

Natural Language Processing

Overview

Daisuke Kawahara

Department of Communications and Computer Engineering,
Waseda University

Language

- Thinking, communication, and recording
- To give a name to a thing and a concept, and describe their relations

Language and Computer

- E-mail
- Text input (kana-kanji conversion), spell checker
- Information retrieval / search
- Machine translation
- Dialogue system / agent system
- ...
- Everywhere in SF

NICT: VoiceTra



<https://youtu.be/1w8yFkoPmxI>

Natural Language Processing (NLP)

- A technology for processing natural language using computers
- Elucidate the mechanisms by which humans understand language
- Also called “Computational Linguistics”

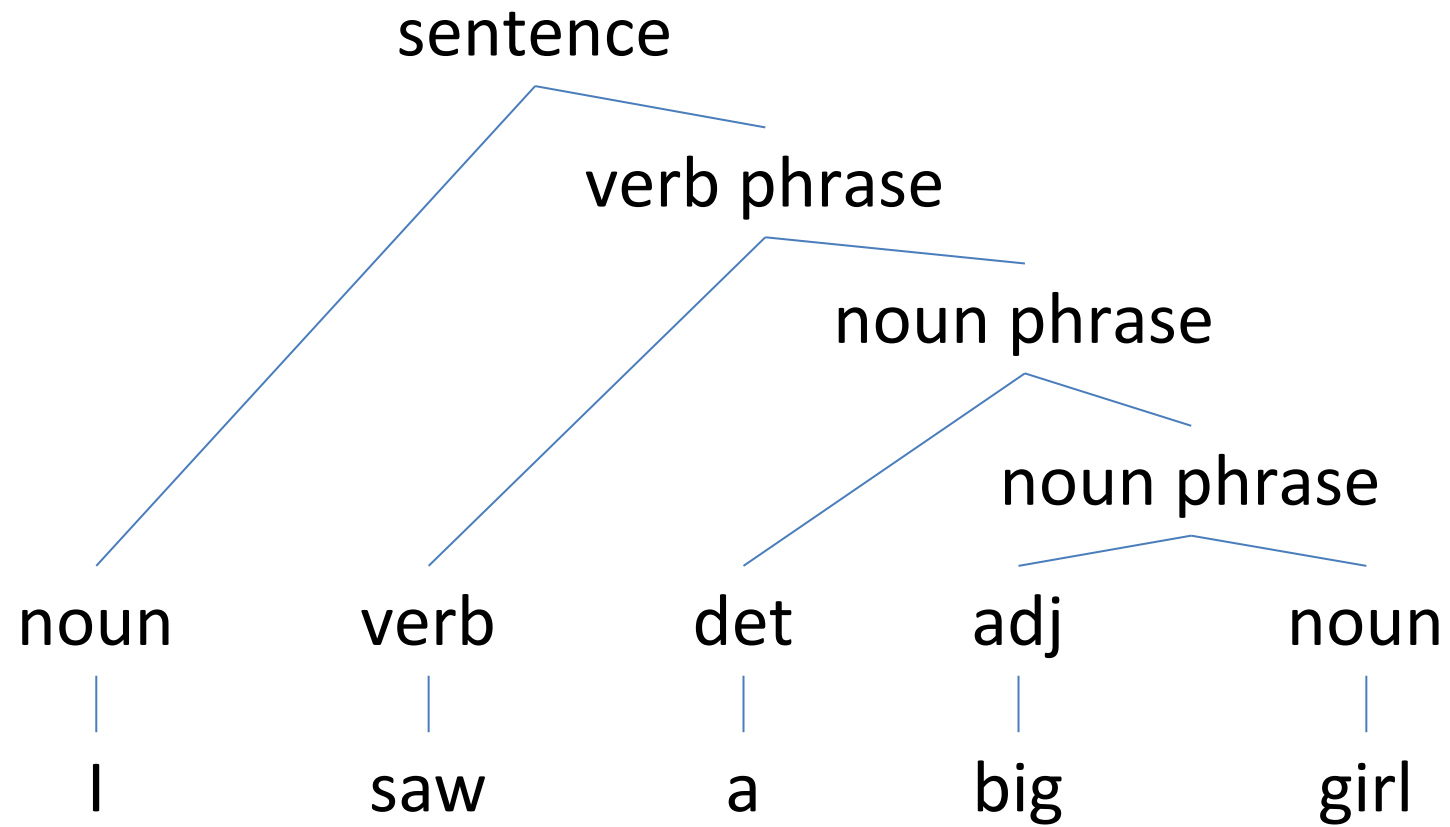
Word

- Content words / open class words
内容語 (自立語)
e.g., nouns, verbs, adjective, adverbs
- Function words / closed class words
機能語 (付属語)
e.g., prepositions, postpositions, determiners, copula (“be”, “だ”)
- Part of speech / 品詞
- Vocabulary / 語彙 (a set of words, a dictionary, a lexicon)

