

Natural Language Processing (11)

IR, QA, and MT

Daisuke Kawahara

Department of Communications and Computer Engineering,
Waseda University

Lecture Plan

1. Overview of Natural Language Processing
2. Formal Language Theory
3. Word Senses and Embeddings
4. Topic Models
5. Collocations, Language Models, and Recurrent Neural Networks
6. Sequence Labeling and Morphological Analysis
7. Parsing (1): Constituency Parsing
8. Parsing (2): Dependency Parsing
9. Transfer Learning
10. Knowledge Acquisition
11. Information Retrieval, Question Answering, and Machine Translation
12. Guest Talk (1): Dr. Chikara Hashimoto (Rakuten Institute of Technology)
13. Guest Talk (2): Dr. Tsubasa Takahashi (LINE Corporation)
14. Project: Survey or Programming (do it yourself)
15. Project Presentation

Table of Contents

- Information Retrieval (IR)
- Question Answering (QA)
- Machine Translation (MT)

History of Information Retrieval (IR)

- IBM: Record and fetch of documents (1950s-)
- MEDLARS started in 1964 (IR system for medical literature)
- SMART IR System in 1960s (term weighting)
- Keyword search → Full text search
- Web search engine (1990s-)

Inverted Index

document	terms
d1	language, computer, problem
d2	computer, problem
d3	language, problem, information
d4	problem, information
d5	information, computer



term	documents
language	d1, d3
computer	d1, d2, d5
problem	d1, d2, d3, d4
information	d3, d4, d5

Term Weighting

- **Term frequency, $tf_{i,j}$** : number of occurrences of w_i in d_j
- Document frequency, df_i : number of documents in the collection in which w_i occurs
- **Inverse document frequency, idf_i** : $\log(N/df_i)$
the occurrence of a term in a document is more important if that term does not occur in many other documents

Importance of Words (TF.IDF)

TF	d1	d2	d3	d4	d5
language	2	0	1	0	0
computer	1	1	0	0	2
problem	2	2	3	1	0
information	0	0	2	1	1

DF	IDF
2	0.40
3	0.22
4	0.10
3	0.22

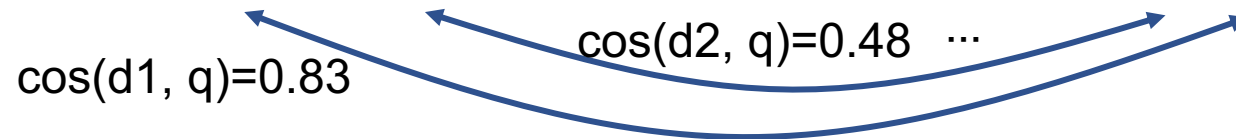
Importance of Words (TF.IDF)

TF.IDF	d1	d2	d3	d4	d5
language	0.80	0.00	0.40	0.00	0.00
computer	0.22	0.22	0.00	0.00	0.44
problem	0.20	0.20	0.30	0.10	0.00
information	0.00	0.00	0.44	0.22	0.22

Vector Space Model

TF.IDF	d1	d2	d3	d4	d5	query
language	0.80	0.00	0.40	0.00	0.00	1
computer	0.22	0.22	0.00	0.00	0.44	0
problem	0.20	0.20	0.30	0.10	0.00	1
information	0.00	0.00	0.44	0.22	0.22	0

$\cos(d1, q)=0.83$ $\cos(d2, q)=0.48 \dots$

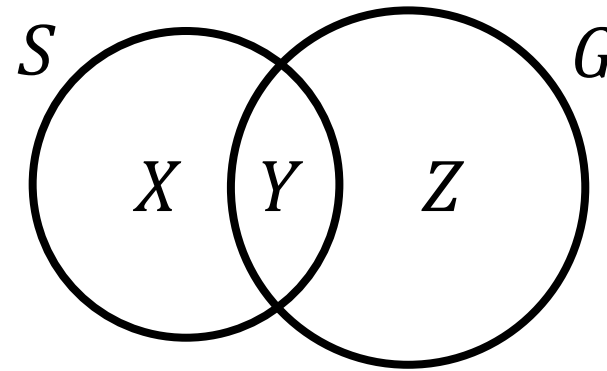


0.83	0.48	0.74	0.30	0.00
------	------	------	------	------

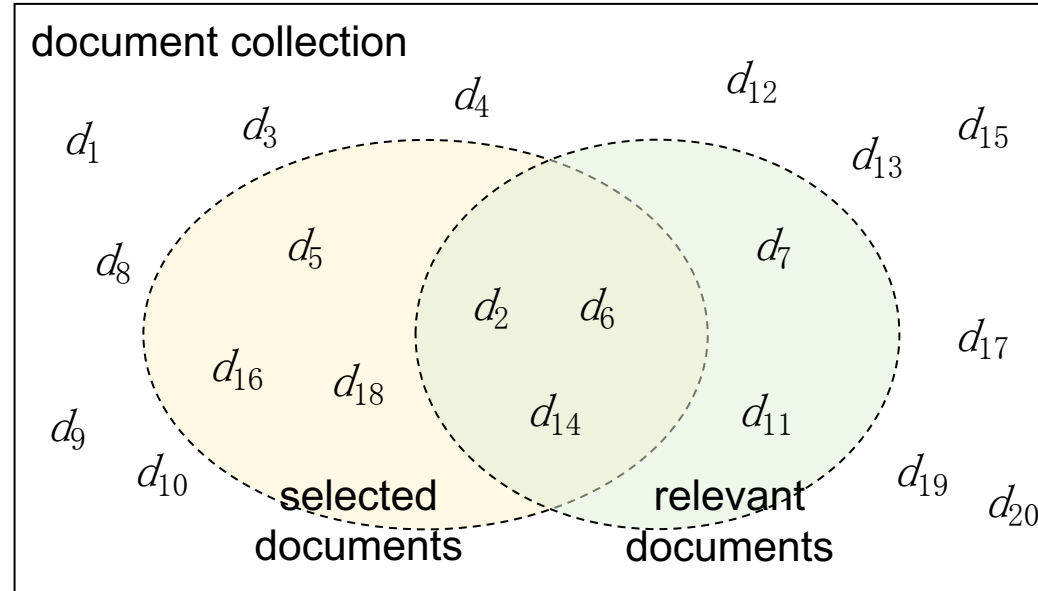
Evaluation Measure

Review: Precision, Recall, and F1

- Precision = $\frac{Y}{X+Y}$
- Recall = $\frac{Y}{Y+Z}$
- $F1 = \frac{1}{\frac{\frac{1}{P} + \frac{1}{R}}{2}} = \frac{2PR}{P+R}$

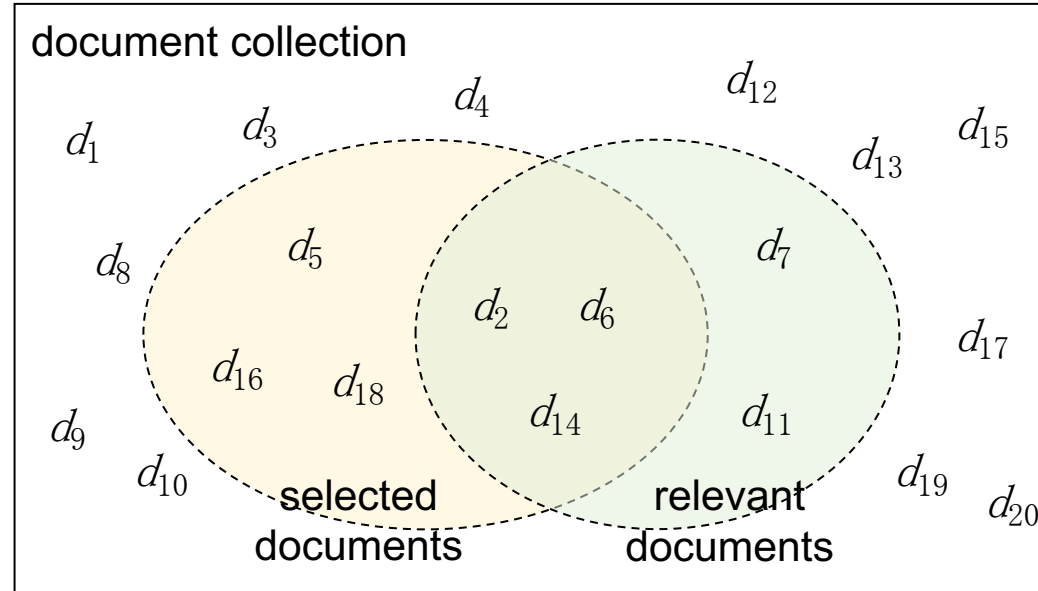


Evaluation Measure: Exercise



1. Calculate precision, recall, and F1.
2. Calculate precision, recall, and F1 when the system selects all the documents.
3. Calculate precision, recall, and F1 when the system selects only d_6 .

Evaluation Measure



$$Accuracy = \frac{3 + 12}{20} = 0.75$$

Is it meaningful?

Evaluation of Ranked Search Results

Ranked search results (d: relevant documents)

d₆ d₁₈ d₁₄ d₅ d₁₆ d₂ d₈ d₁₁ d₁₂ d₁ d₂₀ d₁₇ d₃ d₄ d₇ d₁₉ d₉ d₁₅ d₁₀ d₁₃

Average precision

$$AP(q) = \frac{1}{n} \sum_{k=1}^n \frac{k}{r_k} = \frac{1}{5} \{1/1 + 2/3 + 3/6 + 4/8 + 5/15\} = 0.6$$

n : the number of documents relevant to q

r_k : the rank of the k -th relevant document

Evaluation of an IR System

An evaluation measure for a set of queries $Q=\{q_1, q_2, \dots, q_m\}$ to rank IR systems using a test set collection

MAP (mean average precision)

$$MAP(Q) = \frac{1}{m} \sum_{k=1}^m AP(q_k)$$

Evaluation workshops (test set collections of more than 1M target documents)

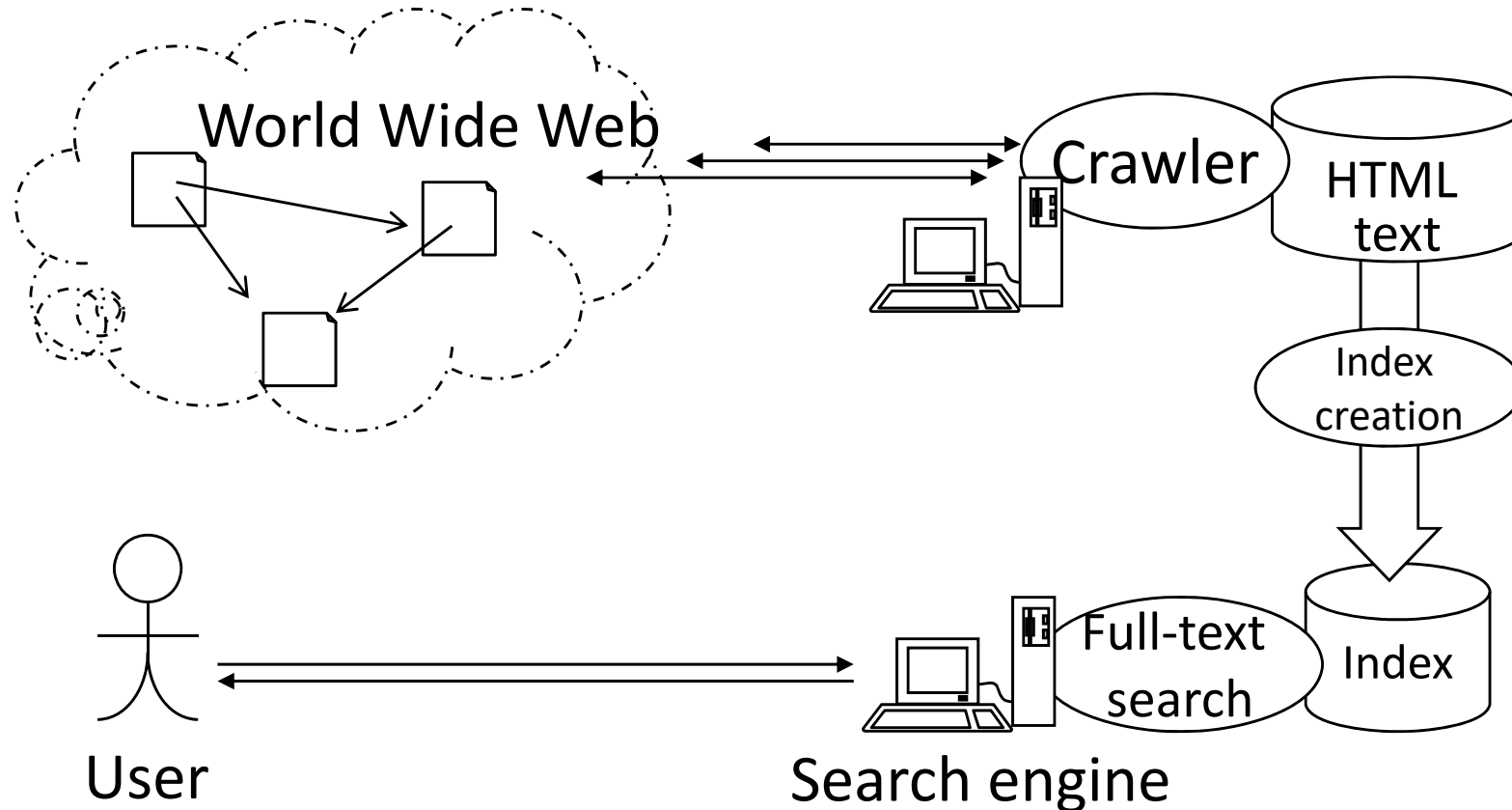
- TREC (1992-), NTCIR (1998-)
- The gold data are created by manually evaluating the union of pages selected by participating systems

The Internet

- Broad sense: the interconnection of multiple networks
- Narrow sense: the global system of interconnected computer networks that use TCP/IP to link billions of devices worldwide
- History
 - 1969: ARPANET developed by the US Department of Defense
 - 1984: JUNET research network developed by Japanese academic organizations
 - 1991: Tim Berners-Lee announced the World Wide Web Project at CERN
- The connected organizations manage each network without a specific centralized authority

