NLP 201: Natural Language Processing 1: NLP Tasks

Jeffrey Flanigan

December 3, 2020

University of California Santa Cruz
jmflanig@ucsc.edu

Plan for next two lectures

Overview of the rest of NLP

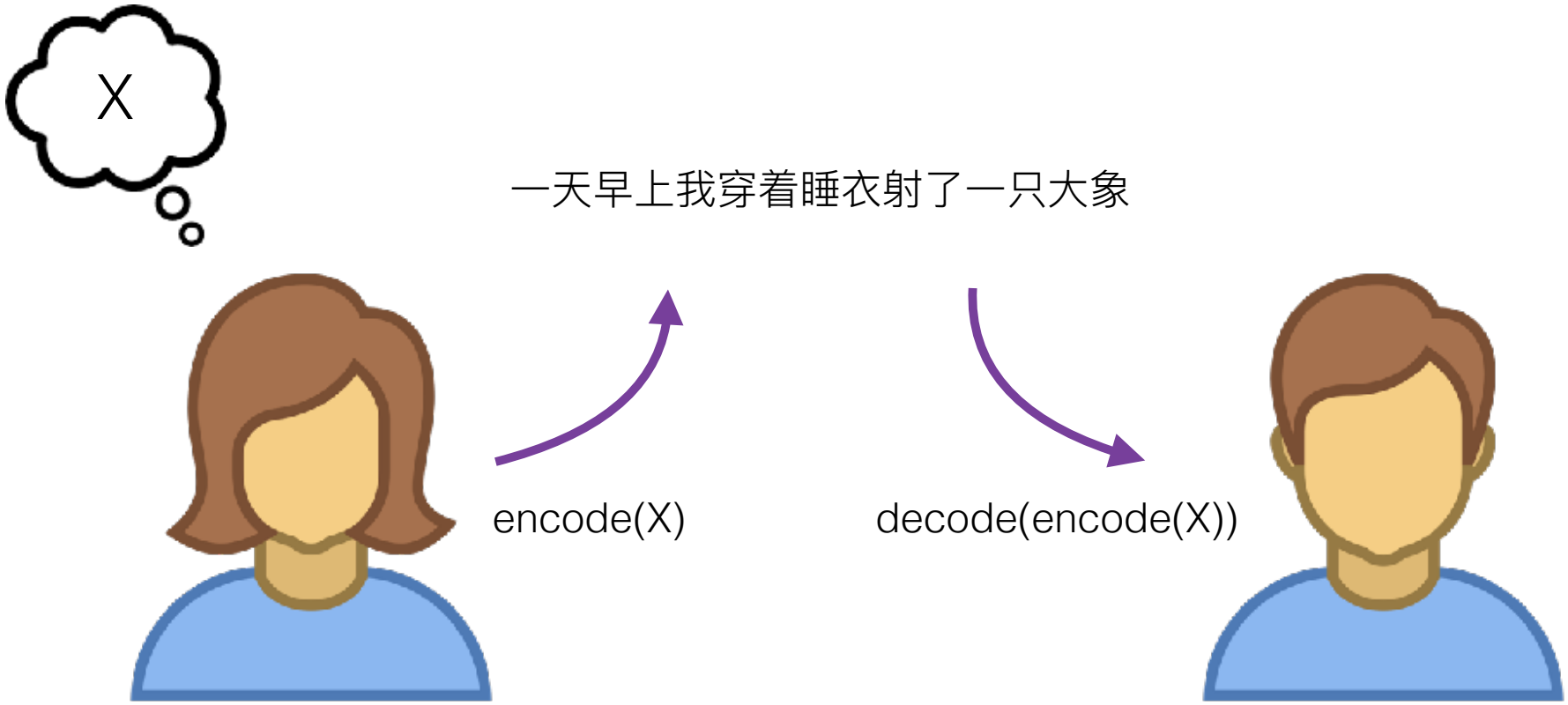
Won't go into much depth, but I want you to be aware of what's out there, and where to find out more.

You might feel a bit like Neo did when he learned Kung-Fu in the movie the Matrix

Plan For Today

- Machine Translation
- Summarization
- Question Answering

Machine Translation

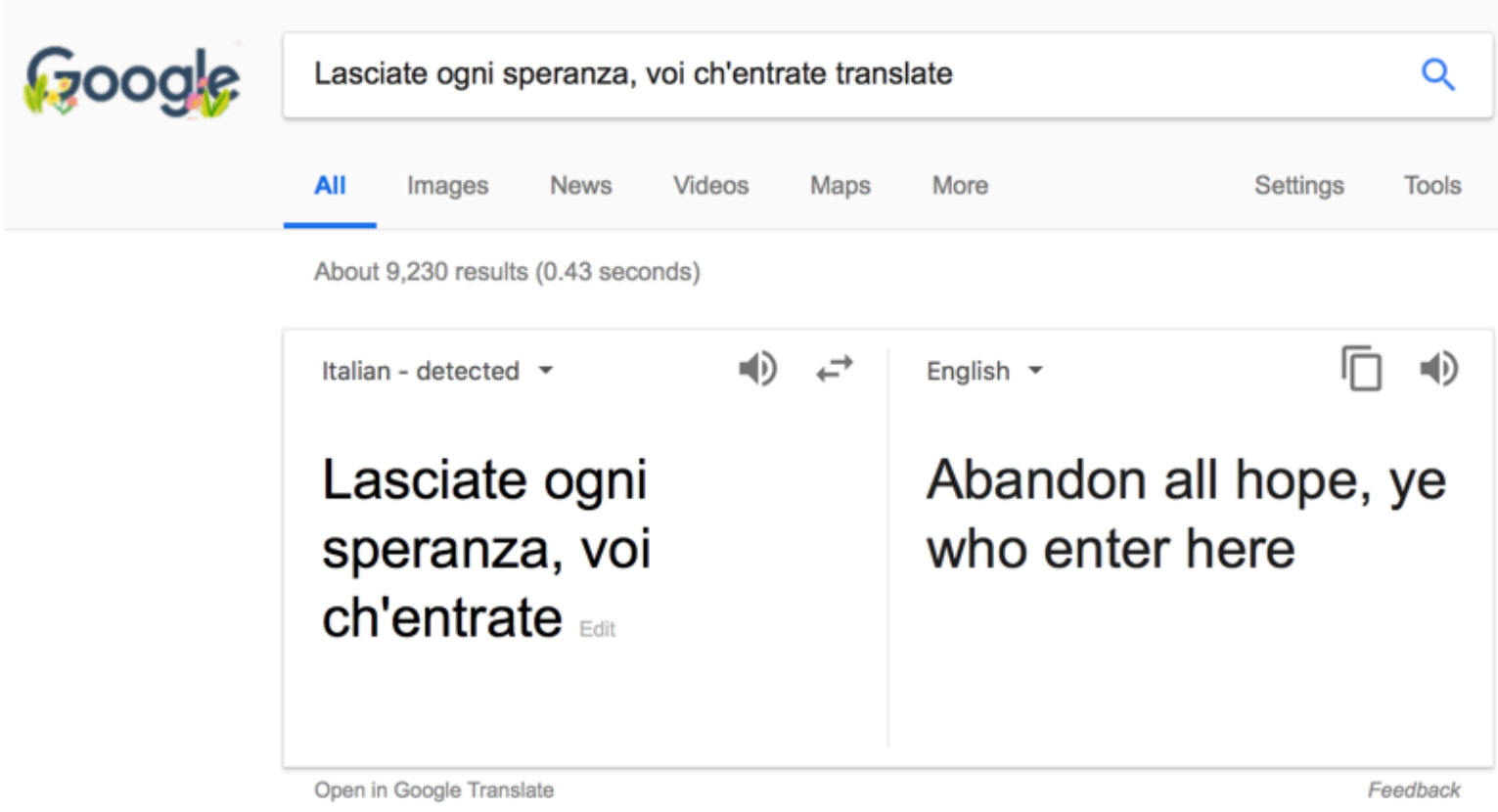


When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'

Weaver 1955

Machine Translation

Task	X	Y
Sentiment analysis	I hate this movie!	negative
POS tagging	I hate this movie!	PRP VB DT NN .
Parsing	I hate this movie!	[tree]
MT	Lasciate ogni speranza, voi ch'entrate	Abandon all hope, you who enter!
Conversational agent	How are you?	I’m great!


Google
Lasciate ogni speranza, voi ch'entrate translate
All
Images
News
Videos
Maps
More
Settings
Tools
About 9,230 results (0.43 seconds)
Italian - detected
English
Lasciate ogni speranza, voi ch'entrate
Edit
Abandon all hope, ye who enter here
Open in Google Translate
Feedback

^ There are many English translations of this famous line. Some examples include

- *All hope abandon, ye who enter here* – Henry Francis Cary (1805–1814)
- *All hope abandon, ye who enter in!* – Henry Wadsworth Longfellow (1882)
- *Leave every hope, ye who enter!* – Charles Eliot Norton (1891)
- *Leave all hope, ye that enter* – Carlyle Okey-Wicksteed (1932)
- *Lay down all hope, you that go in by me.* – Dorothy L. Sayers (1949)
- *Abandon all hope, ye who enter here* – John Ciardi (1954)
- *Abandon every hope, you who enter.* – Charles S. Singleton (1970)
- *No room for hope, when you enter this place* – C. H. Sisson (1980)
- *Abandon every hope, who enter here.* – Allen Mandelbaum (1982)
- *Abandon all hope, you who enter here.* – Robert Pinsky (1993); Robert Hollander (2000)
- *Abandon every hope, all you who enter* – Mark Musa (1995)
- *Abandon every hope, you who enter.* – Robert M. Durling (1996)

Verbatim, the line translates as "Leave (*lasciate*) every (*ogne*) hope (*speranza*), ye (*voi*) that (*ch'*) enter (*intrate*)."

https://en.wikipedia.org/wiki/Inferno_(Dante)#cite_note-18

Data

- Modern machine translation systems are learned from parallel texts: pairs of documents in two languages that have been aligned at the sentence level.

Parallel Corpora

CLASSIC SOUPS

			Sm.	Lg.
清燉雞湯	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯湯	58.	Chicken Rice Soup	1.85	3.25
雞麵湯	59.	Chicken Noodle Soup	1.85	3.25
廣東雲吞	60.	Cantonese Wonton Soup	1.50	2.75
番茄蛋湯	61.	Tomato Clear Egg Drop Soup	1.65	2.95
雲吞湯	62.	Regular Wonton Soup	1.10	2.10
酸辣湯	63.	Hot & Sour Soup	1.10	2.10
蛋花湯	64.	Egg Drop Soup	1.10	2.10
雲蛋湯	65.	Egg Drop Wonton Mix	1.10	2.10
豆腐菜湯	66.	Tofu Vegetable Soup	NA	3.50
雞玉米湯	67.	Chicken Corn Cream Soup	NA	3.50
蟹肉玉米湯	68.	Crab Meat Corn Cream Soup	NA	3.50
海鮮湯	69.	Seafood Soup	NA	3.50

Parallel Corpora (mining parallel data from microblogs Ling et al., 2013)

	ENGLISH	MANDARIN
1	i **wanna** live in a wes anderson world	我想要生活在Wes Anderson的世界里
2	Chicken soup, corn never truly digests. **TMI**.	鸡汤吧，玉米神马的从来没有真正消化过.恶心
3	To DanielVeuleman **yea iknw imma** work on that	对DanielVeuleman说，是的我知道，我正在向那方面努力
4	**msg 4** Warren G his **cday** is today 1 **yr** older.	发信息给Warren G，今天是他的生日，又老了一岁了。
5	Where **the hell** have you been all these years?	这些年你**TMD**到哪去了
	ENGLISH	ARABIC
6	It's **gonna** be a warm week!	الاسبو ع الياي حر
7	onni this gift only **4 u**	أوني هذة الهدية فقط لك
8	sunset in aqaba **:)**	غروب الشمس في العقبة:)
9	RT @MARYAMALKHAWAJA: there is a call for widespread protests in #bahrain **tmrw**	هناك نداء لمظاهرات في عدة مناطق غدا

Table 2: Examples of English-Mandarin and English-Arabic sentence pairs. The English-Mandarin sentences were extracted from Sina Weibo and the English-Arabic sentences were extracted from Twitter. Some messages have been shorted to fit into the table. Some interesting aspects of these sentence pairs are marked in bold.

Data

- Europarl (proceedings of European parliament, 50M words/language) http://www.statmt.org/europarl/
- UN Corpus (United Nations documents, six languages, 300M words/langauge) http://www.euromatrixplus.net/multi-un/
- Common crawl (Web documents, long tail of language pairs)

Evaluating MT

- Fluency: does it sound good in the target language?
- Fidelity/adequacy: does it capture the meaning of the original?
- BLEU score: geometric mean of 1-, 2-, 3-, and 4-gram *precision* vs. a reference, multiplied by brevity penalty (penalizes short translations)

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

- Typically $n = 4$, $w_i = 1/4$

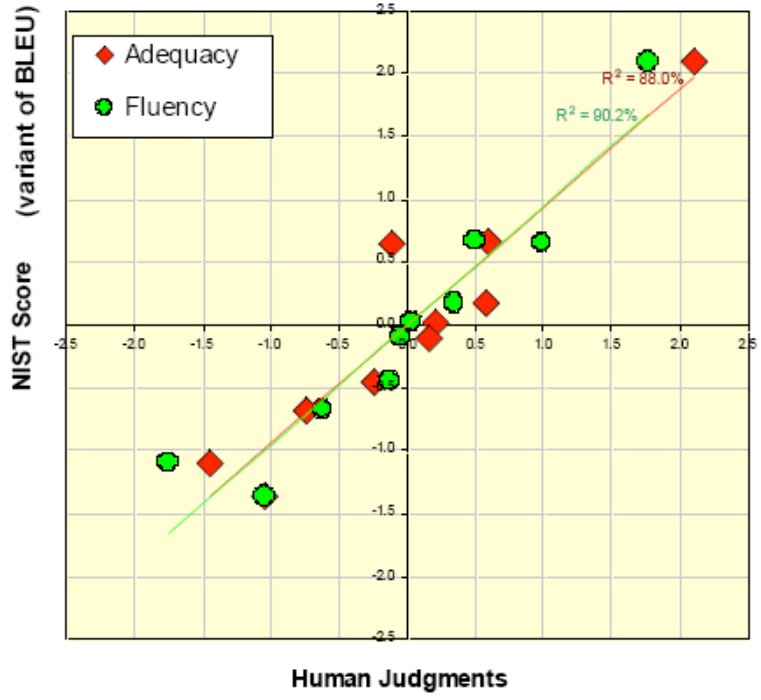
$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

r = length of reference
c = length of prediction

- Does this capture fluency and adequacy?

