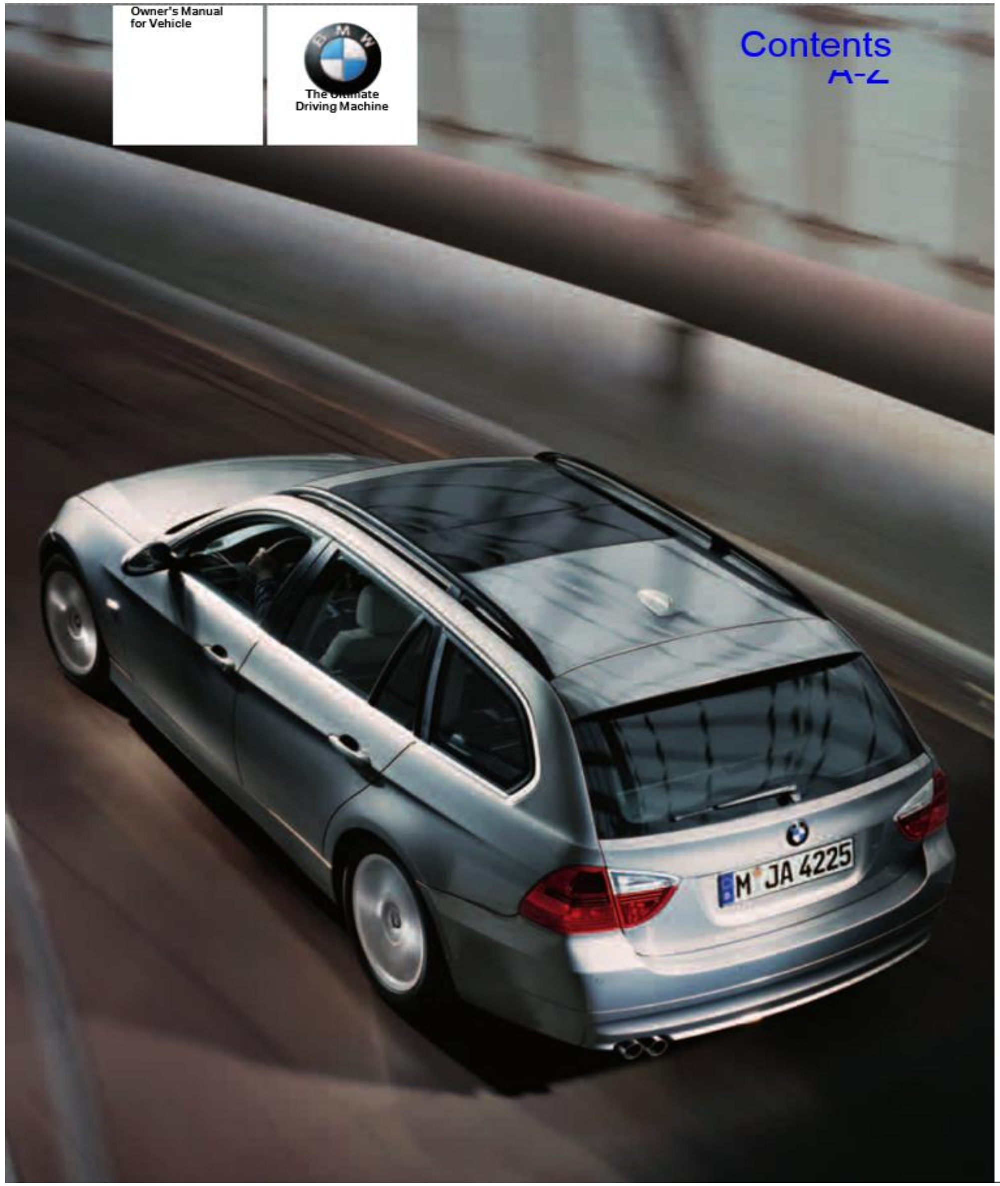
e.g., ELMo – Embeddings from Language Models

Improved snippet

Passage selection

Answer span

Manual search



Contents

The fastest way to find information on a particular topic or item is by using the index, refer to page 146.

