When Cantonese NLP Meets Pre-training: Progress and Challenges

Kam-Fai Wong, Hanzhuo Tan, Rong Xiang, Jing Li

Who are we?

- **Kam-Fai Wong**. Professor at the Chinese University of Hong Kong (CUHK). ACL fellow. He is working on Chinese language processing, NLP for social media, and dialogue systems.

- **Hanzhuo Tan**. Ph.D. student at the Hong Kong Polytechnic University. He is working on pre-training for social media language understanding.

- **Rong Xiang**. Postdoc fellow at the Hong Kong Polytechnic University. He is working on sentiment analysis, Cantonese pre-training, and NLP applications in education.

- **Jing Li**. Assistant Professor at the Hong Kong Polytechnic University. She is working on NLP for noisy and user-generated text and its robust applications in the real world.

Roadmap

- Introduction to Cantonese NLP
- Background and Current Progress in Cantonese NLP
- Pre-training and the state-of-the-art NLP technology
- Preliminary Experiments on Cantonese NLP
- Challenges of Cantonese NLP and the Future Directions

Roadmap

- Introduction to Cantonese NLP
- Background and Current Progress in Cantonese NLP
- Pre-training and the state-of-the-art NLP technology
- Preliminary Experiments on Cantonese NLP
- Challenges of Cantonese NLP and the Future Directions

廣东
東东
話话

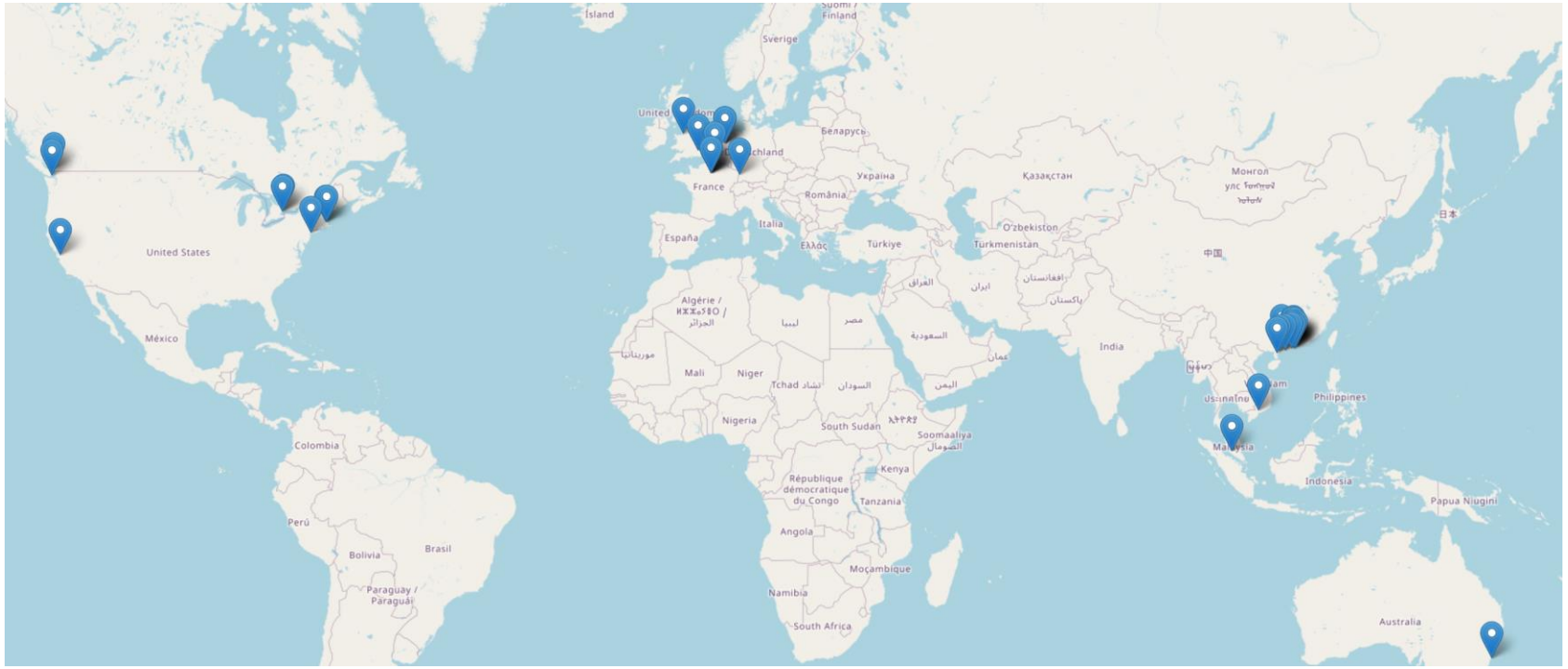
What is Cantonese?

- **Cantonese** is a language variety from the *Chinese* family, just like Mandarin.
- There are over *85 million speakers* in the world, mostly in Guangdong, Hong Kong, Macau, and the Southeast Asia.
- It is primarily used in *colloquial scenarios*, different from standard Chinese (SCN) in formal writings.

History of Cantonese

- In the *Southern Song* period, Guangzhou became the cultural center of the region.
- In the *1700s*, Guangzhou became China's key commercial centre for foreign trade and exchange.
  - Cantonese became the variety of Chinese *interacting most with the Western* World.
- In the *1900s*, the ancestors of most of the population of *Hong Kong and Macau* arrived from Guangzhou and surrounding areas after they were ceded to Britain and Portugal.
  - It allows *Cantonese development in a multilingual environments*.

Chinese dictionary from the Tang dynasty. Modern Cantonese pronunciation preserves almost all terminal consonants.



Spread of Cantonese Language

- In China, Cantonese is *mostly spoken in Guangdong Province, Hong Kong SAR, and Macau SAR*.
- Beyond China, there are speakers in other parts of Asia. Further, because of the popular emigration from China to Canada, the United Kingdom, the United States and Australia, there are large Cantonese-speaking populations in these countries.

Cantonese Impact: from China to the World

Although Cantonese has less speakers than Mandarin, it is *widespread globally* and used in many Chinese communities oversea.

In *Hong Kong and Macau*, Cantonese is the *predominant Chinese variety spoken*, e.g., most public discourse is in Cantonese.

Due to their dominance in Chinese diaspora overseas, *Cantonese is one of the most common Chinese languages one may encounter in the West*.

Many English words are borrowed from Cantonese, e.g., Chow Mein.

Written and Colloquial Cantonese

- In formal writings (e.g., government documents), the written Cantonese is standard Chinese, like other Chinese variants.
- In informal scenarios (e.g., daily chats, social media, etc.), colloquial Cantonese is texted with unique vocabulary, grammar, and pronunciation compared to other Chinese variants.
- **Bilteracy and Trilingualsm** *language policy issued by HKSAR government since 1997.*
  - **Biliteracy** is for *written language* in standard Chinese and English.
  - **Trilingualism** is for *colloquial language* in Cantonese, English and Mandarin.

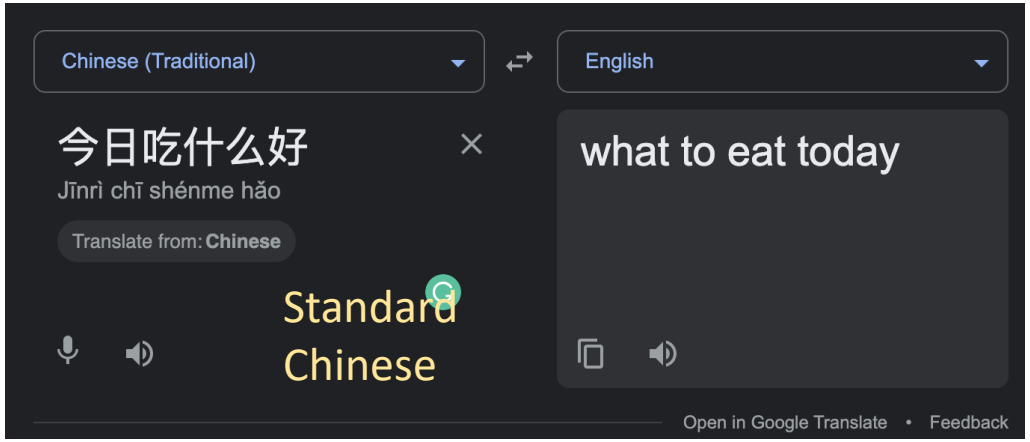
Why Cantonese NLP?