BLEU Score

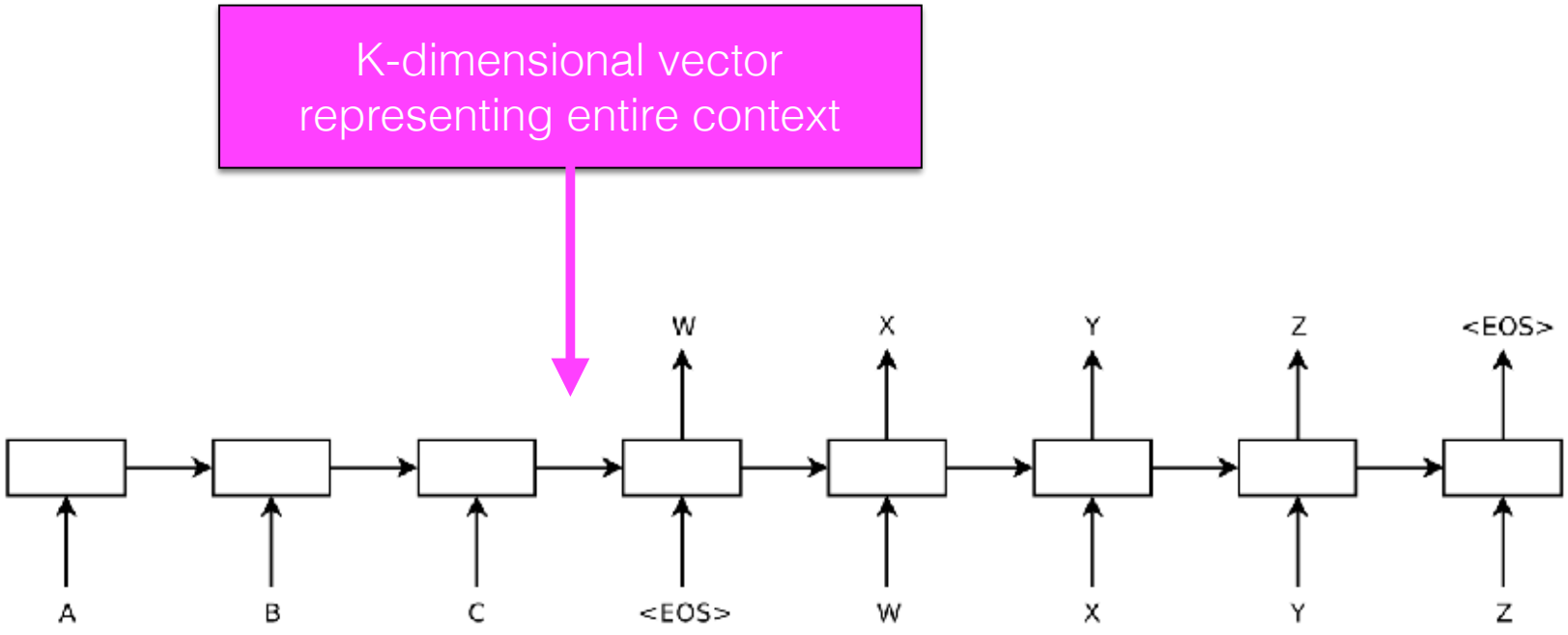
- At a *corpus* level, BLEU correlates pretty well with human judgments
- Better methods with human-in-the-loop
- If you're building real MT systems, you do user studies. In academia, you mostly use BLEU



slide from G. Doddington (NIST)

Method: Sequence to sequence (seq2seq) neural model trained on the parallel text

Encoder-decoder framework



Sutskever et al. (2015);

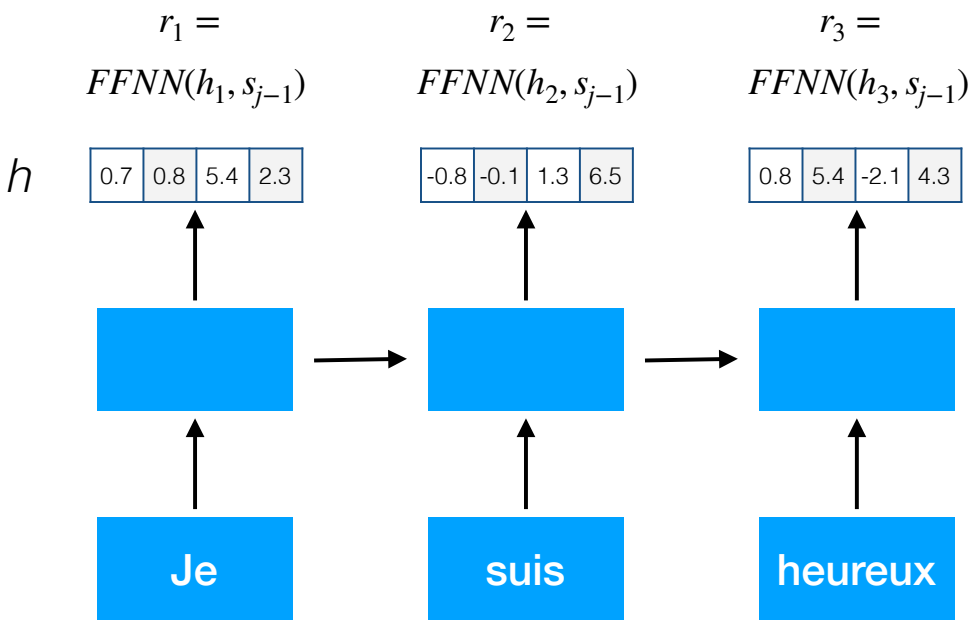
Encoder-decoder with attention

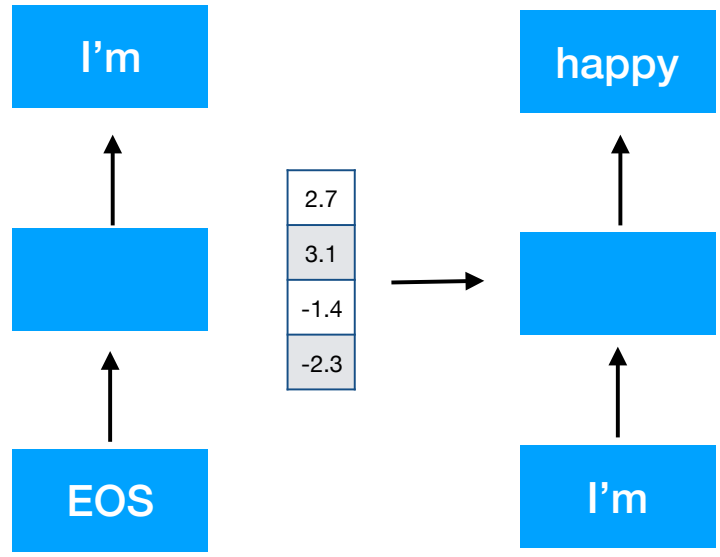
$$c = h_1 a_1 + h_2 a_2 + h_3 a_3$$

$$a = \text{softmax}(r)$$

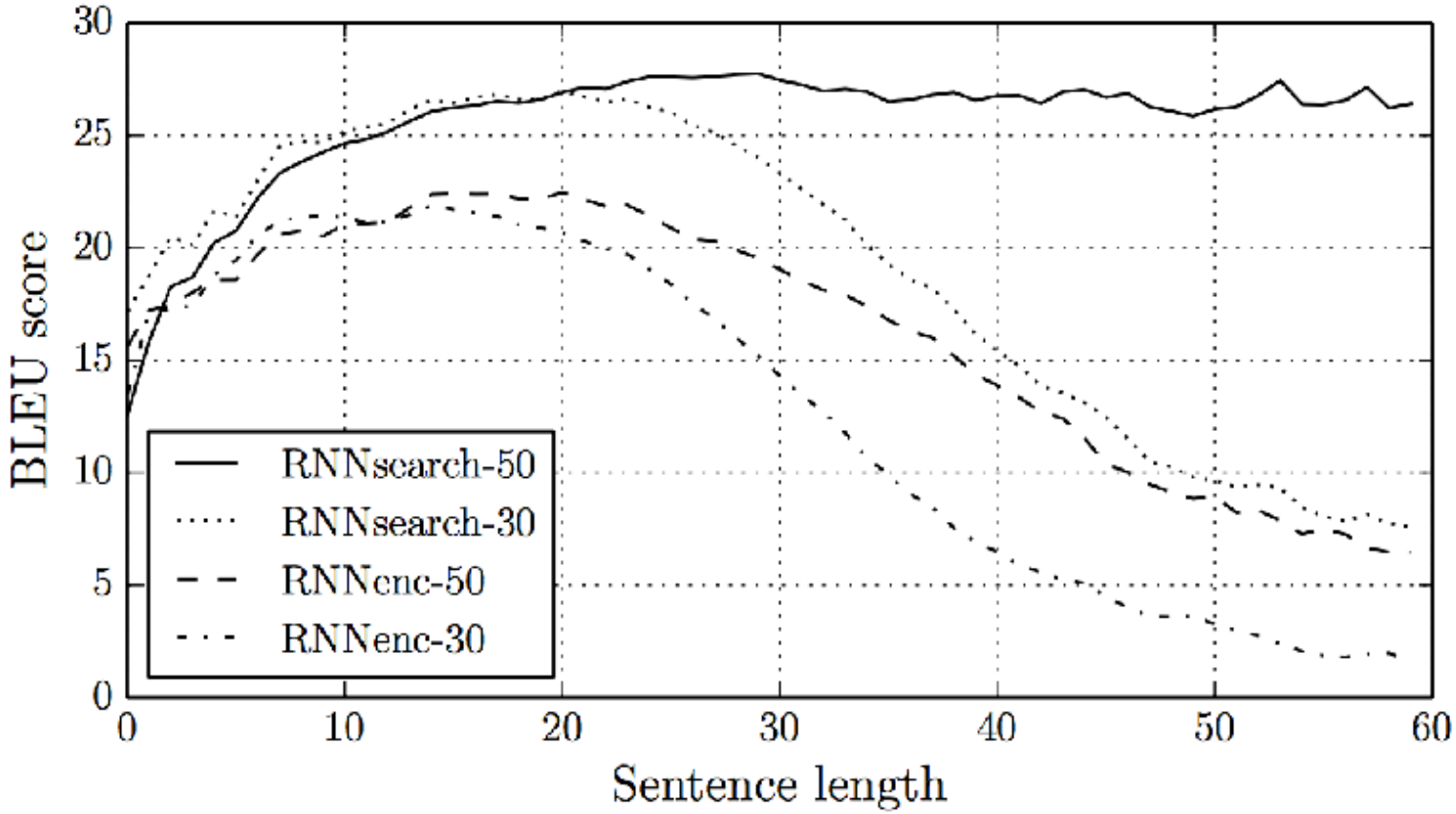
$$r = [r_1, \ldots, r_5]$$

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

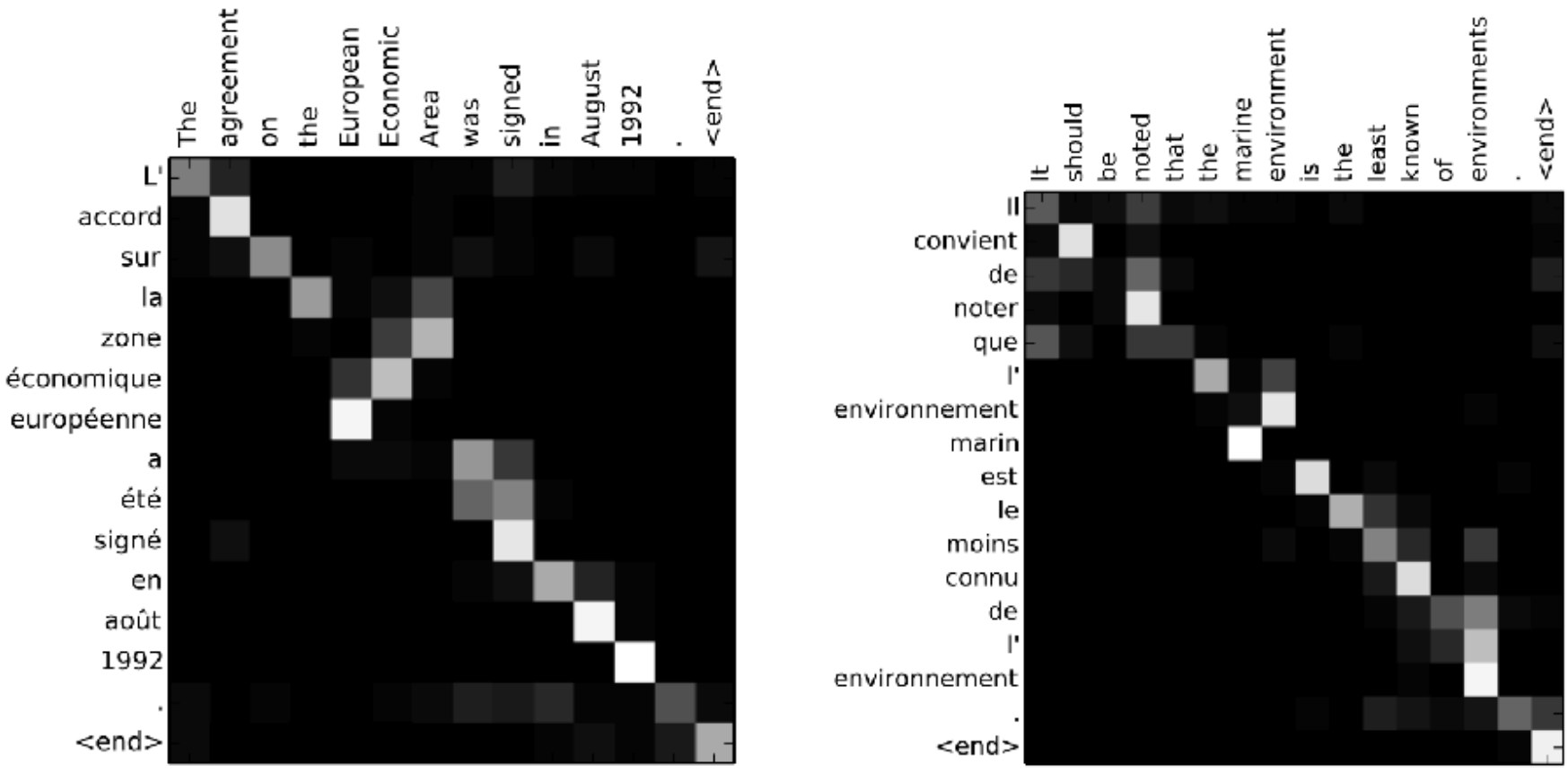




Better performance on long sentences



Bahdanau et al. (2016), “Neural Machine Translation by Jointly Learning to Align and Translate”



Bahdanau et al. (2016), “Neural Machine Translation by Jointly Learning to Align and Translate”

Handling Rare Words

- Words are a difficult unit to work with: copying can be cumbersome, word vocabularies get very large
- Character-level models don't work well
- Compromise solution: use thousands of "word pieces" (which may be full words but may also be parts of words)