Manual search

What is the phone number of BMW north Ameri

Prod Int

You can also call BMW of North America at 1-800-831-1117 or visit the website www.bmwusa.com to obtain this information.
Viscosity ratings

North America, LLC.
To contact NHTSA, you may call the Vehicle Safety Hotline toll-free at 1-800-327-4236 (TTY: 1-800-424-9153); go to

in its vehicles with extensive testing.
Only use approved BMW High Performance Synthetic Oil.

If BMW High Performance Synthetic Oil is not available, you can add small quantities of other synthetic oils in between oil changes. Only use oils with the API SH specification or higher.

Your BMW Center will be glad to answer any questions regarding BMW High Performance Synthetic Oil or approved synthetic oils.◀

You can also call BMW of North America at 1-800-831-1117 or visit the website www.bmwusa.com to obtain this information.

Viscosity ratings

Viscosity is a measure of an oil's flow rating and is categorized in SAE classes.

Selecting the appropriate SAE class depends on the regional climatic conditions in which you normally drive your BMW.

Approved oils belong to the 5W-40 and 5W-30 classes.◀

These oils can be used for driving at all outside temperatures.

Coolant

Do not add coolant to the cooling system when the engine is hot. Escaping coolant can cause burns.◀

Coolant is a mixture of water and an additive. Not all commercially available additives are suitable for your BMW. Ask your BMW Center for suitable additives.

Only use suitable additives, otherwise engine damage may result. The additives are hazardous to your health.◀

Comply with the appropriate environmental protection regulations when dis-

pressure to escape, then continue turning to open.

3. The coolant level is correct if it is between the maximum and minimum marks in the filler neck, refer also to the diagram next to the filler neck.

4. If the coolant is low, slowly add coolant up to the specified level; do not overfill.
5. Turn the cap until there is an audible click.
6. Have the reason for the coolant loss eliminated as soon as possible.