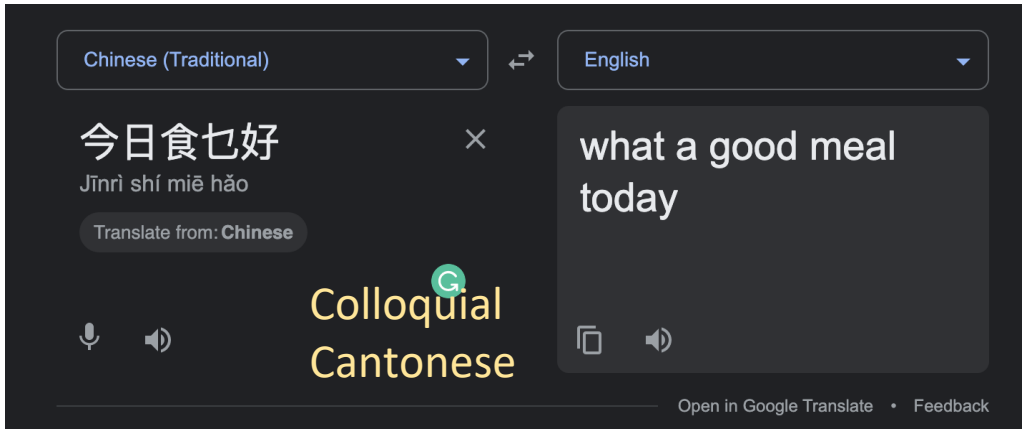
- Digital communication has greatly aided the growth of written colloquial Cantonese.
  - It is widely used across the Internet especially on social media platforms and daily conversations.
- Large volumes of social media data is texted in colloquial Cantonese, especially in Guangdong, Hong Kong, and Macau.
  - E.g., Facebook, Twitter, OpenRice, etc.
- NLP methods should understand colloquial Cantonese to process the data.
  - E.g., social media applications and dialogue systems in Cantonese.

Applying Chinese NLP to Cantonese

- Chinese NLP is well developed and benefited from the state-of-the-art “pre-training and fine-tuning” practice.
- Many pre-trained Chinese models are open-access, e.g., RoBERTa, and ERNIE.
- Colloquial Cantonese diverges substantially from standard Chinese in phonology, orthography, lexicon, and grammar.
  - Chinese NLP methods are trained on standard Chinese data.

An example from Google Translation

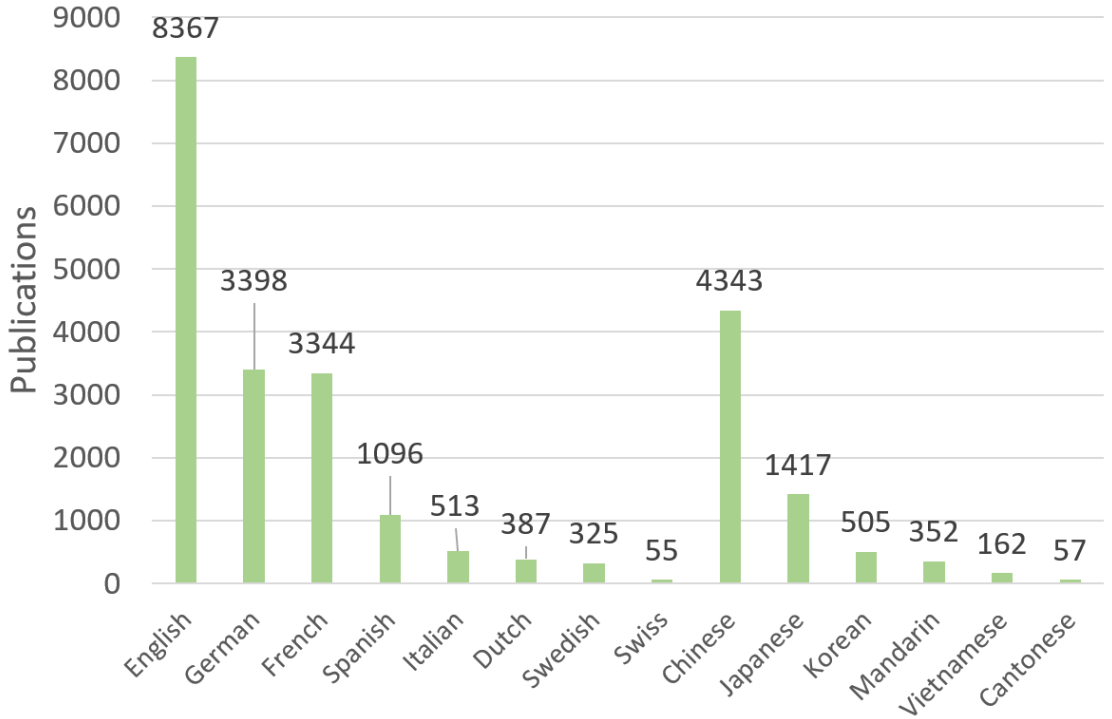




Limited Progress in Cantonese NLP

Statistics about recent publications in ACL anthology.

Observation. There are only 57 papers related to “Cantonese”, compared to 8,367 papers for English, 4,343 for common Chinese, and 352 for Mandarin. (Statistics update to 17th Nov, 2022)

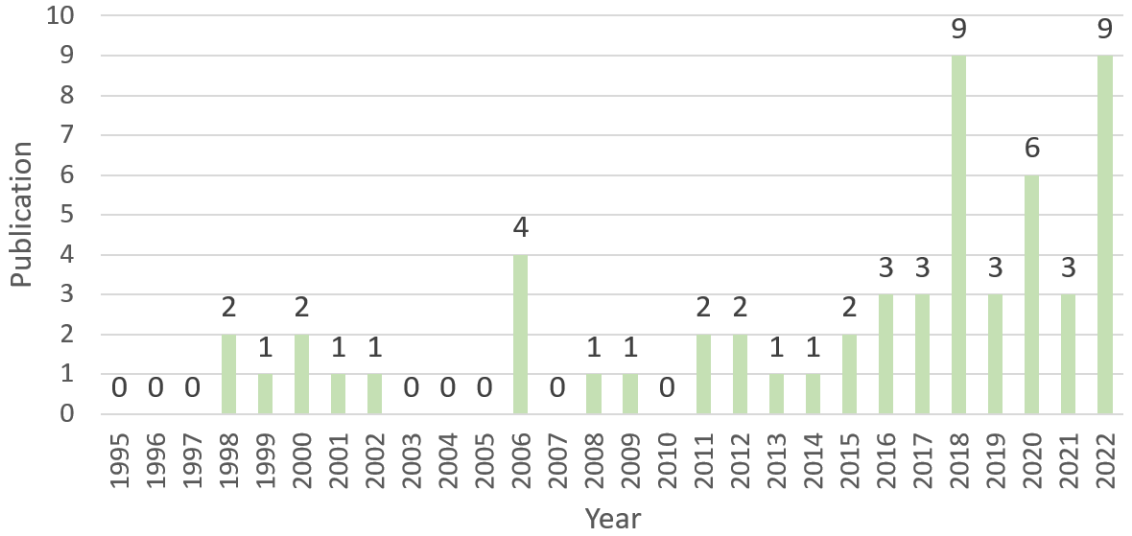


Limited Progress in Cantonese NLP (cont.)

Research Topics	# of Papers
Phonetics&Phonology& Speech Recognition	22
Lexicography&Syntax& Semantics&Morphology	10
NLP Resources	15
NLP Tasks	10
Total	57

Only 15 papers are about NLP resources for Cantonese.

Yearly publications of the 57 papers for Cantonese NLP in ACL Anthology from 1998 to 2022. The distribution is sparse, showing it is underdeveloped!

