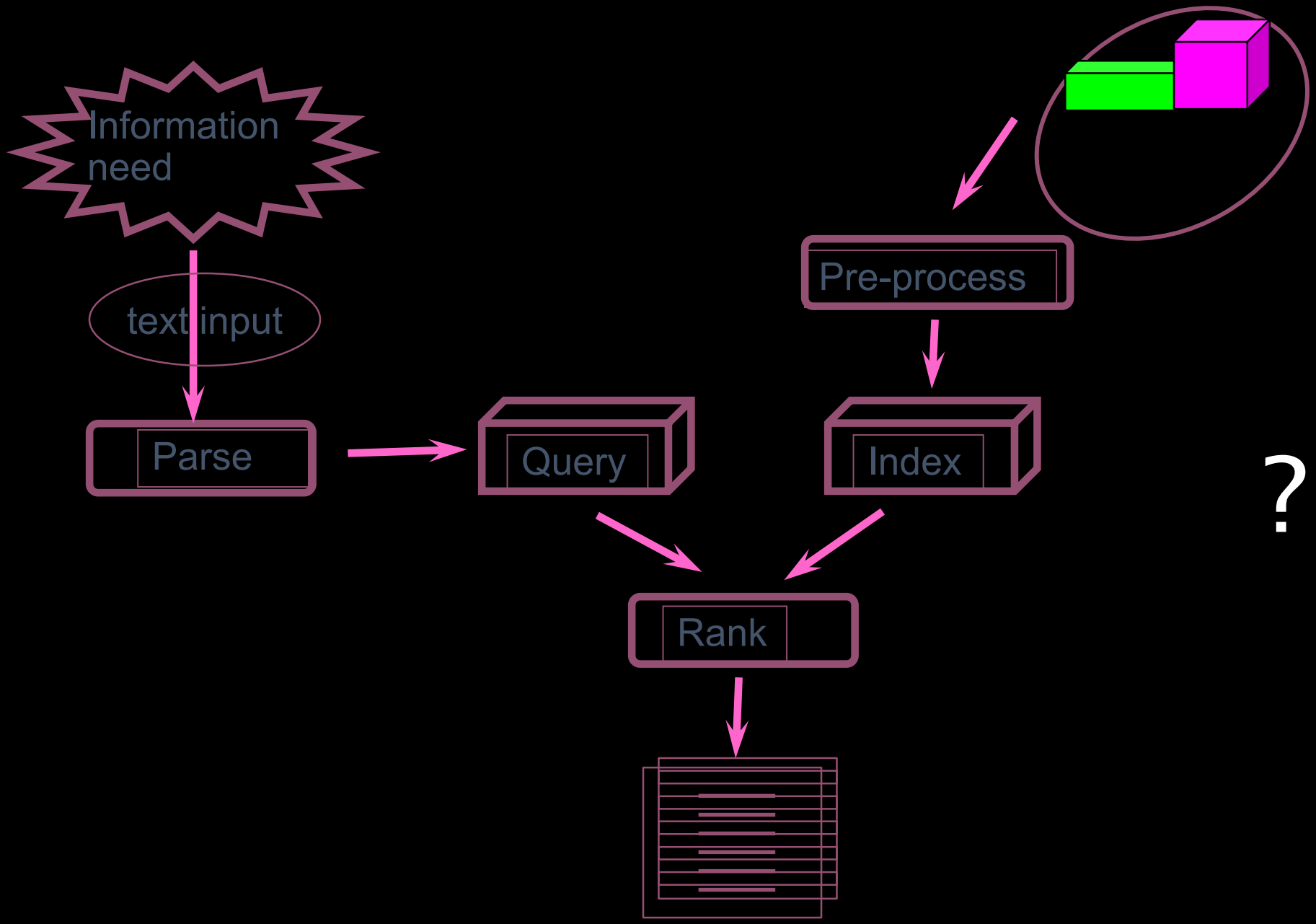
Lecture 6
part 1
Index Construction and Search

Index construction

- How do we construct an index?
- What strategies can we use with limited main memory?


Information need
text input
Parse
Query
Pre-process
Index
Rank
?

Why Index?

- Scan the entire collection
  - Use in early IR
  - Computational cost
  - For small collections only
- Search the indexes for direct access
  - An index associate a document with one or more keys
  - Practical for large collections
- Hybrid method
  - Use small index, then scan a subset of the collection

Overview

- Thesaurus
  - Roget's Thesaurus, Astronomy Thesaurus,…
- Semantic Network
  - WordNet
- Co-occurrence
  - Automatic relevance feedback
  - Local context analysis (LCA)

Thesaurus

- Dictionary(字典):
  - Offer -> presentation, tender, overture, submission, proposal, invitation, ….
  - Refusal -> declining, rejection, denial, …
- Lexicon(語彙典):
  - Asia -> Japan, China, India, Taiwan, …
  - Computer -> software, hardware, disk, operating system, CD-ROM, …

Thesaurus

- Insert query term synonyms into query
  - Automatically
    - Problem: Can introduce words with several unrelated senses
  - Manually
    - Problem: People often find it difficult to select synonyms
- Query expansion with general thesauri has not been consistently useful
- Query expansion with subject-specific thesauri is more successful, especially with trained users
  - Example: MeSH terms

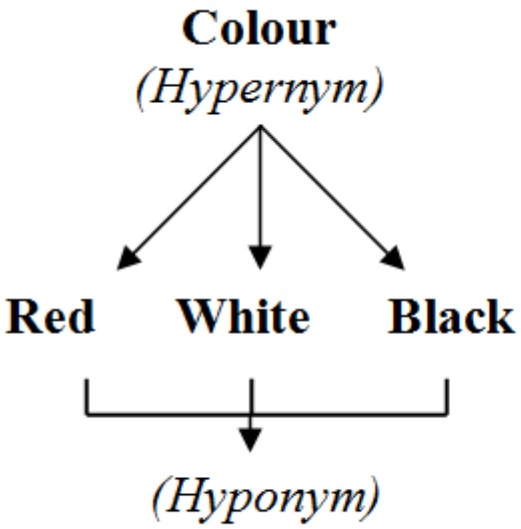
WordNet

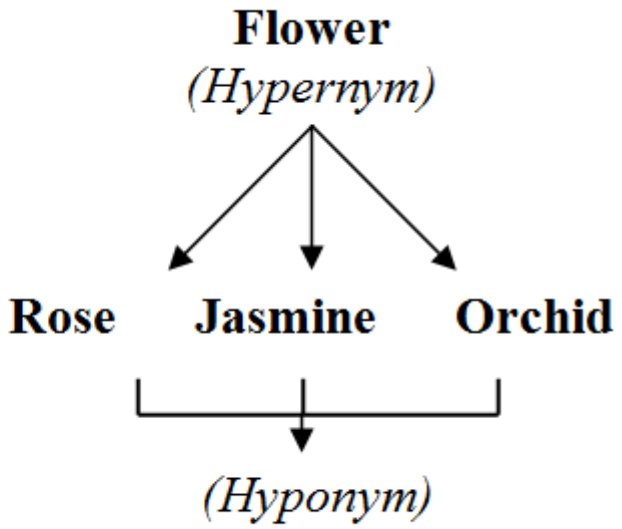
- A lexical thesaurus organized into 4 taxonomies by part of speech
- Created by George Miller & colleagues at Princeton University
- Inspired by psycholinguistic theories of human lexical memory
- English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept

WordNet

- Different relations link the synonym sets
  - Hyponyms: "...is a kind of X" relationships
  - Hypernyms: "X is a kind of ..." relationships
  - Meronyms: "parts of X" relationships
- “air plane”
  - jet is an airplane powered by jet engines
  - airplane is a vehicle that can fly
  - HAS PART: accelerator, accelerator pedal, gas pedal, ...