Brake system

Malfunctions

Brake fluid

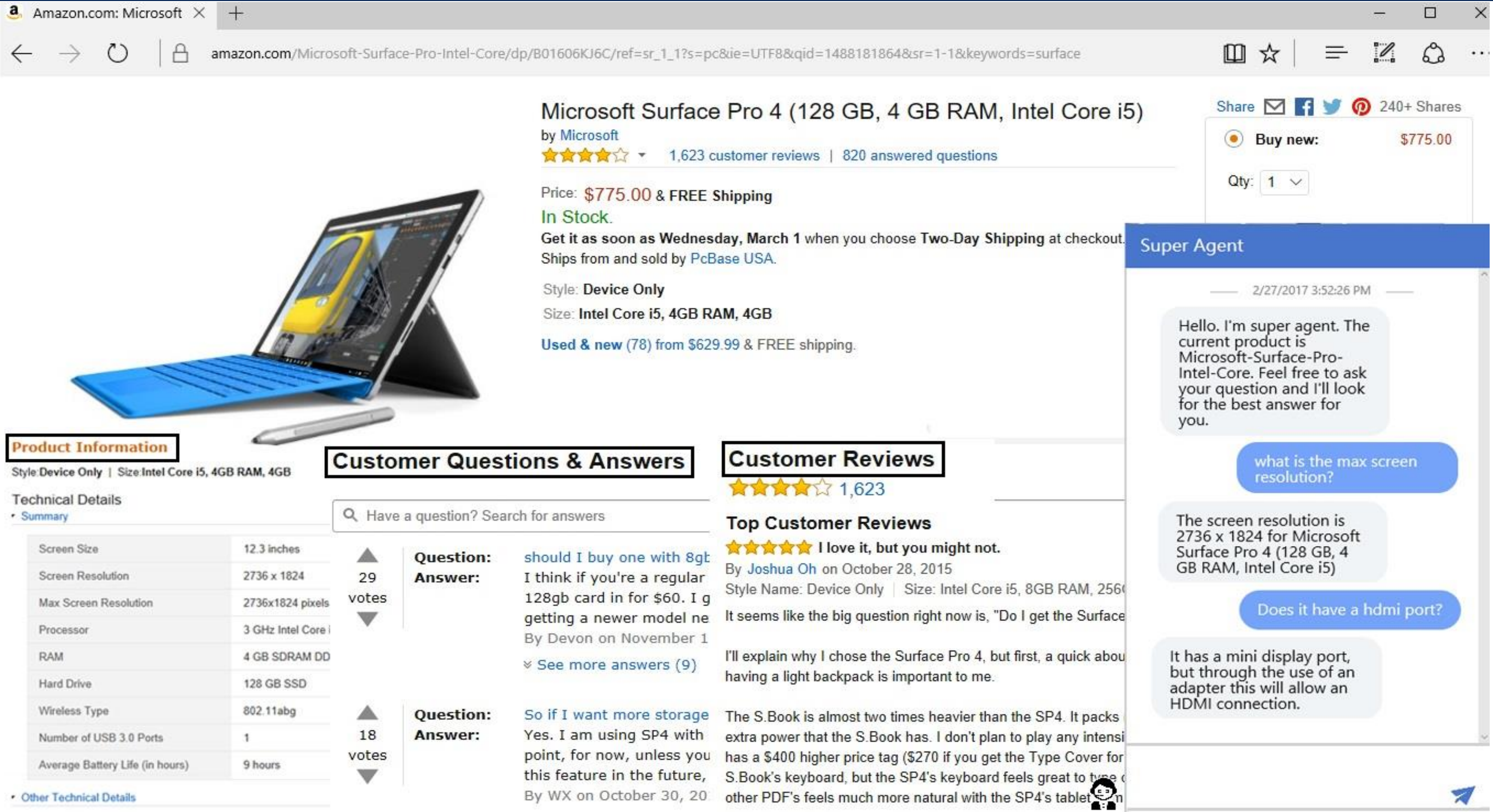
BRAKE

The warning lamps light up in red even though the handbrake has been released. Stop immediately.

The brake fluid in the reservoir has fallen to below the minimum level. At the same time, a considerably longer brake pedal travel may be noticeable. Have the system checked without delay.

Display of this malfunction on Canadian models.