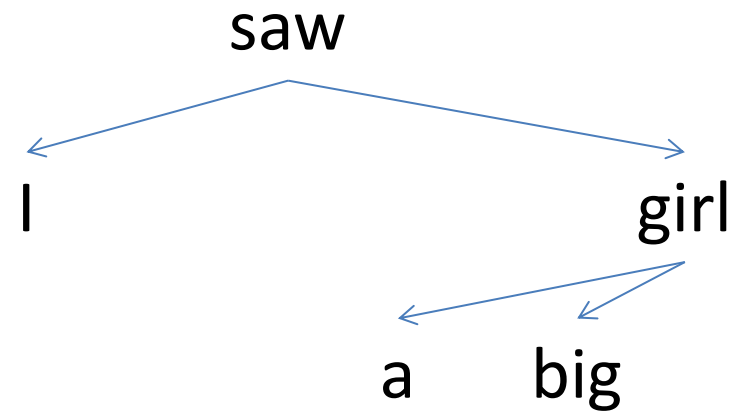
Multi-word Expression / 複合語

- Technical terms / 専門用語
 - e.g., “Natural Language Processing”
“Fast Fourier Transform”
- Proper nouns (proper names, named entities) / 固有名詞
 - e.g., “Waseda University”
“Toyota Motor Corporation”

Phrase Structure / 句構造



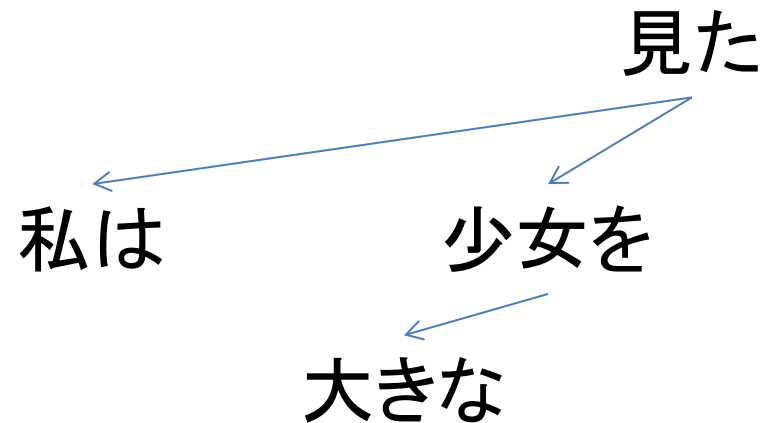
Dependency Structure / 依存構造



governor
head, parent



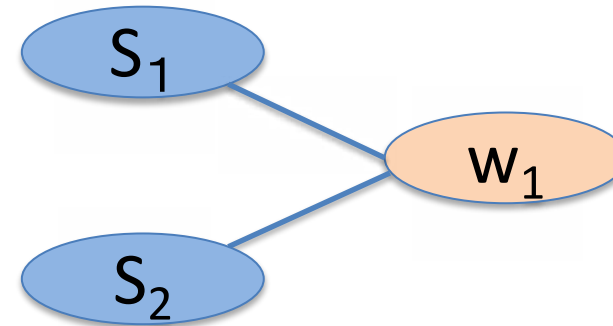
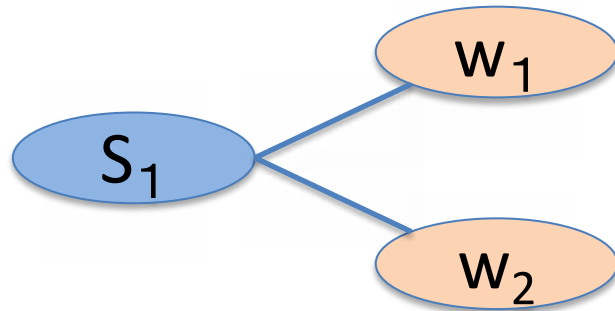
dependant
modifier, child



Problems in NLP

- Problems with clear definitions/answers
 - Ambiguity
 - Synonymy and polysemy
同義性 多義性
 - Syntactic ambiguity

Synonymy and Polysemy



Synonyms

- Spelling variations
 - center, centre
 - 林檎, りんご, リンゴ 
- Different words (synonym ... near synonym)
 - apple, アップル, 林檎 (translation)
 - AI, Artificial Intelligence (acronym)
 - helium, He; meeting, mtg (abbreviation)
 - big, large

Homonyms / Polysemic Words



bank



interest



Homonyms / Polysemic Words

- homonym

- *bank*:

1. The banks of a river, canal, or lake are the **raised areas of ground** along its edge.
2. A bank is an **institution** where people or businesses can keep their money.

- polysemic words

- *interest*:

1. If you have an interest in something, you **want to learn or hear more** about it.
2. Interest is **extra money** that you receive if you have invested a sum of money.