Examples





Meronymy
A term that is used to describe a part-whole relationship between lexical items. A has B means that B is part of A.

– A human has an arm
– An arm has a hand.
– A hand has a finger

So, (arm, leg, body, elbow, hand, finger) are all meronyms of human. Cover, and page are meronyms of book, root and stem are meronyms of a plant.

WordNet for IR

- User selects synsets (synonym sets) for some query terms
  - Add to query all synonyms in synset
  - Add to query all hyponyms ("... is a kind of X")
- Query expansion with WordNet has not been consistently useful
  - Possibly because they are domain-independent
  - Possibly too detailed in some areas, not enough detail in others

中文的世界--知網 HowNet

- 知網是一個以中文和英文的詞語所代表的概念為描述物件，以表達概念與概念之間以及概念所具有的屬性之間的關係為基本內容的知識庫
- 知網是董振東先生、董強先生父子標注的大型語言知識庫，包括中文（英文）的詞彙與概念
- 中央研究院 資訊科學研究所 中文詞知識庫小組
- https://ckip.iis.sinica.edu.tw/CKIP/tr/200901_2016b.pdf
- 廣義知網(E-HowNet) 是中央研究院資訊所詞庫小組於2003年與董振東先生展開建構繁體字知網的合作計畫

Co-occurrence Thesauri

- Observation: Words with related meanings co-occur
  - Example: astronaut, shuttle, space, spacecraft, ...
  - Does capture corpus-specific relationships
  - Does not capture synonymy,
- Hypothesis: It is useful to expand a query with related words
  - Even if the words are not synonyms
  - Even if the words are antonyms

Local Context Analysis

- Retrieve the top n ranked passages using original query q
- Compute similarity sim(q,c) for each concept c in top ranked passages using tf-idf
- Add top m ranked concepts to original query q with weighting values

Text Representation

- Manual vs automatic indexing
  - Controlled vocabularies
  - Domain-specific lexicons
  - Full-text search
- Automatic methods
  - Stemming
  - Stopwords issue
  - Phrases

Index - basic idea

- DB system -- primary and second keys
  - Hybrid method
  - Index provides fast access to a subset of DB records
  - Scan subset to find interest items
- For documents
  - title, authors, id, date,….
- Text IR problem
  - Unable to predict the “keys” in user queries
- Possible solution
  - Index by all keys --> full text indexing

Manual/Automatic Indexing

- **Manual or human indexing**
  - Indexers decide which keywords to assign to document, based on a *controlled vocabulary*
    - » Examples: Libraries, Medline, Yahoo
  - Significant human costs, but no computational costs
- **Automatic indexing:**
  - Indexing program assigns words, phrases, or other features
    - » Example: Controlled vocabulary terms
    - » Example: Words from text of the document
  - Computational costs, but no human costs

Medical document

Review > Radiography (Lond). 2021 May;27(2):682-687. doi: 10.1016/j.radi.2020.09.010.
Epub 2020 Sep 21.

Effectiveness of COVID–19 diagnosis and management tools: A review

W Alsharif [1], A Qurashi [2]

Affiliations + expand

PMID: 33008761 PMCID: PMC7505601 DOI: 10.1016/j.radi.2020.09.010
Free PMC article

Abstract

Objective: To review the available literature concerning the effectiveness of the COVID-19 diagnostic tools.

Background: With the absence of specific treatment/vaccines for the coronavirus COVID-19, the most appropriate approach to control this infection is to quarantine people and isolate symptomatic people and suspected or infected cases. Although real-time reverse transcription-polymerase chain reaction (RT-PCR) assay is considered the first tool to make a definitive diagnosis of COVID-19 disease, the high false negative rate, low sensitivity, limited supplies and strict requirements for laboratory settings might delay accurate diagnosis. Computed tomography (CT) has been reported as an important tool to identify and investigate suspected patients with COVID-19 disease at early stage.

Key findings: RT-PCR shows low sensitivity (60-71%) in diagnosing patients with COVID-19 infection compared to the CT chest. Several studies reported that chest CT scans show typical imaging features in all patients with COVID-19. This high sensitivity and initial presentation in CT chest can be helpful in rectifying false negative results obtained from RT-PCR. As COVID-19 has similar manifestations to other pneumonia diseases, artificial intelligence (AI) might help radiologists to differentiate COVID-19 from other pneumonia diseases.

Conclusion: Although CT scan is a powerful tool in COVID-19 diagnosis, it is not sufficient to detect COVID-19 alone due to the low specificity (25%), and challenges that radiologists might face in differentiating COVID-19 from other viral pneumonia on chest CT scans. AI might help radiologists to differentiate COVID-19 from other pneumonia diseases.

Implication for practice: Both RT-PCR and CT tests together would increase sensitivity and improve quarantine efficacy, an impact neither could achieve alone.

Keywords: Artificial intelligence; CT scan; Consolidation; Crazy-paving; Ground-glass opacification; RT-PCR.

Controlled Vocabulary

MeSH(Medical Subject Headings)

- Anatomy [A]
- Organisms [B]
- Diseases [C]
- Chemicals and Drugs [D]
- Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E]
- Psychiatry and Psychology [F]
- Phenomena and Processes [G]
- Disciplines and Occupations [H]
- Anthropology, Education, Sociology, and Social Phenomena [I]
- Technology, Industry, and Agriculture [J]
- Humanities [K]
- Information Science [L]
- Named Groups [M]
- Health Care [N]
- Publication Characteristics [V]
- Geographicals [Z]

Medical Subject Headings 2023

MeSH 2023 *Preview*: Final MeSH Release Date December 2022

Search MeSH... | FullWord ▾ | Exact Match | All Fragments | Any Fragment

- ○ All Terms
  - ◉ Main Heading (Descriptor) Terms
  - ○ Qualifier Terms
  - ○ Supplementary Concept Record Terms
- ○ MeSH Unique ID
- ○ Search in all Supplementary Concept Record Fields
  - ○ Heading Mapped To
  - ○ Indexing Information
- ○ Pharmacological Action
- ○ Search Related Registry and CAS Registry/EC Number/UNII Code/NCBI Taxonomy ID Number (RN)
  - ○ Related Registry Search
  - ○ CAS Registry/EC Number/UNII Code/NCBI Taxonomy ID Number (RN)
- ○ Search in all Free Text Fields
  - ○ Annotation
  - ○ ScopeNote
  - ○ SCR Note