Site chat with MRC



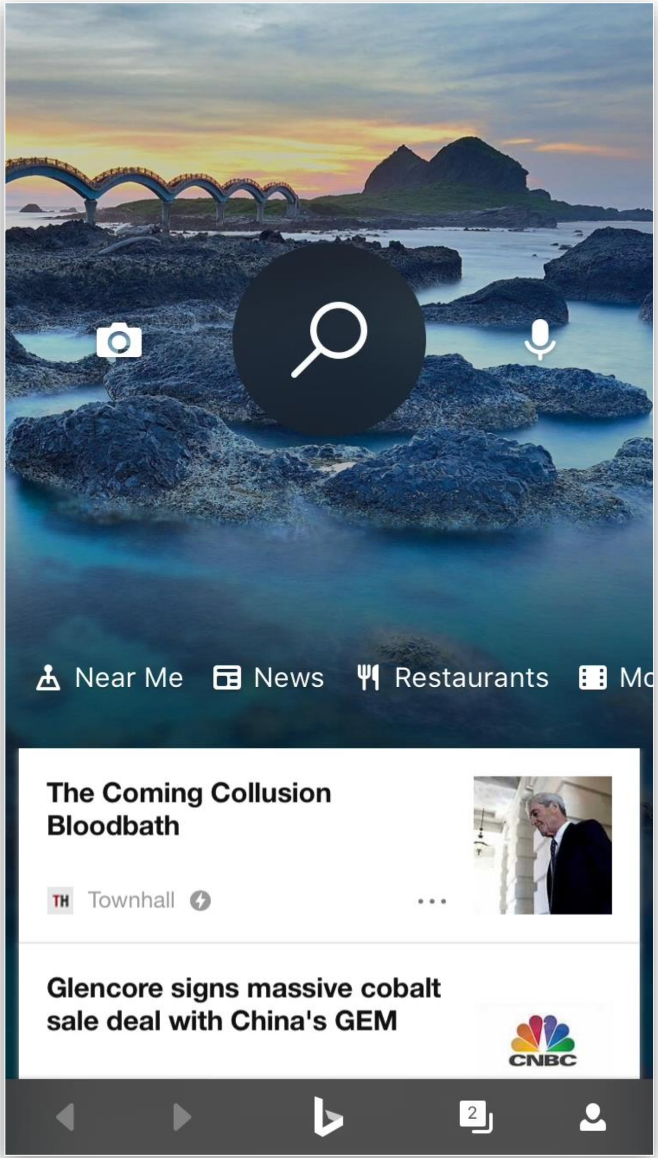
A summary about MRC

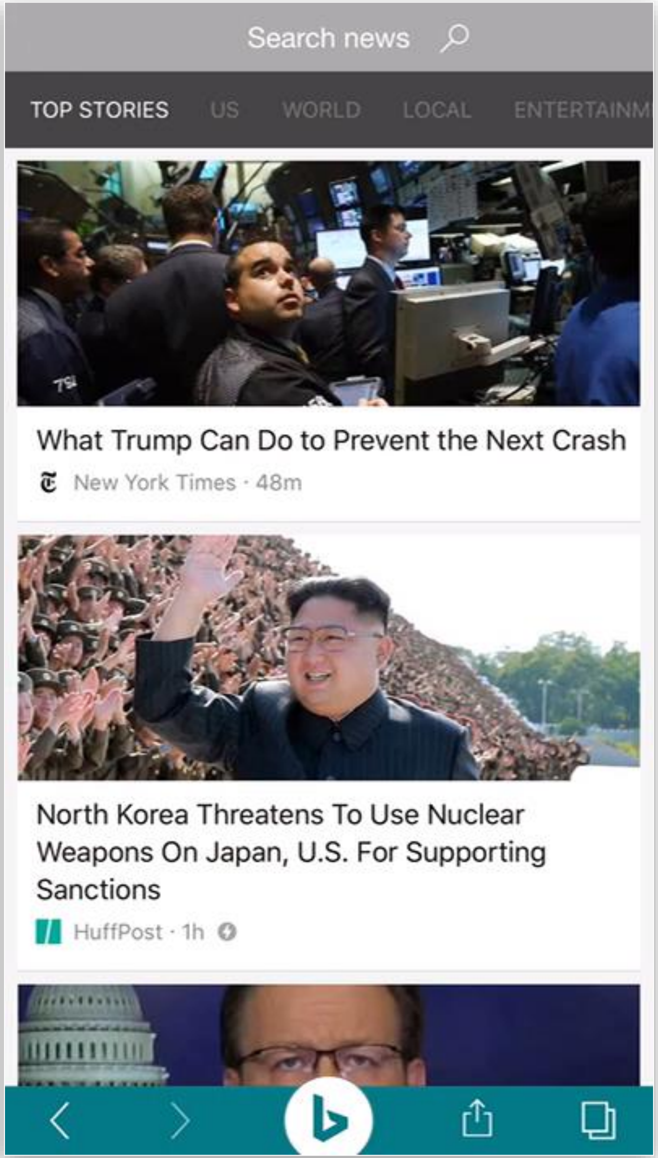
1. Recently we’ve sees a big progress in MRC, such as SQuAD evaluation. End-end training, pre-trained models, contextualized vectors from other tasks contribute to the growth of MRC.
2. MRC will significantly change search engine with better relevance and accuracy. New scenarios such as manual search are promising.
3. Further developing MRC needs contextual inference supported by knowledge and common sense. This would be very challenging and exciting.

New trends brought by NLP

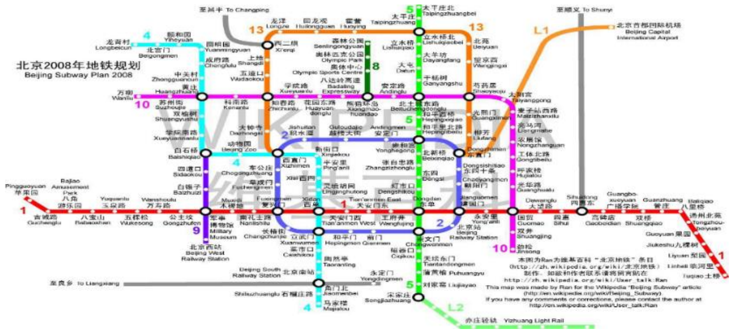
- QA
- Multi-lingual
- Multi-modal
- MRC
- Personalized recommendation