[Collins COBUILD]

Ambiguity in Natural Language

- Time flies like an arrow. (光陰矢のごとし)
noun verb pp det noun
- Time flies like an arrow. (時蠅は矢を好む)
noun noun verb det noun

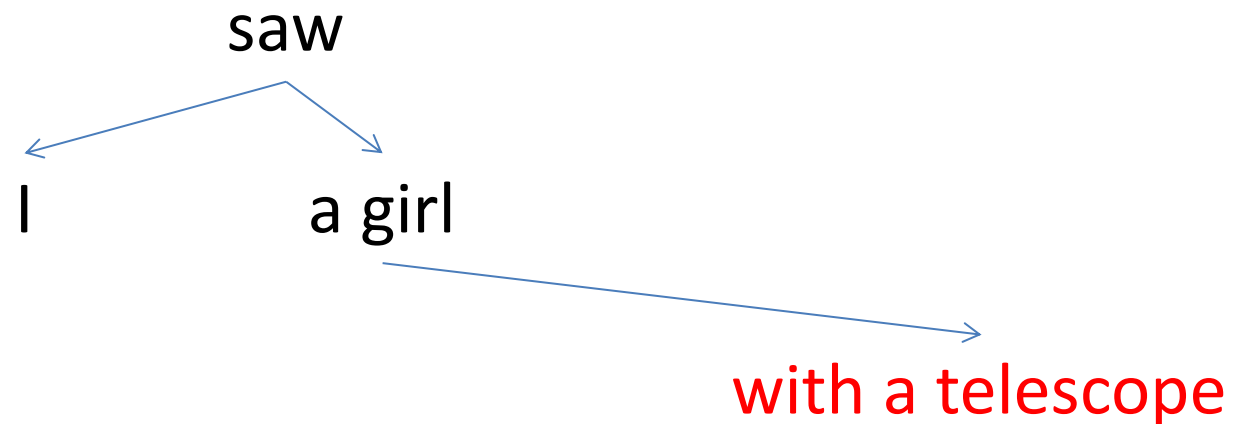
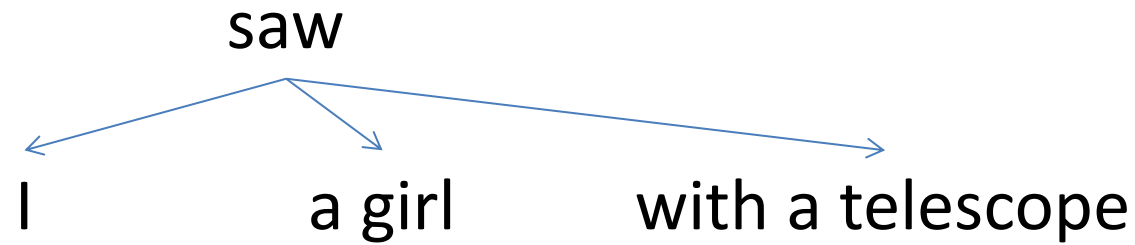
Ambiguity in Natural Language

外国人参政権

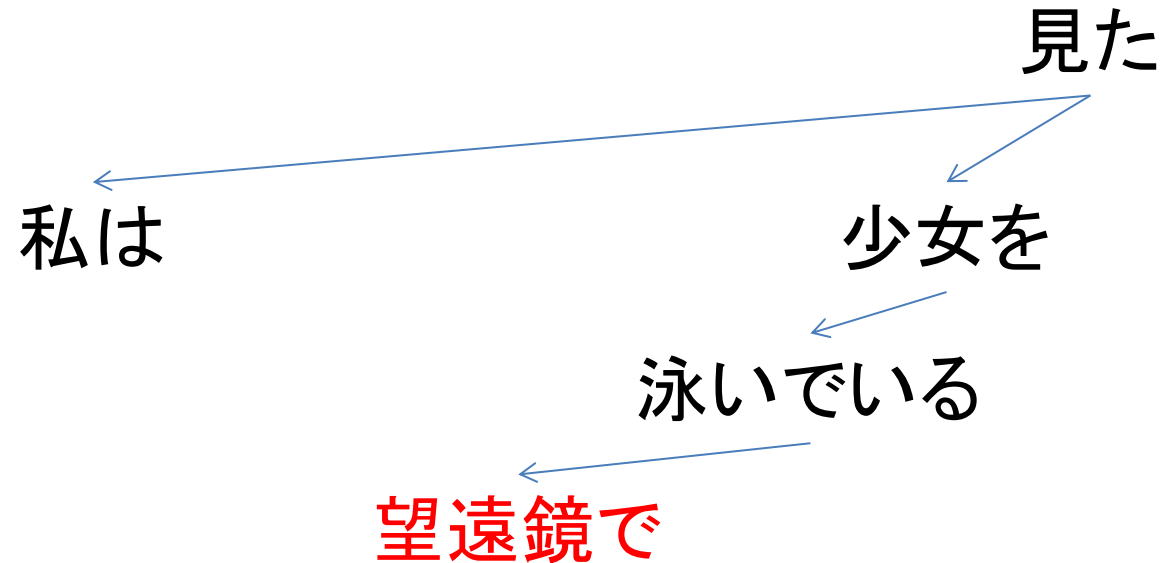
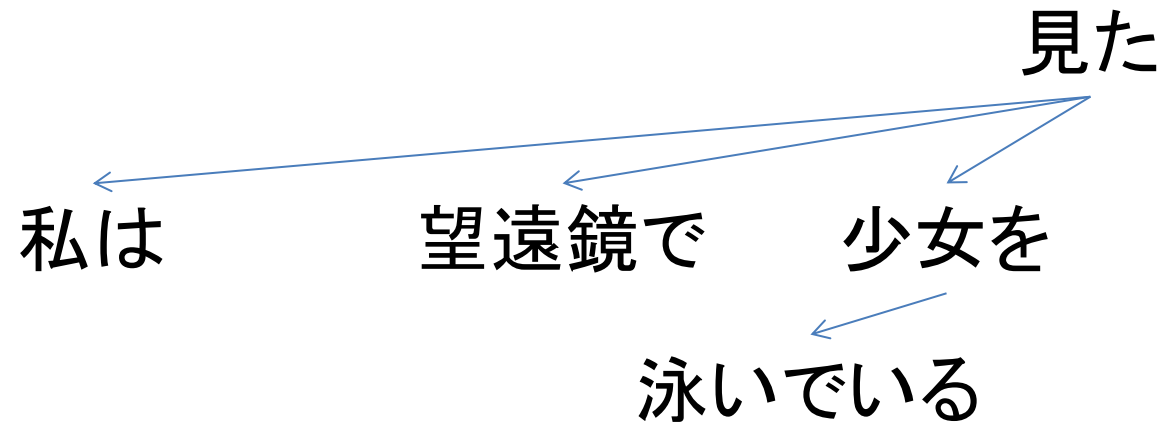