Sort by: Relevance

Results per Page: 20

MeSH Tree Structures

Congenital Abnormalities C16.131
- Abnormalities, Drug Induced C16.131.042
- Abnormalities, Multiple C16.131.077
  - 22q11 Deletion Syndrome C16.131.077.019
    - DiGeorge Syndrome C16.131.077.019.500
  - Alagille Syndrome C16.131.77.65
  - Alstrom Syndrome C16.131.77.80
  - Angelman Syndrome C16.131.77.95

A more complex example, with three Concepts and 12 terms.

```
AIDS Dementia Complex [Descriptor]
     AIDS Dementia Complex                                  [Concept, Preferred]
          AIDS Dementia Complex                               [Term, Preferred]
          Acquired-Immune Deficiency Syndrome Dementia Complex  [Term]
          AIDS-Related Dementia Complex                       [Term]
          HIV Dementia                                        [Term]
          Dementia Complex, Acquired Immune Deficiency Syndrome [Term]
          Dementia Complex, AIDS-Related                      [Term]
     HIV Encephalopathy                                    [Concept, Narrower]
          HIV Encephalopathy                                  [Term, Preferred]
          AIDS Encephalopathy                                 [Term]
          Encephalopathy, HIV                                 [Term, Preferred]
          Encephalopathy, AIDS                                [Term]
     HIV-1-Associated Cognitive Motor Complex              [Concept, Narrower]
          HIV-1-Associated Cognitive Motor Complex            [Term, Preferred]
          HIV-1 Cognitive and Motor Complex                   [Term]
```

Controlled Vocabulary Indexing

- **There are many controlled vocabularies. None is "best".**
  - Library of Congress Subject Headings (LCSH)
  - Medical Subject Headings (MeSH)

    : : : :
- **Tradeoffs:** Coverage vs. Detail
  - Example: LCSH is broad, MeSH is detailed
- **Advantage:** Solves the vocabulary mismatch problem
- **Advantage:** Makes the ontology of a domain explicit
  - Nice for browsing
- **Disadvantage**: Difficult and expensive to create, to use, and to maintain

Full-text indexing

- Medical text

Term	*tf*	Term	*tf*	Term	*tf*	Term	*tf*
the	31	by	6	peptide	4	such	3
of	26	effect	6	several	4	toxic	3
and	22	are	5	toxin	4	who	3
in	21	aspartame	5	also	3	when	3
a	15	exposure	5	countries	3	added	2
to	11	human	5	given	3	africa	2
as	9	with	5	it	3	balkan	2
ota	9	animals	4	preventative	3	be	2
for	8	include	4	rate	3	been	2
is	8	ochratoxin	4	shown	3	compound	2

Manual vs Automatic Indexing

	Manual	Automatic
Controlled Vocabulary	Current practice	Text categorization "Intelligent" IR
Free Text	Current practice	Text search engines "Statistical" IR

Manual vs Automatic Indexing

- **The experimental evidence is that they are about equally effective, on average**
  - Original results were from Cranfield experiments in 1960s
  - Considered counter-intuitive
  - Other results since then support the Cranfield results
- **Experiments also show that a combination of manual and automatic indexing is superior to either alone**
  - "Combination of evidence"
  - Different forms of evidence more likely to agree on relevant documents and more likely to disagree on non-relevant

Full-text Representation

- **Parse documents to recognize structure**
  - E.g., titles, dates, authors, hyperlinks
- **Scan for word tokens**
  - Issues: Numbers, hyphenation, capitalization, special characters
  - Languages such as Chinese and Japanese need segmentation
  - Record positional information for proximity operators
- **Stopword removal**
- **Word stemming**
  - Conflate all morphological variants of a word into a single form
- **Phrase recognition**
- **Concept / feature recognition**

Stopwords

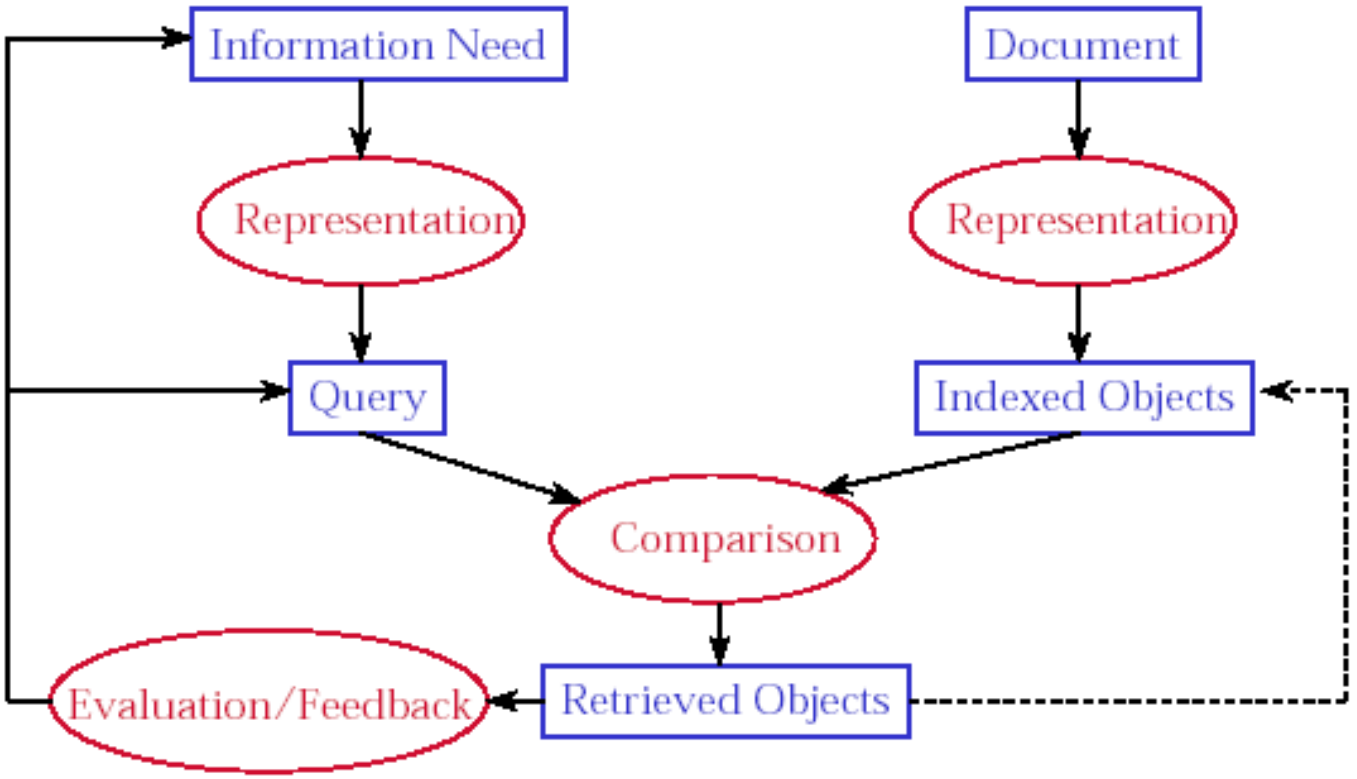
- **Stopwords:** Words that are discarded from a document representation
  - Function words: a, an, and, as, for, in, of, the, to, …
  - Other frequent words: "Lotus" in a Lotus Customer Support db
- **Why remove stopwords?**
  - Reduces the size of the representation
  - May also improve effectiveness of the retrieval algorithm
    - » This implies a weakness in the retrieval algorithm
- **Removing stopwords makes some queries difficult to satisfy:**
  - "To be or not to be", "An eye for an eye", "Sit in", "Take over"
  - Few queries affected, so little effect on *experimental* results
    - » But, very annoying to people