Search Engine = Crawler + Full-text Search

Follow hyperlinks to collect HTML documents and make huge indices to enable full-text search



Web Search

- Informational search
 - e.g., “climate change and rising sea levels”
- Navigational search
 - e.g., “Waseda University”, “Toyota”

We need to **measure the importance of a page or site** independently of its contents

PageRank [Page+ 1998]

$$PR(u) = \frac{1-d}{N} + d \sum_{v \in B_u} \frac{PR(v)}{L_v}$$

d : damping factor (~ 0.85)

N : # of pages

B_u : pages that link to u

L_v : # of outbound links on v

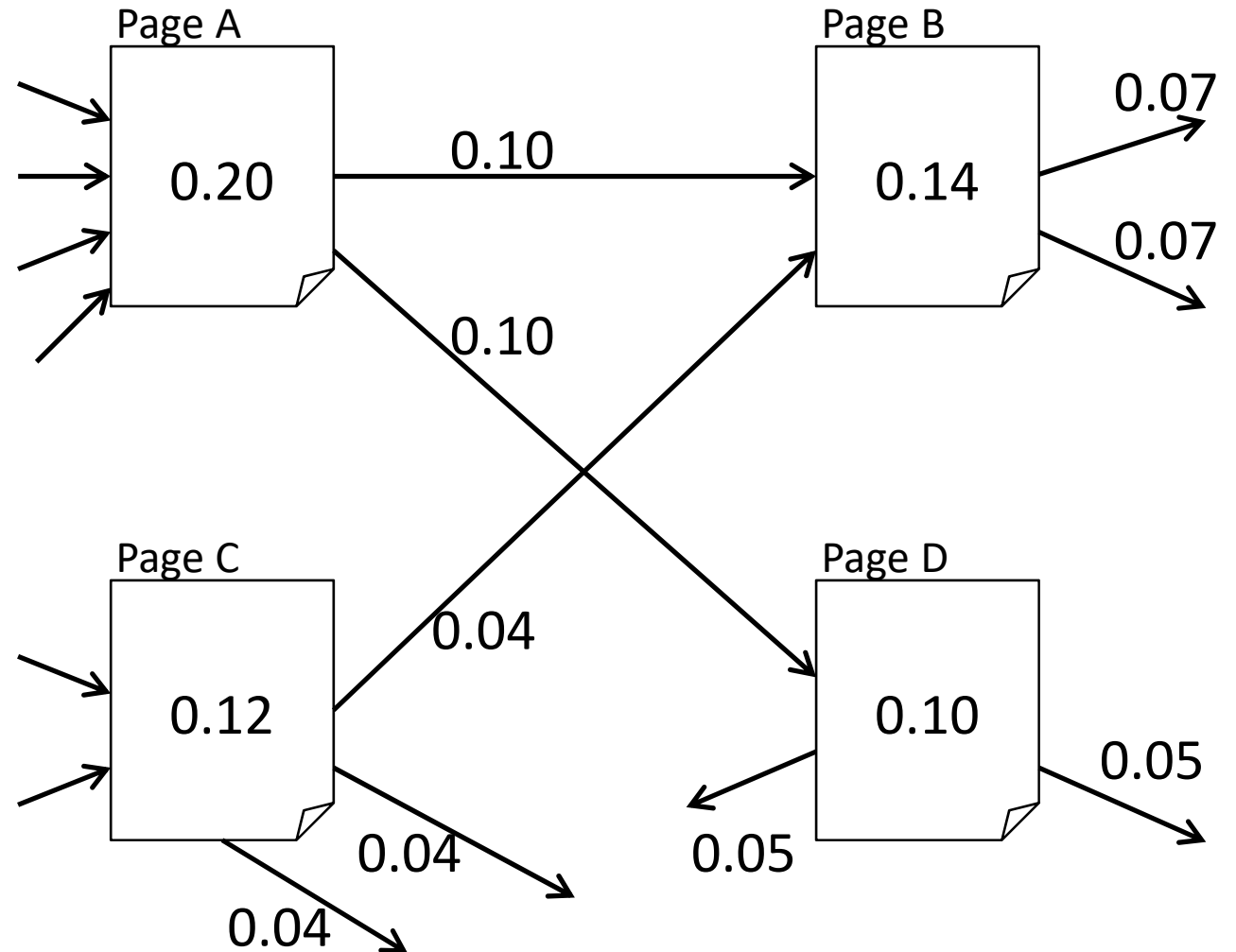


Table of Contents

- Information Retrieval (IR)
- Question Answering (QA)
- Machine Translation (MT)

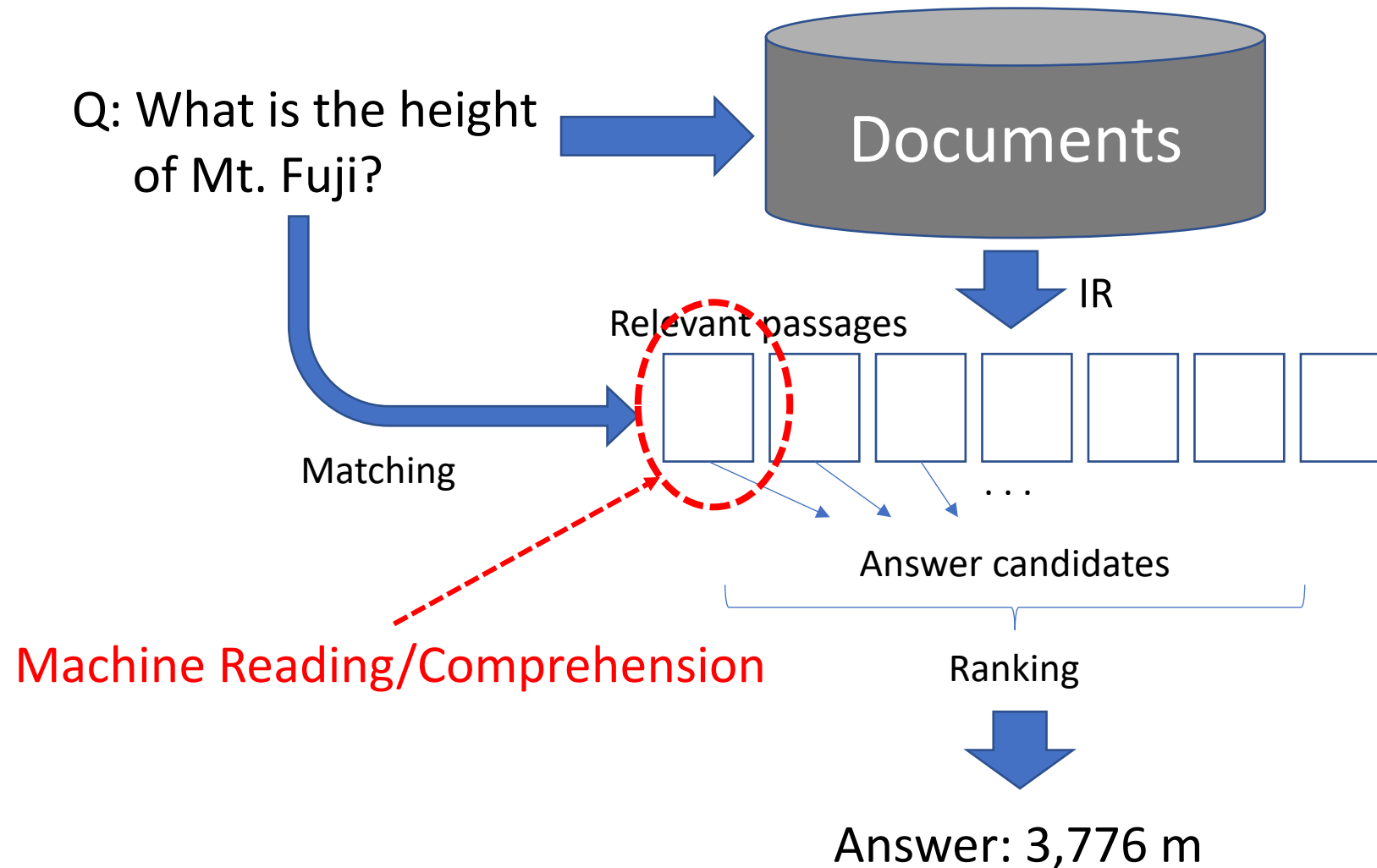
IBM Watson (2011)

- IBM Watson beat human champions in the US quiz show Jeopardy
 - 200M pages (mainly Wikipedia)
 - 2880 CPU cores
 - Handling of metaphor, slang, and sense of humor

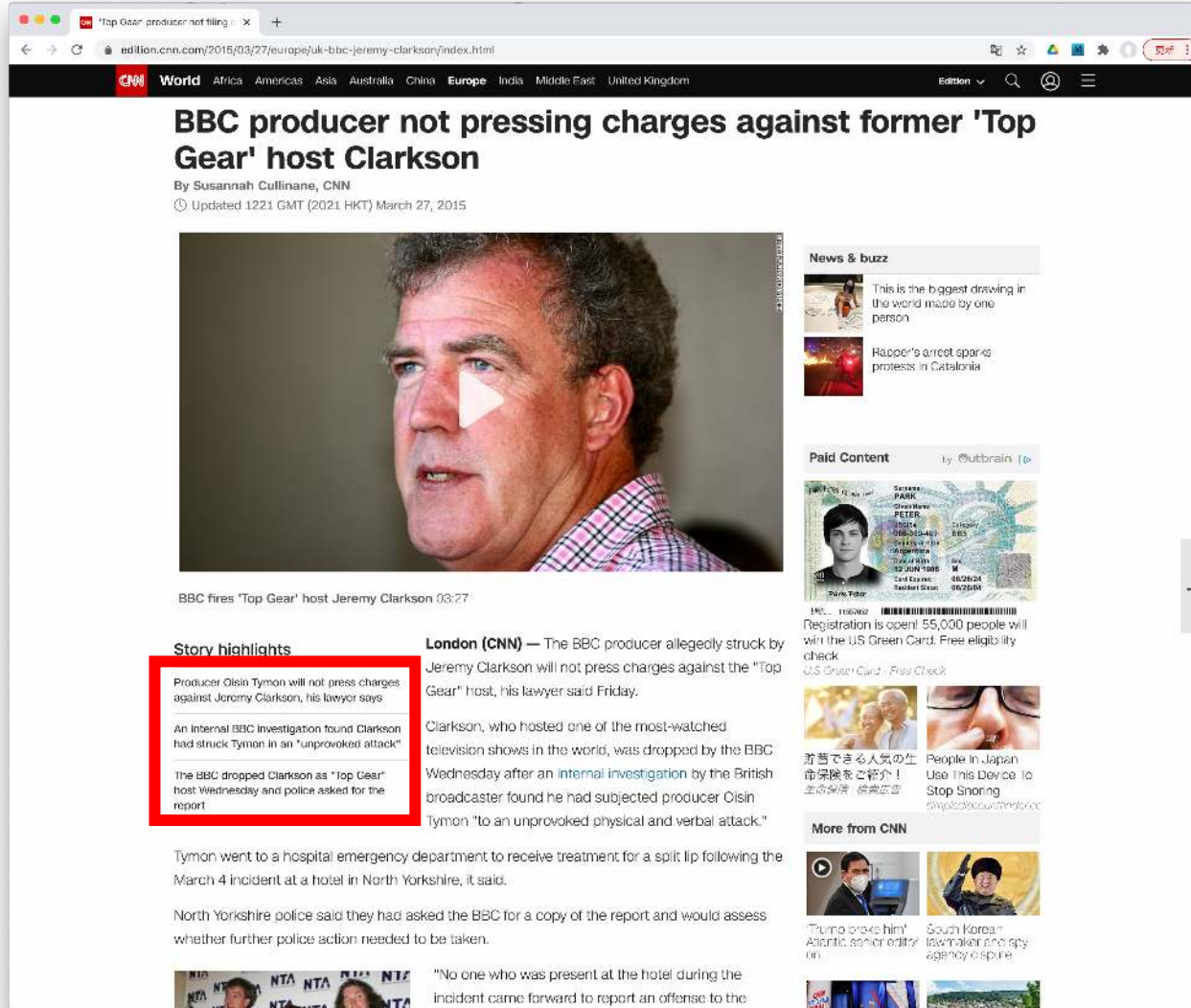
Q: MARILYN MONROE &
BRILLO BOXES WERE 2
OF THIS ARTIST'S
SUBJECTS



Question Answering



CNN / Daily Mail Datasets [Hermann+ 2015]



filling in the correct entity by

low to use world knowledge

Unsanitised Version

ent381 producer allegedly struck by ent212 will press charges against the “ent153” host, his lawyer said Friday. ent212, who hosted one of the most-watched television shows in the world, was dropped by the ent180 broadcaster found he had subjected producer ent193 “to an unprovoked physical and verbal attack.” ...

ent193 will not press charges against ent212, his lawyer says.

CNN:
93K articles
387K queries

Daily Mail:
220K articles
997K queries

SQuAD: 100,000+ Questions for Machine Comprehension of Text [Rajpurkar+ 2016]

536 Wikipedia articles, 23,215 paragraphs

→ 107,785 question-answer pairs (created by crowdsourcing)

Oxygen is a chemical element with symbol O and atomic number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula O₂. Diatomic oxygen gas constitutes 20.8% of the Earth's atmosphere. ...

Q: What is the second most abundant element?

A: helium

Q: The atomic number of the periodic table for oxygen?

A: 8

Q: How many atoms combine to form dioxygen?