※ Current translation is correct.

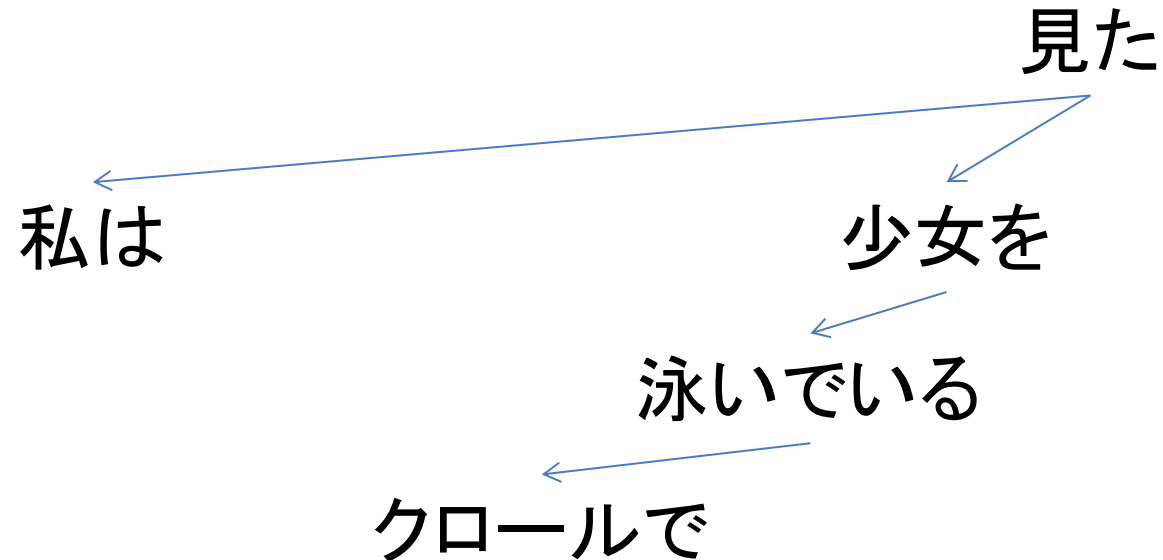
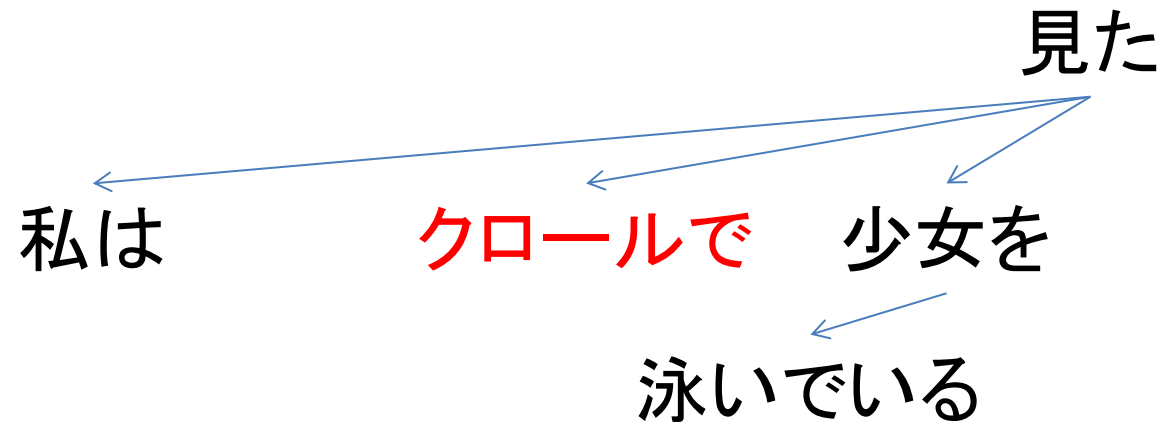
Ambiguity in Natural Language