Input: _the **_eco tax** _port i co _in _Po nt - de - Bu is …

Output: _le _port ique **_éco taxe** _de _Pont - de - Bui s

- Can achieve transliteration with this, subword structure makes some translations easier to achieve

Sennrich et al. (2016)

Byte Pair Encoding (BPE)

- Start with every individual byte (basically character) as its own symbol

```
for i in range(num_merges):
  pairs = get_stats(vocab)
  best = max(pairs, key=pairs.get)
  vocab = merge_vocab(best, vocab)
```

- Count bigram character cooccurrence
- Merge the most frequent pair of adjacent characters

- Do this either over your vocabulary (original version) or over a large corpus (more common version)
- Doing 8k merges => vocabulary of around 8000 word pieces. Includes many whole words
- Most SOTA NMT systems use this on both source + target

Sennrich et al. (2016)

Word Pieces

while voc size < target voc size:

Build a language model over your corpus

Merge pieces that lead to highest improvement in language model perplexity

- Issues: what LM to use? How to make this tractable?
- SentencePiece library from Google: unigram LM
- Result: way of segmenting input appropriate for translation

Schuster and Nakajima (2012), Wu et al. (2016), Kudo and Richardson (2018)

Backtranslation

- Classical MT methods used a bilingual corpus of sentences B = (S, T) and a large monolingual corpus T' to train a language model. Can neural MT do the same?
- Approach 1: force the system to generate T' as targets from null inputs

s_1, t_1

s_2, t_2

…

[null], t'_1

[null], t'_2

…

- Approach 2: generate synthetic sources with a T->S machine translation system (backtranslation)

s_1, t_1

s_2, t_2

…

$MT(t'_1)$, t'_1

$MT(t'_2)$, t'_2

…

Sennrich et al. (2015)

Backtranslation

name	training data	instances	BLEU tst2011	tst2012	tst2013	tst2014
baseline (Gülçehre et al., 2015)			18.4	18.8	19.9	18.7
deep fusion (Gülçehre et al., 2015)			20.2	20.2	21.3	**20.6**
baseline	parallel	7.2m	18.6	18.2	18.4	18.3
$\text{parallel}_{\text{synth}}$	parallel/$\text{parallel}_{\text{synth}}$	6m/6m	19.9	20.4	20.1	20.0
$\text{Gigaword}_{\text{mono}}$	parallel/$\text{Gigaword}_{\text{mono}}$	7.6m/7.6m	18.8	19.6	19.4	18.2
$\text{Gigaword}_{\text{synth}}$	parallel/$\text{Gigaword}_{\text{synth}}$	8.4m/8.4m	**21.2**	**21.1**	**21.8**	20.4

- Gigaword: large monolingual English corpus
- $\text{parallel}_{\text{synth}}$: backtranslate training data; makes additional noisy source sentences which could be useful

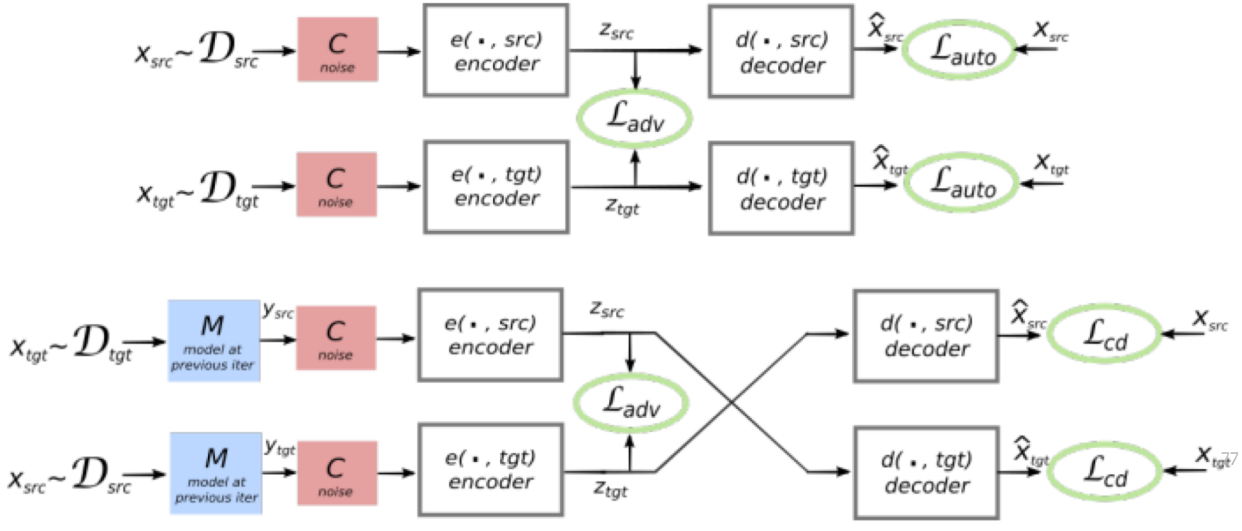
Sennrich et al. (2015)

Unsupervised NMT

UNSUPERVISED MACHINE TRANSLATION
USING MONOLINGUAL CORPORA ONLY

Guillaume Lample † ‡ , **Alexis Conneau** † , **Ludovic Denoyer** ‡ , **Marc'Aurelio Ranzato** †
† Facebook AI Research,
‡ Sorbonne Universités, UPMC Univ Paris 06, LIP6 UMR 7606, CNRS
{gl,aconneau,ranzato}@fb.com,ludovic.denoyer@lip6.fr

encode from another language, using as input the translation produced by the model at the previous iteration (light blue box)



MT Related Resources

EMNLP 2020
FIFTH CONFERENCE ON MACHINE TRANSLATION (WMT20)

November 11-12, 2020
Punta Cana, Dominican Republic

Home

[HOME]
TRANSLATION TASKS: [NEWS] [LIFELONG LEARNING] [ROBUSTNESS] [QUALITY ESTIMATION] [UNSUP AND VERY LOW RES]
OTHER TASKS: [AUTOMATIC POST-EDITING]

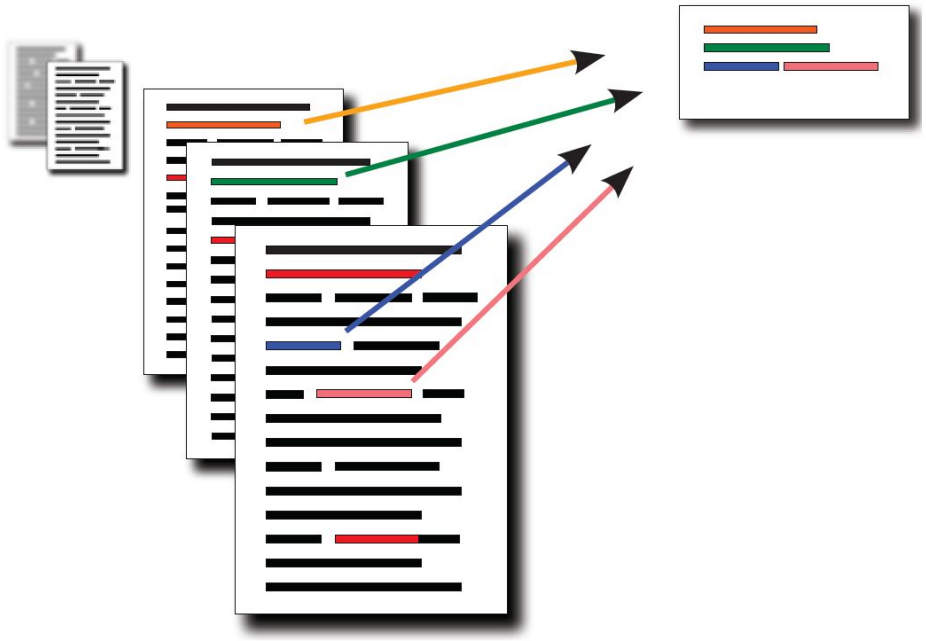
This conference builds on a series of annual workshops and conferences on statistical machine translation, going back to 2006:

- the NAACL-2006 Workshop on Statistical Machine Translation,
- the ACL-2007 Workshop on Statistical Machine Translation,
- the ACL-2008 Workshop on Statistical Machine Translation,
- the EACL-2009 Workshop on Statistical Machine Translation,
- the ACL-2010 Workshop on Statistical Machine Translation
- the EMNLP-2011 Workshop on Statistical Machine Translation,
- the NAACL-2012 Workshop on Statistical Machine Translation,
- the ACL-2013 Workshop on Statistical Machine Translation,
- the ACL-2014 Workshop on Statistical Machine Translation,
- the EMNLP-2015 Workshop on Statistical Machine Translation,
- the First Conference on Machine Translation (at ACL-2016),
- the Second Conference on Machine Translation (at EMNLP-2017),
- the Third Conference on Machine Translation (at EMNLP-2018).
- the Fourth Conference on Machine Translation (at ACL-2019).

Summarization

Text Summarization

- **Goal**: produce an abridged version of a text that contains information that is *important* or *relevant* to a user.



Categories

- **Input**
  - Single-Document Summarization (SDS)
  - Multiple-Document Summarization (MDS)
- **Output**
  - Extractive
  - Abstractive
  - Compressive
- **Focus**
  - Generic
  - Query-focused summarization
- **Machine learning methods:**
  - Supervised
  - Unsupervised

What to summarize? Single vs. multiple documents

- **Single-document summarization**
  - Given a single document, produce
    - abstract
    - outline
    - headline
- **Multiple-document summarization**
  - Given a group of documents, produce a gist of the content:
    - a series of news stories on the same event
    - a set of web pages about some topic or question

Single-document Summarization

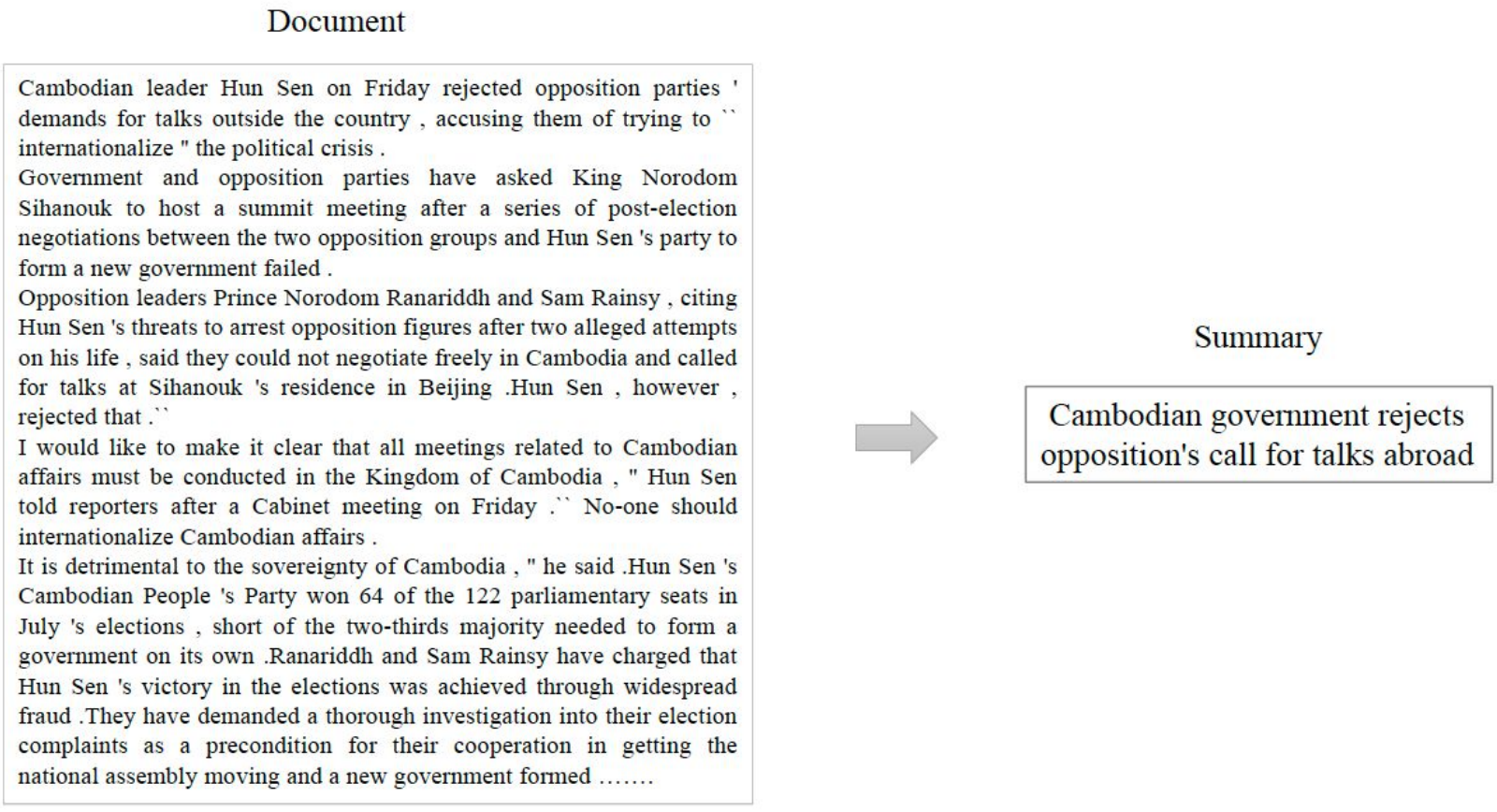


Figure 1: Single-document summarization.