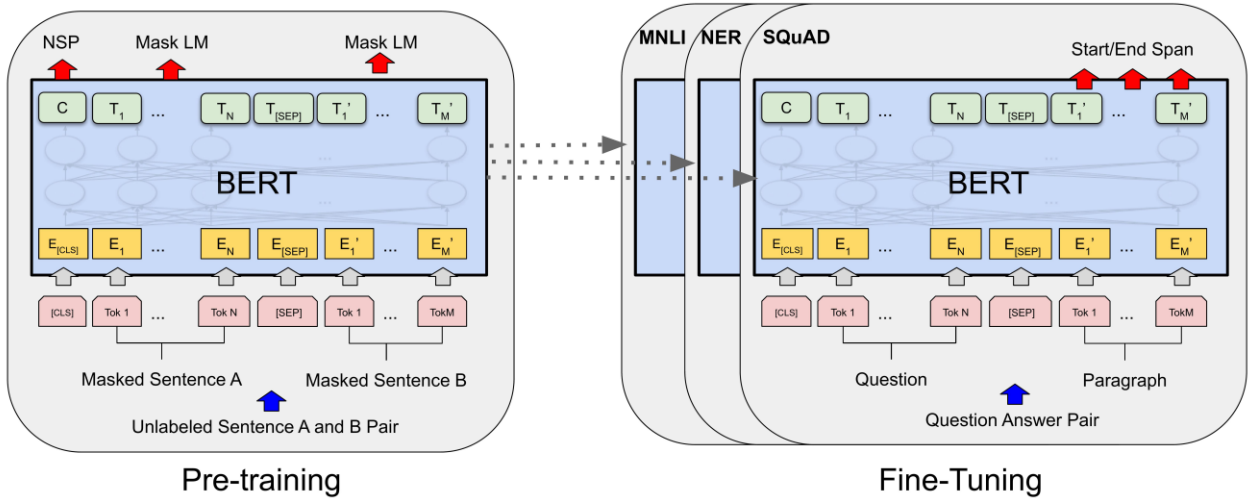


Pre-training: the state-of-the-art NLP

- Pre-training enables models to generally understand languages via examining large-scale text.
- In pre-training, models practiced self-supervised learning tasks for capturing the basic reading skills, e.g., Masked Language Models, Next Sentence Prediction, etc.
- The pre-trained models can then be fine-tuned on specific NLP tasks to narrow the generic reading skills down to handle a certain task, e.g., sentiment analysis.

Advantages:

- State-of-the-art fine-tuned results on many NLP tasks.
- Utilize large-scale text without human labels.



Question: what if we are handling a low-resource language like Cantonese?

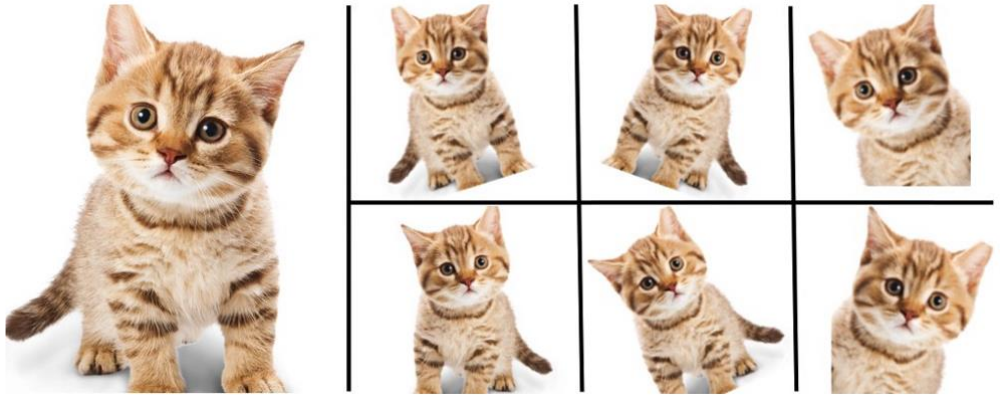
Cantonese NLP Challenges: **Colloquialism**

- Cantonese is mainly derived from *pronunciation* (this is evident from some identifiable *colloquial phonetic features*, e.g., Ham Baang Laang 冚棒吟 ‘entire/all’).
- It is unlike standard Chinese, whose character standardization dates back to the Qin dynasty (221 BC).
- Colloquial Cantonese is mostly used in *informal communications and social media discussions*. The data is, therefore, full of *colloquial features*, such as non-standard spelling, local lang, neologisms, emoji, etc.

Cantonese NLP Challenges: **Multilingualism**

- The Cantonese language historically evolved in *multilingual* environments.
- This is especially true for HK Cantonese, as shown by the large inventory of English *loanwords borrowed through phonetic transliteration* (e.g., sido 士多 'store').
- More generally, *Cantonese differs from Mandarin in vocabulary by 30-50%*.
  - It demonstrates systematic differences from those Chinese varieties used in SCN in various linguistic aspects.

Future Directions (generating more data)



For example, we can do data augmentation to generate the data from data with heuristic rules.

- It is to mitigate low-resource problems via scaling up the Cantonese dataset for NLP model training.
- We might need to distinguish standard Chinese from Cantonese in the data generation process. Though both are encoded in the Chinese language system, standard Chinese dominates the Chinese resources while Cantonese (the data we want) is a minority.

Future Directions (customizing NLP models)

- **Cross-lingual Learning**: it is to borrow knowledge from standard Chinese and mandarin.
- **Cross-modal Learning**: Cantonese origins from pronunciations, future work may consider injecting phonetic knowledge into language learning or developing multimodal understanding across text and speech modalities.

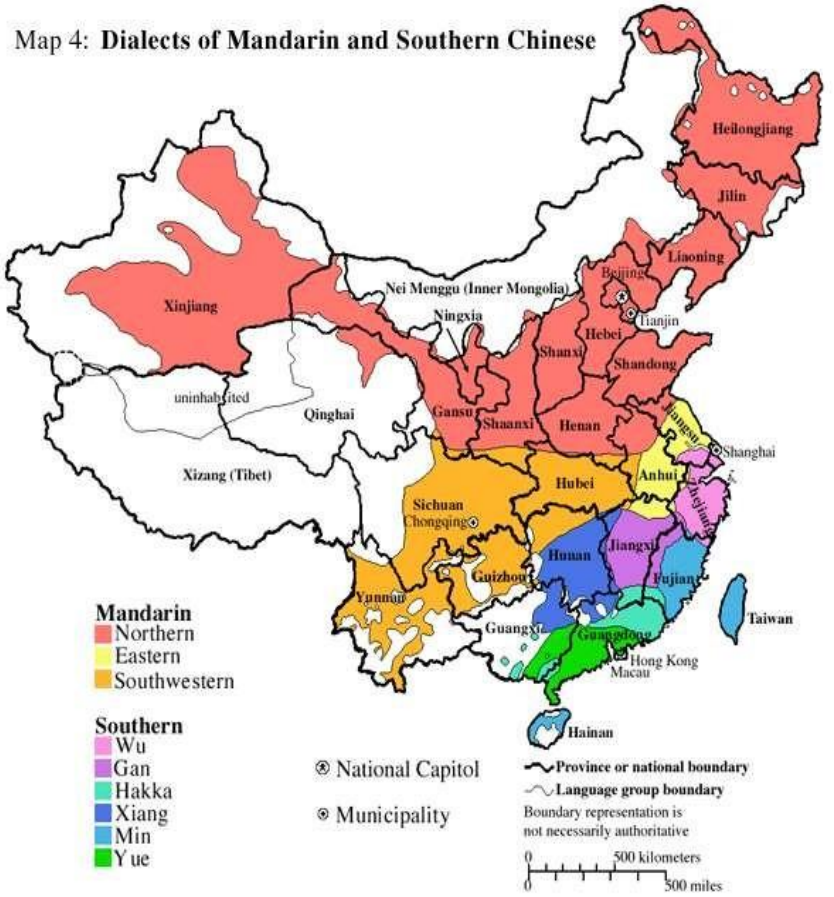
Questions?

Roadmap

- Introduction to Cantonese NLP
- Background and Current Progress in Cantonese NLP
- Pre-training and the state-of-the-art NLP technology
- Preliminary Experiments on Cantonese NLP
- Challenges of Cantonese NLP and the Future Directions

A. Background

There are seven main dialects of Chinese — Mandarin, **Cantonese**, Hakka, Wu, Min, Xiang and Gan. Pǔtōnghuà, the type of Mandarin based on the speech in the capital Beijing, is the official national language of mainland China. Cantonese is the 2nd most used dialect of Chinese.