A: two

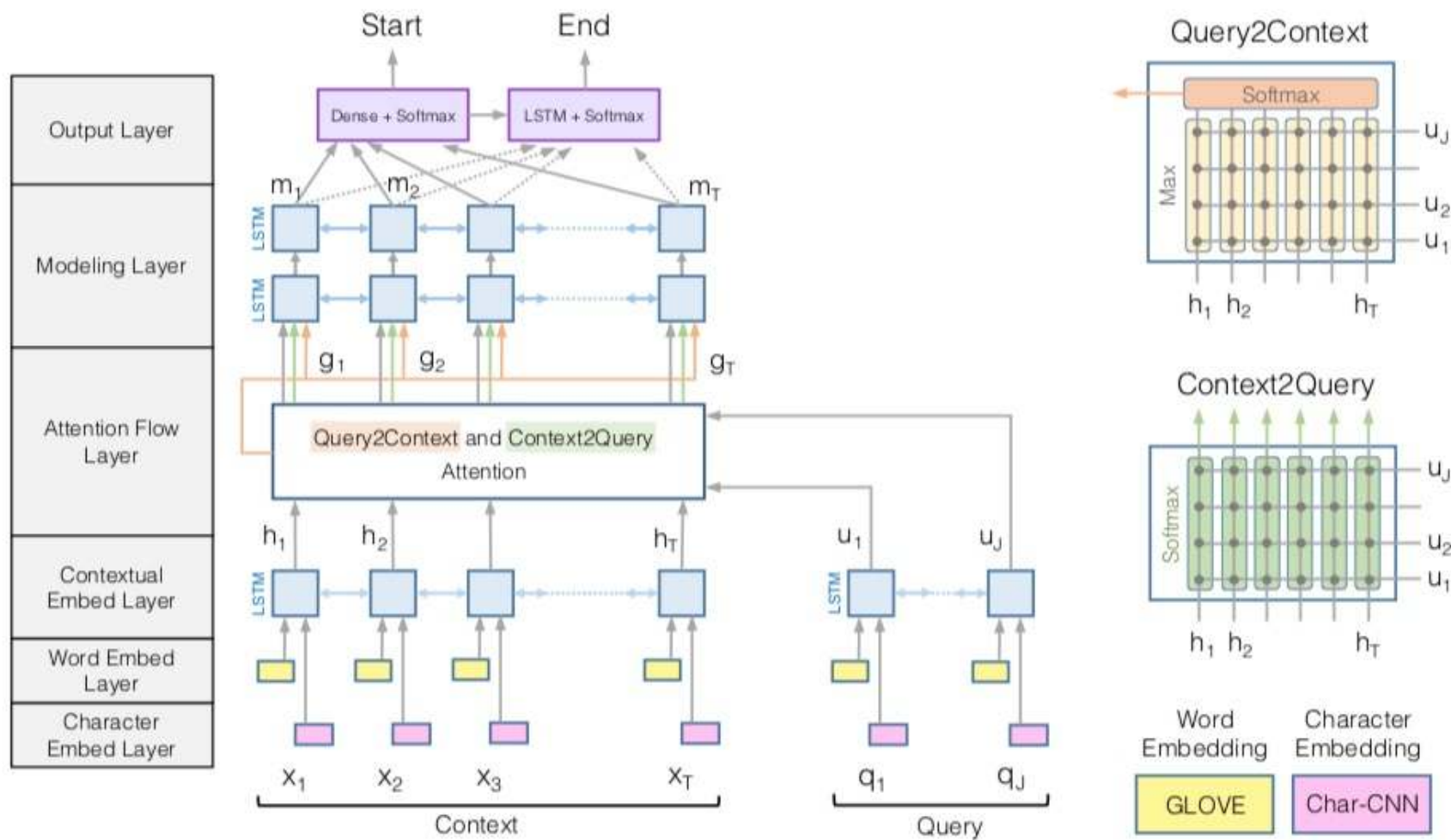
Q: Which gas makes up 20.8% of the Earth's atmosphere?

A: Diatomic oxygen gas, Diatomic oxygen

Human: 86.8% F1

Logistic Regression: 51.0% F1

Bi-Directional Attention Flow Model [Seo+ 2017]

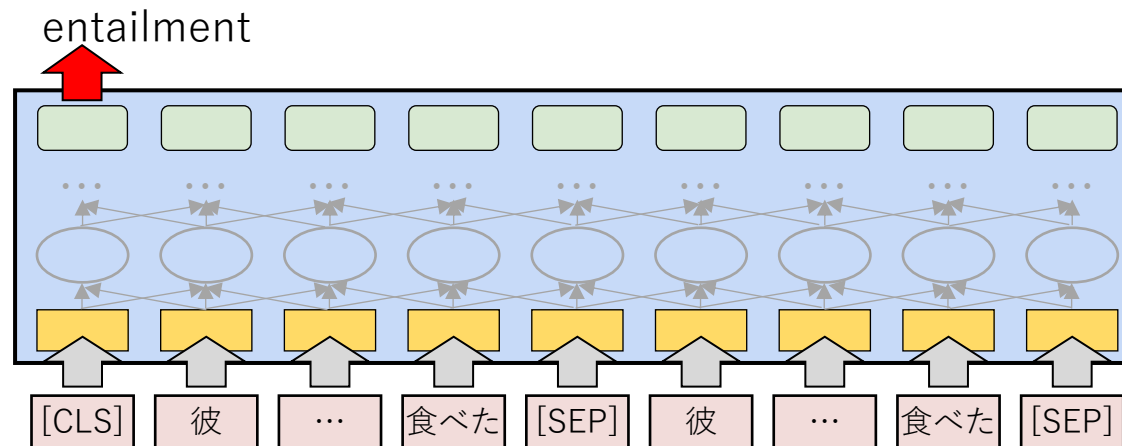


SQuAD: 81.1% F1 (ensemble)

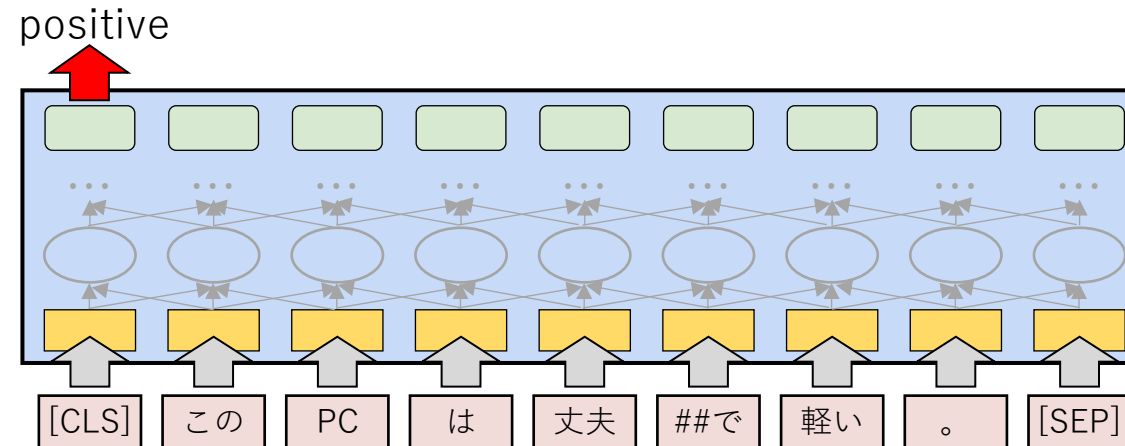
BERT [Devlin+ 2019]

SQuAD: 93.2% F1 (ensemble)

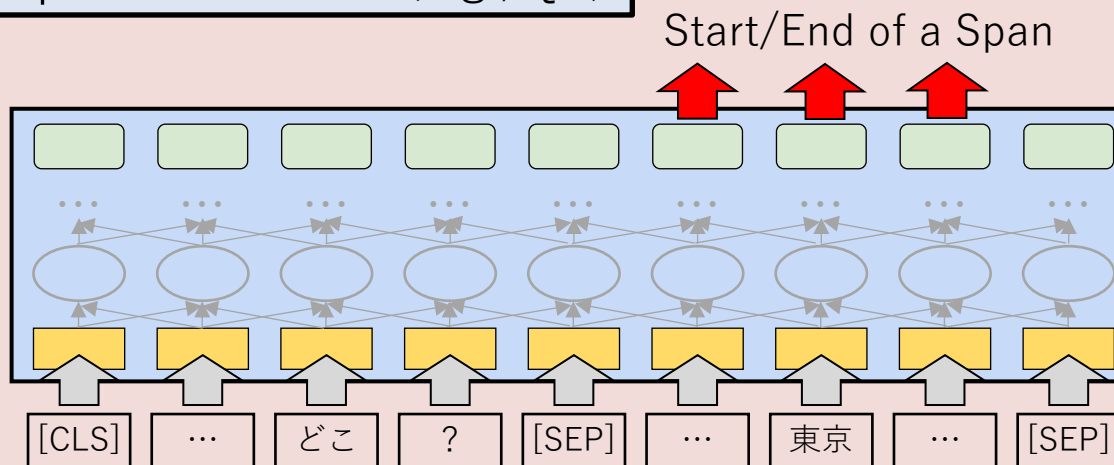
Sentence pair classification (e.g., RTE)



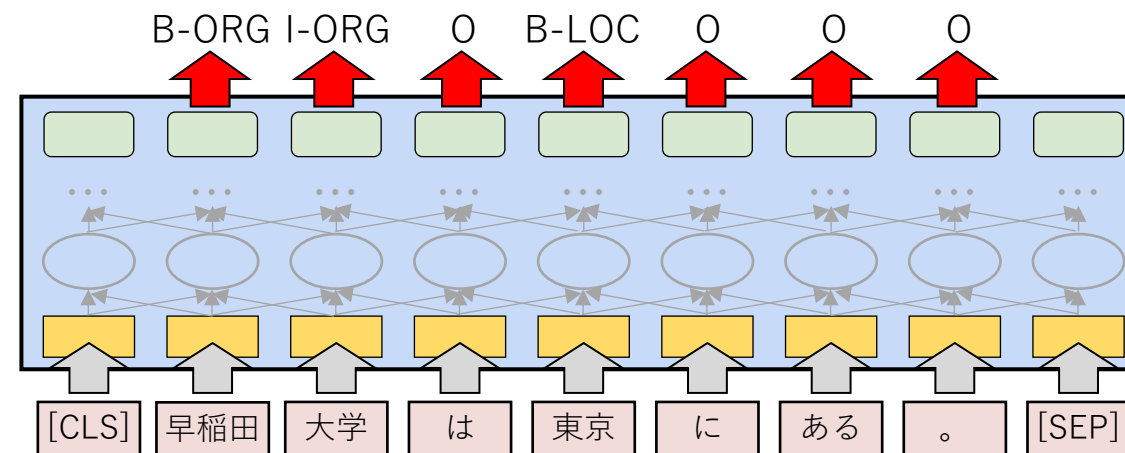
Sentence classification (e.g., sentiment analysis)



Span extraction (e.g., QA)



Sequence labeling (e.g., NER)



SQuAD1.1 Leaderboard

Here are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

Rank	Model	EM	F1
	Human Performance <i>Stanford University</i> (Rajpurkar et al. '16)	82.304	91.221
1 <div>Apr 10, 2020</div>	LUKE (single model) <i>Studio Ousia & NAIST & RIKEN AIP</i>	90.202	95.379
2 <div>May 21, 2019</div>	XLNet (single model) <i>Google Brain & CMU</i>	89.898	95.080
3 <div>Dec 11, 2019</div>	XLNET-123++ (single model) <i>MST/EOI</i> http://tia.today	89.856	94.903
3 <div>Aug 11, 2019</div>	XLNET-123 (single model) <i>MST/EOI</i>	89.646	94.930
4 <div>Sep 25, 2019</div>	BERTSP (single model) <i>NEUKG</i> http://www.techkg.cn/	88.912	94.584
4 <div>Jul 21, 2019</div>	SpanBERT (single model) <i>FAIR & UW</i>	88.839	94.635
5 <div>Jul 03, 2019</div>	BERT+WWM+MT (single model) <i>Xiaoi Research</i>	88.650	94.393
6 <div>Jul 21, 2019</div>	Tuned BERT-1seq Large Cased (single model) <i>FAIR & UW</i>	87.465	93.294
7 <div>Oct 05, 2018</div>	BERT (ensemble) <i>Google AI Language</i> https://arxiv.org/abs/1810.04805	87.433	93.160

Know What You Don't Know: Unanswerable Questions for SQuAD [Rajpurkar+ 2018]

Added 53,775 unanswerable questions created by crowdsourcing

Oxygen is a chemical element with symbol O and atomic number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula O₂. Diatomic oxygen gas constitutes 20.8% of the Earth's atmosphere. ...

Q: What is the second most abundant element?

A: helium

Q: What elements follow oxygen as the most abundant in the universe?

A: N/A

86% F1 on SQuAD 1.1

→ 66% F1 on SQuAD 2.0

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

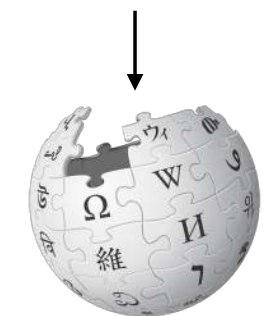
Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) QIANXIN	90.724	93.011
2 May 05, 2020	SA-Net-V2 (ensemble) QIANXIN	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) Shanghai Jiao Tong University http://arxiv.org/abs/2001.09694v2	90.578	92.978
3 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) SRCB_DML	90.442	92.839
4 Jun 21, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) SRCB_DML	90.420	92.799
5 Mar 12, 2020	ALBERT + DAAF + Verifier (ensemble) PINGAN Omni-Sinitic	90.386	92.777
6 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) Shanghai Jiao Tong University http://arxiv.org/abs/2001.09694v2	90.115	92.580

Open-domain QA [Chen+ 2017]

Open-domain QA

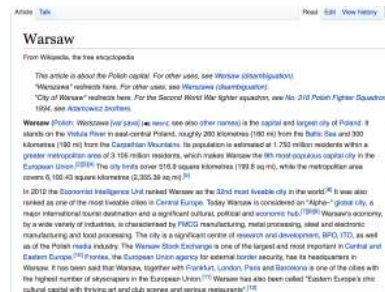
SQuAD, TREC, WebQuestions, WikiMovies

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



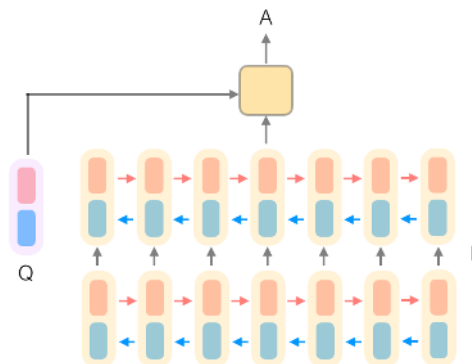
WIKIPEDIA
The Free Encyclopedia

Document
Retriever



Document
Reader

833,500



The system needs to search a relevant document to the question from 5M documents in Wikipedia.