Ambiguity in Natural Language



Ambiguity in Natural Language



Problems in NLP

- Problems with clear definitions/answers
 - Ambiguity
 - Synonymy and polysemy
同義性 多義性
 - Syntactic ambiguity
 - How to find and combine clues from the context
文脈

CHAT

THAE

.....

.....

文脈 (context)

Problems in NLP

- Problems with clear definitions/answers
 - Ambiguity
 - Synonymy and polysemy
同義性 多義性
 - Syntactic ambiguity
 - How to find and combine clues from the context
文脈
- Problems without clear definitions
 - Meaning
 - What is understanding?
 - Creative usage: metaphor, metonymy
隱喻 換喻

Metaphor / Metonymy

- Metaphor

- How can I kill a process? [Martin 88]
- My car drinks gasoline. [Wilks 78]
- He shot down all of my arguments. [Lakoff & Johnson 80]
- He is a big star.



- Metonymy

- Washington and Tokyo agree on ...
- The ham sandwich is waiting for his check. [Lakoff & Johnson 80]
- Japanese people often eat nabe in winter.



History of Natural Language Processing

Early Days

- 1947
English-French machine translation (MT)
(W. Weaver, 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**."*

- 1954
Russian-English MT
(Georgetown Univ. and IBM)
- 1955-
 - English-Japanese machine translation (電気試験所)
 - English-German-Japanese MT (Kyushu Univ.)

Dark Ages of MT

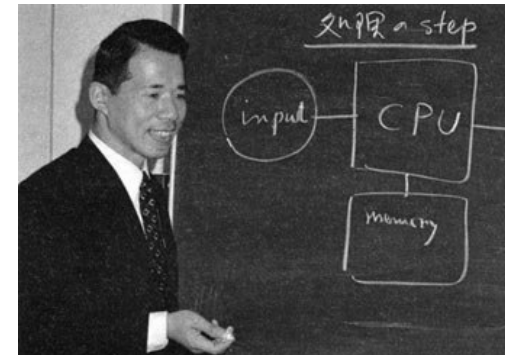
- ALPAC Report (Automatic Language Processing Advisory Committee, 1966)
 - High-quality MT is not easy to develop soon
 - More fundamental research is needed, such as computational linguistics (計算言語学)

Introduction of Meaning

- C. Fillmore: Case Grammar (格文法), 1968
 - John broke the window with a hammer.
 - The window broke.
 - A hammer broke the window.
 - Agent: John
 - Object: window
 - Instrument: hammer
- Selectional restriction by semantic features

MT Projects

- EUROTRA Project (1982-1991)
 - By the European Commission
 - An MT system for the seven (later nine) official languages of the European Community
- Mu Project (1982-1985)
 - By Science and Technology Agency (科学技術庁) and Kyoto Univ.
 - Japanese-English and English-Japanese MT



Prof. Makoto Nagao

Difficulty in Translation

- Lexicon
 - put on X \Leftrightarrow 帽子をかぶる, 服を着る, 靴を履く
- Word order
 - He wears a suit (SVO) \Leftrightarrow 彼はスーツを着ている (SOV)
- Structure
 - She has a long hair \Leftrightarrow 彼女は髪が長い
- Number, gender, definiteness

Data-driven machine translation

- Example-based machine translation [Nagao 1984]
- Statistical machine translation [Brown+ 1993]

Artificial Intelligence (1955-)

- ELIZA [J. Weizenbaum 1966]
 - Dialogue system
 - No deep understanding of input texts
- SHRDLU [T. Winograd 1972]
 - Operation of blocks in a virtual world
 - Detailed knowledge about the (small) domain

ELIZA: DOCTOR Script (M-x doctor on Emacs)

User: Men are all alike.

ELIZA: In what way?

User: They're always bugging us about something or other.

ELIZA: Can you think of a specific example?

User: Well, my boyfriend made me come here.

ELIZA: Your boyfriend made you come here.

ELIZA: DOCTOR Script

Word:Rank

Outputs

alike:10

In what way?

What resemblance do you see?

always:5

Can you think of a specific example?

Really, always?

what:2

Why do you ask?

Does that interest you?

SHRDLU (1972)

Person: PICK UP A BIG RED BLOCK.