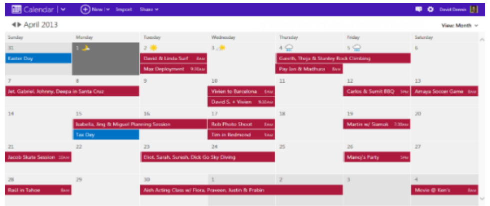
Personalized feed: go beyond search

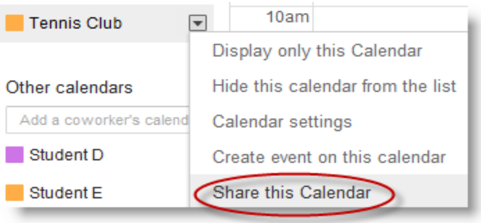


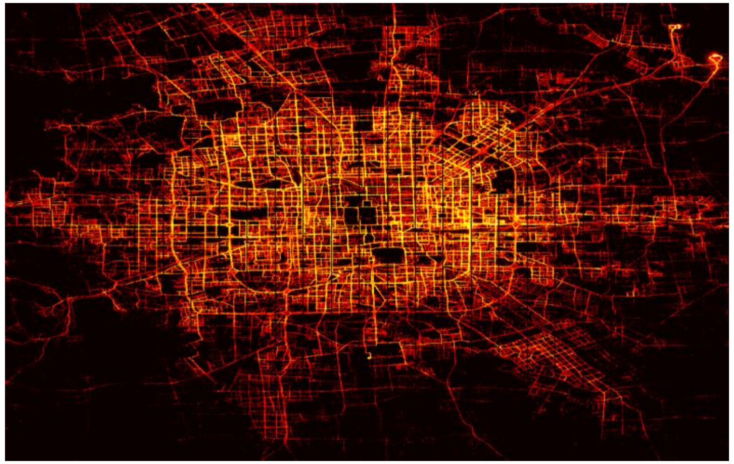


User modeling with behavioral data









Explicit/implicit user modeling

- gifts for classmates
- cool math games