Natural Questions [Kwiatkowski+ 2019]

- 323k real anonymized questions issued to the Google search engine
 - Annotated with
 - Wikipedia article
 - Long answer (paragraph)
 - Human: 87%
 - Best system: 79%
 - Short answer
 - Human: 76%
 - Best system: 64%

Q: when are hops added to the brewing process?



The screenshot shows the Wikipedia article for "Brewing". At the top, there's a navigation bar with "Article" and "Talk" tabs, and a search bar. The article title "Brewing" is prominently displayed. Below the title, it says "From Wikipedia, the free encyclopedia" and includes disambiguation notes. The main text describes the production of beer, mentioning ingredients like starch sources (cereal grains, malted barley) and yeast. It also discusses the history of brewing, from ancient times to the modern brewing industry. A table of contents is visible at the bottom left, listing sections like History, Ingredients, Brewing process, Mashing, and Boiling. On the right side, there is an illustration of a 16th-century brewery with the caption "A 16th-century brewery".

VQA: Visual Question Answering [Agrawal+ 2015]

614K questions for 205K images

150K questions for 50K abstract scenes (created by crowdsourcing)



Does this man have children?	yes	yes
	yes	yes
	yes	yes
Is this man crying?	no	no
	no	yes
	no	yes



Has the pizza been baked?	yes	yes
	yes	yes
	yes	yes
What kind of cheese is topped on this pizza?	feta	mozzarella
	feta	mozzarella
	ricotta	mozzarella



How many pickles are on the plate?	1	1
	1	1
	1	1
What is the shape of the plate?	circle	circle
	round	round
	round	round



What does the sign say?	stop	stop
	stop	stop
	stop	yield
What shape is this sign?	octagon	diamond
	octagon	octagon
	octagon	round



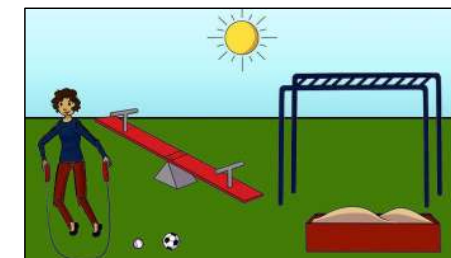
How many glasses are on the table?	3	2
	3	2
	3	6
What is the woman reaching for?	door	fruit
	handle	glass
	wine	remote



Do you think the boy on the ground has broken legs?	yes	no
	yes	no
	yes	yes
Why is the boy on the right freaking out?	his friend is hurt	ghost
	other boy fell down	lightning
	someone fell	sprayed by hose



Are the kids in the room the grandchildren of the adults?	probably	yes
	yes	yes
	yes	yes
What is on the bookshelf?	nothing	books
	nothing	books
	nothing	books



How many balls are there?	2	1
	2	2
	2	3
What side of the teeter totter is on the ground?	right	left
	right	left
	right side	right side

Table of Contents

- Information Retrieval (IR)
- Question Answering (QA)
- Machine Translation (MT)

MT at the Beginning

- Started right after computers have appeared
- 1947: English-French, W. Weaver and A. D. Booth

*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**."*

- Basically, replacement of words
- 1954: Russian-English, Georgetown Univ. and IBM
- 1955-: English-Japanese: Electro Technical Laboratory
English-German-Japanese: Kyusyu Univ.
- 1966: ALPAC Report
 - MT is too difficult for the time being
 - More fundamental research is needed, such as computational linguistics

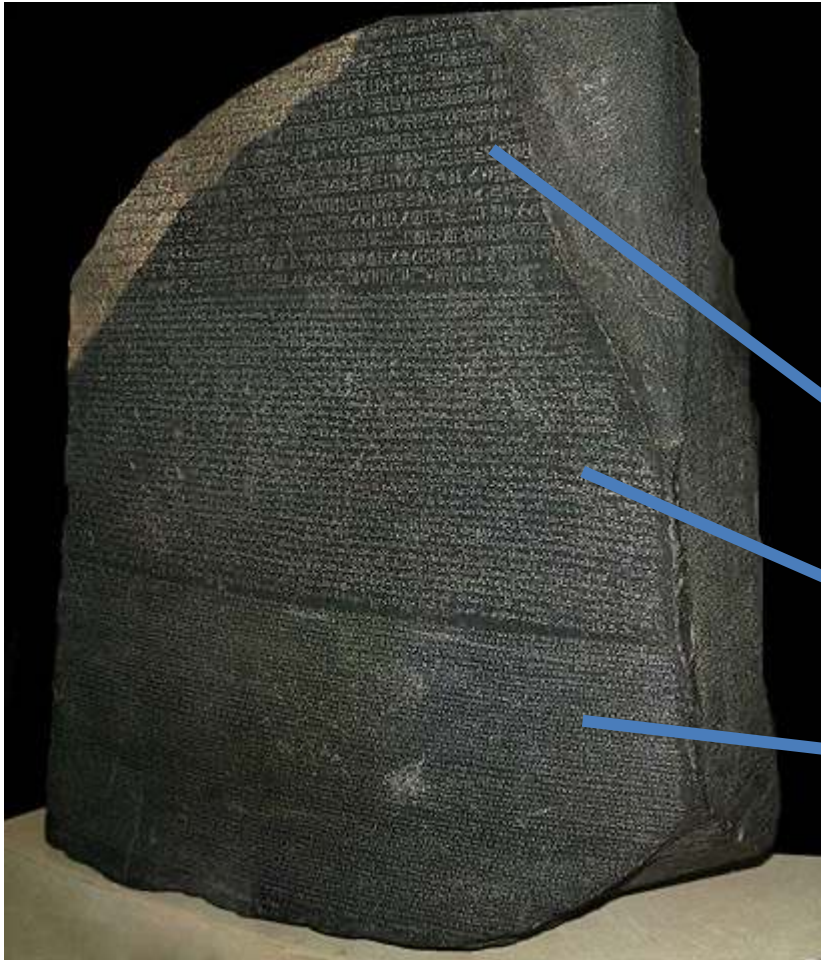
1980s: MT Projects

- Syntactic analysis-based translation
- 1982-1991: EUROTRA Project
 - Bidirectional translations between seven (later nine) languages in European Commission
- 1982-1985: Mu Project
 - Science and Technology Agency and Kyoto Univ.
 - Japanese-English, English-Japanese translation

1990s-: Corpus-based MT

- Example-based MT (EBMT)
 - [Nagao 1981]
- Statistical MT (SMT)
 - [Brown+ 1991]
- Neural MT (NMT)
 - [Sutskever+ 2014]
 - [Bahdanau+ 2014]
- Google's Neural MT System
 - 2016.9: Bridging the gap between human and MT

Rosetta Stone



https://en.wikipedia.org/wiki/Rosetta_Stone

- A stele discovered in 1799 in Rosetta, Egypt
- A decree issued during the Ptolemaic dynasty

Hieroglyphic text in ancient Egyptian

Demotic text in ancient Egyptian

Greek text

Corpus-based MT =
Learning with bilingual text

Machine Translation

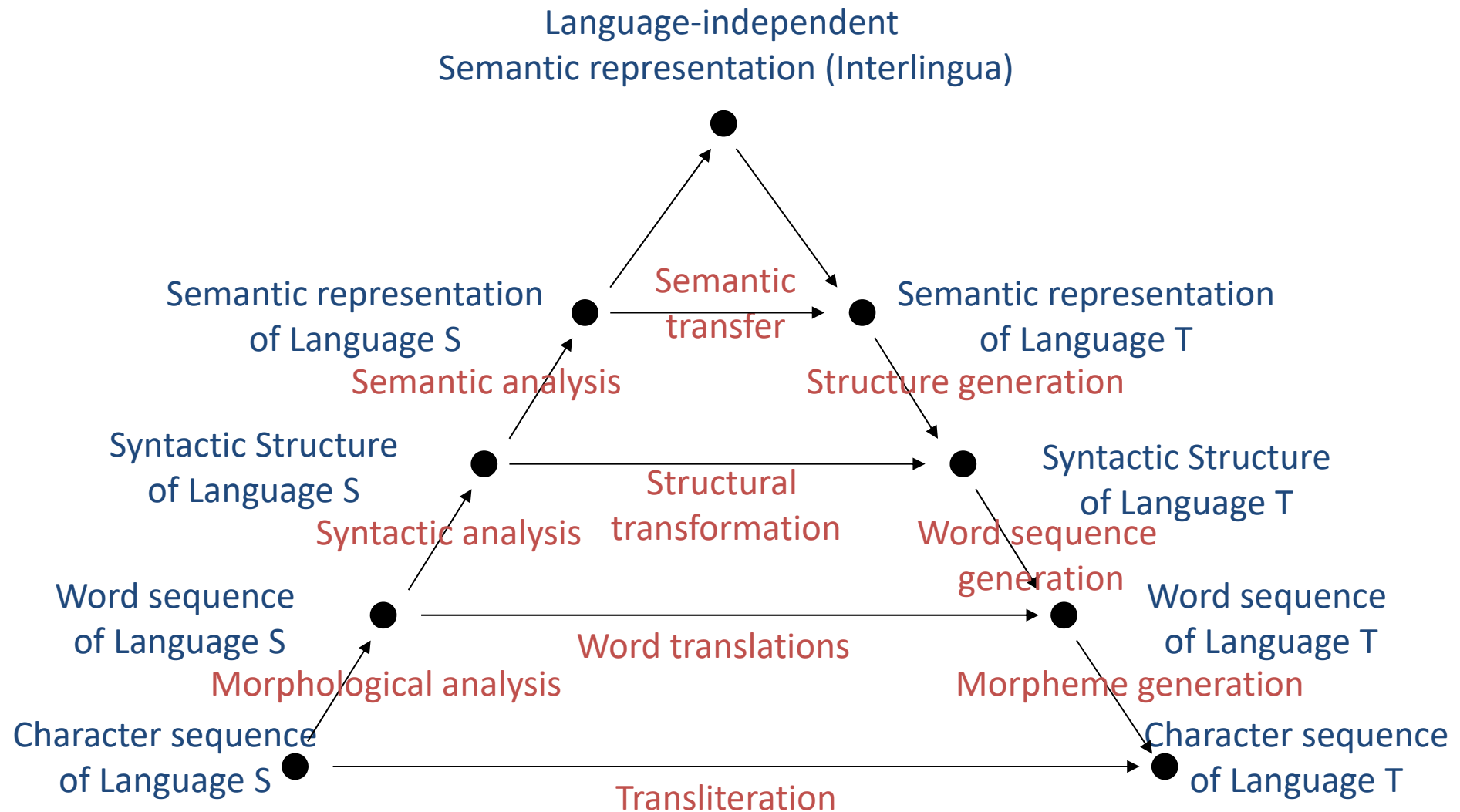
Replace text in one language (source language) with **equivalent** text in another language (target language)

1. Equivalence in words and syntax
 - literal translation
2. Equivalence in semantics
 - liberal translation
3. Equivalence in effects
 - sophisticated translation, creative translation

Difficulty in Translation

- Lexicon
 - put on X ⇔ 帽子をかぶる, 服を着る, 靴を履く
- Word order
 - He wears a suit (SVO) ⇔ 彼はスーツを着ている (SOV)
- Structure
 - She has a long hair ⇔ 彼女は髪が長い
- Explicit expressions (number, gender, definiteness, ...)
 - No plural/singular distinction nor article (a, the) in Japanese
 - No counter in English
 - e.g., 1個, 2本


MT Pyramid



Statistical Machine Translation (SMT)

For a given sentence j in the source language, take e as j 's translation that maximizes $P(e|j)$, which is the probability that j is translated to e .

Noisy channel model

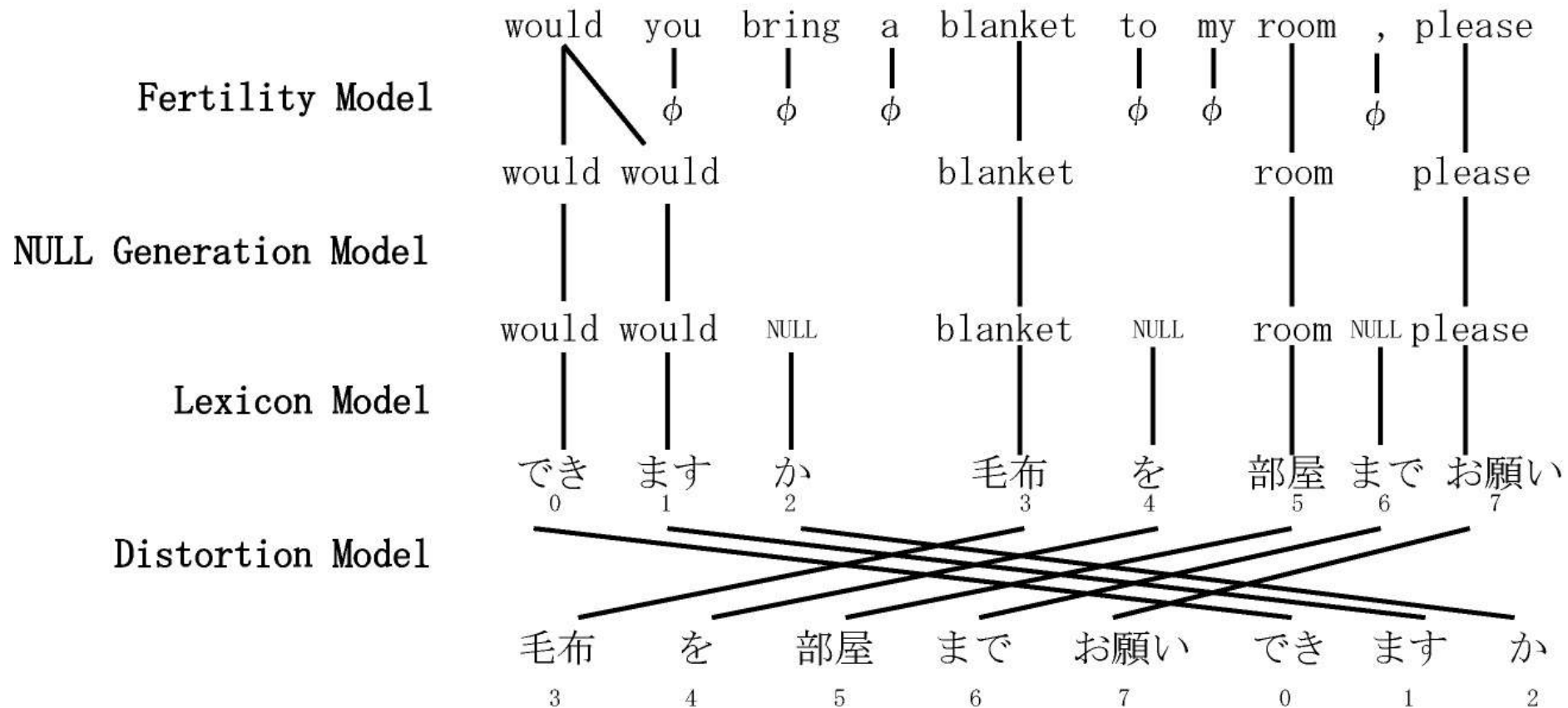
$$\begin{aligned}\hat{e} &= \arg \max_e P(e|j) \\ &= \arg \max_e \frac{P(j|e)P(e)}{P(j)} \\ &= \arg \max_e P(j|e)P(e)\end{aligned}$$