Words/Phrases/Concepts

- **Simple indexing is based on words and word stems**
- **More complex indexing includes phrases or thesaurus classes**
- **Index term: General name for any indexing feature**
  - Word, phrase, person name, company name
- **Concept:** Features generated by recognition rules, tables, etc.
- **Concept-based retrieval:** Something beyond word indexing
- **Words, phrases, synonyms, linguistic relations can all be evidence used to infer presence of a concept**
  - Example: Concept "Carnegie Mellon" can be inferred from words "Carnegie" and "Mellon", the phrase "Carnegie Mellon", the acronym "CMU", and maybe the acronym "LTI".

Indexing Techniques
Part II Introduction to BERT and Transformer

Basic IR Processes



Index Implementation

- Bag of words
- Inverted files
- Signature files
- Hashing
- …

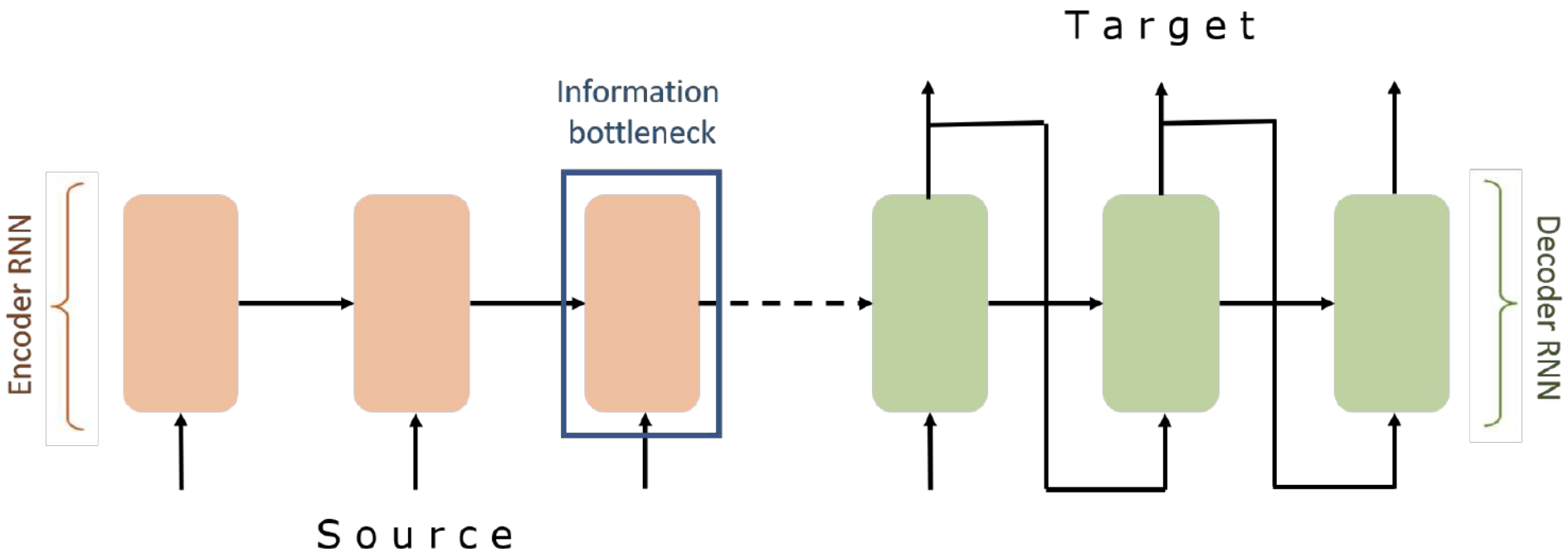
Inverted Files

- Each document is assigned a list of keywords or attributes.
- Each keyword (attribute) is associated with operational relevance weights.
- An inverted file is the sorted list of keywords (attributes), with each keyword having links to the documents containing that keyword.

General language representations

- Feature-based approaches
  - Non-neural word representations
  - Neural embedding
    - Word embedding: Word2Vec, Glove, ···
    - Sentence embedding, paragraph embedding, ···
  - Deep contextualised word representation (ELMo, Embeddings from Language Models) (Peters *et al.*, 2018)
- Fine-tuning approaches
  - OpenAI GPT (Generative Pre-trained Transformer) (Radford *et al.*, 2018a)
  - **BERT** (Bi-directional Encoder Representations from Transformers) (Devlin *et al.*, 2018)

Avoiding Information bottleneck



Pre-trained self-attention models

- ELMo (Peters *et al.*, 2018)
- OpenAI GPT (Radford *et al.*, 2018a)
- Transformer (especially **self-attention**) (Vaswani *et al.*, 2017)
- **BERT** (Devlin *et al.*, 2018)

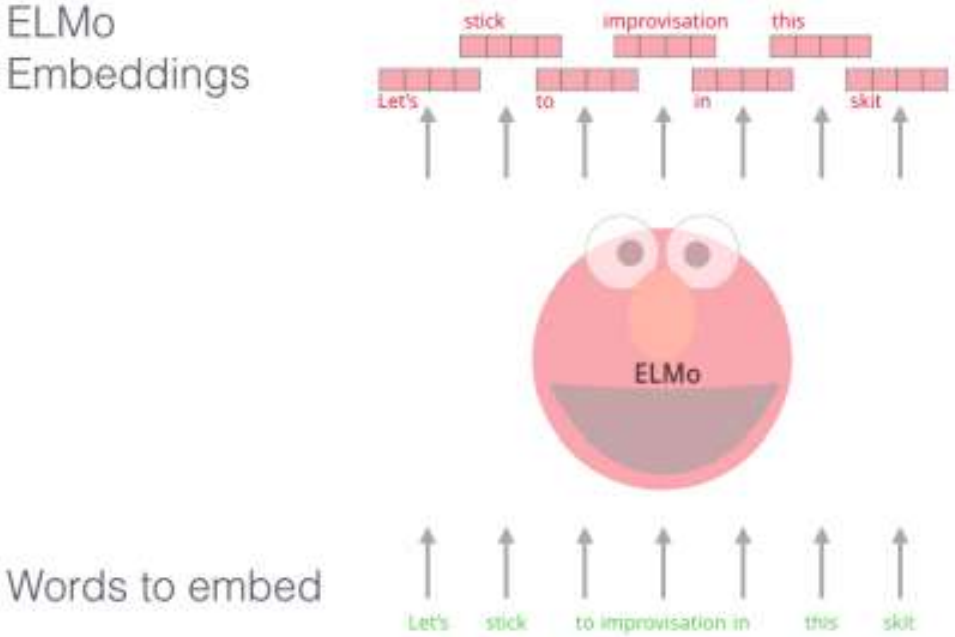
BERT pretraining

ELMo: Bidirectional training (LSTM)
Transformer: Although used things from left, but still missing from the right.
GPT: Use Transformer Decoder half.
BERT: Switches from Decoder to Encoder, so that it can use both sides in training and invented corresponding training tasks:
masked language model

ELMo: deep contextualised word representation

(Peters *et al.*, 2018)

- "Instead of using a fixed embedding for each word, ELMo looks at the entire sentence before assigning each word in it an embedding."