- mickey mouse cartoon

- shower chair for elderly
- presbyopic glasses
- costco hearing aids

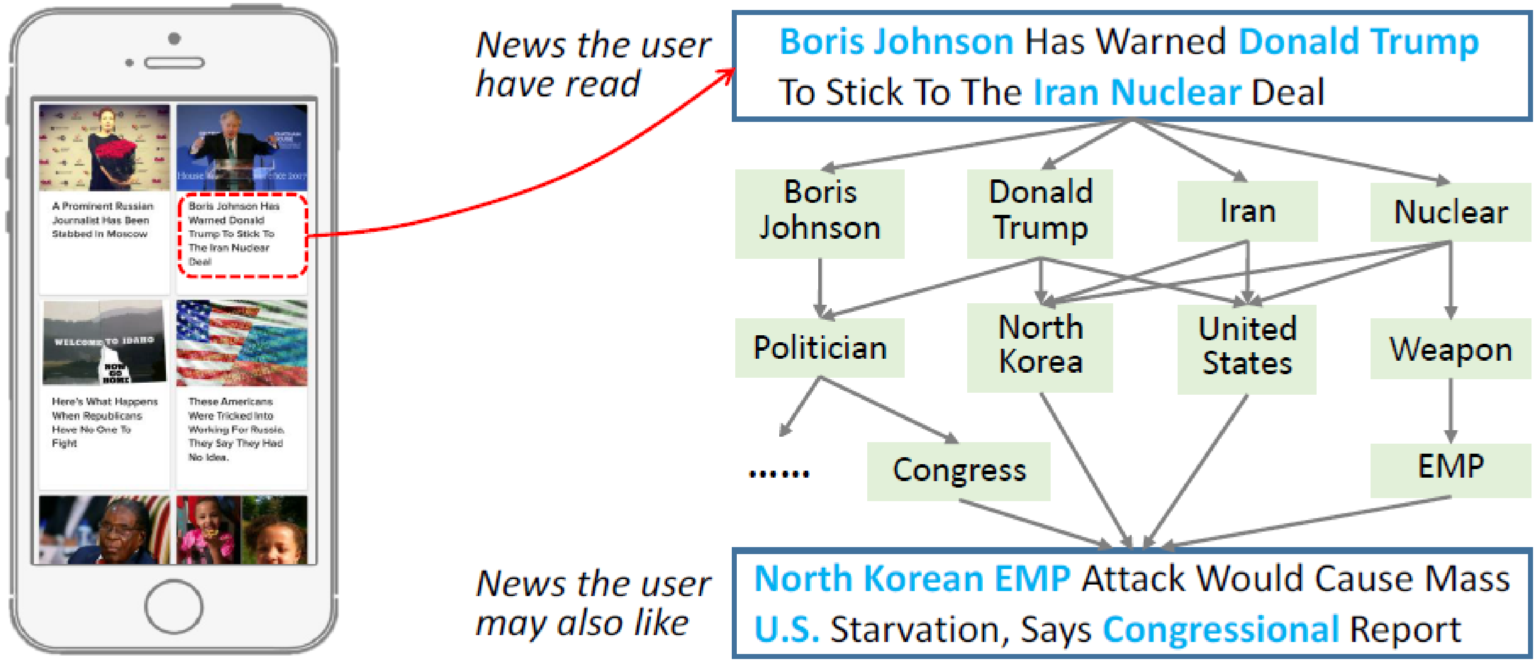
Age Inference

- groom to bride gifts
- tie clips
- philips shaver

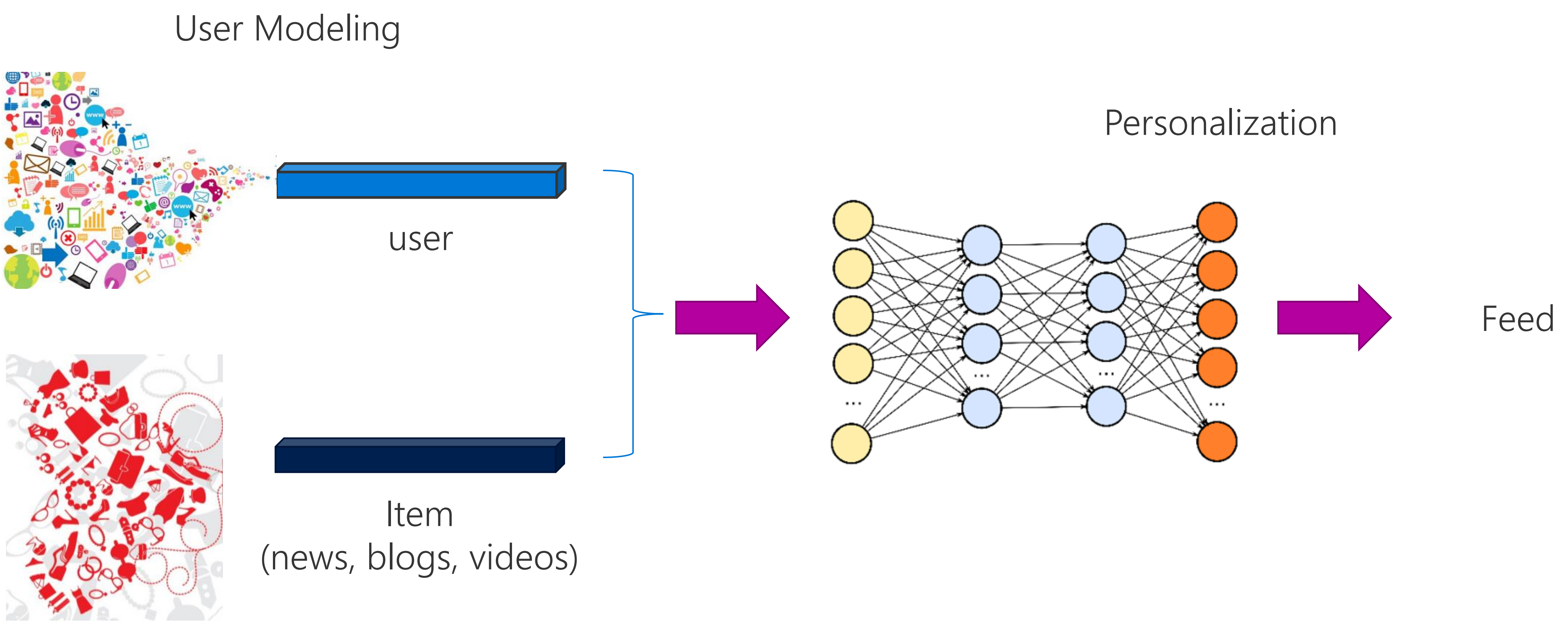
- lipstick color chart
- womans ana blouse
- Dior Makeup