Multiple-document Summarization

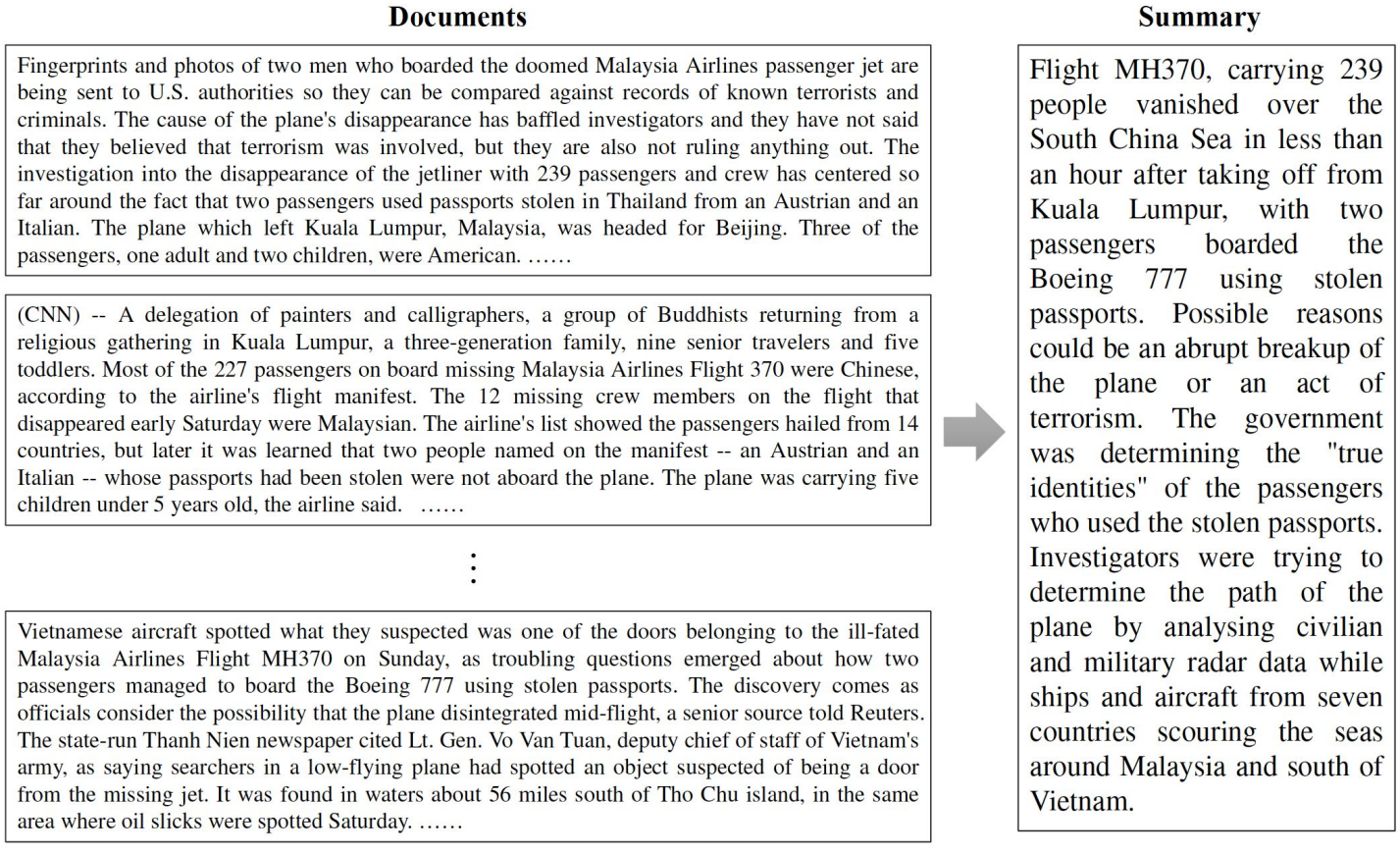


Figure 2: Multi-document summarization for the topic "Malaysia Airlines Disappearance".

Extractive summarization & Abstractive summarization

- Extractive summarization:
  - create the summary from phrases or sentences in the source document(s)
- Abstractive summarization:
  - express the ideas in the source documents using (at least in part) different words

Evaluating Summaries: ROUGE

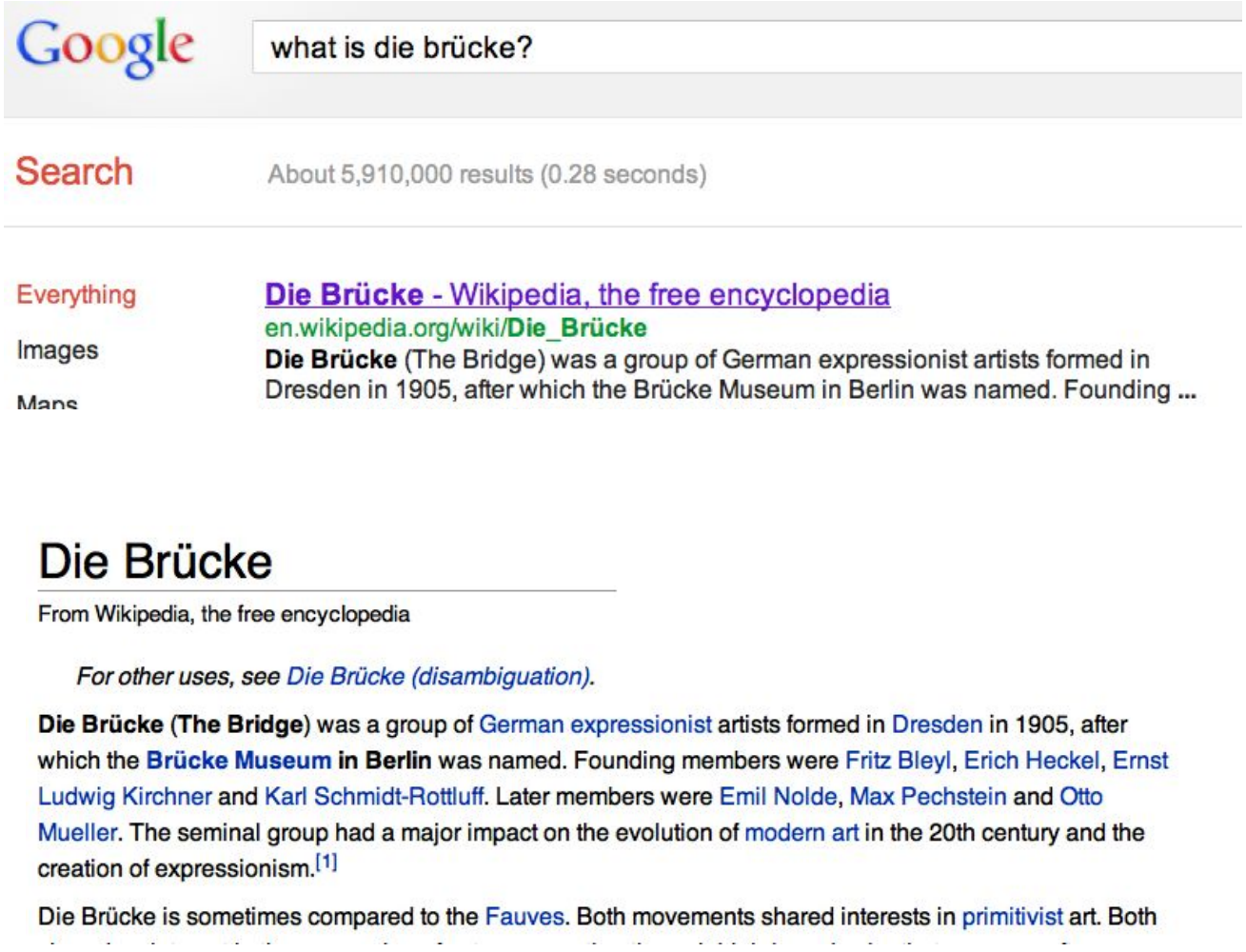
ROUGE (Recall Oriented Understudy for Gisting Evaluation) Lin and Hovy 2003

- Intrinsic metric for automatically evaluating summaries
  - Based on BLEU (a metric used for machine translation)
  - Not as good as human evaluation ("Did this answer the user's question?")
  - But much more convenient
- Given a document D, and an automatic summary X:
  1. Have N humans produce a set of reference summaries of D
  2. Run system, giving automatic summary X
  3. What percentage of the bigrams from the reference summaries appear in X?

$$ROUGE-2 = \frac{\sum_{s \in \{\text{RefSummaries}\}} \sum_{\text{bigrams } i \in S} \min(count(i,X), count(i,S))}{\sum_{s \in \{\text{RefSummaries}\}} \sum_{\text{bigrams } i \in S} count(i,S)}$$

Summarization: Methods

Simple baseline: take the first sentence



Extractive Summarization: MMR

- Given some articles and a length budget of *k* words, pick some sentences of total length <= *k* and make a summary
- Pick important yet diverse content: maximum marginal relevance (MMR)

While summary is < *k* words

Calculate $$MMR \stackrel{\text{def}}{=} Arg \max_{D_i \in R \setminus S} \left[\lambda (Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j)) \right]$$

"max over all sentences not yet in the summary"

"make this sentence similar to a query"

"make this sentence maximally different from all others added so far"

Add highest MMR sentence that doesn't overflow length

Carbonell and Goldstein (1998)

Extractive Summarization: Bigram Recall

- Count number of *documents* each bigram occurs in to measure importance

score(*massive earthquake*) = 3 score(*magnitude 7.3*) = 2

score(*six killed*) = 2 score(*Iraqi capital*) = 1

- Find summary that maximizes the score of bigrams it covers
- ILP formulation: *c* and *s* are indicator variables indexed over bigrams ("concepts") and sentences, respectively

$$\text{Maximize: } \sum_i w_i c_i \qquad s_j Occ_{ij} \leq c_i, \quad \forall i, j$$

$$\text{Subject to: } \sum_j l_j s_j \leq L \qquad \sum_j s_j Occ_{ij} \geq c_i \quad \forall i$$

"set c_i to 1 iff some sentence that contains it is included"

sum of included sentences' lengths can't exceed L

Gillick and Favre (2009)

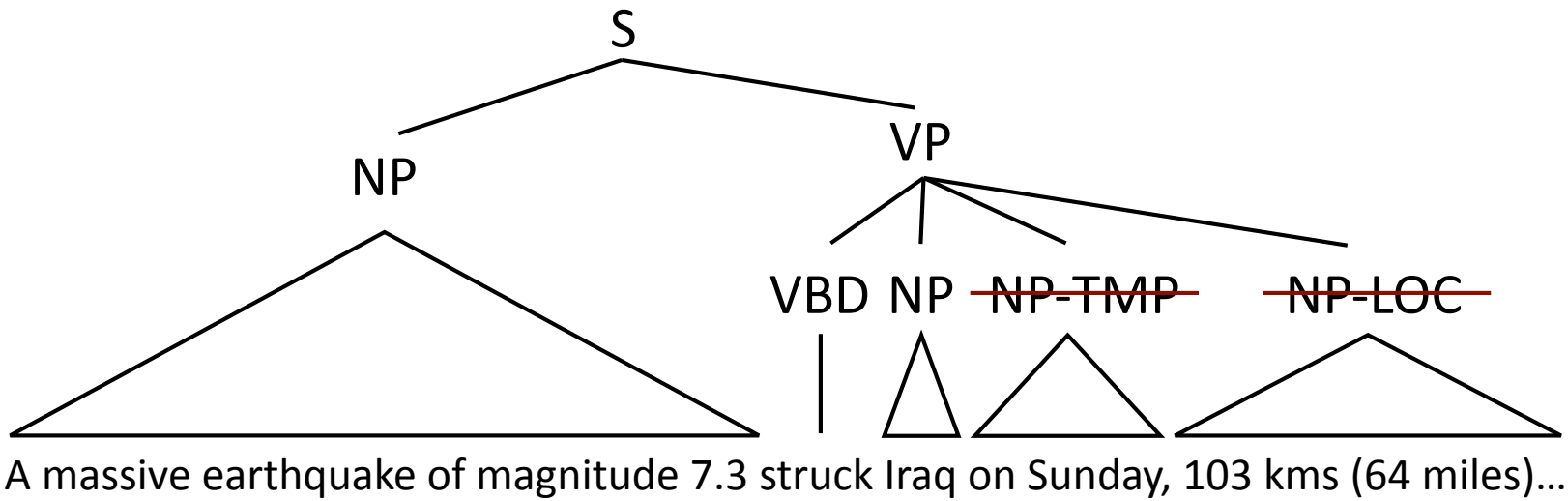
Compressive Summarization

Indian Express — A massive earthquake of magnitude 7.3 struck Iraq on Sunday, 103 kms (64 miles) southeast of the city of As-Sulaymaniyah, the US Geological Survey said, reports Reuters. US Geological Survey initially said the quake was of a magnitude 7.2, before revising it to 7.3.

- Sentence extraction isn't aggressive enough at removing irrelevant content
- Want to extract sentences and also delete content from them

Syntactic Cuts

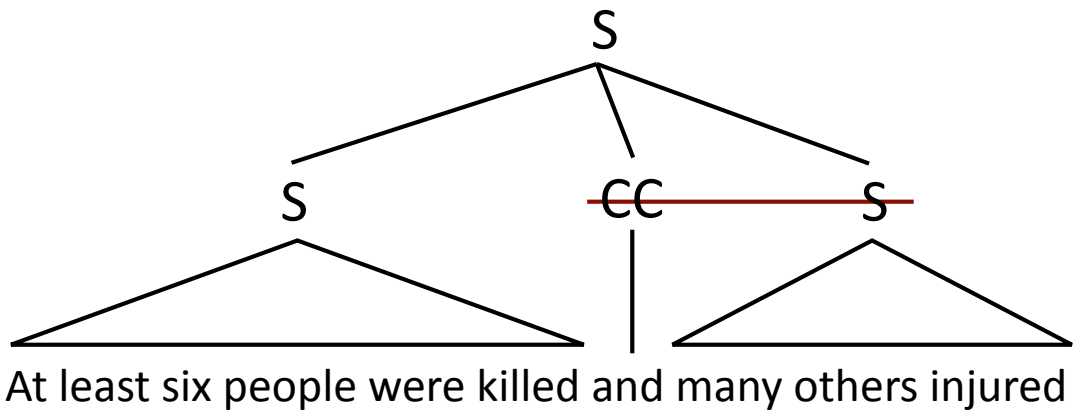
- Use syntactic rules to make certain deletions
- Delete adjuncts



Berg-Kirkpatrick et al. (2011)

Syntactic Cuts

- Use syntactic rules to make certain deletions
- Delete second parts of coordination structures



Berg-Kirkpatrick et al. (2011)

Compressive ILP

- Recall the Gillick+Favre ILP: Berg-Kirkpatrick et al. (2011)

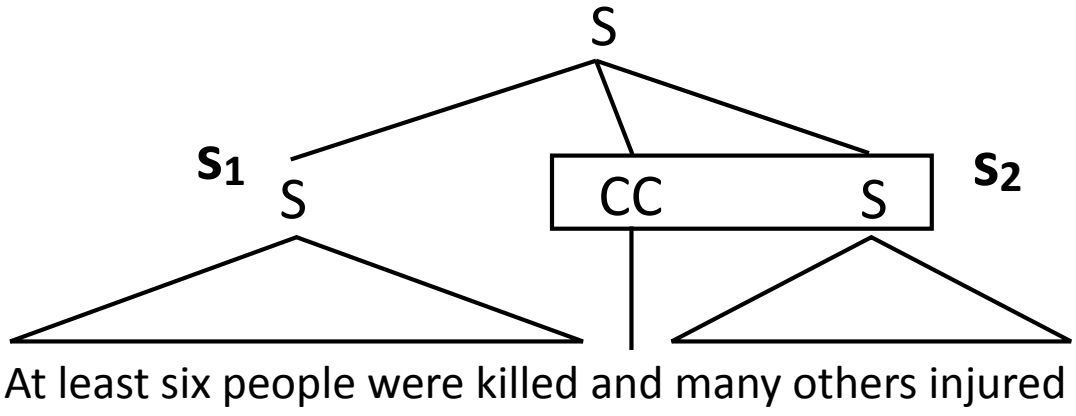
$$\text{Maximize: } \sum_i w_i c_i \qquad s_j Occ_{ij} \leq c_i, \quad \forall i, j$$

$$\text{Subject to: } \sum_j l_j s_j \leq L \qquad \sum_j s_j Occ_{ij} \geq c_i \quad \forall i$$