Map of China Dialects

Written and Spoken Form

Simplified vs Traditional

- Traditional Chinese for HK, Macau; spoken Cantonese
- Simplified Chinese for Chinese mainland; Spoken Mandarin
- Traditional Chinese for Taiwan; spoken Mandarin

Translation
HK 翻譯= Traditional (faan1 yik6)
CN 翻译= Simplified (fān yì)
TW 翻譯= Traditional (fān yì)

Quality
HK 質量= Traditional (jat1 leung6)
CN 质量= Simplified (zhì liàng)
TW 質量= Traditional (zhí liàng)

Is Cantonese a language?

YES and NO

A language ——

Chinese is a family of languages, including Mandarin, Cantonese, Wu, Min and Hakka (Wikipedia)

A dialect ——

There are a large number of differences between Cantonese and Mandarin (Wikipedia)

Cantonese / Standard Chinese (Mandarin)

Similarity and Differences

Same:

- **Multiple variants**

Cantonese:

Examples:

Chinese mainland	落班/上班	
Hong Kong SAR	收工/返工	撳錢
Macao SAR		暗錢

South-east Asia, North America, ……

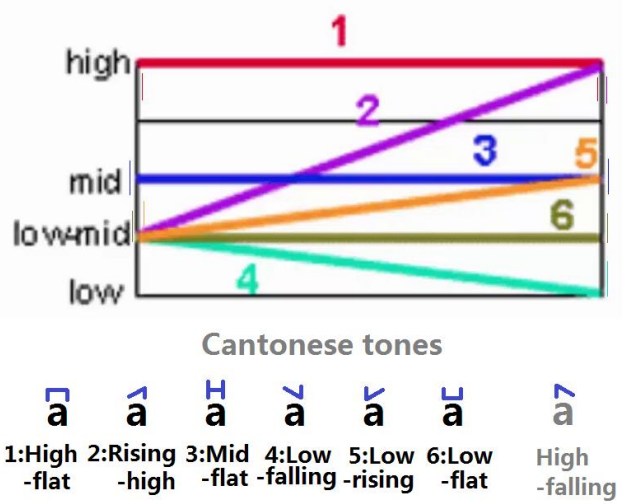
Similarity and Differences

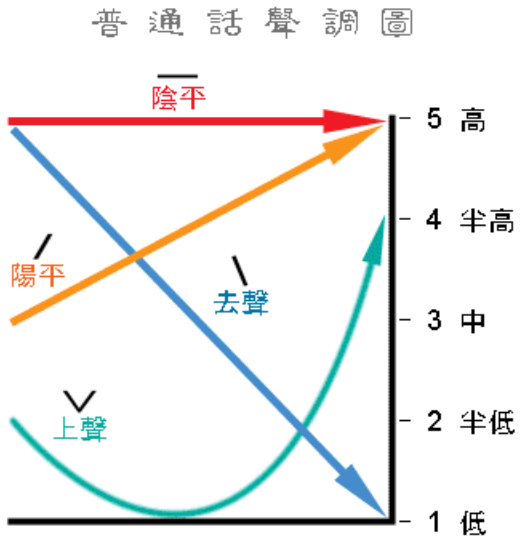
Same:

- **Tone based pronunciation**
- Different tones carry different meanings

Different:

- Cantonese with 9 tones and 6 modes （九聲六調）
- Mandarin with 4 tones





Similarity and Differences

Same:

- **Romanized letter based pronunciation system**
- Help the non-Chinese people to easily understand the Chinese letters and its pronunciation

Different:

- Cantonese Pinyin systems: Yale Romanization system, Jyutping (粵拼), Cantonese Pinyin, etc.
- Standard Chinese with (Hanyu) Pinyin system (汉语拼音)

春曉 (Chunxiao)　**孟浩然** (Meng Haoran)	**Tsoen1 Hiu2　Maang6 Hou6jin4**
春眠不覺曉，(Sleeping past sunrise in springtime.)	Tsoen1 min4 bat7 gok8 hiu2,
處處聞啼鳥。(Everywhere one hears birdsong.)	Tsy3 tsy3 man4 tai4 niu5.
夜來風雨聲，(Night brings the sound of wind and rain,)	Je6 loi4 fung1 jy5 sing1,
花落知多少？(I wonder how many flowers fell?)	faa1 lok9 dzi1 do1 siu2?

Traditional characters:

mā *má* *mǎ* *mà* *·ma*

媽 麻 馬 罵 嗎

Simplified characters:

mā *má* *mǎ* *mà* *·ma*

妈 麻 马 骂 吗

Similarity and Differences

Same:

- **Stroke based characters.**
- Number of characters to learn to get basic knowledge.

Different:

- Cantonese uses traditional Chinese characters
- Mandarin uses simplified Chinese characters

Similarity and Differences

Same:

- It **doesn’t have alphabet**, the character itself delivers the word.

- **Weak grammar rules**

- Doesn’t need grammar like verbs, tenses, etc.…

Similarity and Differences

Different vocabulary:

Bulb

Hair dryer

Ice cream

Cantonese unique vocabulary

“咗”（了，already, done, finished）

“喎”（”啊，啦”，already）

“咁”（”这样”，such, so）

“咩”（”乜嘢”，what? Is it?）

“喳”（”只是”，well）

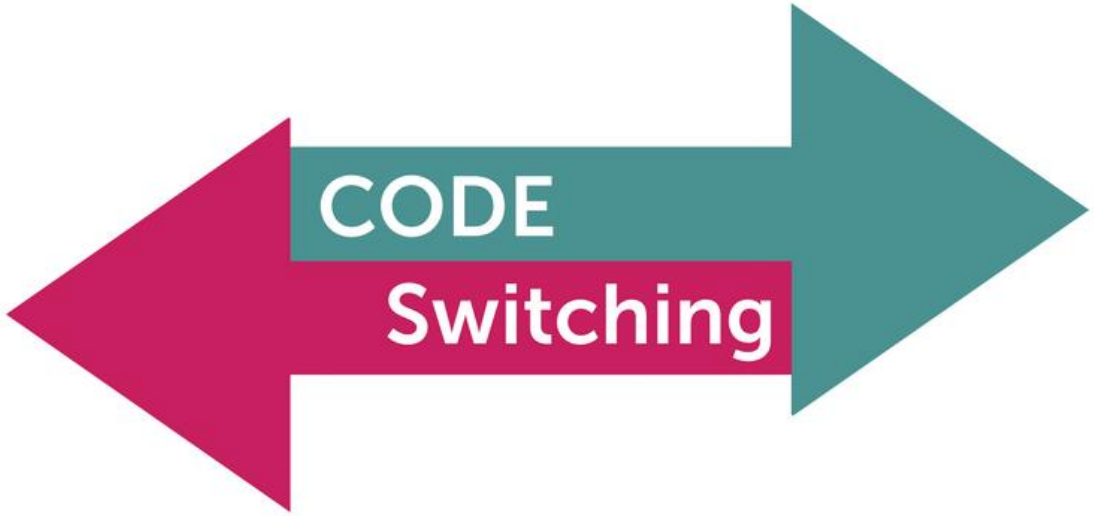
“啲”（”吧，我猜的”，well, I guess)

Similarity and Differences

Different Grammar:

	Cantonese	Standard Chinese
定语后置 post-attributive	鸡公 Rooster 椰青 Green Coconut	公鸡 青椰
状语后置 post-adverbial	唔该晒你 Thank you so much 食多啲 Eat much	多谢你 多吃点
倒装 inverted sentence	我行先 I will leave first 唔怪得 No wonder	我先走 怪不得

Similarity and Differences



Same:

- **Code-switch/Cross-lingual**

  friend 嚟㗎 (This is my friend.)

  今天好happy (I am very happy today.)

- **Loanwords**

  Cantonese: More loanwords

pie	p^hay^{55}	game	kem^{55}
boxing	pok^{55} $siŋ^{35}$	sexy	sek^{55} si^{35}
bumper	pam^{55} pa^{35}	cancer	k^hen^{55} sa^{35}
captain	k^hep^{55} $t^hön^{35}$	engine	en^{55} $tsin^{35}$
cast	k^ha si	dockyard	tok^{55} ya^{55}
band	pen^{55}	pump	pam^{55}
sink	$siŋ^{55}$	cement	si man

Short Summary

- Cantonese differs from Standard Chinese in Vocabulary, Grammar and Multi-cultural and multi-lingual.

- The **colloquial and multilingual** features of Cantonese pose a serious challenge to models developed primarily for SCN.