Translation model

Language model

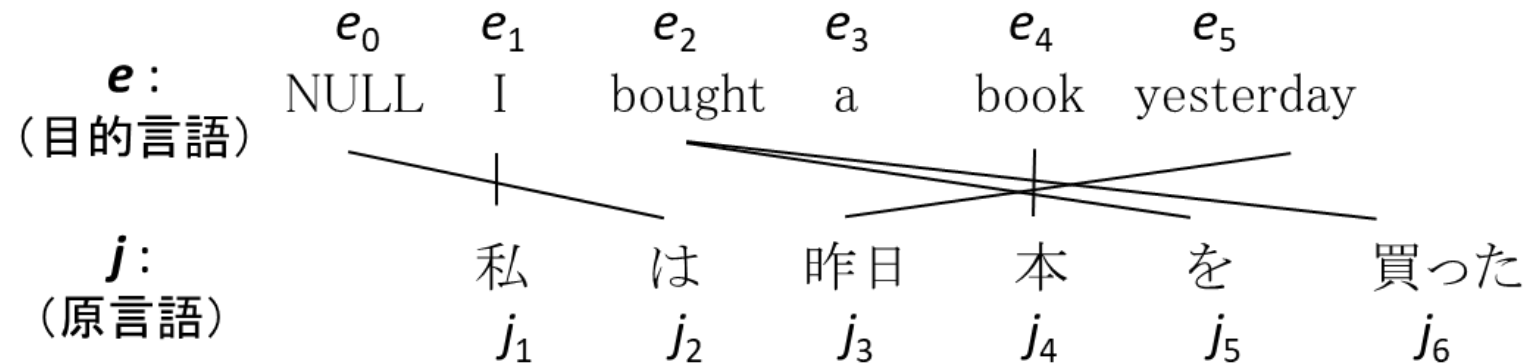
Statistical Machine Translation (SMT)

Calculation of $P(j|e)$ with the IBM model



Statistical Machine Translation (SMT)

Word alignment

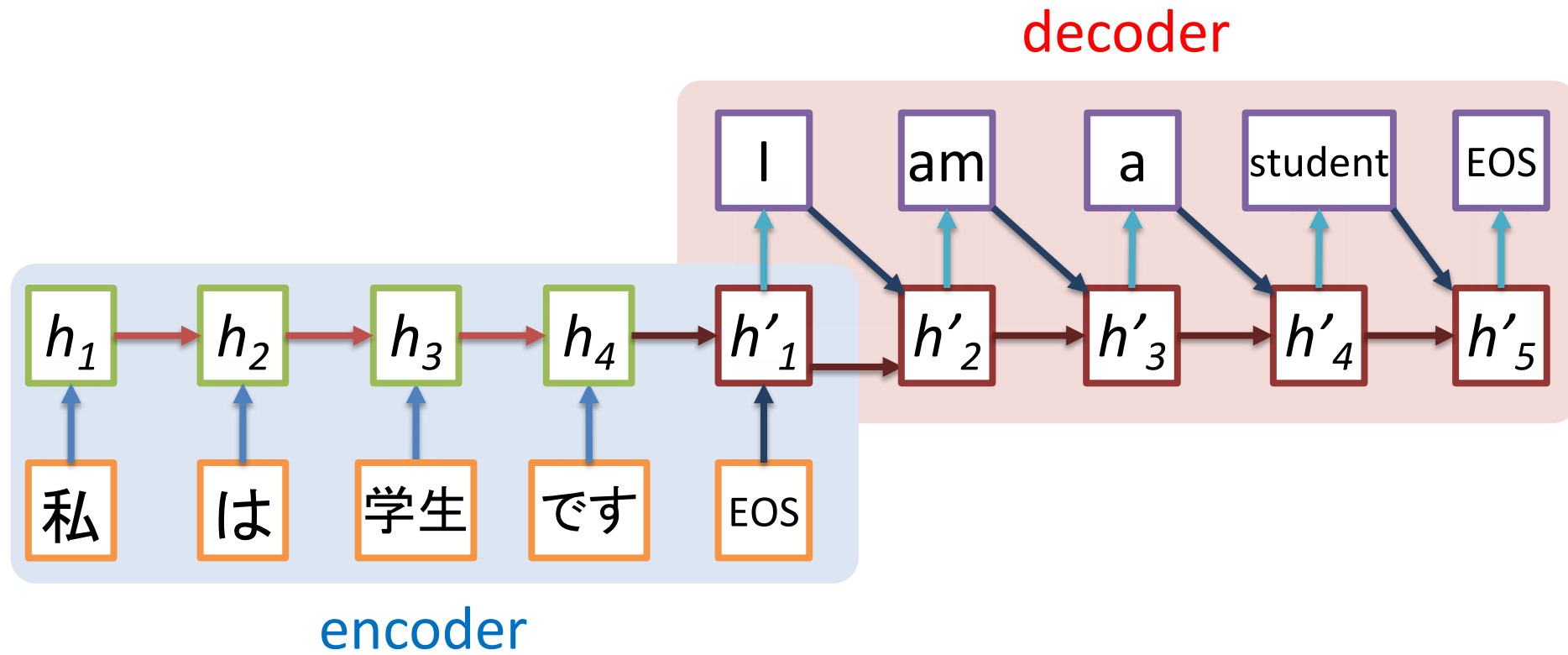


Parameters (probabilities) and word alignment are simultaneously estimated from a bilingual corpus without a bilingual lexicon

⇒ EM algorithm

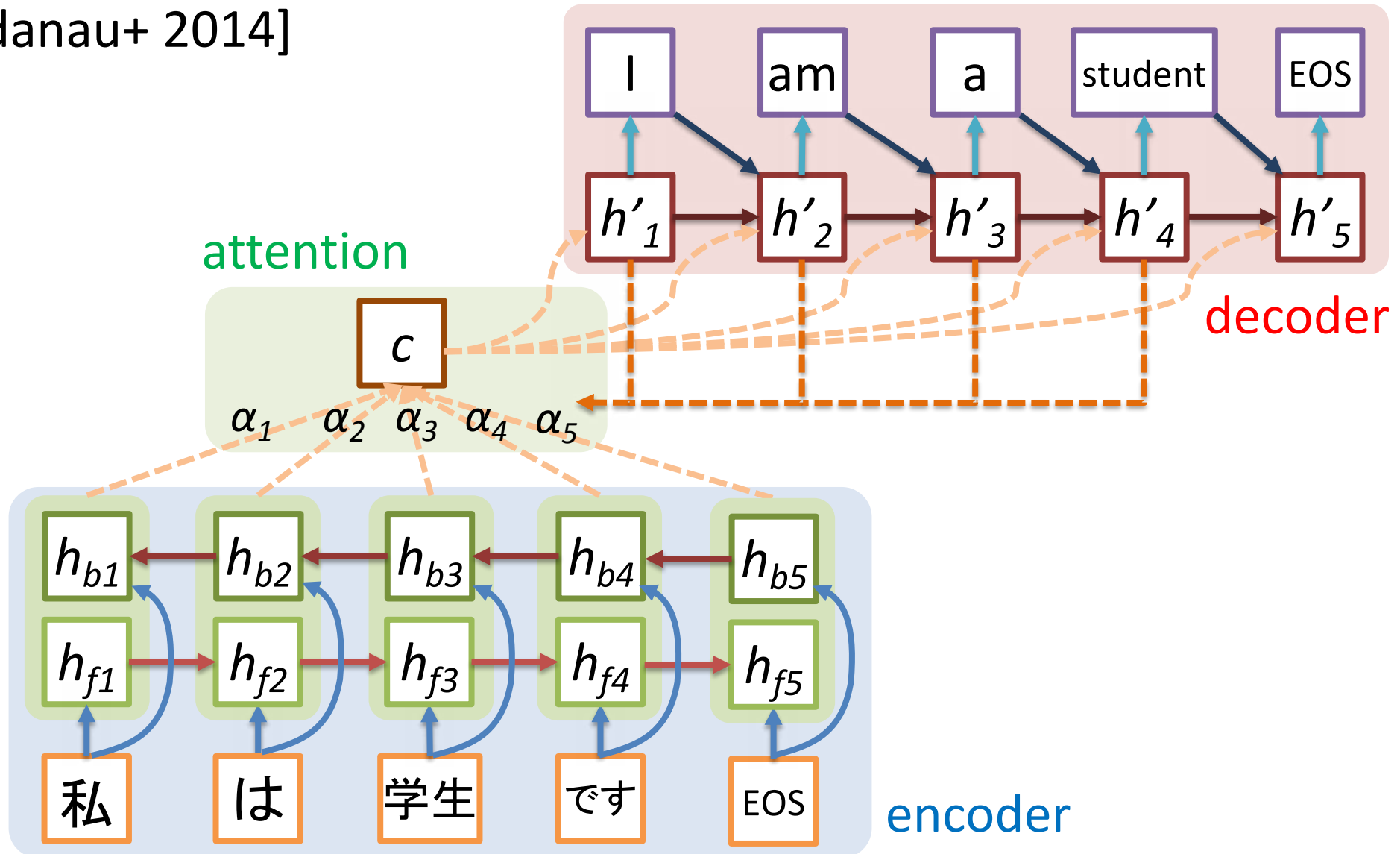
Neural Machine Translation (NMT)

[Sutskever+ 2014]

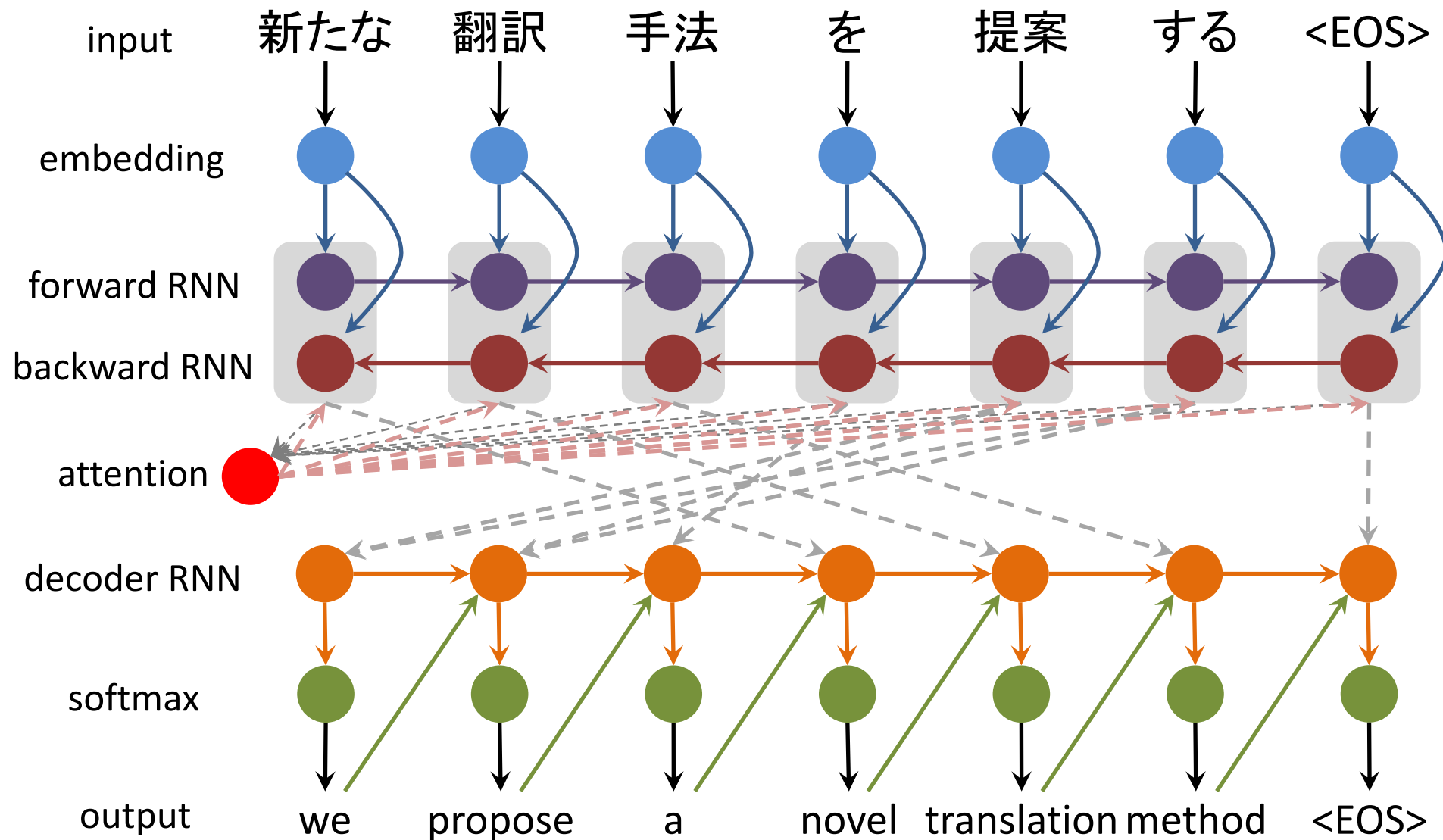


Attention-based NMT

[Bahdanau+ 2014]