- Now s_j variables are nodes or sets of nodes in the parse tree

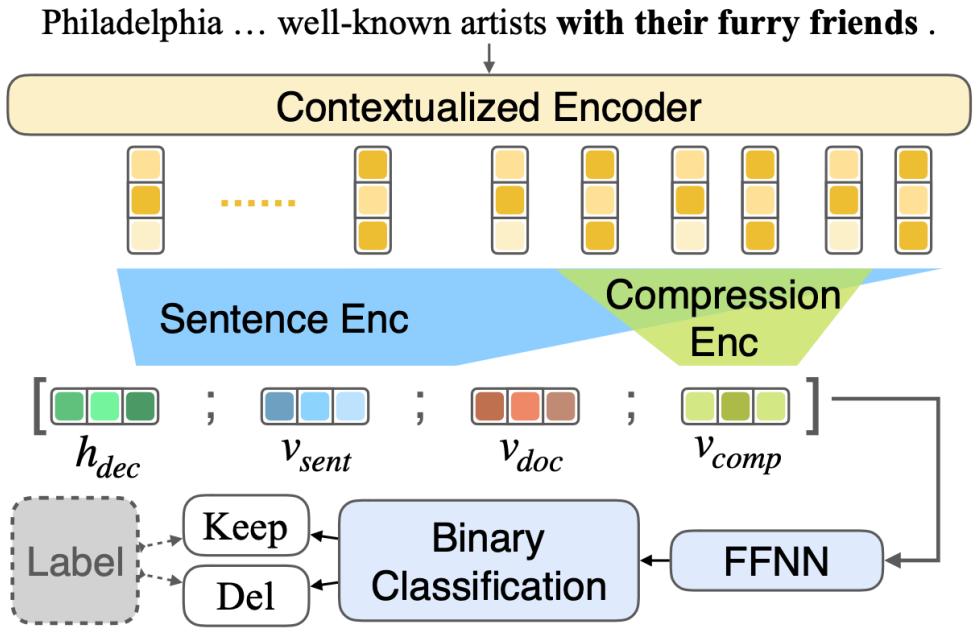
- New constraint: $s_2 \leq s_1$

  "s_1 is a prerequisite for s_2"



Neuralizing this Model

- Model is now a neural model that scores sentences and compression options
- Decoding is done by beam search (not ILP), length constraint is not enforced as strongly anymore



- Stronger results on NYT and on CNN/Daily Mail

Xu and Durrett (2019)

A neural attention model for abstractive sentence summarization

Rush et al., EMNLP 2015

- Inspired by attention-based seq2seq models (Bahdanau, 2014)

Input $(\mathbf{x}_1, \ldots, \mathbf{x}_{18})$. First sentence of article:
russian defense minister ivanov called sunday for the creation of a joint front for combating global terrorism

Output $(\mathbf{y}_1, \ldots, \mathbf{y}_8)$. Generated headline:
russia calls for joint front against ***terrorism*** $\Leftarrow$ $g(terrorism, \mathbf{x}, for, joint, front, against)$

Figure 2: Example input sentence and the generated summary. The score of generating $\mathbf{y}_{i+1}$ (`terrorism`) is based on the context $\mathbf{y}_c$ (`for ... against`) as well as the input $\mathbf{x}_1 \ldots \mathbf{x}_{18}$. Note that the summary generated is abstractive which makes it possible to *generalize* (`russian defense minister` to `russia`) and *paraphrase* (`for combating` to `against`), in addition to *compressing* (dropping `the creation of`), see Jing (2002) for a survey of these editing operations.

Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond

Nallapati et al., CoNLL 2016

- Implements many tricks (nmt, copy, coverage, hierarchical, external knowledge)

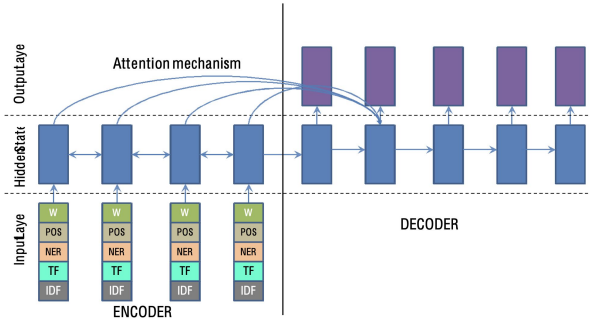


Figure 1: Feature-rich-encoder: We use one embedding

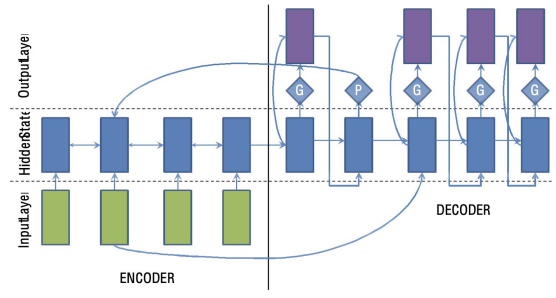


Figure 2: Switching generator/pointer model: When the

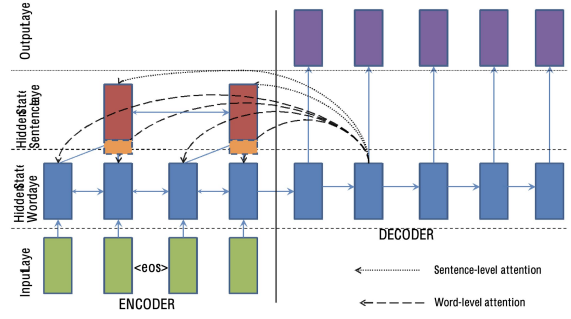


Figure 3: Hierarchical encoder with hierarchical attention:

End

We stopped here.

Question Answering

Google

us congress

All | News | Images | Videos | Maps | More | Settings | Tools

About 793,000,000 results (0.57 seconds)

Ad · www.ballotpedia.org/116th/Congress

116th U.S. Congress | Full List | ballotpedia.org

Find out who holds leadership positions in **Congress** and see a list of members. Find out who is part of the 116th **Congress** including membership and leadership. Candidate Facts. Encyclopedic information. Highlights: Professional Staff Available, Founded In 2007.

About Ballotpedia
Digital Encyclopedia
Covering U.S. Elections

Support Our Work
Grow The Encyclopedia
Of American Politics

www.congress.gov

Congress.gov | Library of Congress

U.S. Congress legislation, Congressional Record debates, Members of Congress , legislative process educational resources presented by the Library of ...

Show all members · Roll Call Votes · Browse · Contact Us

www.house.gov

House.gov: Home

Speaker Nancy Pelosi addresses the Members of the 116th **Congress**. Watch Live House Proceedings. The Office of the Clerk gives you a front-row seat. U.S. ...

Find Your Representative · Representatives · The House Explained · Committees

en.wikipedia.org › wiki › United_States_Congress

United States Congress - Wikipedia

The United States **Congress** is the bicameral legislature of the federal government of the United States, and consists of two chambers: the House of Representatives and the Senate. The **Congress** meets in the United States Capitol in Washington, D.C.

House of Representatives last election: Nove... | Houses: Senate; House of Representatives
House of Representatives next election: Nove... | Speaker of the House: Nancy Pelosi (D); since ...

United States Congress

congress.gov

The United States Congress is the bicameral legislature of the federal government of the United States, and consists of two chambers: the House of Representatives and the Senate. The Congress meets in the United States Capitol in Washington, D.C. Wikipedia

President of the Senate: Mike Pence (R); since January 20, 2017

President pro tempore of the Senate: Chuck Grassley (R); since January 3, 2019

Senate political groups: Republican (53); Democratic (45); Independent (2);

Senate last election: November 6, 2018

Senate next election: November 3, 2020

Number of members: Congress has 535 voting members: 435 representatives and 100 senators. wikipedia.org

Google

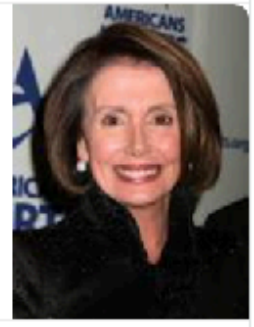
who is the speaker of the house

All | News | Images | Books | Shopping | More | Settings | Tools

About 1,350,000,000 results (0.69 seconds)

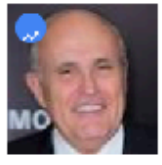
United States House of Representatives / **Speaker of the House**

Nancy Pelosi

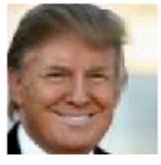


People also search for

View 15+ more



Rudy Giuliani
Trending



Donald Trump

Dianne Feinstein
Trending

Adam Schiff

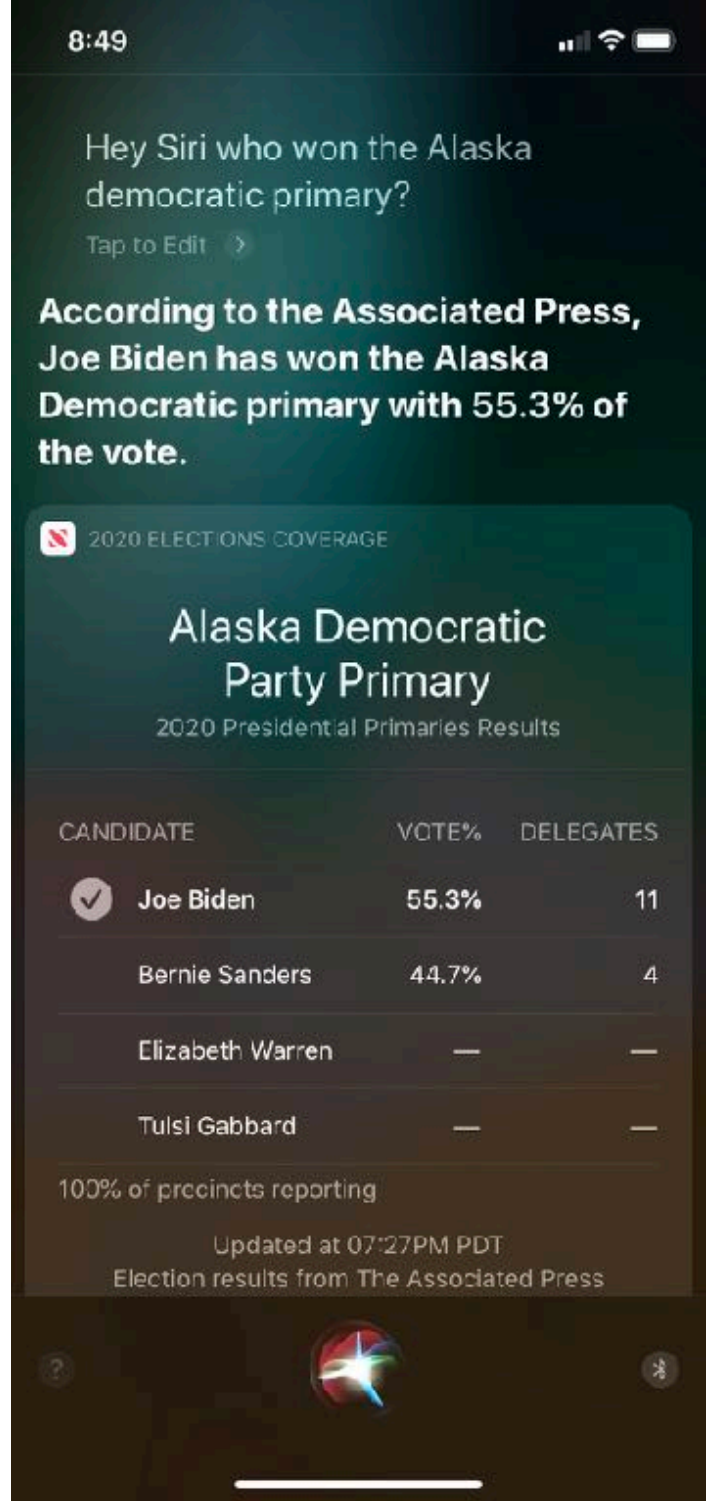
Devin Nunes
Trending

Chuck Schumer

Alexar Ocasi

Claim this knowledge panel

Feedback


8:49
Hey Siri who won the Alaska democratic primary?
Tap to Edit
According to the Associated Press, Joe Biden has won the Alaska Democratic primary with 55.3% of the vote.
2020 ELECTIONS COVERAGE
Alaska Democratic Party Primary
2020 Presidential Primaries Results
CANDIDATE VOTE% DELEGATES
Joe Biden 55.3% 11
Bernie Sanders 44.7% 4
Elizabeth Warren — —
Tulsi Gabbard — —
100% of precincts reporting
Updated at 07:27PM PDT
Election results from The Associated Press

Watson


THINK
$24,000
Who is Stoker?
((FOR ONE WELCOME OUR NEW COMPUTER OVERLORDS))
$1,000
$77,147
Who is Bram Stoker?
$ 17,973
$21,600
WHO IS BRAM STOKER?
$5600

Question Answering: Tasks

Textual Question Answering (Reading Comprehension)

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

(Rajpurkar et al, 2016): SQuAD: 100,000+ Questions for Machine Comprehension of Text

Textual Question Answering

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?

A) Fries

B) Pudding

C) James

D) Jane

2) What did James pull off of the shelves in the grocery store?