Acknowledgement to Figure from http://jalammar.github.io/illustrated-bert/

ELMo

Many linguistic tasks are improved by using ELMo

	Task	Previous SOTA		Our baseline	ELMo + baseline	Increase (absolute/ relative)
Q&A	SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
Textual entailment	SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
Semantic role labelling	SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coreference resolution	Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
Named entity recognition	NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
Sentiment analysis	SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

Table 1: Test set comparison of ELMo enhanced neural models with state-of-the-art single model baselines across six benchmark NLP tasks. The performance metric varies across tasks – accuracy for SNLI and SST-5; F_1 for SQuAD, SRL and NER; average F_1 for Coref. Due to the small test sizes for NER and SST-5, we report the mean and standard deviation across five runs with different random seeds. The "increase" column lists both the absolute and relative improvements over our baseline.

TRANSFORMER
變形金剛

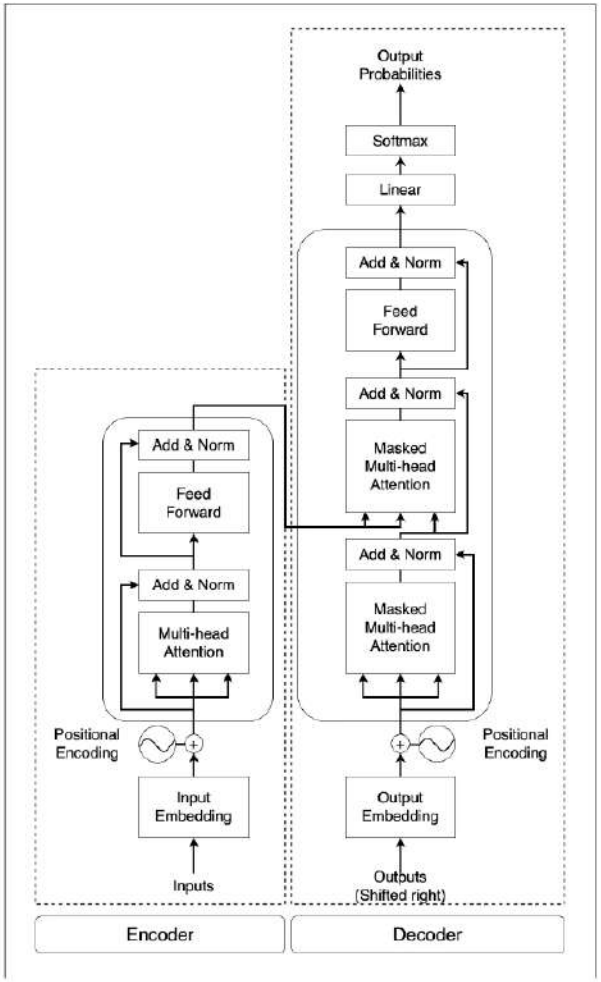
優點：

看過去的歷史

同時考慮所有的輸入資訊

訓練時下一個時間點的輸入不需等待上一個時間的輸出

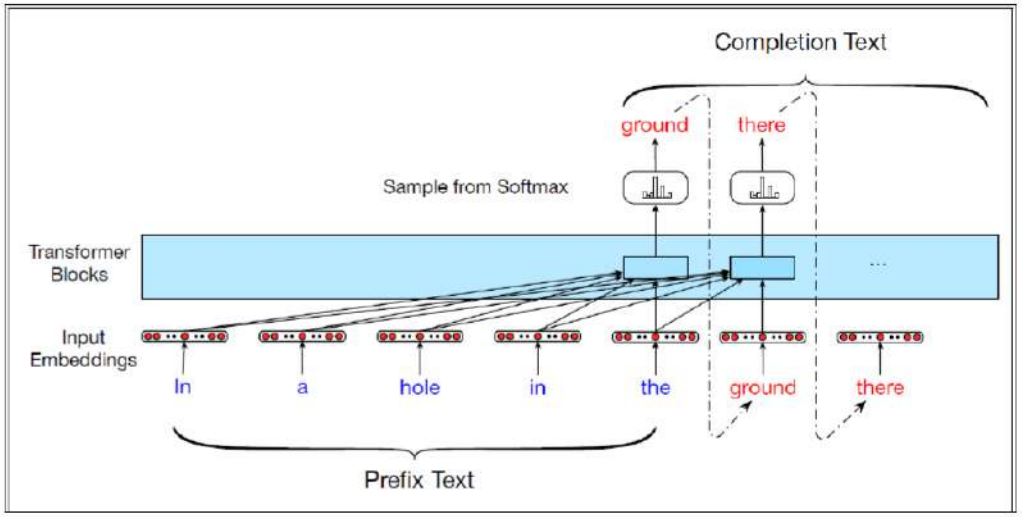
是一個任務型的模型



架構圖: 參考論文https://arxiv.org/abs/1706.03762

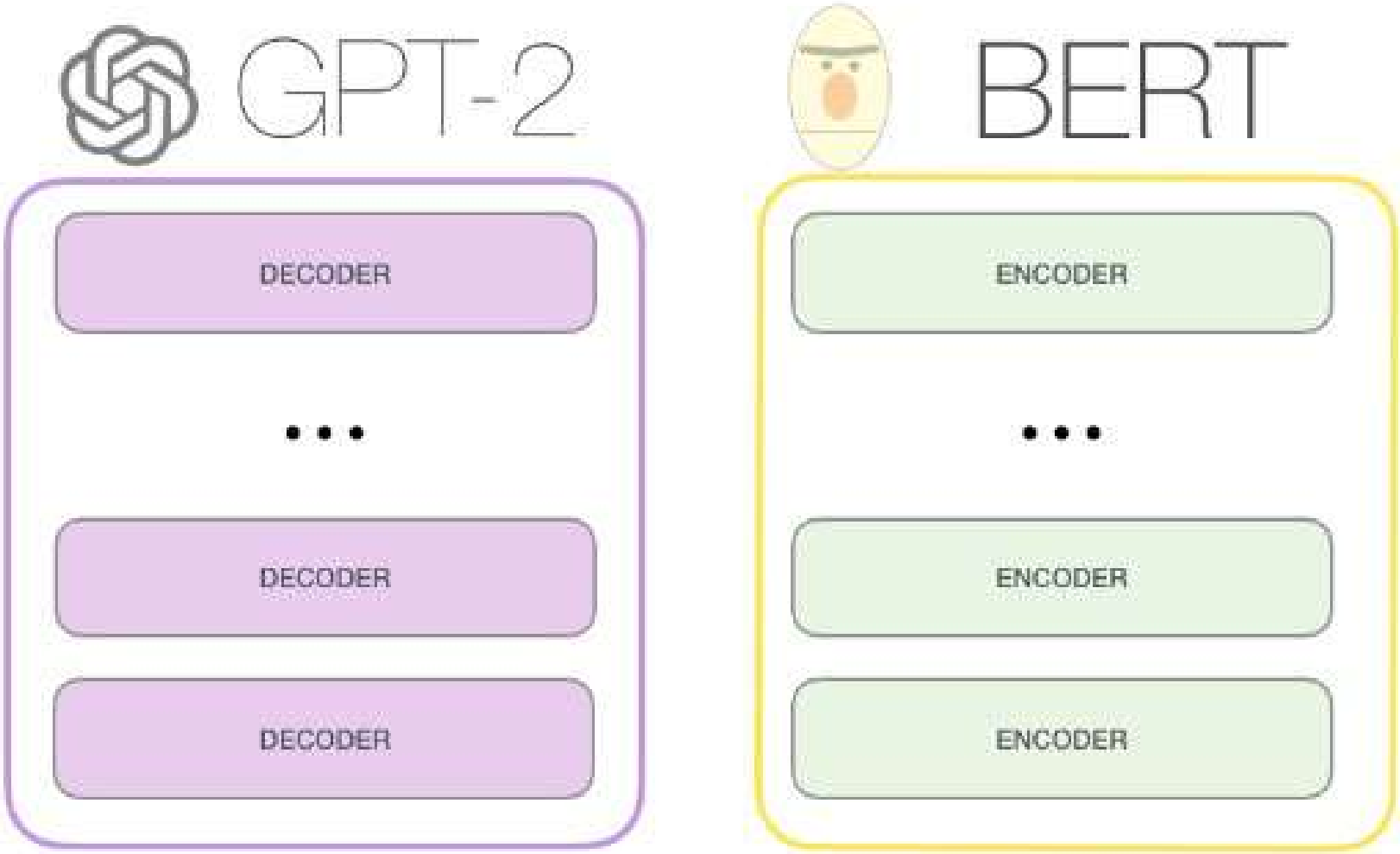
TRANSFORMER

自回歸模型
(Autoregressive)
文字生成



Transformers, GPT-2, and BERT

1. A transformer uses Encoder stack to model input, and uses Decoder stack to model output (using input information from encoder side).
2. But if we do not have input, we just want to model the "next word", we can get rid of the Encoder side of a transformer and output "next word" one by one. This gives us GPT.
3. If we are only interested in training a language model for the input for some other tasks, then we do not need the Decoder of the transformer, that gives us BERT.


GPT-2
DECODER
DECODER
DECODER
BERT
ENCODER
ENCODER
ENCODER

What is BERT (Bidirectional Encoder Representations from Transformers)?

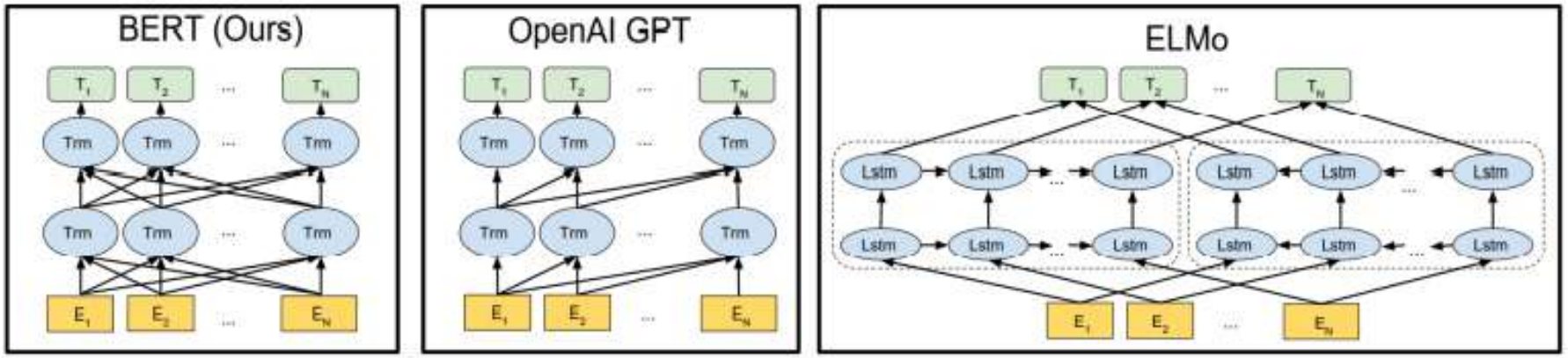
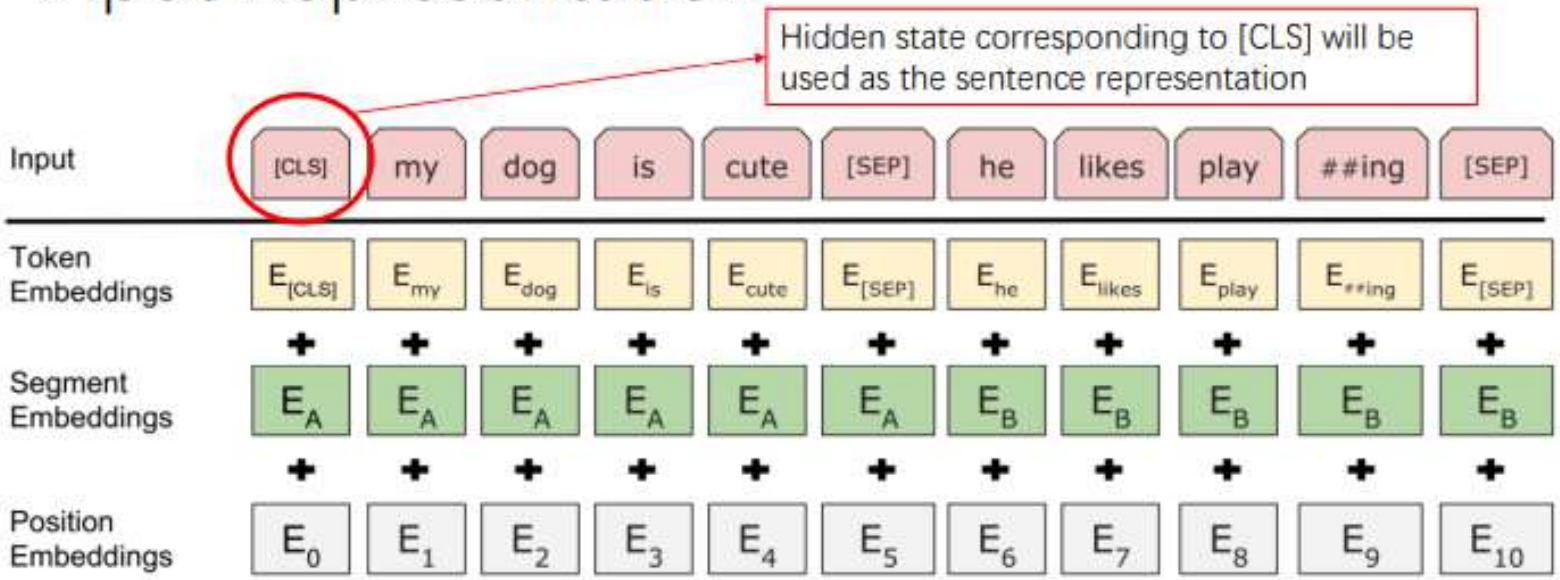