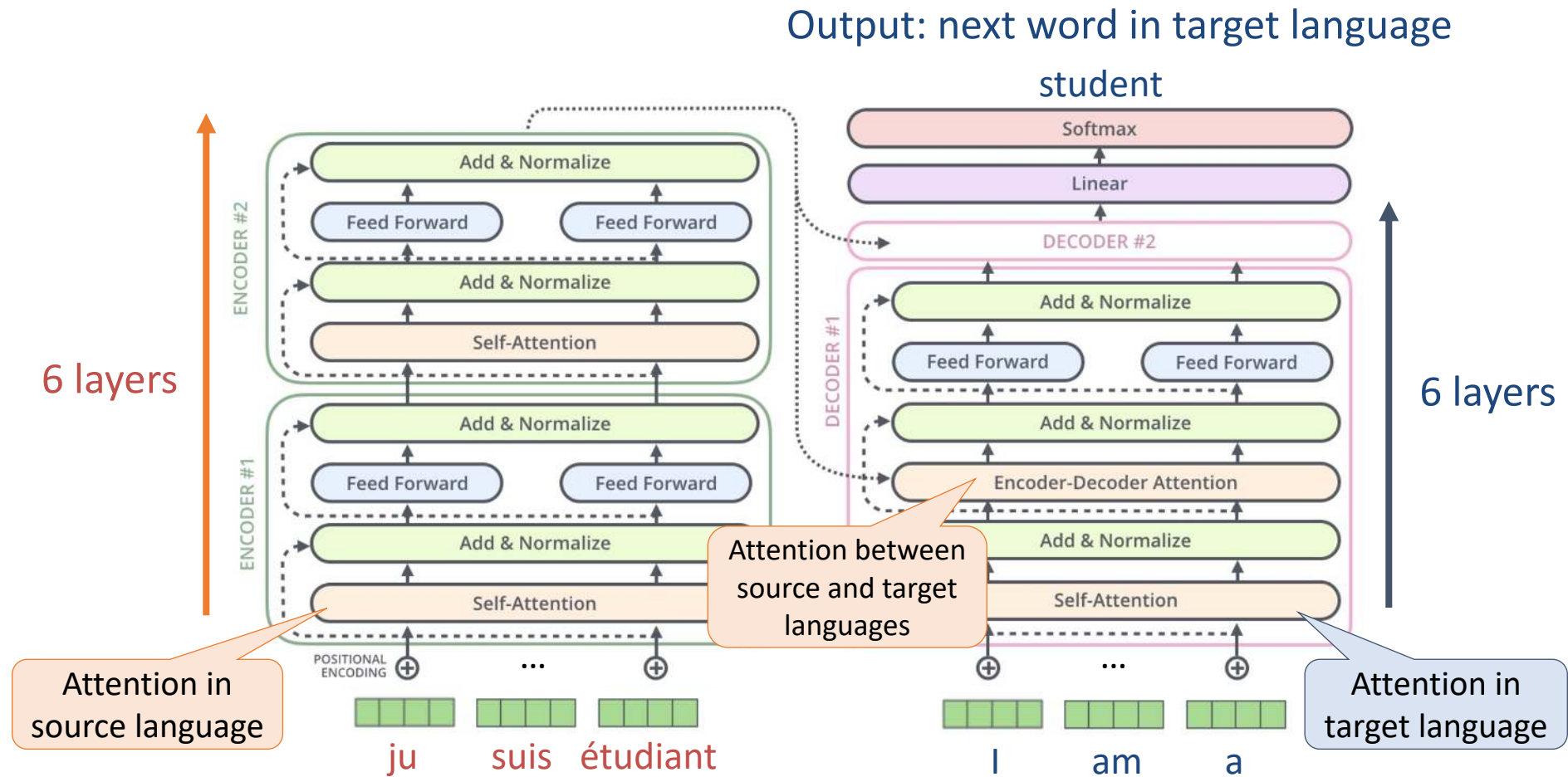


Attention-based NMT [Bahdanau+ 2014]



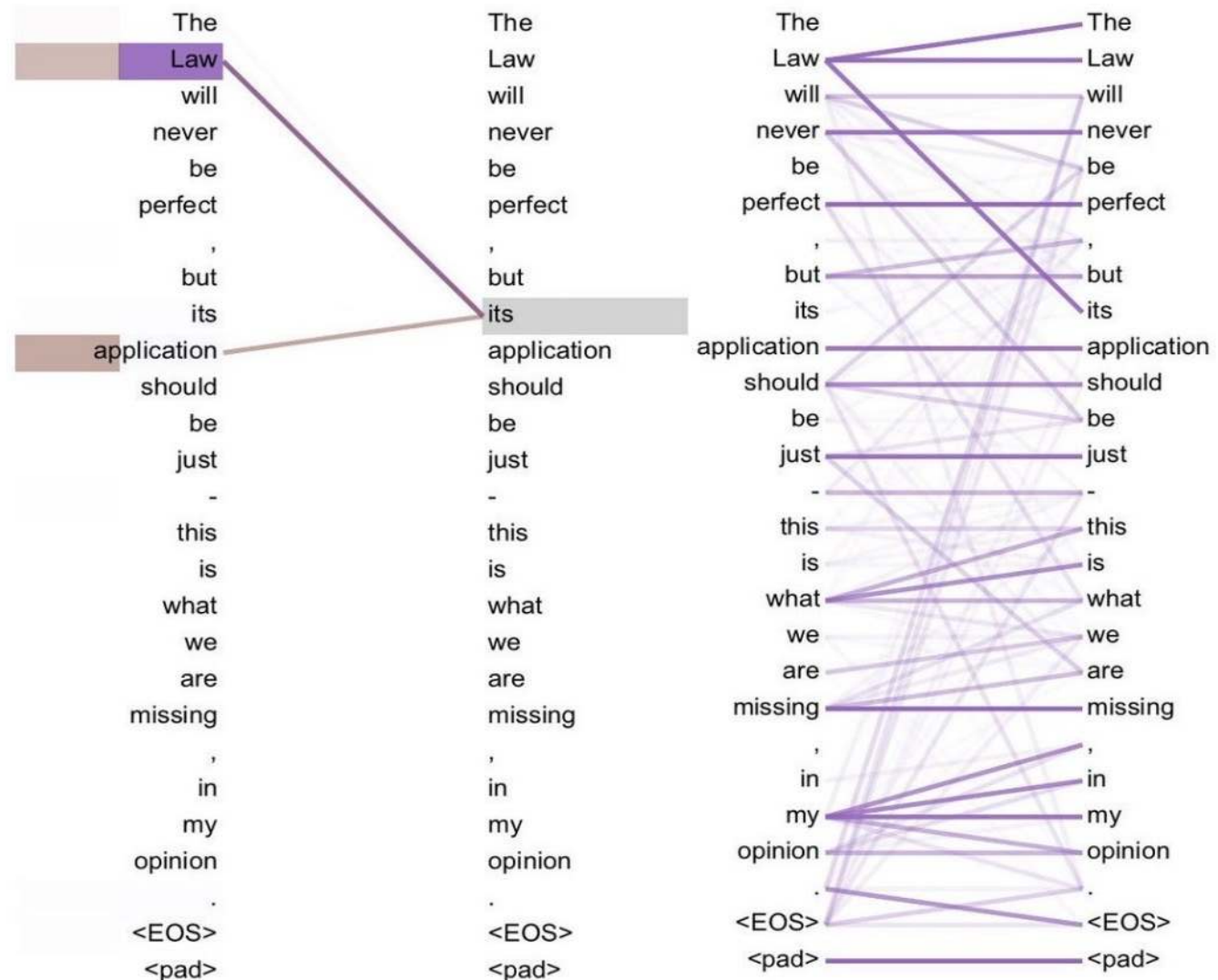
Transformer: “Attention is All You Need”

[Vaswani+ 2017]



Self Attention = Anaphora Resolution?

- Some attention heads look at long distance relationships
 - Anaphora?



Evaluation of MT: Automatic Evaluation

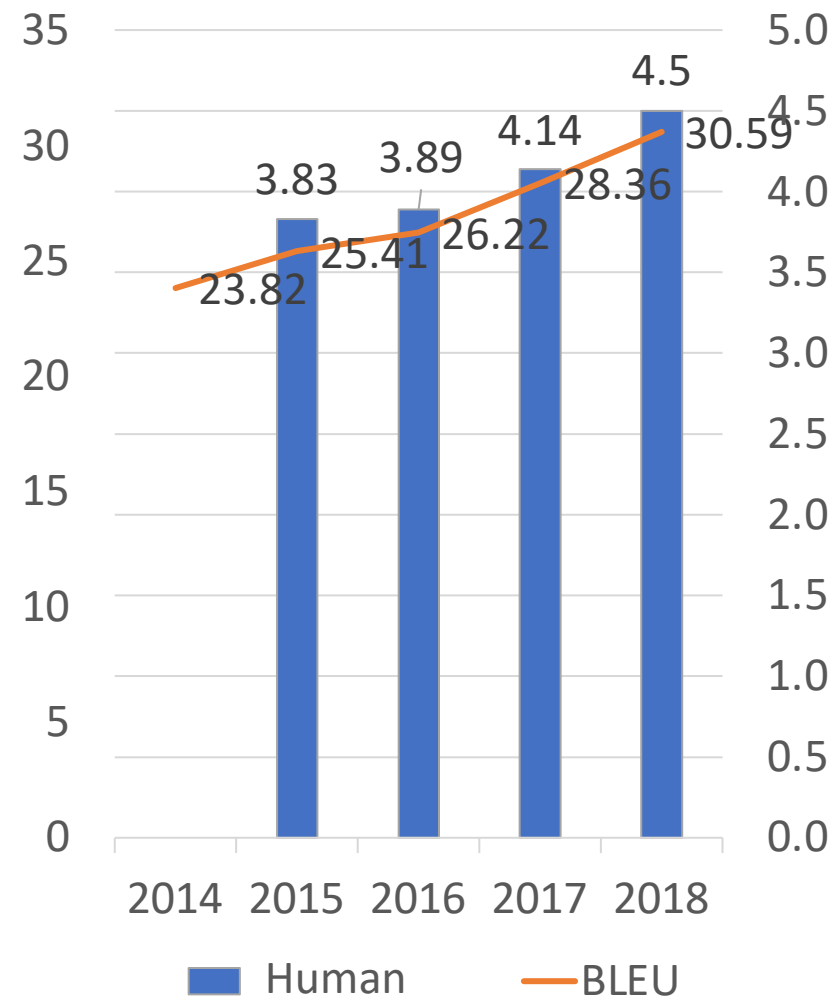
- **BLEU = BP · (p₁ · p₂ · p₃ · p₄)^{1/4}** [Papineni+ 2002]
 - p_n = common n-grams between MT and reference / n-grams in MT
 - Geometric mean of ratios that 1-grams, 2-grams, 3-grams, and 4-grams in MT are included in the reference
- **BP = e^{1 - length of reference / length of MT}**
 - Brevity penalty: penalty for short MT compared to the reference
 - Only applied to the case that the length of MT is shorter than that of the reference

Evaluation of MT: Human Evaluation

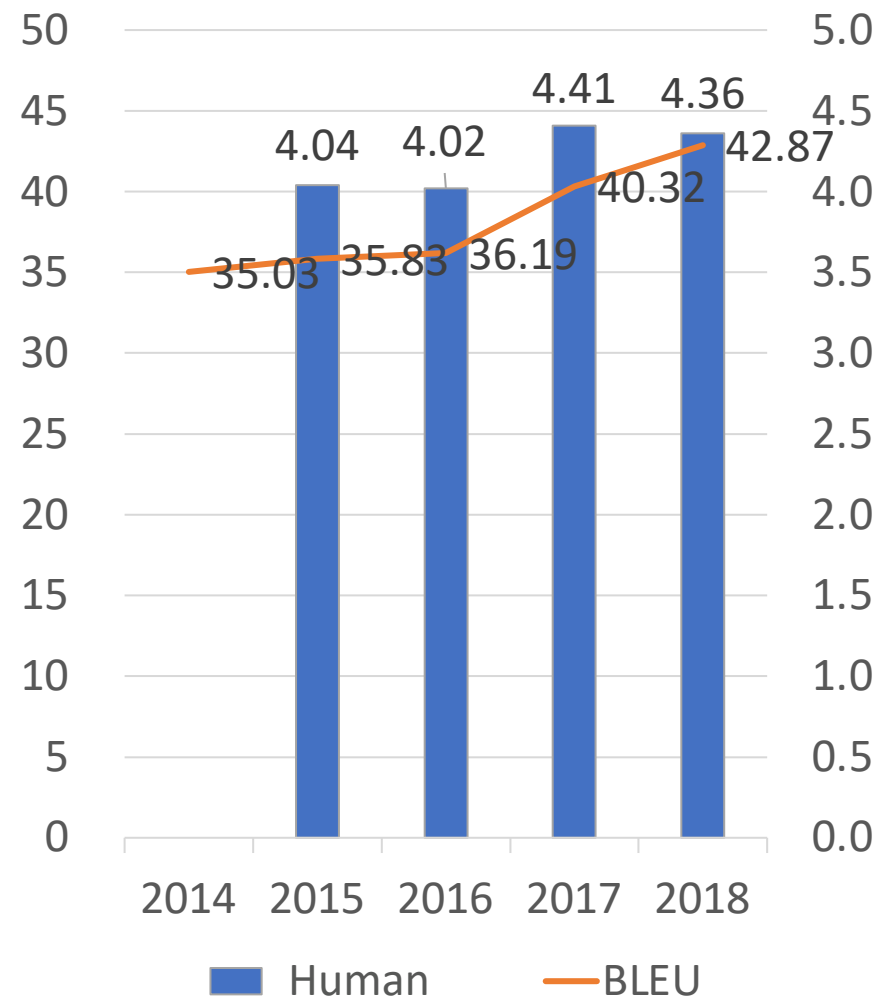
Accuracy evaluation by the Japan Patent Office: evaluation of the communication level of contents