A) pudding

B) fries

C) food

D) splinters

(Richardson et al, 2013): MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text

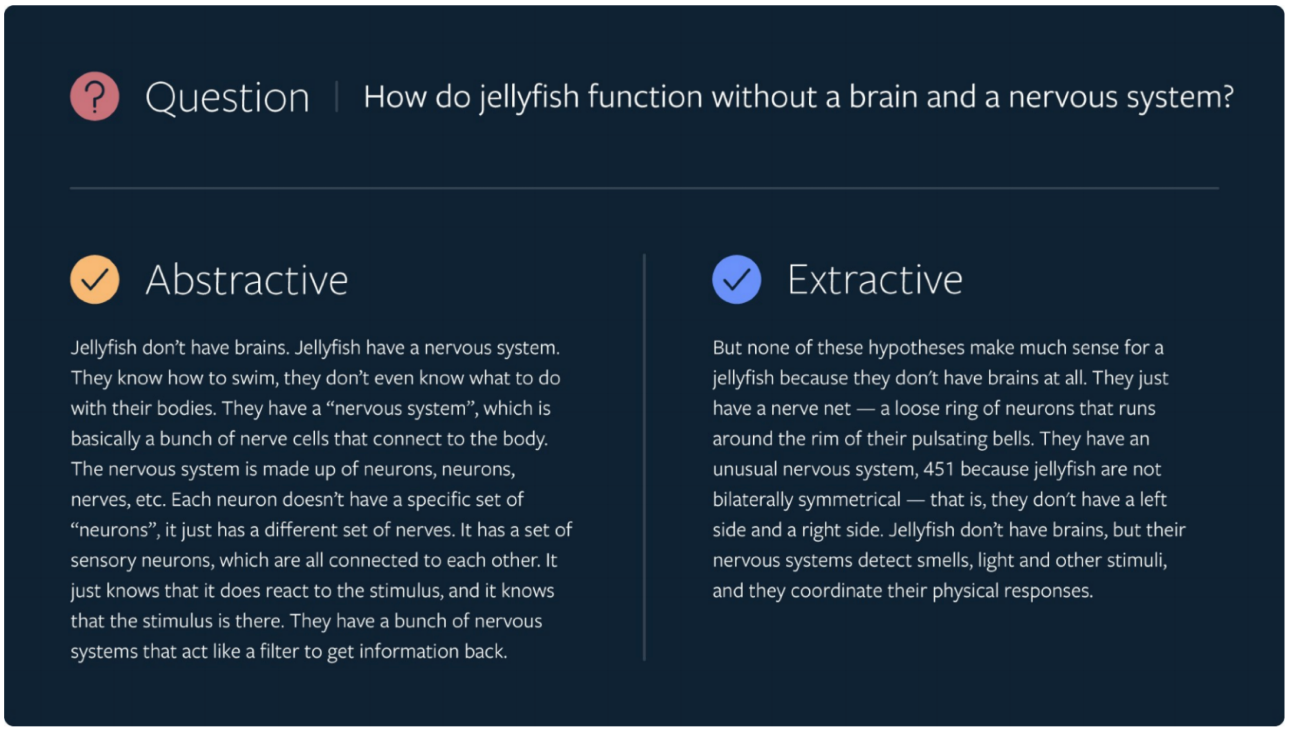
Conversational Question Answering

The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Q: What are the candidates **running** for?	A: Governor
Q: **Where?**	A: Virginia
Q: Who is the democratic candidate?	A: Terry McAuliffe
Q: Who is **his** opponent?	A: Ken Cuccinelli
Q: What party does **he** belong to?	A: Republican
Q: Which of **them** is winning?	

(Reddy et al, 2019): CoQA: A Conversational Question Answering Challenge

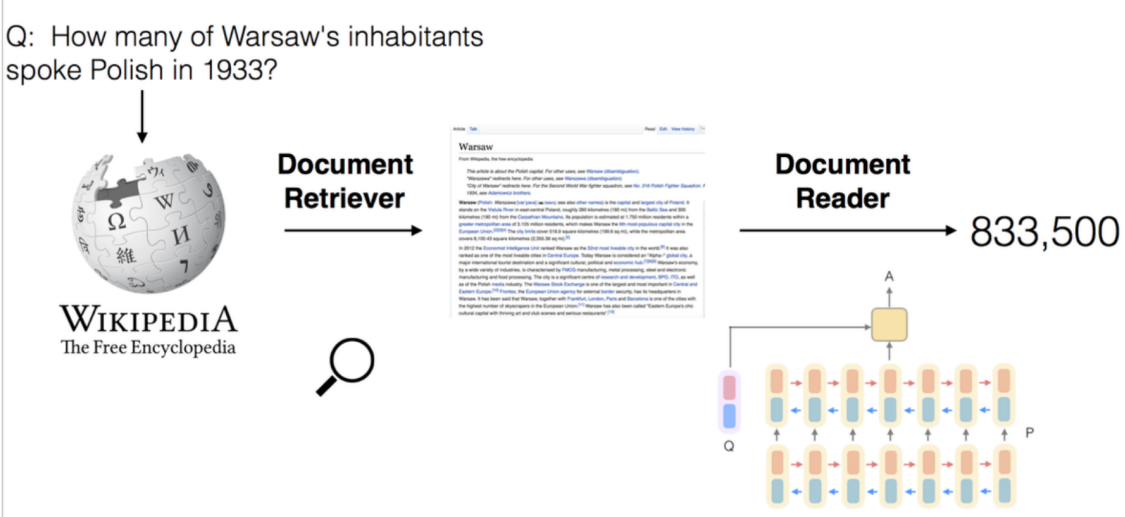
Long-form Question Answering



https://ai.facebook.com/blog/longform-qa/
(Fan et al, 2019): ELI5: Long Form Question Answering

Open-domain Question Answering

DrQA

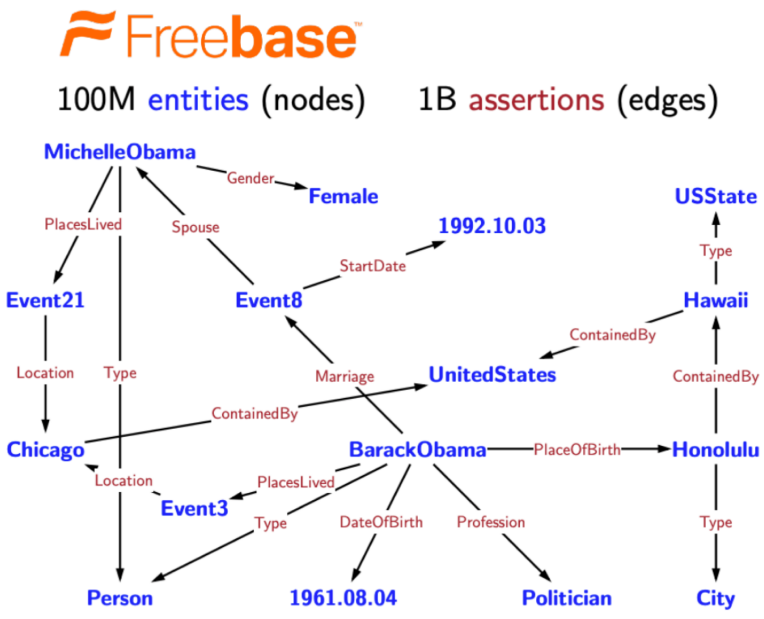



```
>>> process('What is the answer to life, the universe, and everything?')

Top Predictions:
+------+--------+-----------------------------------------------------+--------------+-----------+
| Rank | Answer |                         Doc                         | Answer Score | Doc Score |
+------+--------+-----------------------------------------------------+--------------+-----------+
|  1   |   42   | Phrases from The Hitchhiker's Guide to the Galaxy |    47242     |   141.26  |
+------+--------+-----------------------------------------------------+--------------+-----------+
```

(Chen et al, 2017): Reading Wikipedia to Answer Open-Domain Questions

Knowledge Base Question Answering



Which states' capitals are also their largest cities by area?

semantic parsing

$\mu x.\text{Type.USState} \sqcap \text{Capital.argmax}(\text{Type.City} \sqcap \text{ContainedBy}.x, \text{Area})$

execute

Arizona, Hawaii, Idaho, Indiana, Iowa, Oklahoma, Utah

(Berant et al, 2013): Semantic Parsing on Freebase from Question-Answer Pairs

Table-based Question Answering

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...	...	...	...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

x = Greece held its last Summer Olympics in which year?

y = 2004

(Pasupat and Liang, 2015): Compositional Semantic Parsing on Semi-Structured Tables.

Visual Question Answering

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

(Antol et al, 2015): Visual Question Answering

Question Answering Datasets

- **Reading Comprehension**

  CNN/Daily Mail, CoQA, HotpotQA, QuAC, RACE, SQuAD, SWAG, Receipt QA, NarrativeQA, DROP, Story Cloze Test ……

- **Open-domain question answering**

  DuReader, Quasar, SearchQA, …

- **Knowledge base question answering**

Check out more datasets: http://nlpprogress.com/english/question_answering.html

Reading Comprehension: Methods

SQuAD

- Single-document, single-sentence question-answering task where the answer is always a substring of the passage
- Predict start and end indices of the answer in the passage

One of the most famous people born in Warsaw was Maria Skłodowska-Curie, who achieved international recognition for her research on radioactivity and was the first female recipient of the Nobel Prize. Famous musicians include Władysław Szpilman and Frédéric Chopin. Though Chopin was born in the village of Żelazowa Wola, about 60 km (37 mi) from Warsaw, he moved to the city with his family when he was seven months old. Casimir Pulaski, a Polish general and hero of the American Revolutionary War, was born here in 1745.

What was Maria Curie the first female recipient of?
Ground Truth Answers: Nobel Prize | Nobel Prize | Nobel Prize

What year was Casimir Pulaski born in Warsaw?
Ground Truth Answers: 1745 | 1745 | 1745

Who was one of the most famous people born in Warsaw?
Ground Truth Answers: Maria Skłodowska-Curie | Maria Skłodowska-Curie | Maria Skłodowska-Curie

Rajpurkar et al. (2016)

Feature Based Methods

- Generate a list of candidate answers $(a_1, a_2, \ldots, a_M)$
  - Considered only the constituents in parse trees
- Define a feature vector $\phi(p, q, a_i) \in R^d$:
  - Word/bigram features
  - Parse tree matches
  - Dependency labels, length, part-of-speech tags
- Apply a multi-class logistic regression model

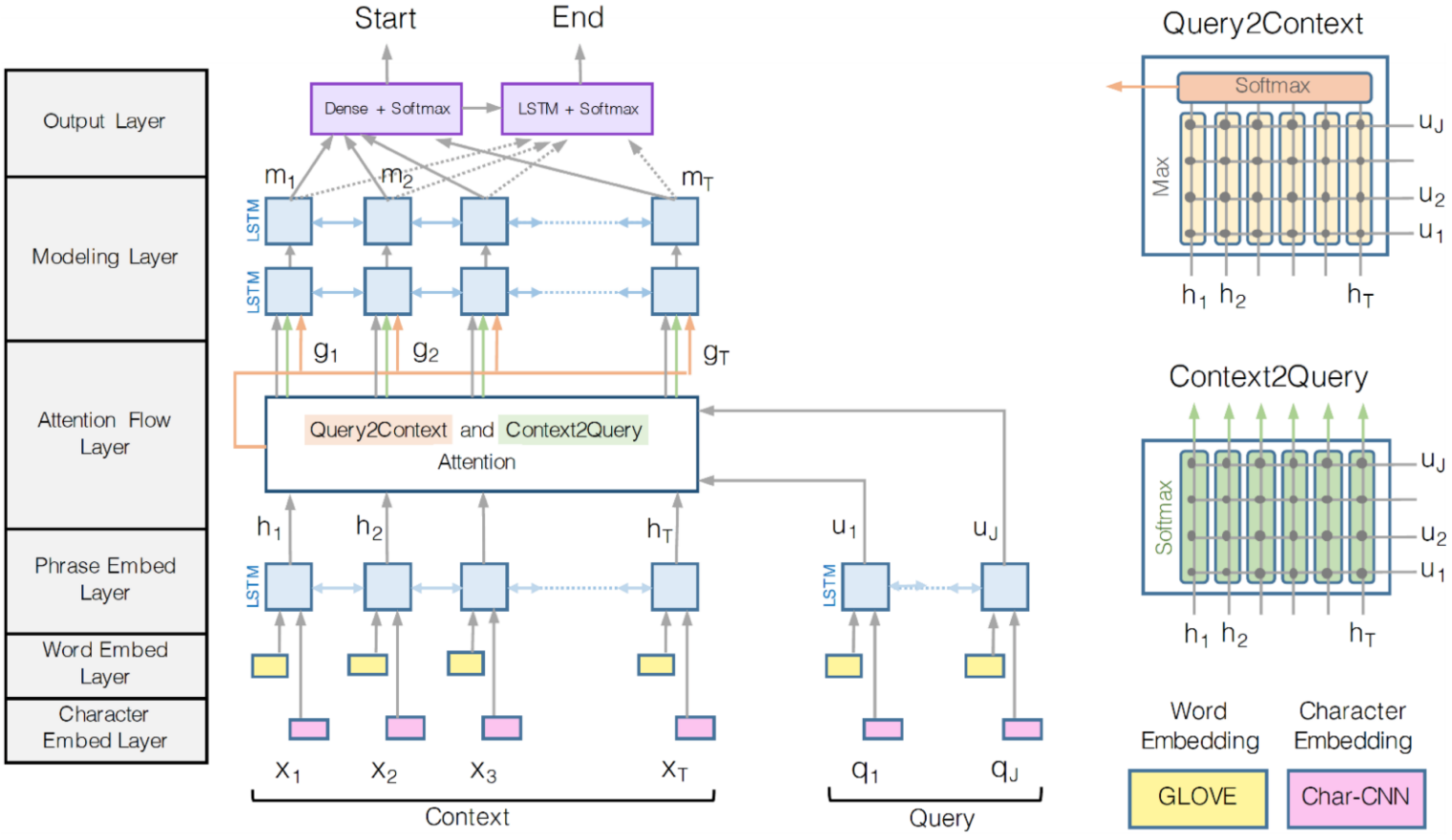
(Rajpurkar et al, 2016): SQuAD: 100,000+ Questions for Machine Comprehension of Text

BiLSTM-based Models

- Encode the question using word/char embeddings; pass on an biLSTM encoder
- Encode the passage similarly
- Passage-to-question and question-to-passage attention
- Modeling layer: another BiLSTM layer
- Output layer: two classifiers for predicting start and end points
- The entire model can be trained in an end-to-end way

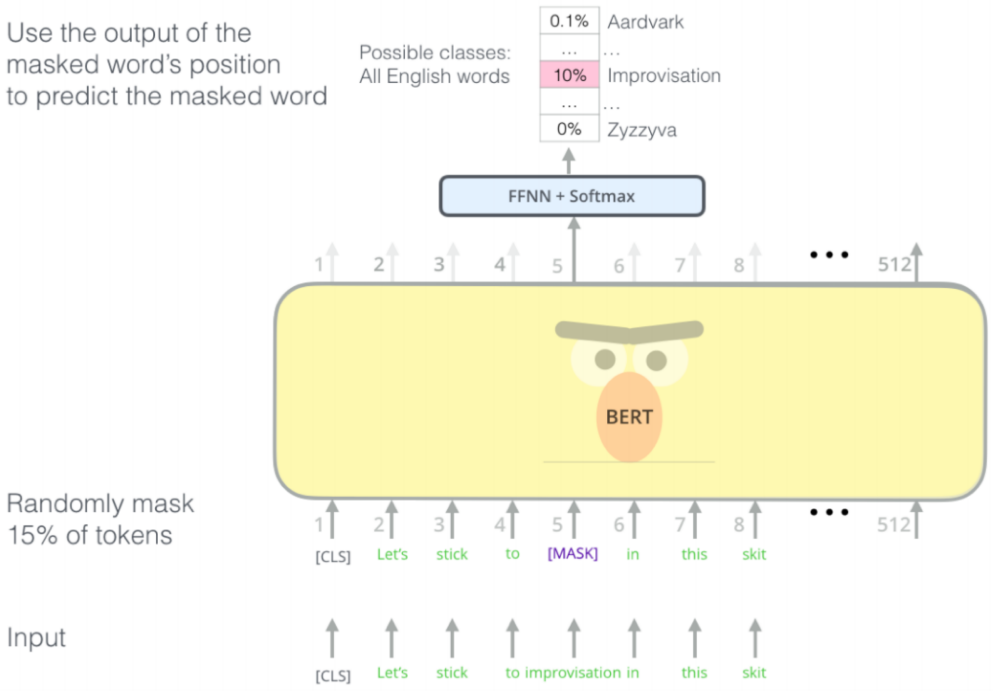
(Seo et al, 2017): Bidirectional Attention Flow for Machine Comprehension