B. Current Progress in Cantonese NLP

1 Corpora

2 Benchmark

3 Expert resource

4 Natural language understanding (NLU)

5 Natural language generation (NLG)

6 Phonetics & Phonology & Speech (Speech)

Corpora

Corpora	Reference	Information
CANCORP	Lee and Wong 1998	1 M characters, hk children
HKCAC	Leung and Law, 2001	adult, speech record
HKUCC	Wong, 2000	university, speech record, word segmentation
Parallel corpus	Lee, 2011	TV, mandarin subtitles 36k characters
Parallel treebank	Wong et al., 2017	569 aligned sentences
Political corpus	Pan, 2019	Machine translation
Counselling Corpus	Lee et al.,2020, 2021, 2022	12.6k post-restatement pairs and 9k post-question pairs

Benchmark

Task	Reference	Information
Spelling check	Fung et al., 2017	6,800 sentences
Sentiment analysis	Xiang et al., 2019	Openrice, 60k sentences, 5-level sentiment degree
Rumor detection	Chen et al., 2020	Twitter, 27,328 instances
Topic classification	ToastyNews project	LIHKG, 11k sentences, 20 classes
Intention classification	Wang et al., 2020	Facebook, Openrice, 10 HK restaurants
Machine translation	Liu, 2022	35.9k sentence pairs

Expert resource

Resource	Reference	Information
Digital dictionary	Cheung et al., 2018	Modern Cantonese, 12k entries
Sentiment lexicon	Klyueva et al., 2018	2.7k Cantonese words
Cantonese Wordnet	Sio and Costa, 2019 Sio and Costa, 2022	3.5k concepts and 12k senses
Cifu	Lai and Winterstein, 2020	Cantonese words with definition, frequency, strokes and structure
CantoMap	Winterstein et al., 2020	768 minutes of recordings and transcripts
NER	Community	https://github.com/CanCLID/words
jyut6 din2 粵典	Lau et al., 2022	Cantonese dictionary, 55.6k words
PyCantonese	Lee et al., 2022	https://pycantonese.org/ Stopwords, sentence segmentation, POS, etc.

NLU

Task	Reference	Information
Sentiment analysis	Zhang et al, 2011 Chen et al, 2013, 2015 Ngai et al, 2018 Xiang et al., 2019	Naïve Bayes and SVM PoS tagging based HMM Machine learning & lexical knowledge Lexicon enhanced LSTM
Emotion detection	Lee, 2019	Leverage Mandarin emotion resources
Rumor detection	Chen et al., 2020 Ke et al, 2020	XLNet-based Bidirectional GRU, Pre-trained BERT and Bi-LSTM

NLG

Task	Reference	Information
Dialogue Summarization	Liu and Lapata, 2019 Lee et al., 2021	Fine-tuning of BertSum Text summarization & question generation
Machine Translation	Zhang, 1998 Wu et al, 2006 Huang et al, 2016 Wong and Lee, 2018 Liu, 2022	Handcrafted rules Parallel corpora Parallel corpora & statistical methods Refined with lexical mappings and syntactic transformations BPE-level LSTM

Speech

Data/Task	Reference	Information
Corpus	Kwong, 2015	Spoken Cantonese from TV and radio
Corpus	Luke and Wong, 2015	HKCanCor, 150k words
Corpus	Liesenfeld, 2018	MYCanCor, 20 hours of Cantonese speech
Corpus	Johnson et al., 2020	Bilingual speech conversation, Google ASR annotation.
Phonetic	Kirby, 2021	Trigram, RNN
ASR	Yu et al., 2022	Fairseq S2T transformer
WeCanTalk	Jones et al., 2022	Speaker recognition

The full reference list can be downloaded at:

https://www4.comp.polyu.edu.hk/~jing1li/talks/aacl2022-can-pretrain/reference-list.txt

Questions?

Roadmap

- Introduction to Cantonese NLP
- Background and Current Progress in Cantonese NLP
- Pre-training and the state-of-the-art NLP technology
- Preliminary Experiments on Cantonese NLP
- Challenges of Cantonese NLP and the Future Directions

What's your first thought on Pre-Training?

BERT: Bidirectional Encoder Representations from Transformers

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

ELMo: Embeddings from Language Models

Sesame Street (Children's TV show)

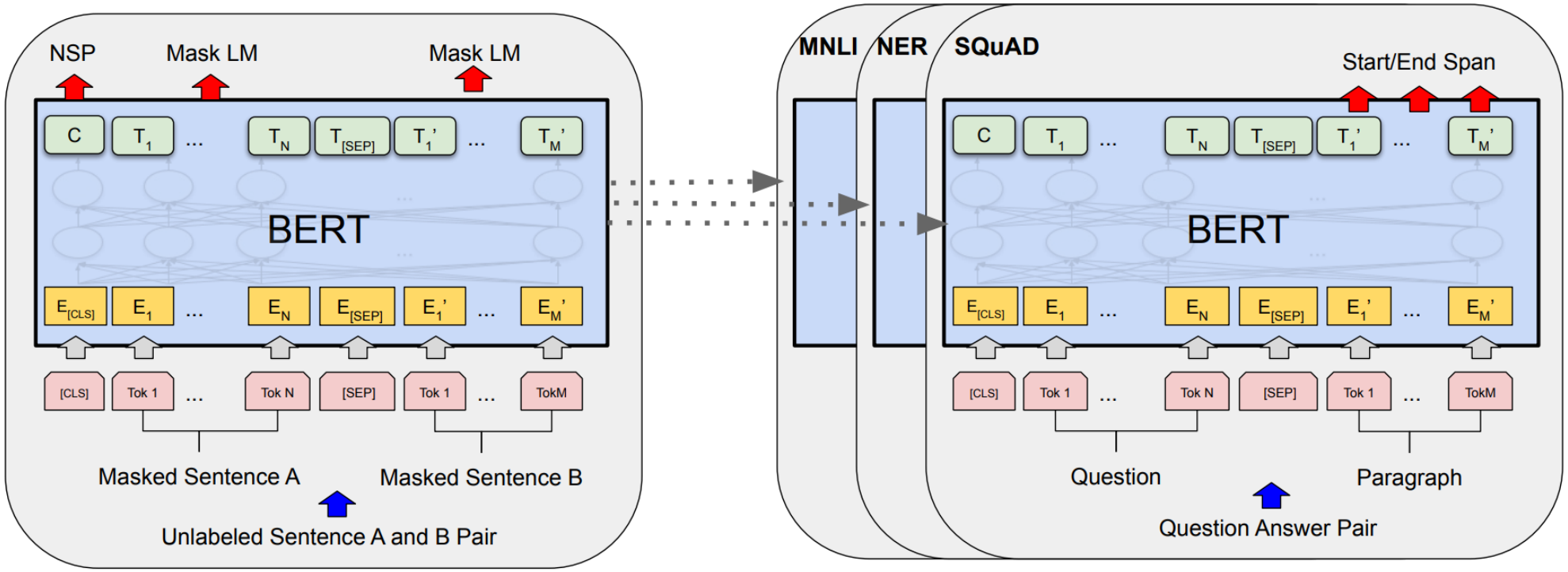
ERNIE: Enhanced Representation through Knowledge Integration

Agenda

- Introduction
- Background
- Transformers
- Prompt

Introduction

What's Pre-training?