Figure 1: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks. Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

Figure in (Devlin *et al.*, 2018)

Input Representation



- Token Embeddings: WordPiece embedding (Wu *et al.*, 2016)
- Segment Embeddings: randomly initialized and learned; single sentence input only adds E_A
- Position embeddings: randomly initialized and learned

Figure in (Devlin *et al.*, 2018)

Training tasks (1) - Masked Language Model

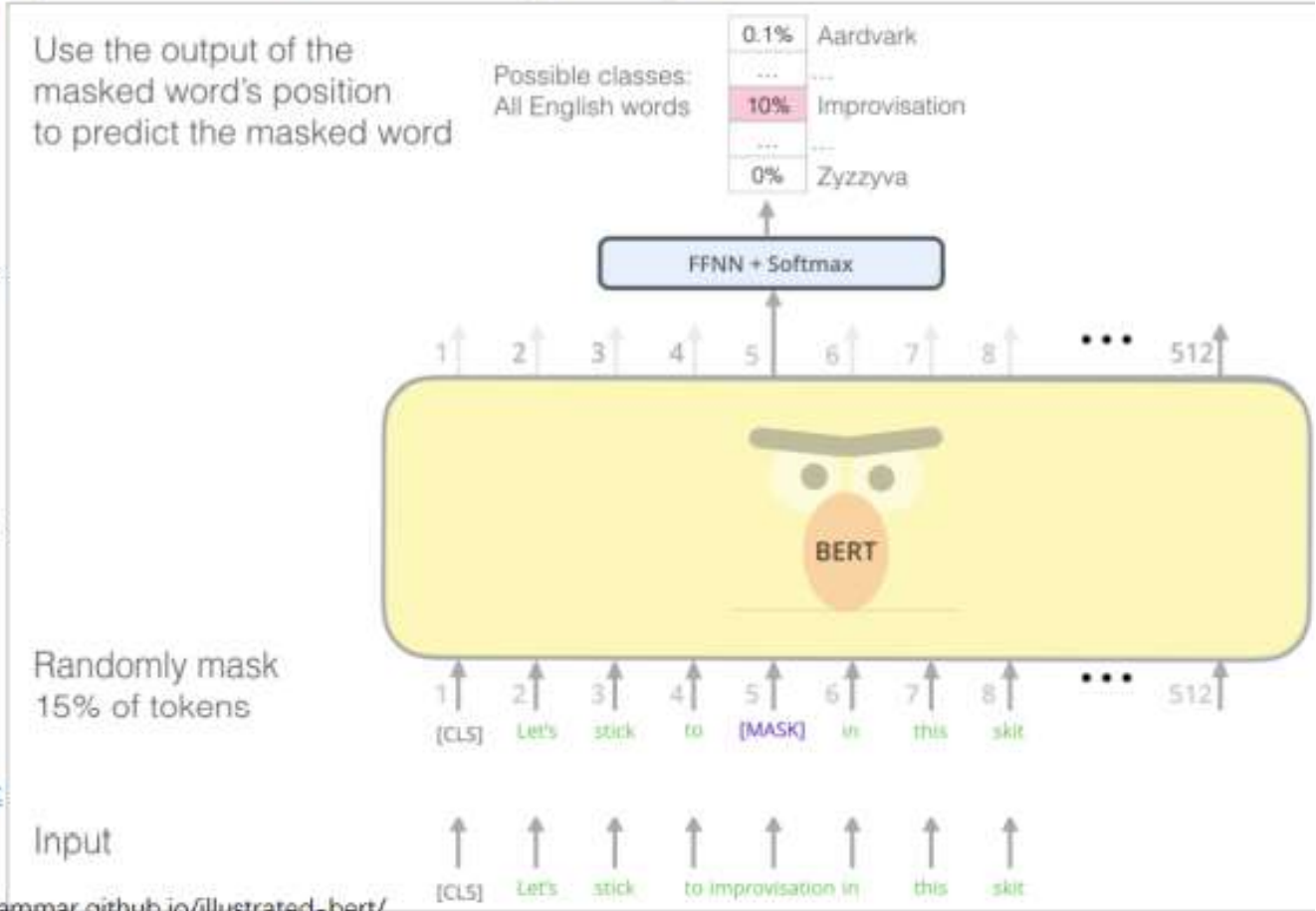
- Masked Language Model: Cloze Task

- Masking(input_seq):

  For every input_seq :

  - Randomly select 15% of tokens (not more than 20 per seq)
    - For 80% of the time:
      - Replace the word with the [MASK] token.
    - For 10% of the time:
      - Replace the word with a random word
    - For 10% of the time
      - Keep the word unchanged..

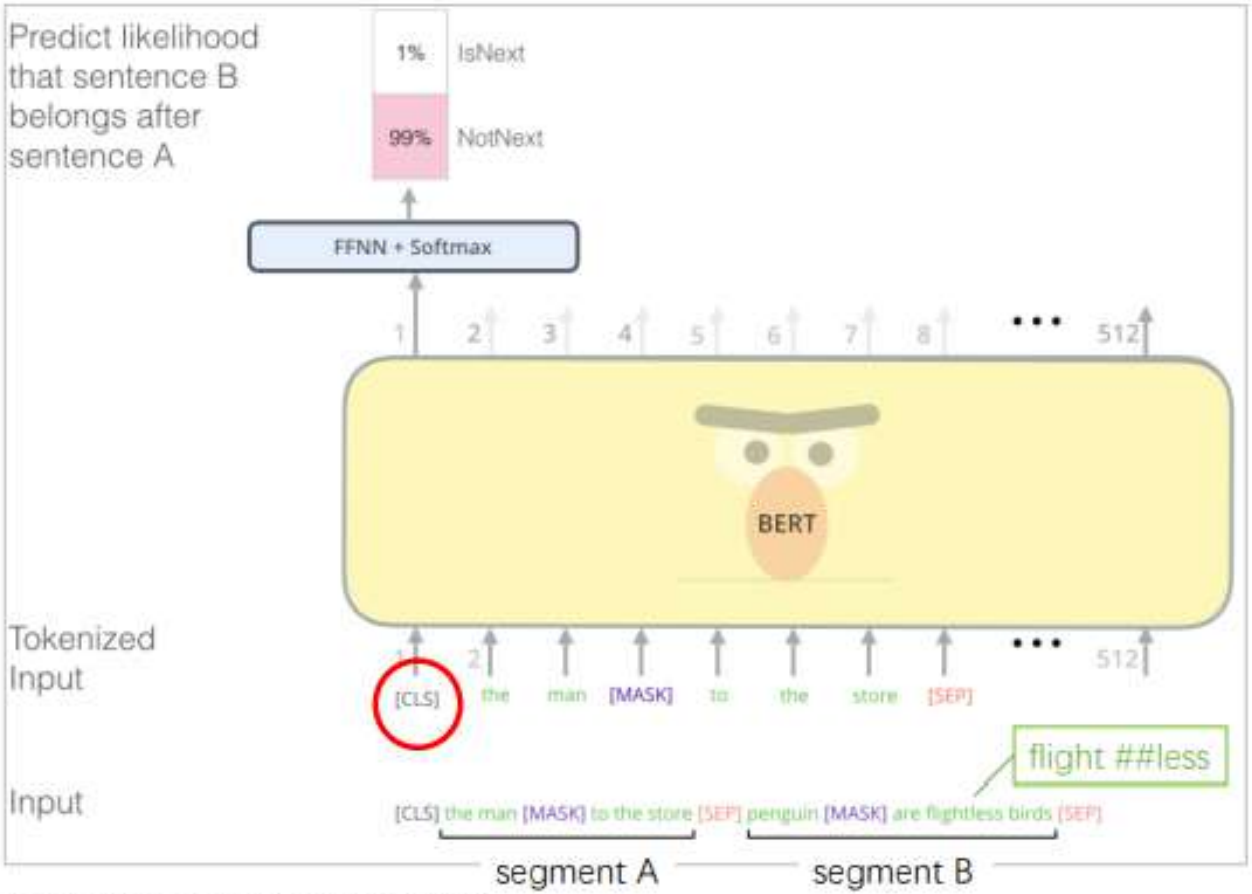
- For related code see *def create_masked_lm_predictions(…)* in https://github.com/google-research/bert/blob/master/create_pretraining_data.py