Computer: OK. (does it)

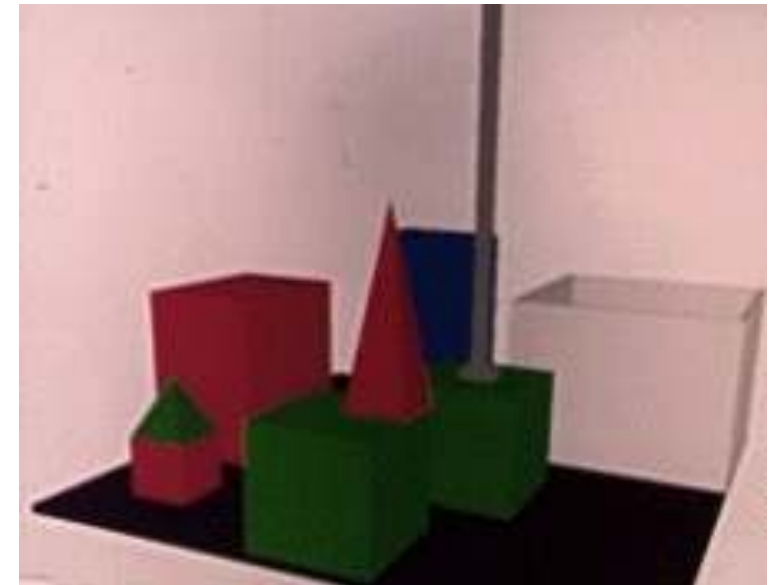
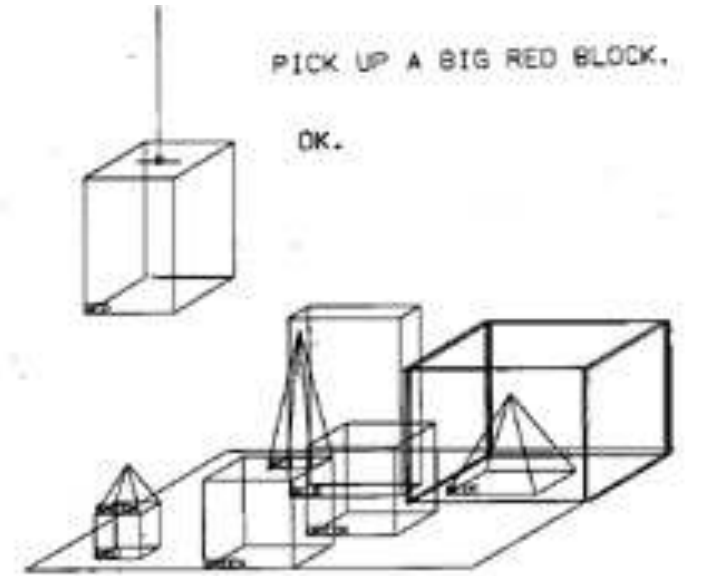
Person: GRASP THE PYRAMID.

Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.

Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.

Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.

Computer: OK. (does it)



Dialogue Systems / Chatbots

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



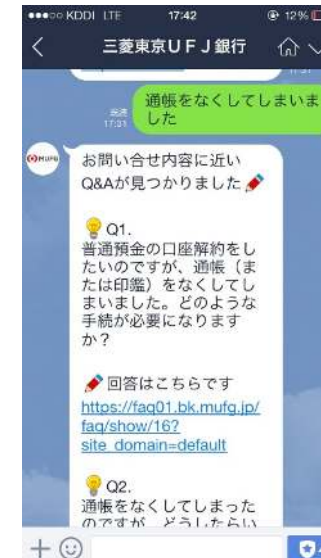
Siri (Apple)



Talking Concierge
(NTT Docomo)



Voice Assist
(Yahoo! JAPAN)



Mitsubishi Tokyo
UFJ BOT (LINE)



Recruit BOT
(LINE)

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



Machine Comprehension

The Stanford Question Answering Dataset (SQuAD) [Rajpurkar+ 2016]

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, ...

Q: What is the second most abundant element?