BiDAF



(Seo et al, 2017): Bidirectional Attention Flow for Machine Comprehension

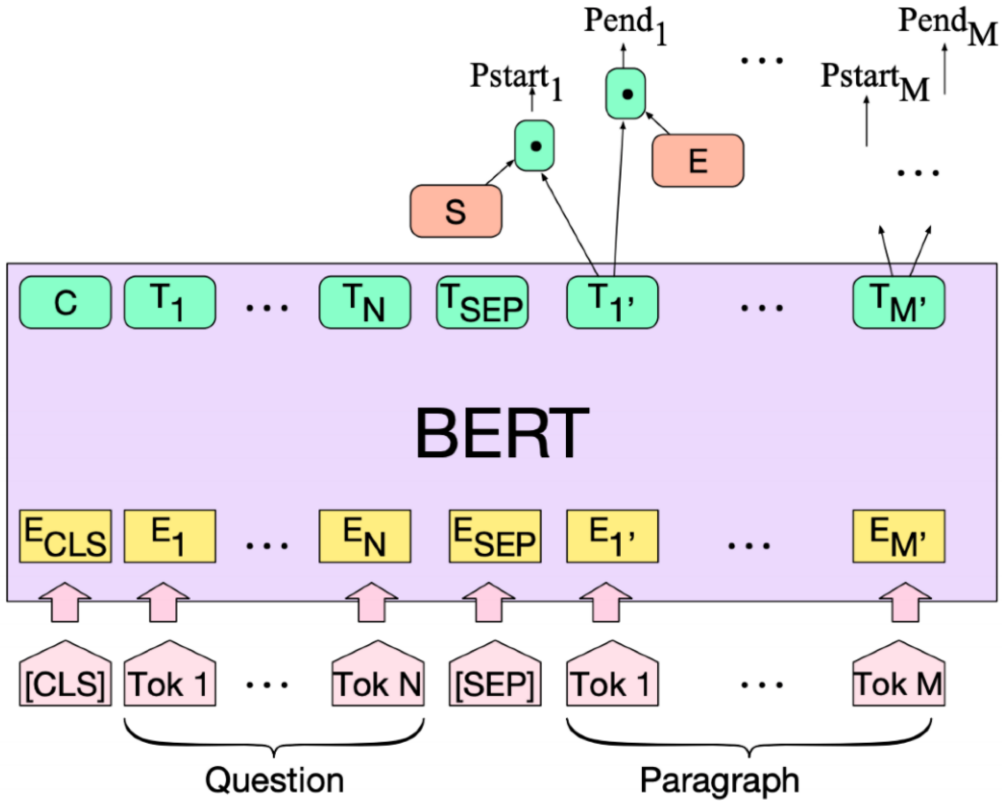
BERT-based Models



Pre-training

$$Pstart_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}$$

$$Pend_i = \frac{e^{E \cdot T_i}}{\sum_j e^{E \cdot T_j}}$$



- Concatenate question and passage as one single sequence separated with a [SEP] token, then pass it to the BERT encoder
- Train two classifiers on top of passage tokens

SQuAD SOTA: Today

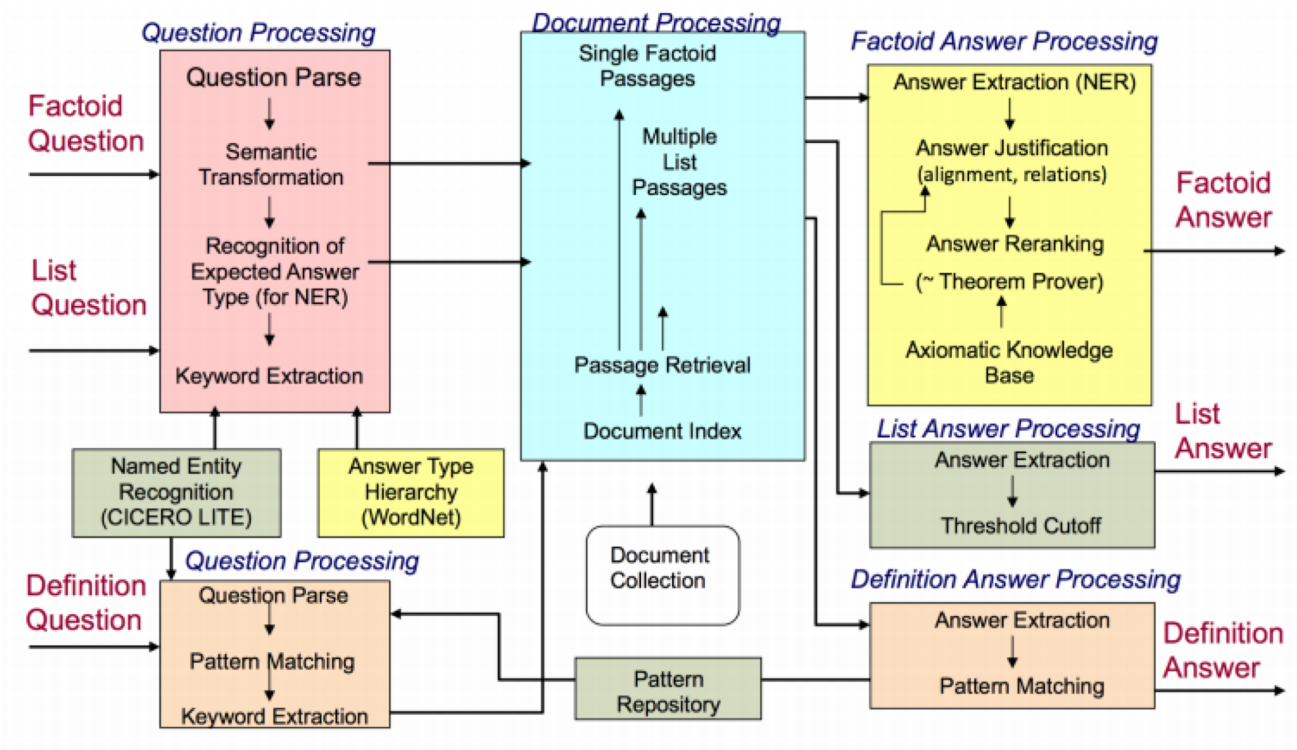
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	**89.731**	**92.215**
2 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
2 Sep 16, 2019	ALBERT (single model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	88.107	90.902
2 Jul 26, 2019	UPM (ensemble) *Anonymous*	88.231	90.713
3 Aug 04, 2019	XLNet + SG-Net Verifier (ensemble) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	88.174	90.702
4 Aug 04, 2019	XLNet + SG-Net Verifier++ (single model) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	87.238	90.071

- Performance is very saturated
- Harder QA settings are needed!

Open-Domain QA: Methods

IBM’s Watson and Jeopardy! Challenge

Traditional QA System



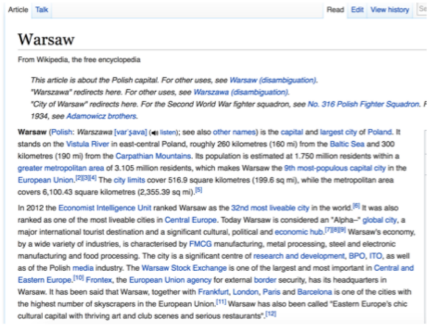
Architecture of LCC (Harabagiu/Moldovan) QA system, circa 2003

Open-domain QA

SQuAD, TREC, WebQuestions, WikiMovies

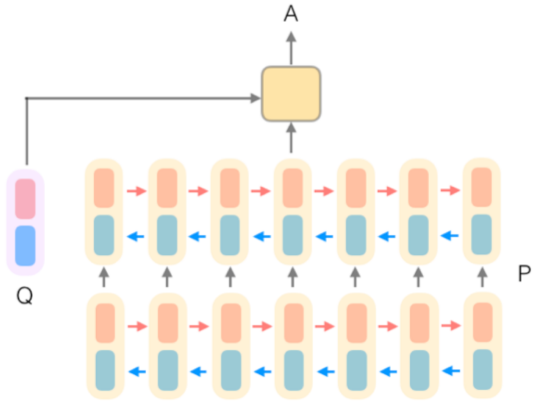
Q: How many of Warsaw's inhabitants spoke Polish in 1933?

Document Retriever



Document Reader

833,500

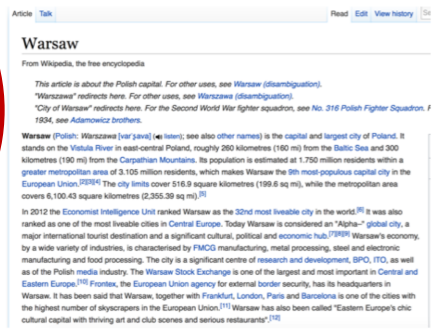


Open-domain QA

SQuAD, TREC, WebQuestions, WikiMovies

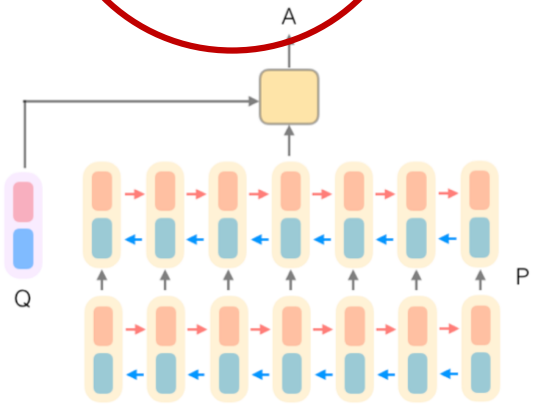
Q: How many of Warsaw's inhabitants spoke Polish in 1933?

Document Retriever

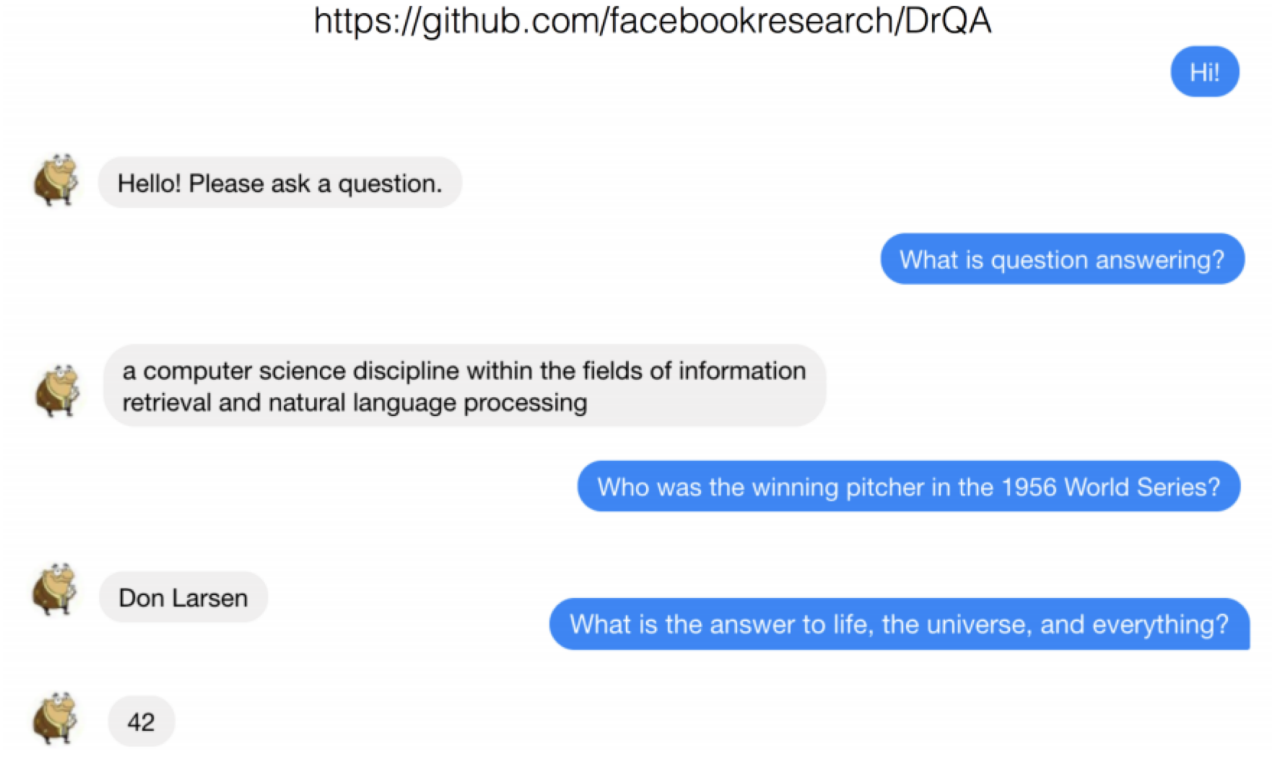


Document Reader

833,500



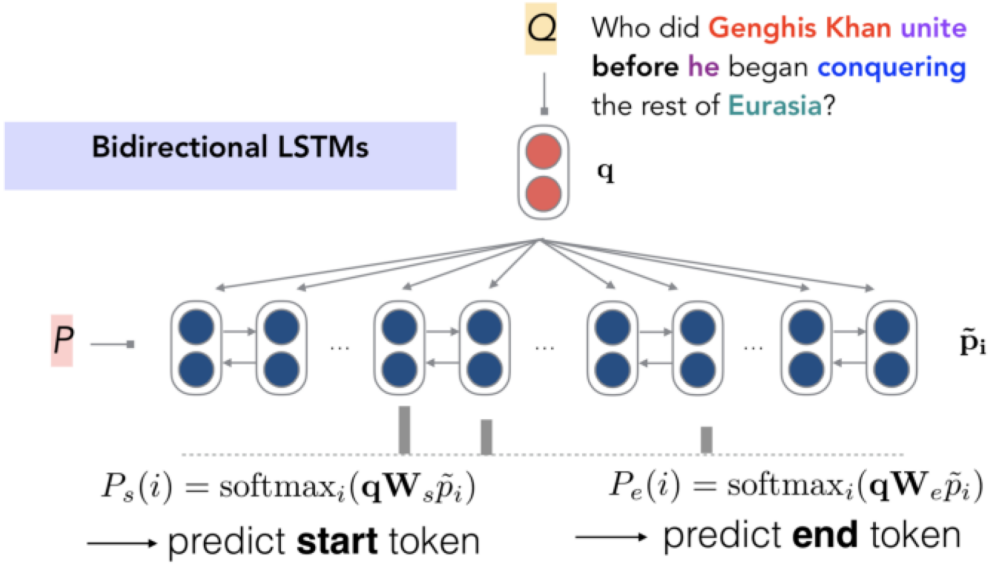
DrQA: Two-Stage Retriever and Reader



Document Retriever: Two Steps

1. TF-IDF bag-of-words vectors
2. Efficient bigram hashing (Weinberger et al., 2009)
   - Map the bigram to 2^{24} bins with an unsigned murmur3 hash
     - Preserving speed and memory efficiency
   - Murmur3: map a word or string to a 32-bit or 128 bit value
     - Online: http://murmurhash.shorelabs.com/