Pre-training

Large Corpus

Language understanding/generation objectives

Fine-Tuning

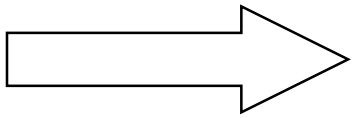
Downstream tasks

Specific losses

Why Pre-training

Get trained at University

Programming
Algorithm
Database
…



Go to work

Java
PHP
Python
…

Background

Background-Word2Vec

- Pre-training is not a new concept!
- Word2Vec (Mikolov et al, 2014)

Efficient Estimation of Word Representations in Vector Space

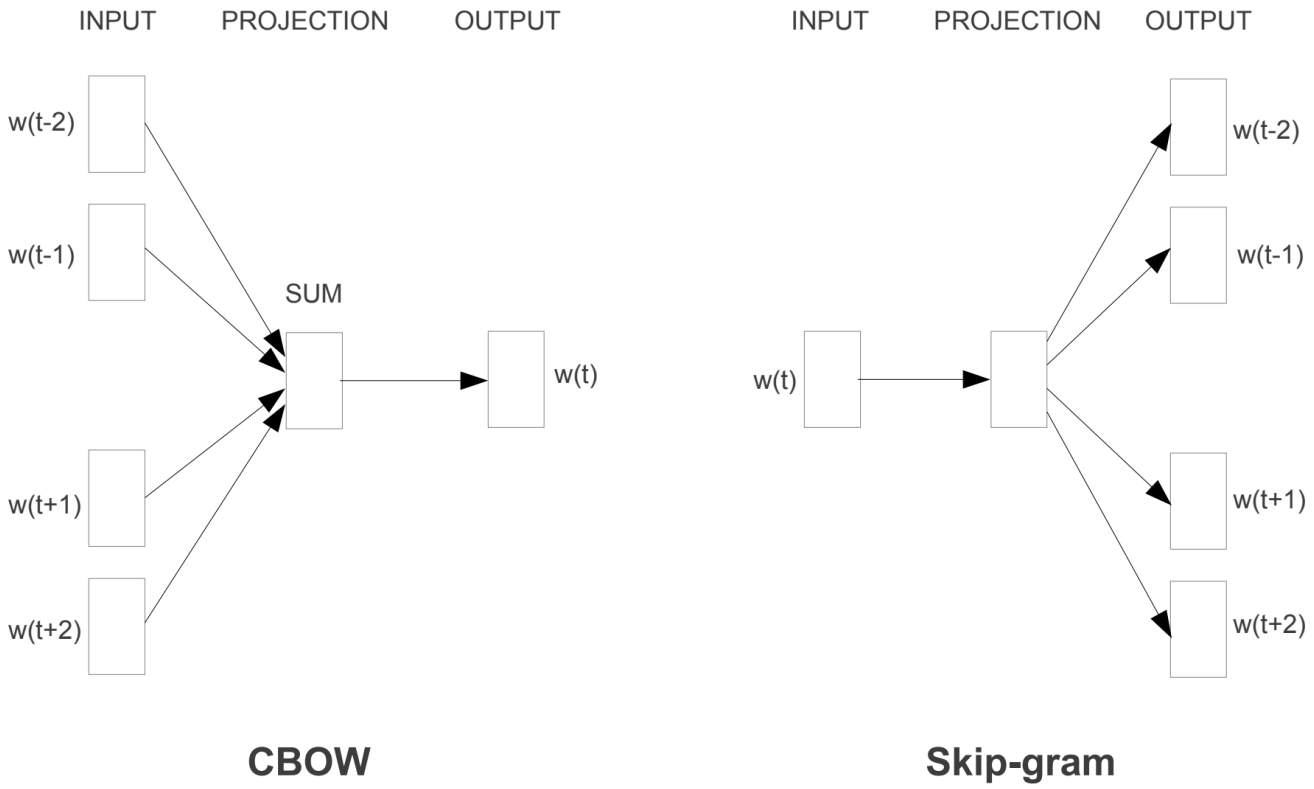
Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

Greg Corrado
Google Inc., Mountain View, CA
gcorrado@google.com

Jeffrey Dean
Google Inc., Mountain View, CA
jeff@google.com

Background-Word2Vec



Model	Vector Dimensionality	Training words	Accuracy [%]			Training time [days]
			Semantic	Syntactic	Total	
3 epoch CBOW	300	783M	15.5	53.1	36.1	1
3 epoch Skip-gram	300	783M	50.0	55.9	53.3	3
1 epoch CBOW	300	783M	13.8	49.9	33.6	0.3
1 epoch CBOW	300	1.6B	16.1	52.6	36.1	0.6
1 epoch CBOW	600	783M	15.4	53.3	36.2	0.7
1 epoch Skip-gram	300	783M	45.6	52.2	49.2	1
1 epoch Skip-gram	300	1.6B	52.2	55.1	53.8	2
1 epoch Skip-gram	600	783M	56.7	54.5	55.5	2.5

Google News corpus (most frequent 30k words)

Feed-Forward Neural Network without non-linear layers

Background-LSTM

- ULMFiT

Universal Language Model Fine-tuning for Text Classification

Jeremy Howard*
fast.ai
University of San Francisco
j@fast.ai

Sebastian Ruder*
Insight Centre, NUI Galway
Aylien Ltd., Dublin
sebastian@ruder.io

3-layer LSTM, Language Modeling (predict the next word), Wikitext-103 consisting of 28,595 preprocessed Wikipedia articles and 103 million words.

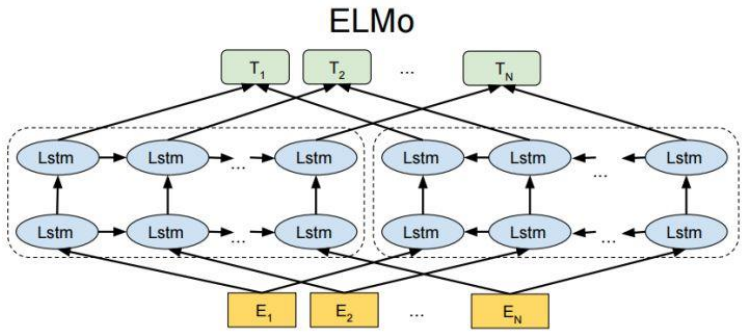
- ELMo

Deep contextualized word representations

Matthew E. Peters†, Mark Neumann†, Mohit Iyyer†, Matt Gardner†,
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark*, Kenton Lee*, Luke Zettlemoyer†*
{csquared,kentonl,lsz}@cs.washington.edu

†Allen Institute for Artificial Intelligence
*Paul G. Allen School of Computer Science & Engineering, University of Washington



Bi-directional LSTM, Language Modeling (predict the next word), 10 epochs on the 1B Word Benchmark

Background-Transformer

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

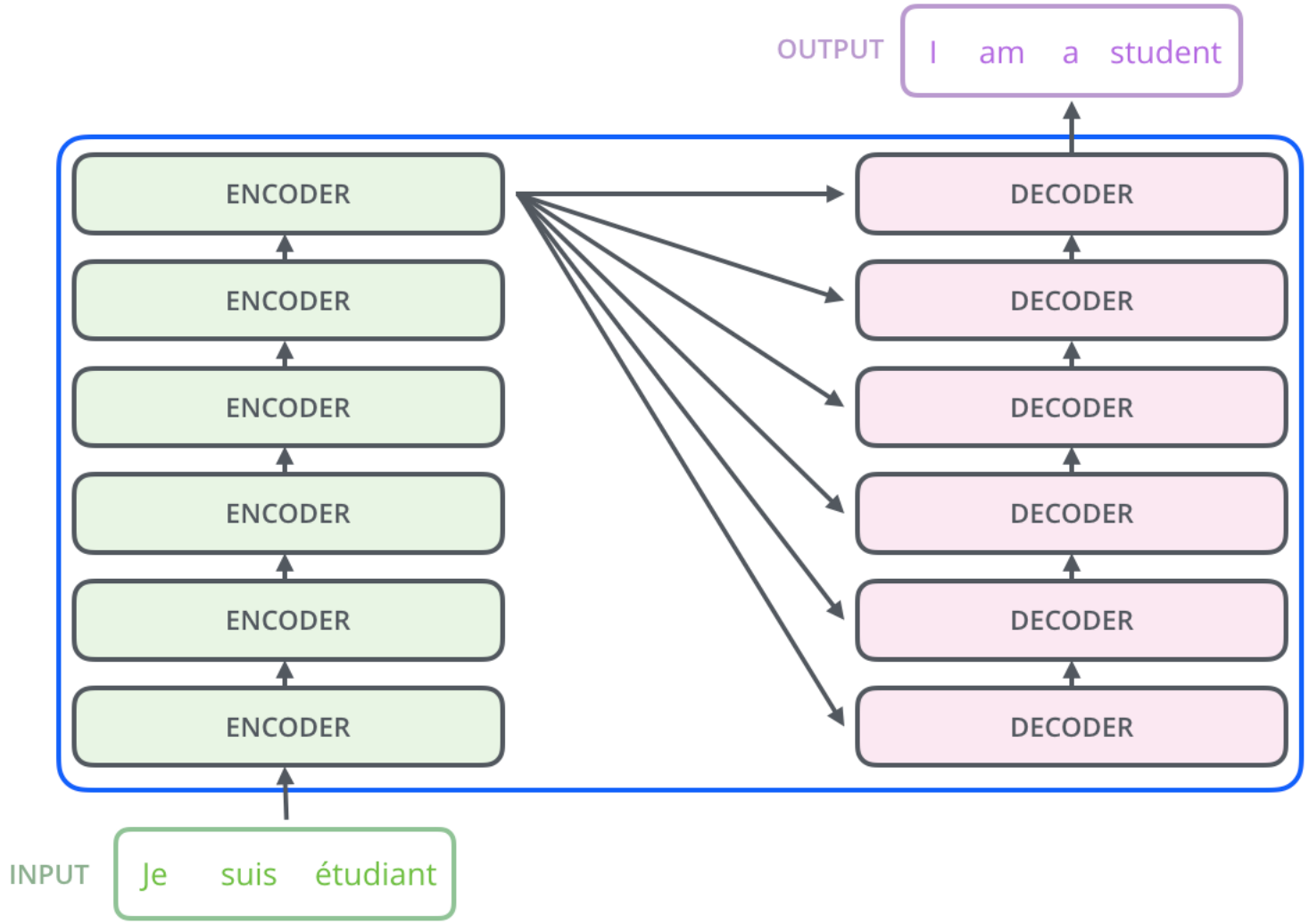
Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