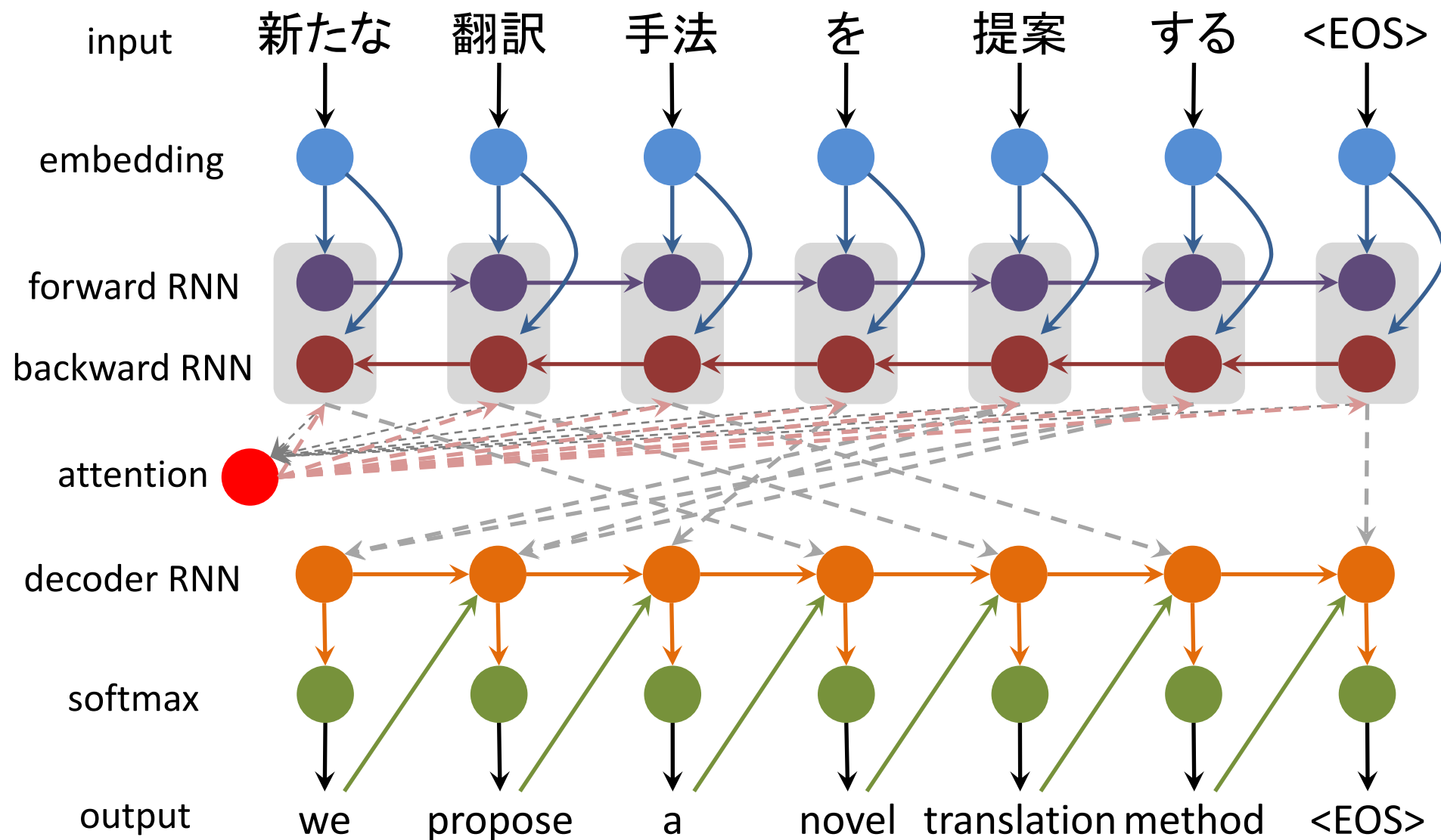
A: helium

SQuAD1.1 Leaderboard <https://rajpurkar.github.io/SQuAD-explorer/>

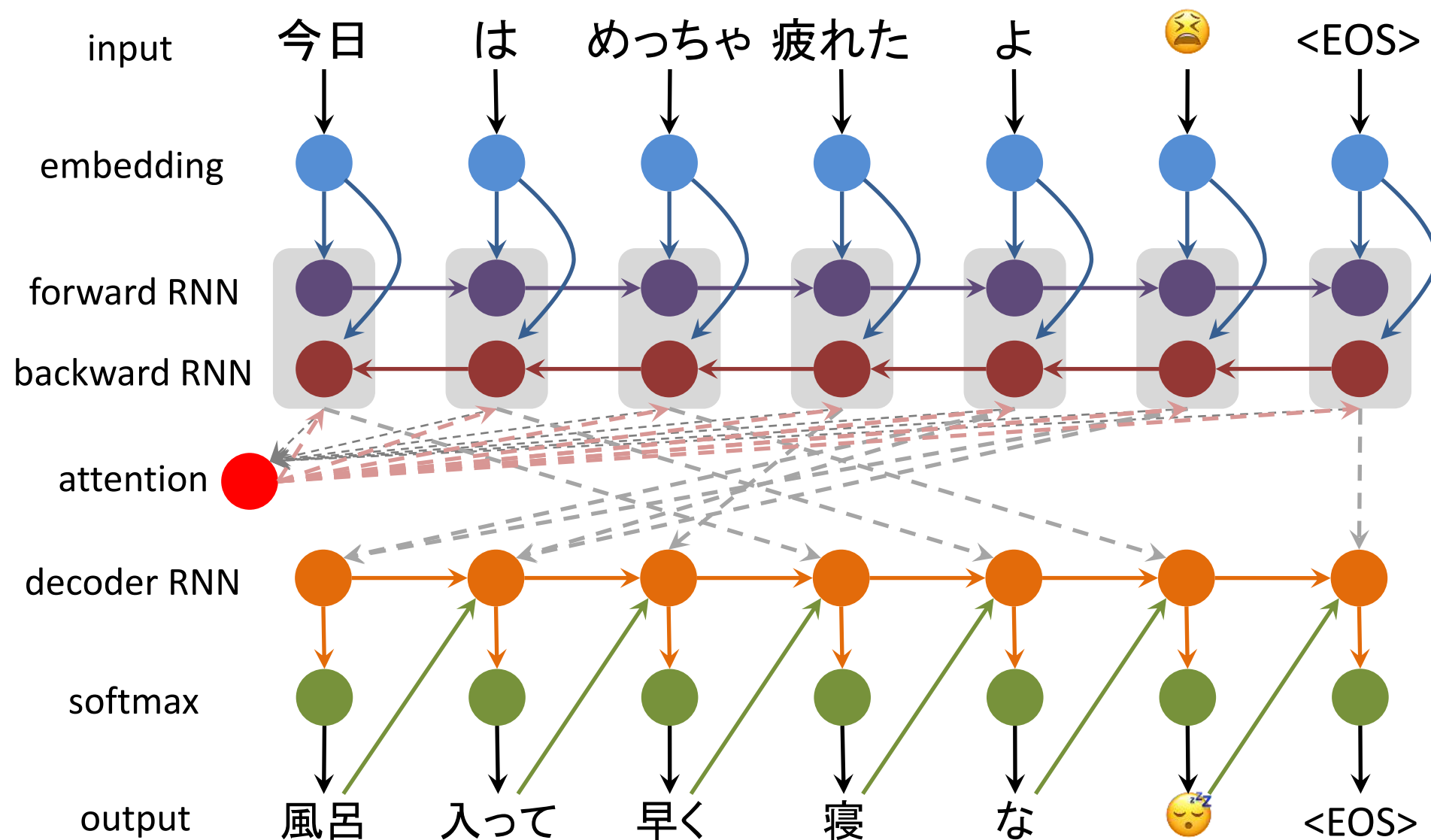
Here are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

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

End-to-end MT [Bahdanau+ 2014]



Neural Chatbot



BERT [Devlin+ 2019]

- Model: Transformer [Vaswani+ 2017]
- Training: two steps (pre-training and fine-tuning)

1. Pre-training

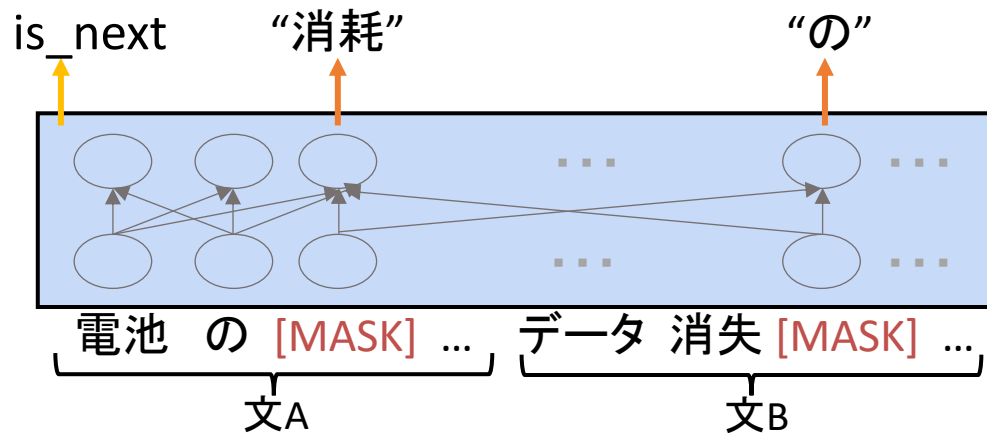
Learn vector representations using a large-scale raw corpus

2. Fine-tuning

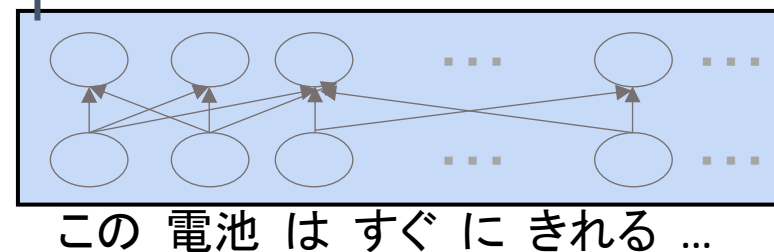
Fine-tune parameters for a specific task (e.g., sentiment analysis)

Next sentence prediction

Masked language model



negative



NLP is exciting now!

- 20 years ago
 - Humans wrote grammars and analysis rules
 - Real texts contain many exceptional (ungrammatical) expressions
 - Deadlock between language understanding and knowledge acquisition
- Now
 - Empirical / data-driven
 - No knowledge
 - Big text data (~ 10 -20 years)
 - No method for using knowledge
 - Deep neural networks (~ 5 years)