Gender Inference

Candidate news detection



Personalized recommendation

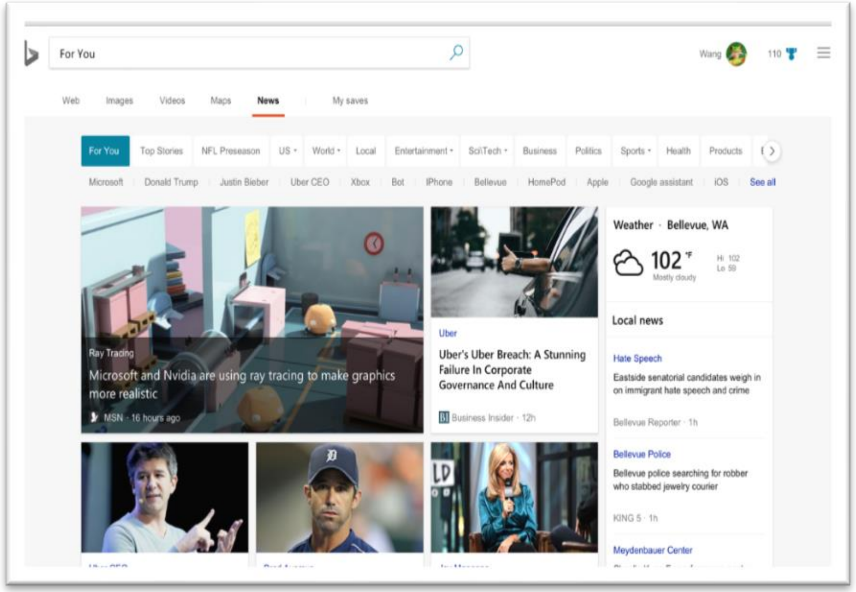


Text Understanding in User and Item Modeling

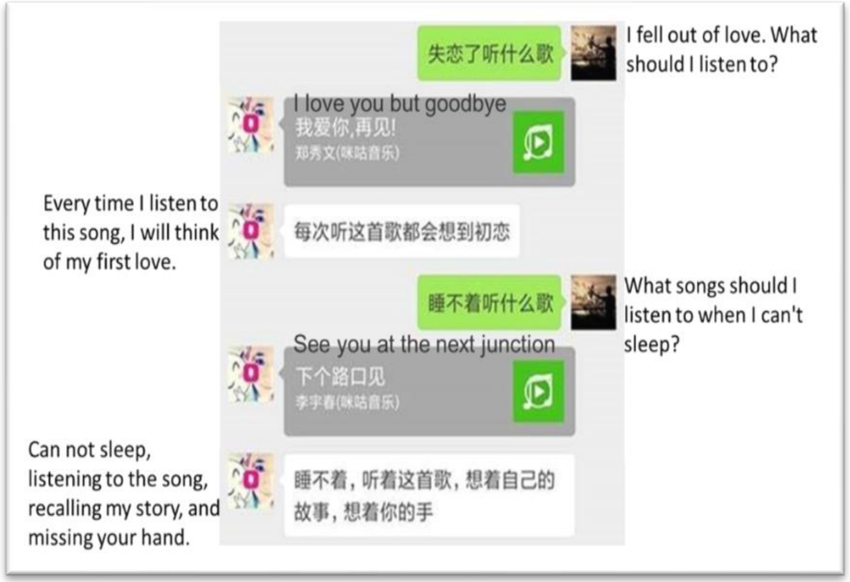
Knowledge Aware Recommendation

Text Generation

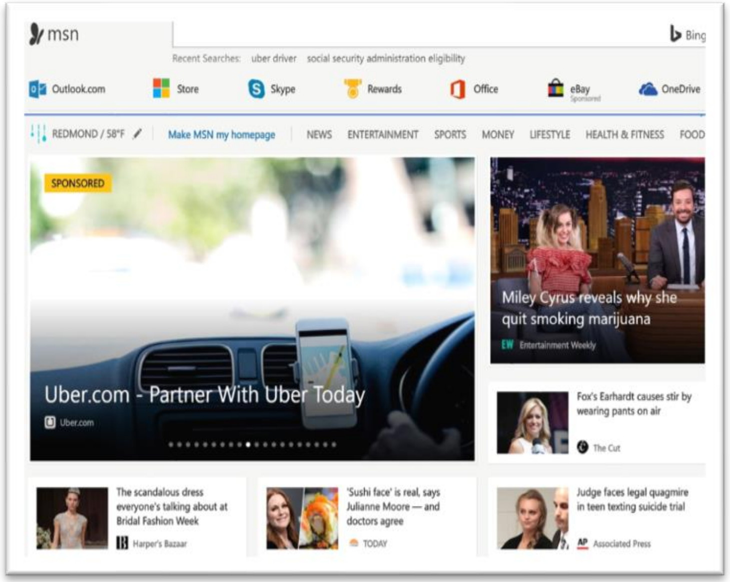
Applications



Personalized News Feed



XiaoIce Music Recommendation



Native Ads Ranking

- Xiting Wang, Yiru Chen, Jie Yang, etc. A Reinforcement Learning Framework for Explainable Recommendation, ICDM 2018
- Hongwei Wang, Fuzheng Zhang, Jialin Wang, etc. Ripple Network: Propagating User Preferences on the Knowledge Graph for Recommender Systems, CIKM 2018
- Jianxun Lian, Xiaohuan Zhou, etc., xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems, KDD 2018
- Zheng Liu, Xing Xie, Lei Chen, Context-aware Academic Collaborator Recommendation, KDD 2018
- Jianxun Lian, etc. Towards Better Representation Learning for Personalized News Recommendation: a Multi-Channel Deep Fusion Approach, IJCAI 2018
- Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, etc. DRN: A Deep Reinforcement Learning Framework for News Recommendation, WWW 2018
- Hongwei Wang, Fuzheng Zhang, Xing Xie, Minyi Guo, DKN: Deep Knowledge-Aware Network for News Recommendation, WWW 2018

A summary about personalized recommendation

1. Personalized recommendation has become the strategically important for search companies.
2. There are many interesting research topics in user modelling with various user behavior data, timely recommendation of user-like contents, providing personalized recommendation comments.

Summary

With new advancements of NLP, what new changes will be brought to search?

- QA
- Multi-lingual
- Multi-modal
- MRC
- Personalized recommendation

However, there are many challenges to each of these tasks and there are still long way to make them convincingly become the integral parts of search engine to reach good user experience. We should have enough patience in pushing forward these directions.

Future research topics

- Knowledge acquisition and representation
    - word/sentence embedding, commonsense knowledge, specific domain knowledge, knowledge graph
- New learning methods
    - Multi-task and transfer learning, reinforcement learning, semi and unsupervised learning for low-resource tasks, reasoning for MRC
- Context modeling
    - Multi-turn modeling, context-aware semantic parser, dialogue system
- New search modality
    - Conversational search, multi-modal search
- Search results generation and summarization
    - Auto-generation of a comprehensive report for certain type of queries
- Feeds
    - User modelling, content generation, recommendation, comments

Acknowledgment

A big thanks go to my colleagues and students in MSRA who are working on various NLP tasks mentioned in this talk, especially Nan Duan, Shujie Liu, Dongdong Zhang, Furu Wei and Xing Xie.