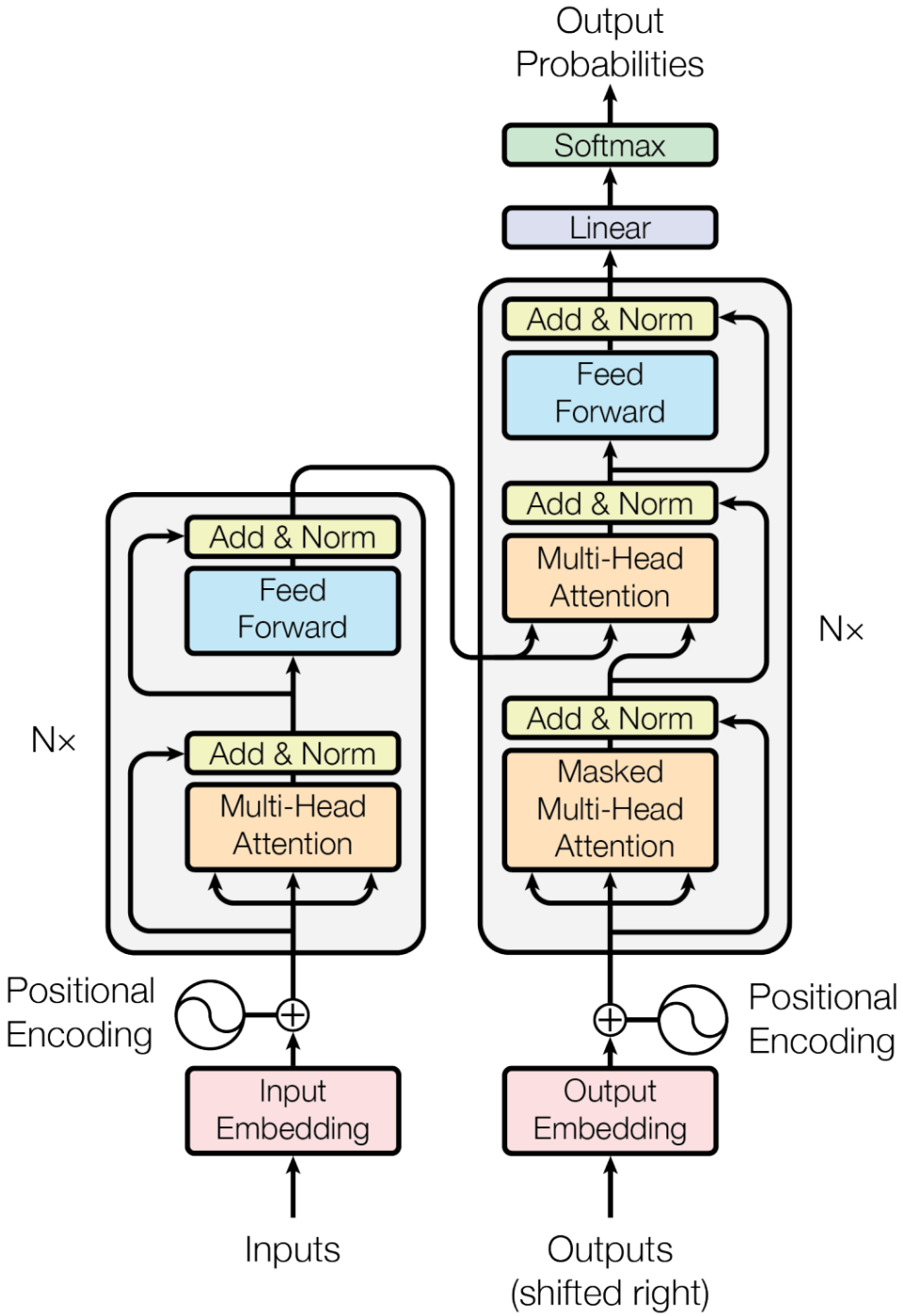
Transformers

Transformer

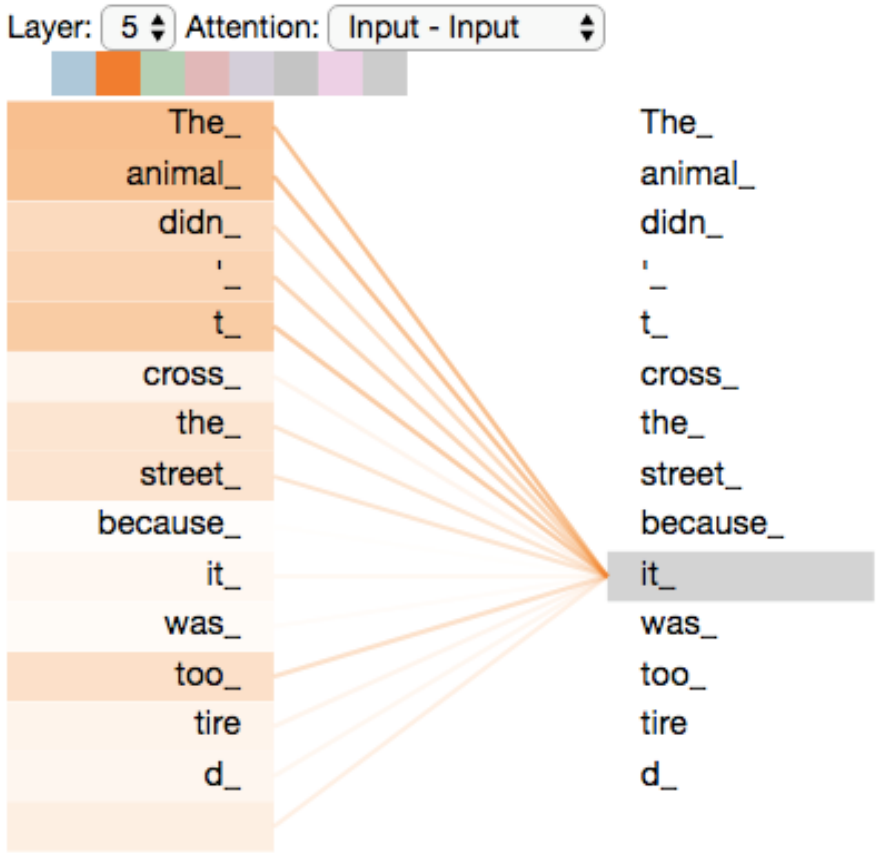


https://jalammar.github.io/illustrated-transformer/

Transformer

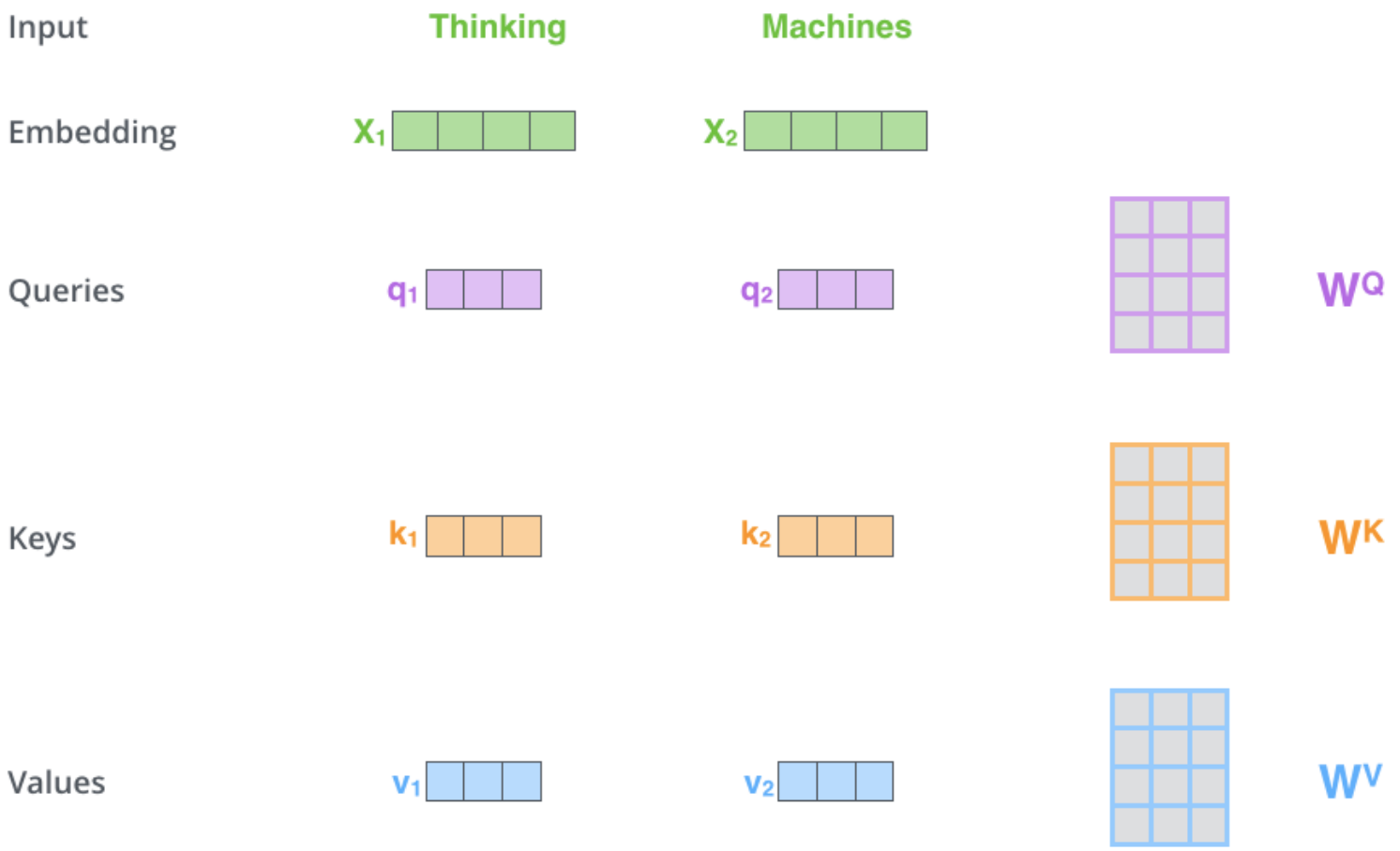


Transformer-Attention



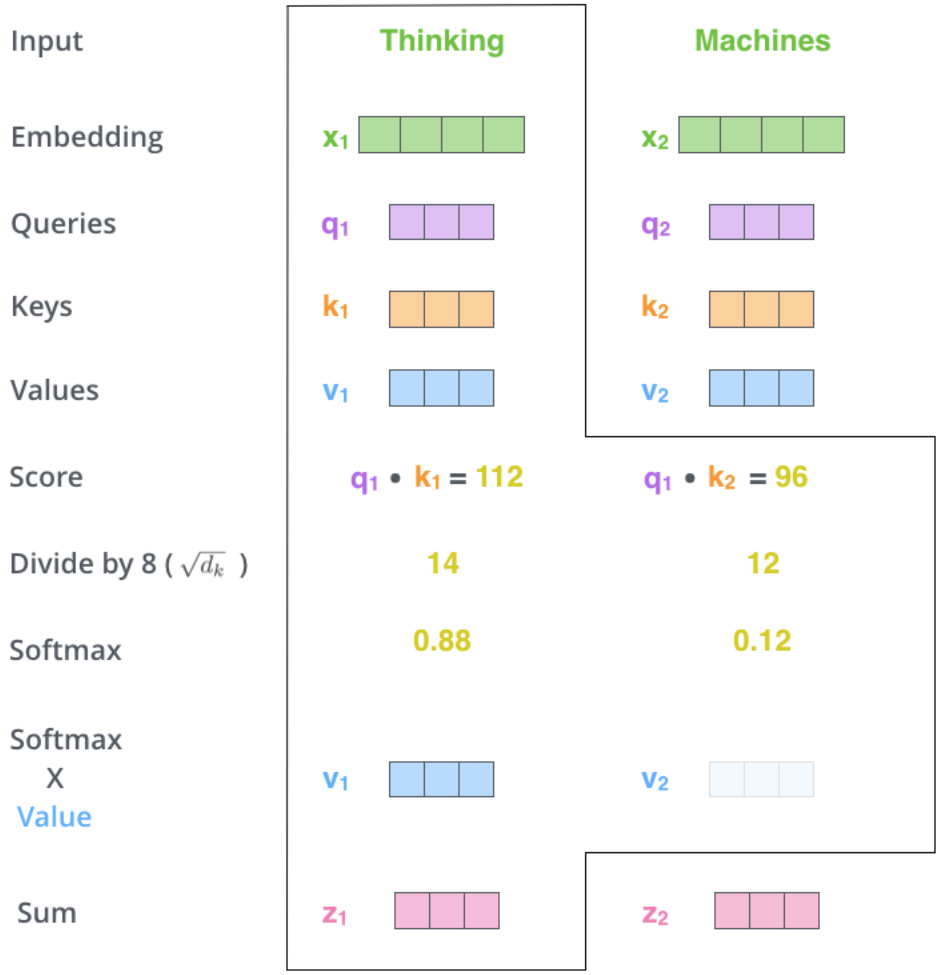