Score	Evaluation Criteria
5	All the important information is accurately conveyed. (100%)
4	Most of the important information is accurately conveyed. (80%~)
3	More than half of the important information is accurately conveyed. (50%~)
2	Some of the important information is accurately conveyed. (20%~)
1	Hard to understand the meaning, or little important information is accurately conveyed. (~20%)

J→E Paper



E→J Paper

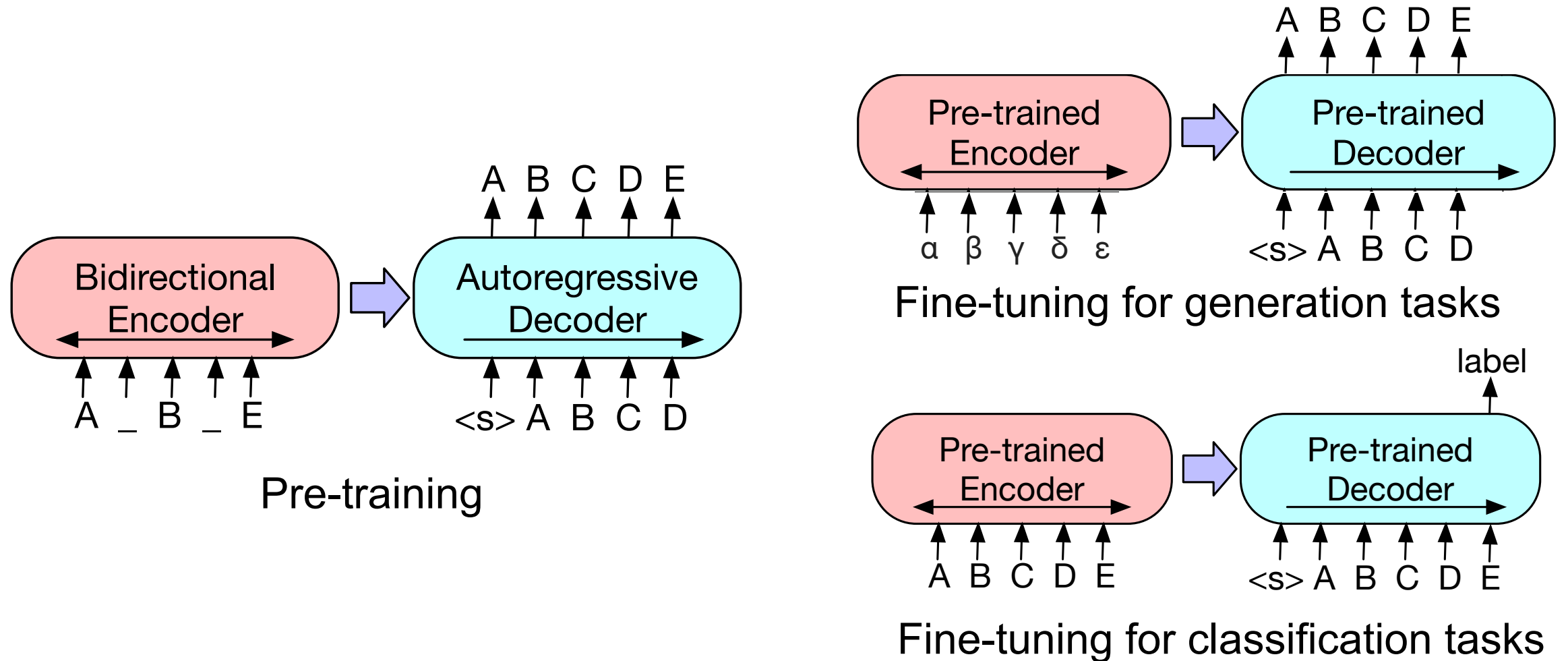


Recent Topics on NMT

- Back translation [Sennrich+ 2016]
 - To augment bilingual corpora, a monolingual corpus of the target language is automatically translated to the source language
- Unsupervised NMT [Lample+ 2018] [Artetxe+ 2018]
 - Trained only on large-scale monolingual corpora using crosslingual embeddings
 - Denoising autoencoders and back translation are used
- Multilingual NMT [Johnson+ 2016]
 - An NMT model is trained on large-scale bilingual corpora of more than 100 languages
- NMT based on pre-trained models [Lample+ 2019] [Lewis+ 2019] [Raffel+ 2019]

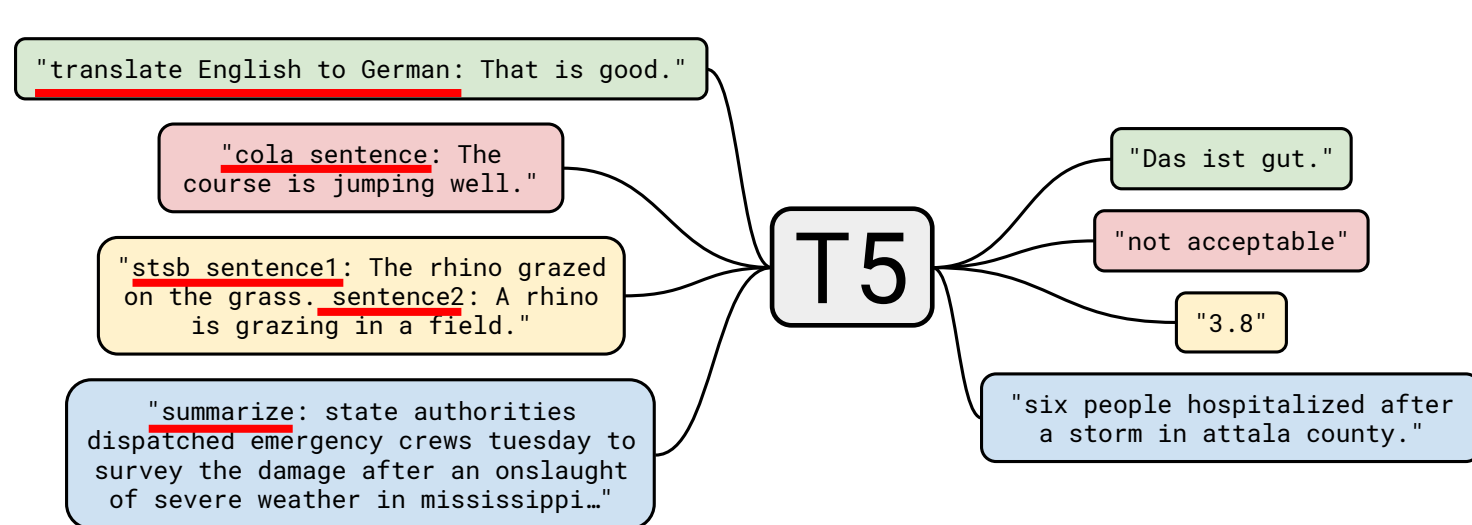
BART (Bidirectional and Auto-Regressive Transformers)

[Lewis+ 19]



Text-to-Text Transfer Transformer (T5) [Raffel+ 19]

- Classification and generation tasks are handled in a single model through a text-to-text format



Pre-training (on English)

Original text

Thank you ~~for inviting~~ me to your party last week.

Inputs

Thank you <X> me to your party <Y> week.

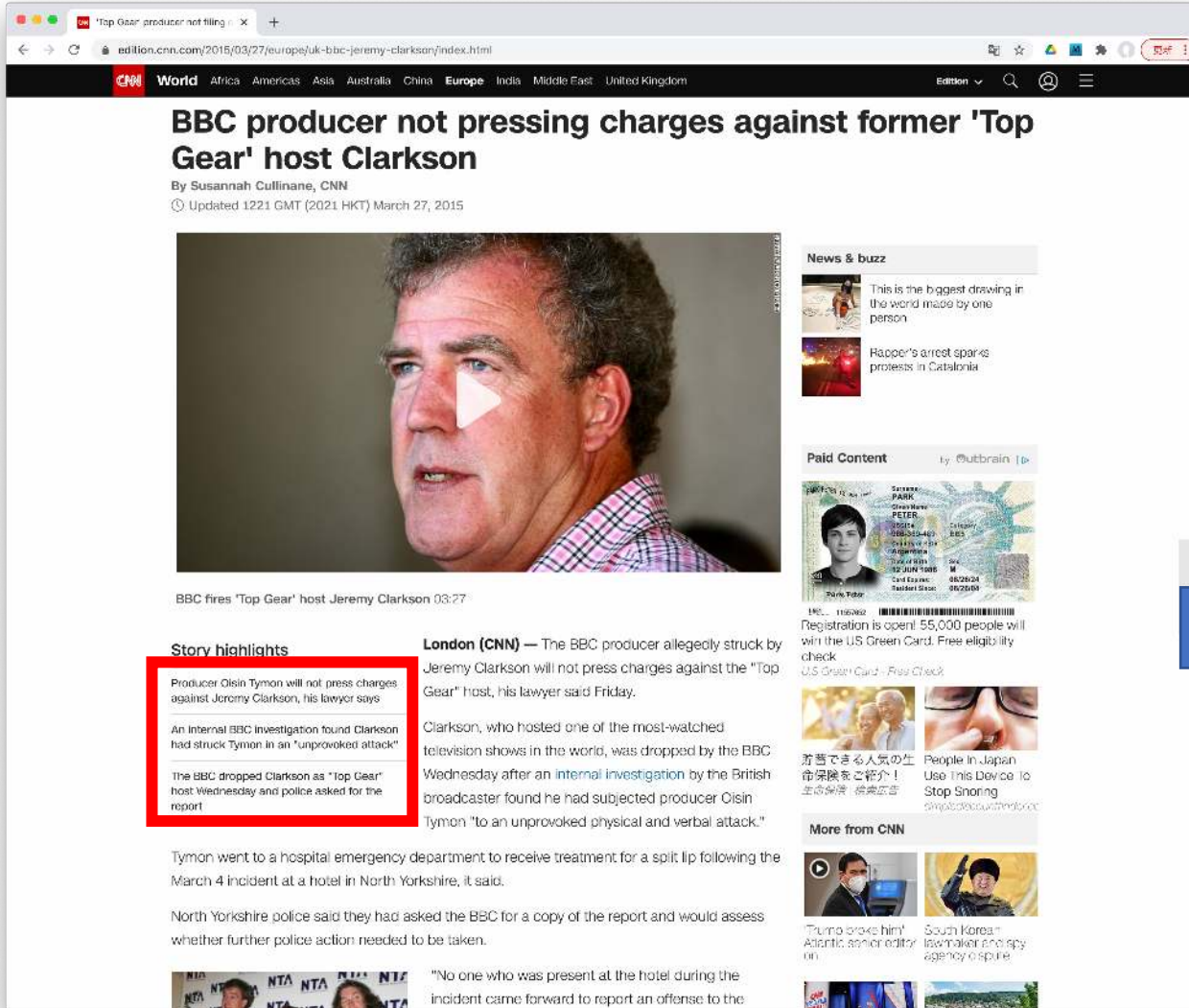
Targets

<X> for inviting <Y> last <Z>

Other Topics

- Automatic summarization
- Dialogue systems

CNN / Daily Mail [Hermann+ 2015] [Nallapati+ 2016] [See+ 2017]



Article

The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the “Top Gear” host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon “to an unprovoked physical and verbal attack.” ...

Summary

Producer Oisin Tymon will not press charges against Jeremy Clarkson, his lawyer says. An internal BBC investigation found Clarkson had struck Tymon in an “unprovoked attack”. The BBC dropped Clarkson as “Top Gear” host Wednesday and police asked for the report.

Privacy Policy | Feedback  Like 3.4M

Thursday, Feb 18th 2021 12PM 4°C  3PM 5°C  5-Day Forecast

MailOnline



Tennis

Home | News | U.S. **Sport** | TV&Showbiz | Australia | Femall | Health | Science | Money | Video | Travel | DailyMailTV | Discounts

Football | Premier League | FA Cup | Champions League | Transfer News | Boxing | Rugby | Golf | Cricket | F1 | **Tennis** | MMA | Racing | More [Login](#)

Andy Murray survives early scare against young hotshot Dominic Thiem to reach Miami Open semi finals

- British No 1 defeated Dominic Thiem in Miami Open quarter finals
- Andy Murray celebrated his 500th career win in the previous round
- Third seed will play the winner of Tomas Berdych and Juan Monaco in the semi finals of the ATP Masters 1000 event in Key Biscayne

Site ☐ Web ☐ Enter your search  Search

Like Daily Mail Sport Follow Daily Mail Sport

Follow @mailsport Follow Daily Mail

Follow @MailSport Follow Daily Mail

Get the Tennis RSS feed

Automatic summarization [Seo+ 2017]

andy murray defeated dominic thiem 3-6 6-4 , 6-1 in an hour and three quarters .

murray was awaiting the winner from the last eight match between tomas berdych and argentina 's juan monaco .

prior to this tournament thiem lost in the second round of a challenger event to soon-to-be new brit aljaz bedene .