Acknowledgement to the Figure from http://jalammar.github.io/illustrated-bert/

Training tasks (2) – Next Sentence Prediction

- Next sentence prediction – Binary classification
- For every input document as a sentence-token 2D list:
  - Randomly select a split over sentences:
    - Store the segment A
    - For 50% of the time:
      - Sample random sentence split from *another* document as segment B.
    - For 50% of the time:
      - Use the actual sentences as segment B.
  - Masking (Truncate([segment A, segment B]))
- For related code see *def create_instances_from_document (…)* in https://github.com/google-research/bert/blob/master/create_pretraining_data.py

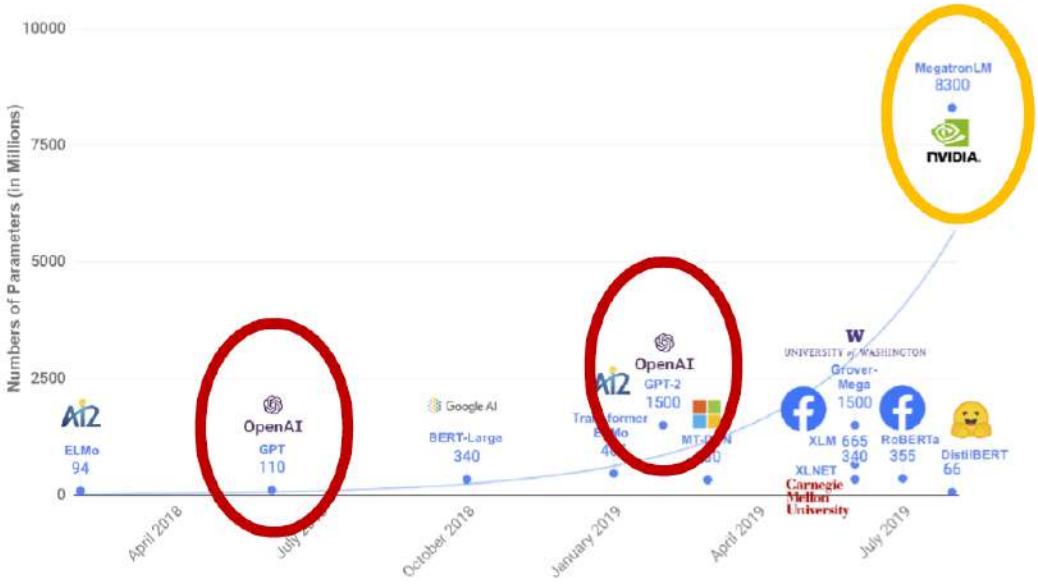


Acknowledgement to the Figure adapted from http://jalammar.github.io/illustrated-bert/

GPT系列

GPT
GPT2
GPT3: 175 billion
ChatGPT: 1 trillion

GPT4: 100 trillion



(圖片來源: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter (2020)

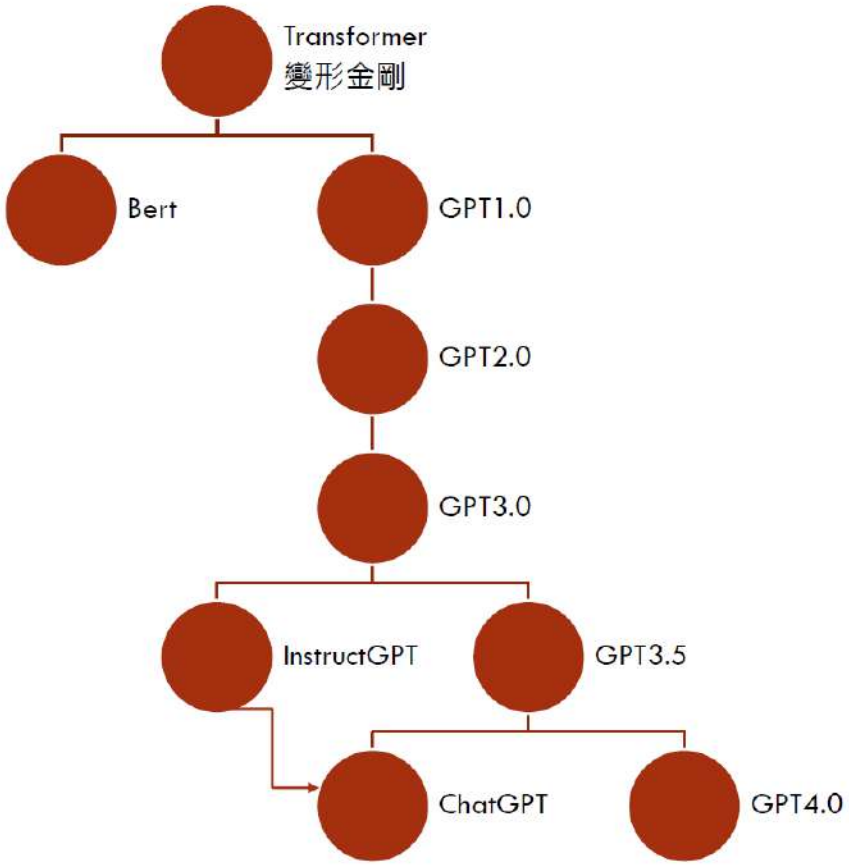
GPT 4.0 (+影像)

多模式模型

可輸入圖像與文字 – 具圖像理解能力

十倍的參數量，有更強的文字運用能力

GPT 家族


Transformer
變形金剛
Bert
GPT1.0
GPT2.0
GPT3.0
InstructGPT
GPT3.5
ChatGPT
GPT4.0