Document Reader

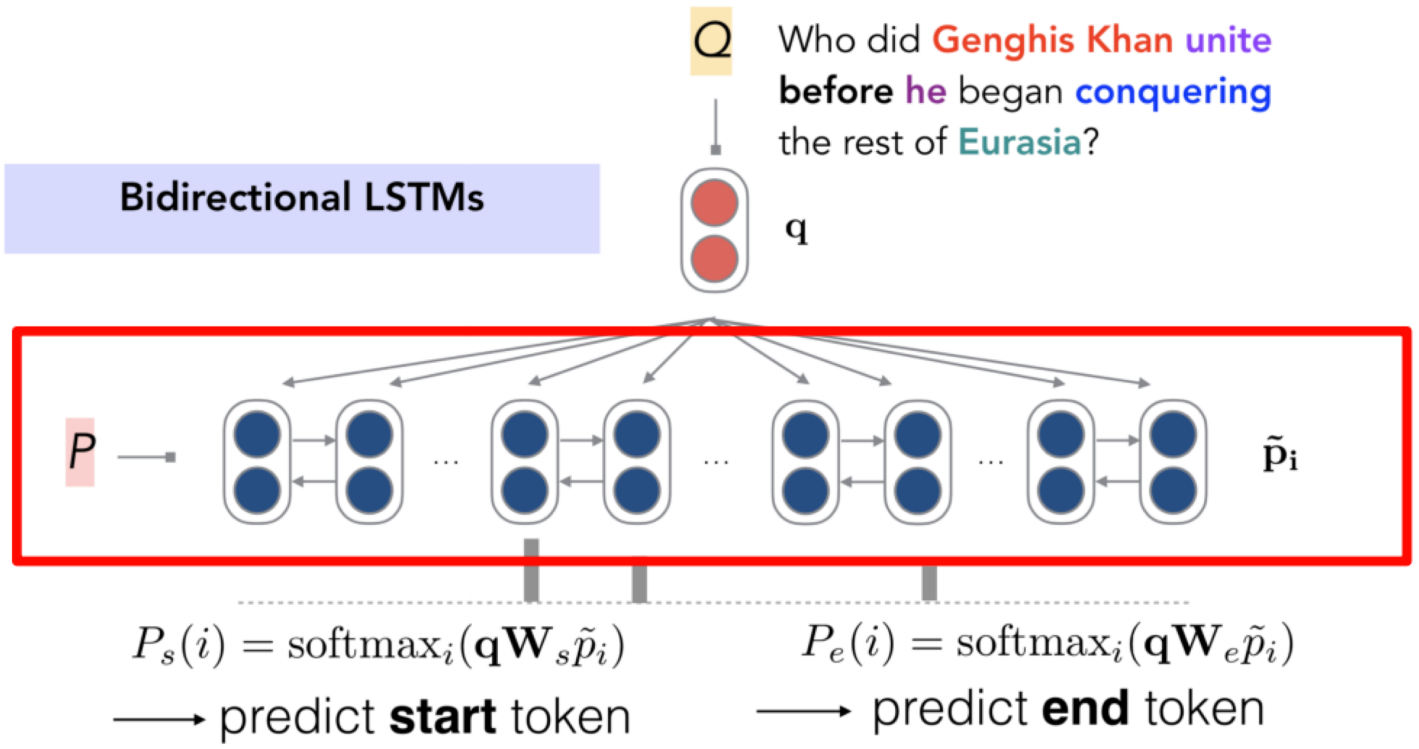


Three steps:

1. Paragraph encoding
2. Question encoding
3. Prediction

similar to AttentiveReader (Hermann et al, 2015; Chen et al, 2016)

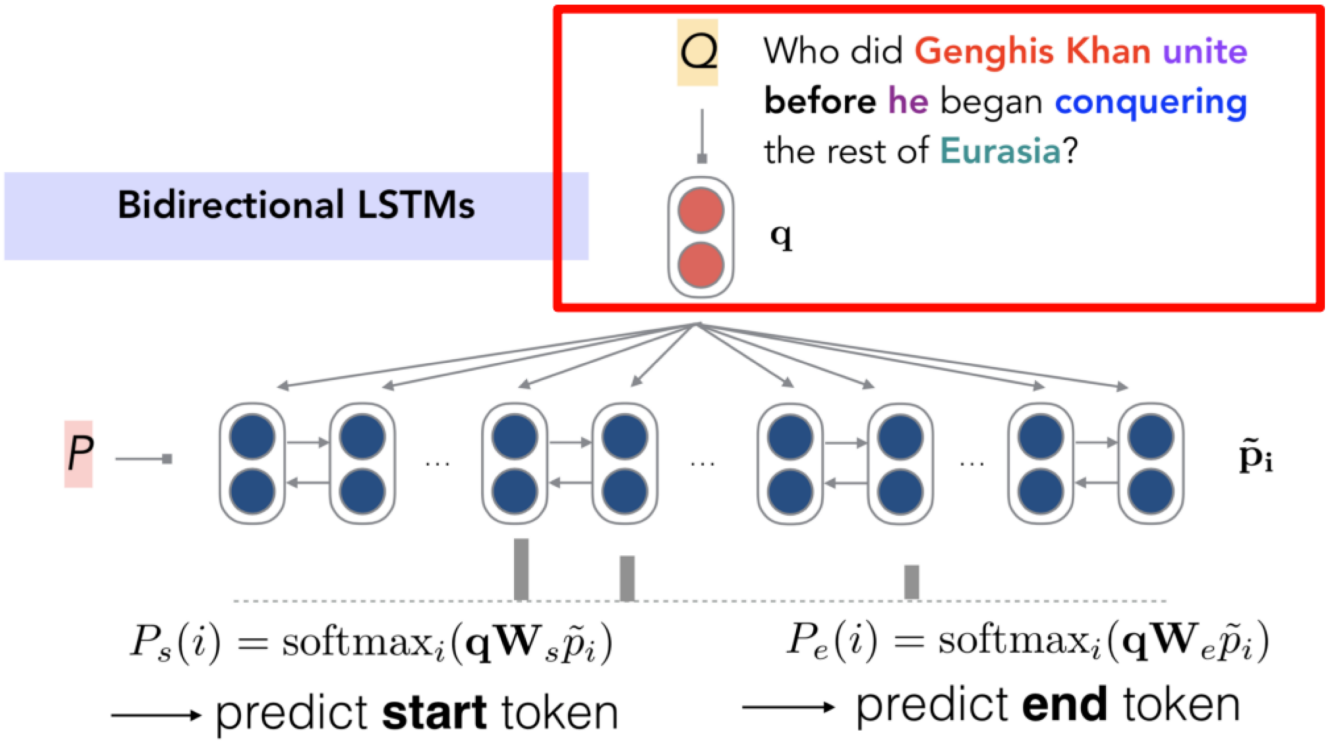
Document Reader: Paragraph Encoding



Document Reader: Paragraph Encoding

1. Represent tokens in a paragraph as a sequence of feature vectors
   - Word embedding
   - Exact match
   - Token features
   - Aligned question embedding
2. Pass features as the input to a RNN (multi-layer Bidirectional LSTM)

Document Reader: Question Encoding



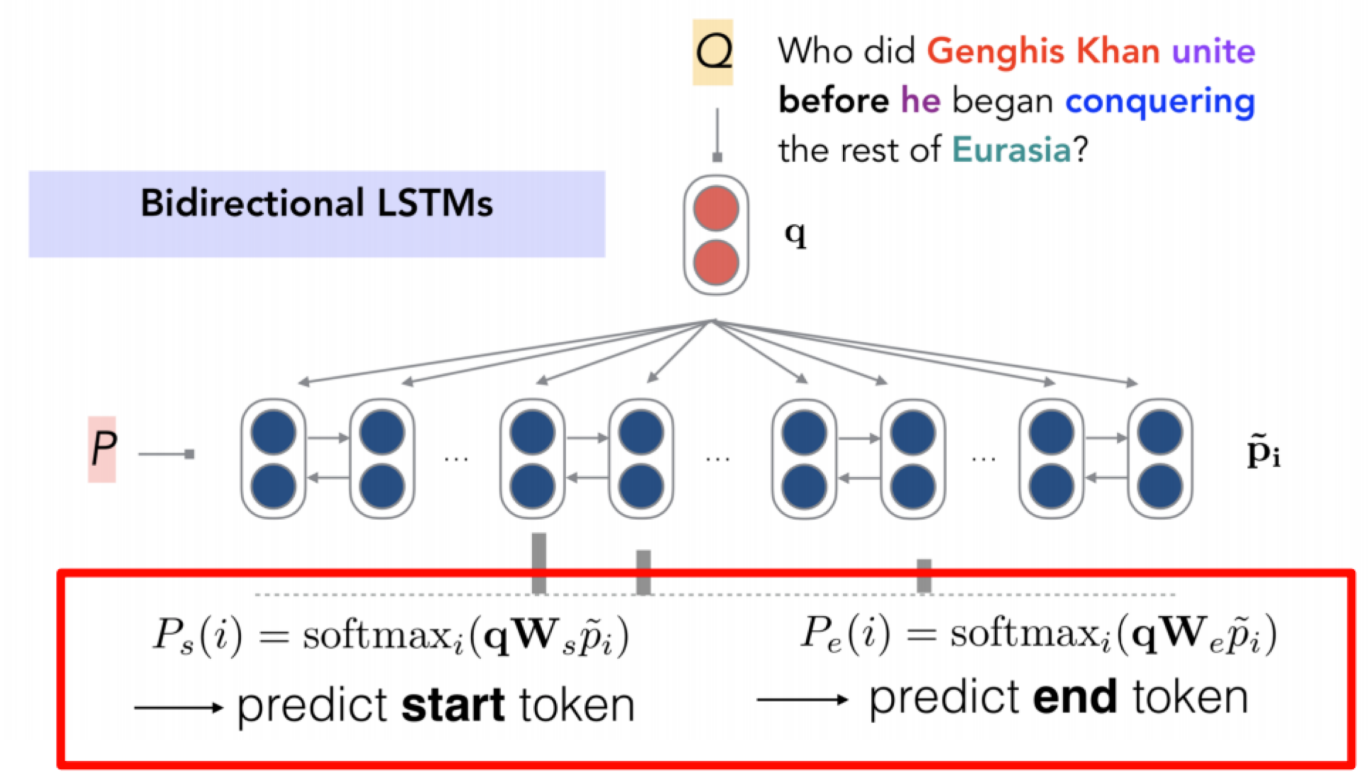
Document Reader: Question Encoding

1. Apply another RNN to top of word embeddings of q_i and get $\mathbf{q_j}$
2. Combining the resulting units into one single vector

$$\boldsymbol{q} = \sum_j b_j \boldsymbol{q_j}$$

Here $b_j = \frac{\exp(\boldsymbol{w} \cdot q_j)}{\sum_{j'} \exp(\boldsymbol{w} \cdot q_{j'})}$ and **w** is a weight vector to learn

Document Reader: Prediction



Document Reader: Prediction

- Goal: predict the span of tokens that is most likely the correct answer
- Method: train two classifiers independently for predicting ends of span

$$\max_{i,j} P_{start}(i) \times P_{end}(j)$$

such that $i \leq j \leq i + 15$ and $P_{start}(i)/P_{end}(\text{j})$ is probability of each token being start/end

$$P_{start}(i) \propto \exp(p_i W_s \boldsymbol{q})$$

$$P_{end}(i) \propto \exp(p_i W_e \boldsymbol{q})$$

Dataset and Example Training Data

Dataset	Example	Article / Paragraph
SQuAD	**Q**: How many provinces did the Ottoman empire contain in the 17th century? **A**: 32	**Article**: Ottoman Empire **Paragraph**: ... At the beginning of the 17th century the empire contained 32 provinces and numerous vassal states. Some of these were later absorbed into the Ottoman Empire, while others were granted various types of autonomy during the course of centuries.
CuratedTREC	**Q**: What U.S. state's motto is "Live free or Die"? **A**: New Hampshire	**Article**: Live Free or Die **Paragraph**: "Live Free or Die" is the official motto of the U.S. state of New Hampshire, adopted by the state in 1945. It is possibly the best-known of all state mottos, partly because it conveys an assertive independence historically found in American political philosophy and partly because of its contrast to the milder sentiments found in other state mottos.
WebQuestions	**Q**: What part of the atom did Chadwick discover?† **A**: neutron	**Article**: Atom **Paragraph**: ... The atomic mass of these isotopes varied by integer amounts, called the whole number rule. The explanation for these different isotopes awaited the discovery of the neutron, an uncharged particle with a mass similar to the proton, by the physicist James Chadwick in 1932. ...
WikiMovies	**Q**: Who wrote the film Gigli? **A**: Martin Brest	**Article**: Gigli **Paragraph**: Gigli is a 2003 American romantic comedy film written and directed by Martin Brest and starring Ben Affleck, Jennifer Lopez, Justin Bartha, Al Pacino, Christopher Walken, and Lainie Kazan.

Result

Dataset	YodaQA	DrQA		
		SQuAD	+Fine-tune (DS)	+Multitask (DS)
SQuAD *(All Wikipedia)*	n/a	27.1	28.4	29.8
CuratedTREC	31.3	19.7	25.7	25.4
WebQuestions	39.8	11.8	19.5	20.7
WikiMovies	n/a	24.5	34.3	36.5

Table 6: Full Wikipedia results. Top-1 exact-match accuracy (in %, using SQuAD eval script). +Fine-tune (DS): Document Reader models trained on SQuAD and fine-tuned on each DS training set independently. +Multitask (DS): Document Reader single model trained on SQuAD and all the distant supervision (DS) training sets jointly. YodaQA results are extracted from https://github.com/brmson/yodaqa/wiki/Benchmarks and use additional resources such as Freebase and DBpedia, see Section 2.

Quick Summary: DrQA

- DrQA was the first attempt to scale up reading comprehension to open-domain question answering, by **combining IR techniques and neural reading comprehension models**.
- Although we achieved good accuracy on SQuAD in 2017 (EM = 70.. vs state-of-the-art EM = 90 in 2020), the final QA accuracy still remains low: 20.7 - 36.5.
- Distant supervision + multi-task learning helps!