Article

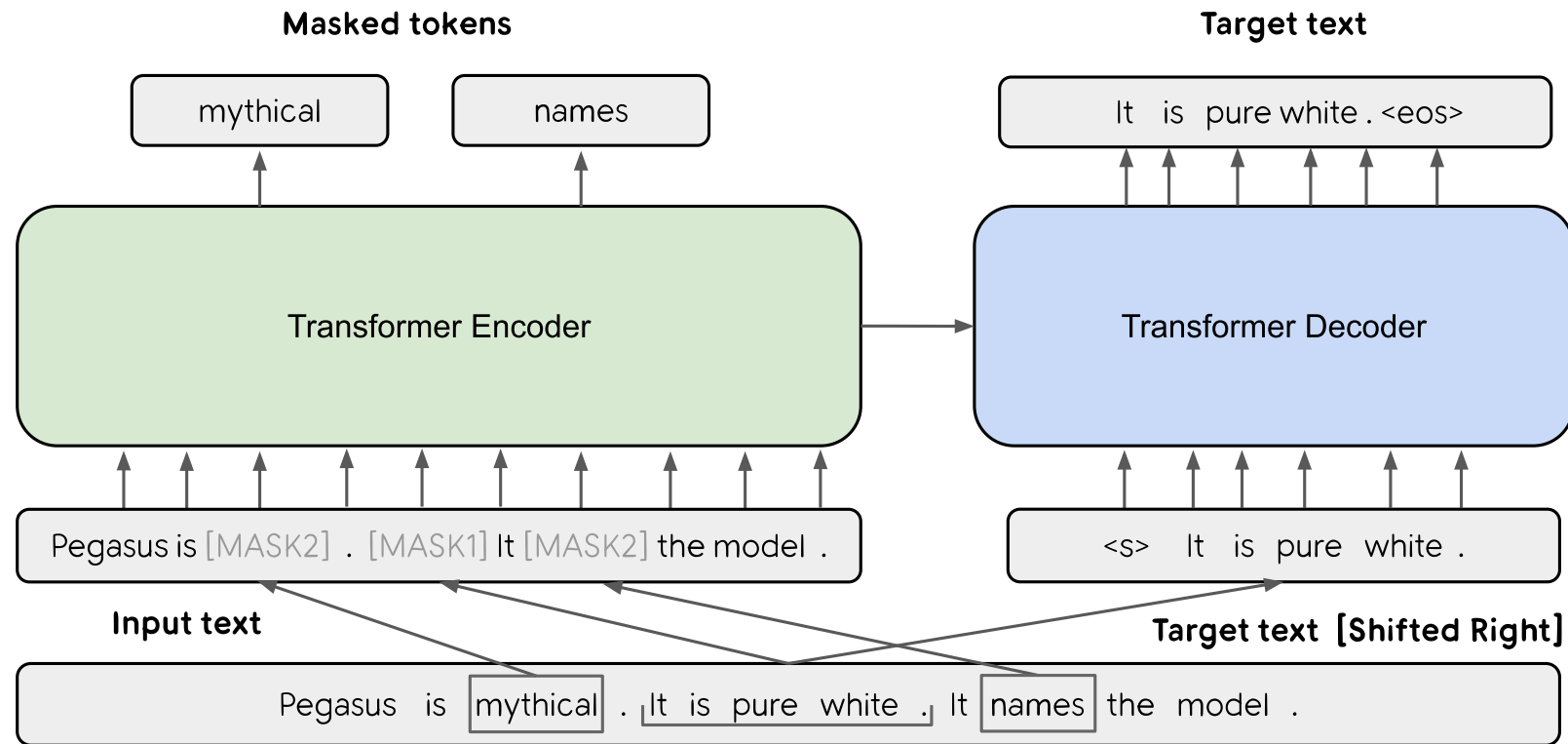
Andy Murray came close to giving himself some extra preparation time for his wedding next week before ensuring that he still has unfinished tennis business to attend to. The world No 4 is into the semi-finals of the Miami Open, but not before getting a scare from 21 year-old Austrian Dominic Thiem, who pushed him to 4-4 in the second set before going down 3-6 6-4, 6-1 in an hour and three quarters. Murray was awaiting the winner from the last eight match between Tomas Berdych and Argentina's Juan Monaco. Prior to this tournament Thiem lost in the second round of a Challenger event to soon-to-be new Brit Aljaz Bedene. ...

Summary

British No 1 defeated Dominic Thiem in Miami Open quarter finals. Andy Murray celebrated his 500th career win in the previous round. Third seed will play the winner of Tomas Berdych and Juan Monaco in the semi finals of the ATP Masters 1000 event in Key Biscayne.

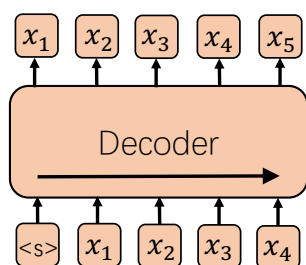
PEGASUS [Zhang+ 19]

- Pre-training to predict masked important sentences

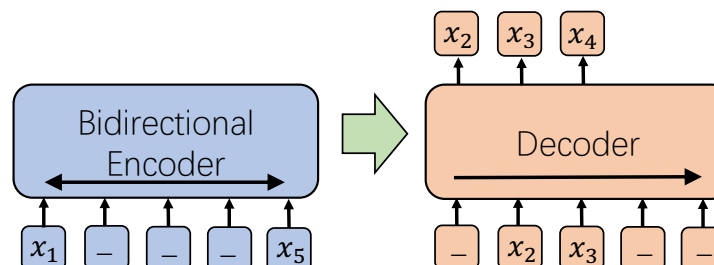


PALM [Bi+ 20]

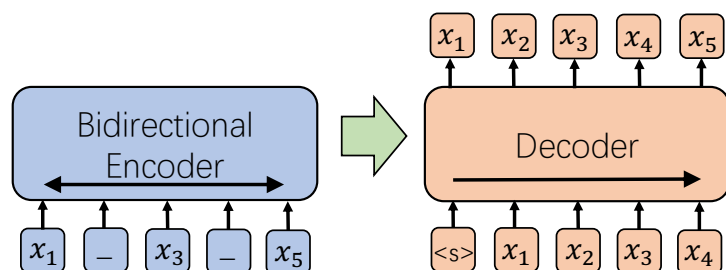
- Pre-training to predict subsequent text to the input text



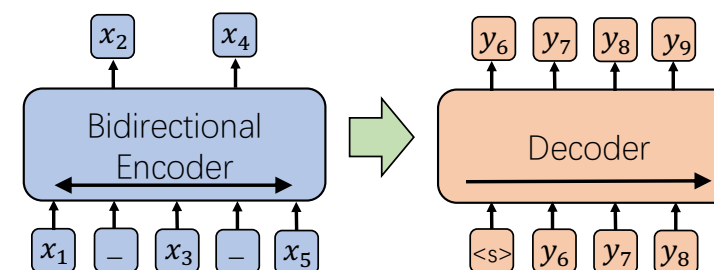
(a) GPT: Tokens are predicted autoregressively, meaning that GPT can be used for generation. However, it lacks an encoder to condition generation on context.



(b) MASS: It is based on the encoder-decoder architecture, but the decoder predicts only the tokens that are masked out in the text input to the encoder.



(c) BART: Rather than masked tokens, the decoder reconstructs the original full sentence from the corrupted input to the encoder. However, it mismatches with most downstream generation which is more than reconstructing original input.



(d) PALM: The encoder predicts masked tokens by encoding context bidirectionally, and the decoder predicts the text segment subsequent to the context. It forces the model to learn to comprehend the context for generating relevant text.

	CNN/DailyMail		
	RG-1	RG-2	RG-L
BERTSUMABS (Liu and Lapata, 2019)	41.72	19.39	38.76
MASS (Song et al., 2019)	42.12	19.50	39.01
UniLM _{LARGE} (Dong et al., 2019)	43.33	20.21	40.51
T5 _{LARGE} (Raffel et al., 2019)	42.50	20.68	39.75
BART _{LARGE} (Lewis et al., 2019)	44.16	21.28	40.90
PEGASUS (Zhang et al., 2019)	44.17	21.47	41.11
ERNIE-GEN _{LARGE} (Xiao et al., 2020)	44.02	21.17	41.26
PALM	42.71	19.97	39.71
PALM_{LARGE}	44.30	21.12	41.41

Summarization accuracy on the CNN/DailyMail test set

Dialogue Systems / Chatbots

- Intuitive easy-to-use interface for everyone
- Efficient information exchange



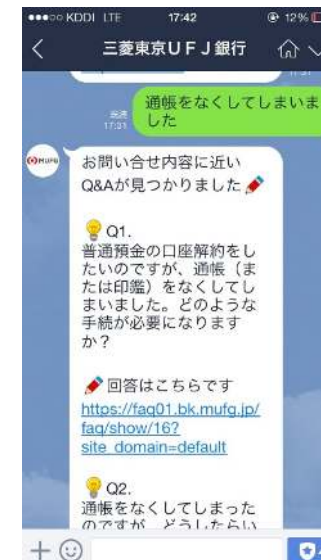
Siri (Apple)



Talking Concierge
(NTT Docomo)



Voice Assist
(Yahoo! JAPAN)



Mitsubishi Tokyo
UFJ BOT (LINE)



Recruit BOT
(LINE)

Neural Dialogue Systems

- Based on the encoder-decoder framework (cf. MT and summarization)
- Require a large volume of dialogue corpora
 - Reddit
 - Datasets for building human-like dialogue systems
 - Persona-Chat [Zhang+ 2018]
 - Empathetic Dialogues [Rashkin+ 2019]
 - Wizard of Wikipedia [Dinan+ 2019]
 - BlendedSkillTalk [Smith+ 2020]

Persona-Chat [Zhang+ 2018]

- Crowdsourced 1,155 personas
- Crowdsourced 10,981 dialogues based on personas (164,356 utterances)
- A larger corpus is available (ConvAI2)

Persona 1	Persona 2
I like to ski My wife does not like me anymore I have went to Mexico 4 times this year I hate Mexican food I like to eat cheetos	I am an artist I have four children I recently got a cat I enjoy walking for exercise I love watching Game of Thrones

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?

Empathetic Dialogues [Rashkin+ 2019]

- Crowdsourced 25k conversations grounded in emotion situations
 1. A speaker selects an emotion label and describes a situation when they felt that way
 2. The speaker has a conversation with a listener about the situation

Label: Afraid

Situation: Speaker felt this when...

“I’ve been hearing noises around the house at night”

Conversation:

Speaker: I’ve been hearing some strange noises around the house at night.

Listener: oh no! That’s scary! What do you think it is?

Speaker: I don’t know, that’s what’s making me anxious.

Listener: I’m sorry to hear that. I wish I could help you figure it out

Label: Proud

Situation: Speaker felt this when...

“I finally got that promotion at work! I have tried so hard for so long to get it!”

Conversation:

Speaker: I finally got promoted today at work!

Listener: Congrats! That’s great!

Speaker: Thank you! I’ve been trying to get it for a while now!

Listener: That is quite an accomplishment and you should be proud!

Wizard of Wikipedia [Dinan+ 2019]

- Crowdsourced 22k conversations grounded with knowledge retrieved from Wikipedia (202k utterances)

Topic:	Ice cream
Wizard:	I just love ice cream. I love the types with fruits and flavours. Do you like ice cream?
Apprentice:	I love Ice cream as much as any one. I especially like Gelato, foreign ice cream!
Knowledge	<p>Ice cream is a sweetened frozen food typically eaten as a snack or dessert.... It is usually made from dairy products, such as milk and cream, and ...</p> <p>...</p> <p>Bacon ice cream (or bacon-and-egg ice cream) is an ice cream generally created by adding bacon to egg custard and freezing the mixture.</p>
Wizard:	Me too. There are some strange combinations though, have you heard of bacon ice cream? where they add bacon and even egg custard to the freezing mixture!
Apprentice:	Surprisingly bacon ice cream doesn't surprise me. That doesn't sound appealing to me, but perhaps it could be delicious...

BlendedSkillTalk [Smith+ 2020]

- Building a dialogue system equipped with three skills: persona, empathy, and use of knowledge
- Crowdsourced 5k conversations for evaluation and fine-tuning

Persona for Unguided Speaker:

My son plays on the local football team.
I design video games for a living.

Persona for Guided Speaker:

My eyes are green.
I wear glasses that are cateye.

Wizard of Wikipedia topic: Video game design

Previous utterances (shown to speakers):

U: What video games do you like to play?

G: all kinds, action, adventure, shooter, platformer, rpg, etc. but video game design requires both artistic and technical competence AND writing skills. that is one part many people forget

Actual utterances:

U: Exactly! I think many people fail to notice how beautiful the art of video games can be. (PB)

(G selected the WoW suggestion: "Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics.")

G: Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics. (K)

U: Video games are undervalued by many and too easily blamed for problems like obesity or violence in kids (K)

G: Indeed, Just last week my son was playing some Tine 2 and it was keeping him so calm.

Games are therapeutic to some. (S)

U: I use games to relax after a stressful day, the small escape is relaxing. (PB)

(G selected the ED suggestion: "I enjoy doing that after a hard day at work as well. I hope it relaxes you!")

BlenderBot [Roller+ 2020]

- Transformer with 2.7B parameters
 - Encoder layers: 2
 - Decoder layers: 24
 - Attention heads: 32
 - Embedding dimension: 2,560
- Pre-training: Reddit 1.5B comments
- Fine-tuning: ConvAI2, ED, WoW, BST
- Engagingness evaluation:

	Model vs. Human	
Meena (Adiwardana et al., 2020)	28 **	72 **
BST Generative (2.7B) std. beam	21 **	79 **
pushshift.io Reddit Generative (2.7B)	36 **	64 **
BST RetNRef (256M/90M)	37 **	63 **
BST Generative (90M)	42	58
BST Generative (9.4B)	45	55
BST RetNRef (622M/2.7B)	46	54
Wiz Generative (2.7B)	47	53
BST Unlikelihood (2.7B)	48	52
BST Generative (2.7B)	49	51

Crowdworker

Generative BST 2.7B

Hi!

What kind of music do you like? I like all kinds of music, but country is my favorite.

I enjoy hard rock

That's cool. Do you have a favorite band? Mine is led zeppelin.

Led Zeppelin is one of my favorites too! Also Pink Floyd

I like them too. My favorite album of theirs is dark side of the moon.

Thats one of the greatest albums of all time

It really is. I can listen to it over and over and never get tired of it.

Ever watched it with Wizard of Oz?

No, I haven't. Is it good? I know it was made in 1939.

The music syncs up with the movie, pretty cool

I'll have to check it out. Thanks for the recommendation. Do you have a favorite movie?

Tombstone, how about you?

My favorite movie is the godfather part ii. Have you seen it? It's one of my favorites.

Project

- You can choose one from:
 - Survey
 - Make a survey of very recent trends in an NLP topic
 - e.g., parsing, transfer learning, machine comprehension, question answering
 - Programming
 - Make a model for an NLP task (e.g., BERT-based models)
 - Implement a baseline model proposed in a paper and add one or more ideas to improve the baseline model
 - Use Google Colab, which provides free use of a GPU/TPU for 12 hours, if you need computing power
- You can do it by yourself or make a team consisting of up to 3 members
 - If you make a team, it is necessary to clarify the role of each member.
- Final presentation on July 22: 3 minutes for each person
- Report: due on July 31 23:59