https://jalammar.github.io/illustrated-transformer/

Transformer-Attention



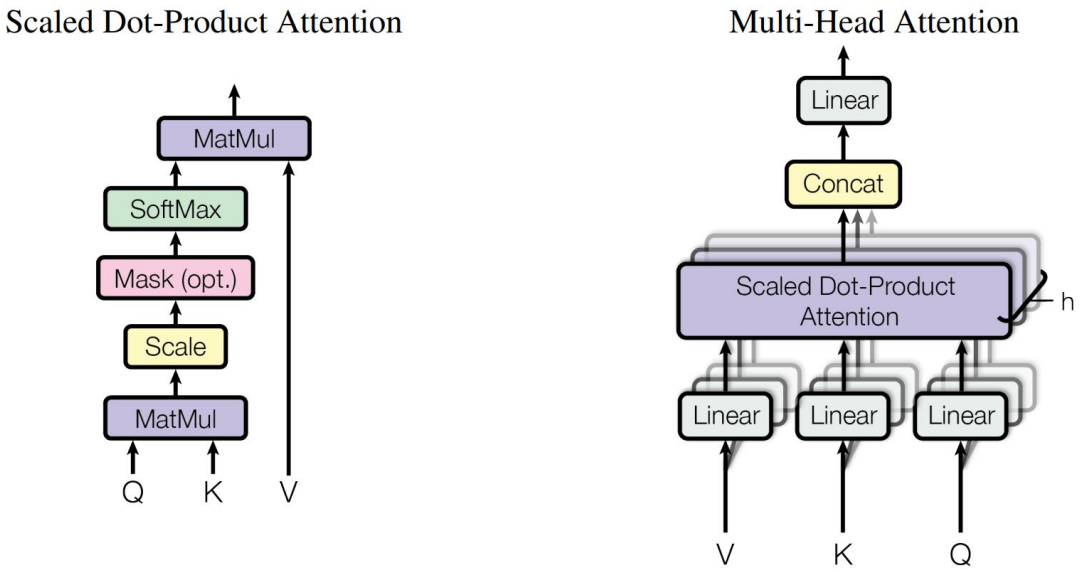
https://jalammar.github.io/illustrated-transformer/

Transformer-Attention



https://jalammar.github.io/illustrated-transformer/

Transformer



$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Transformer

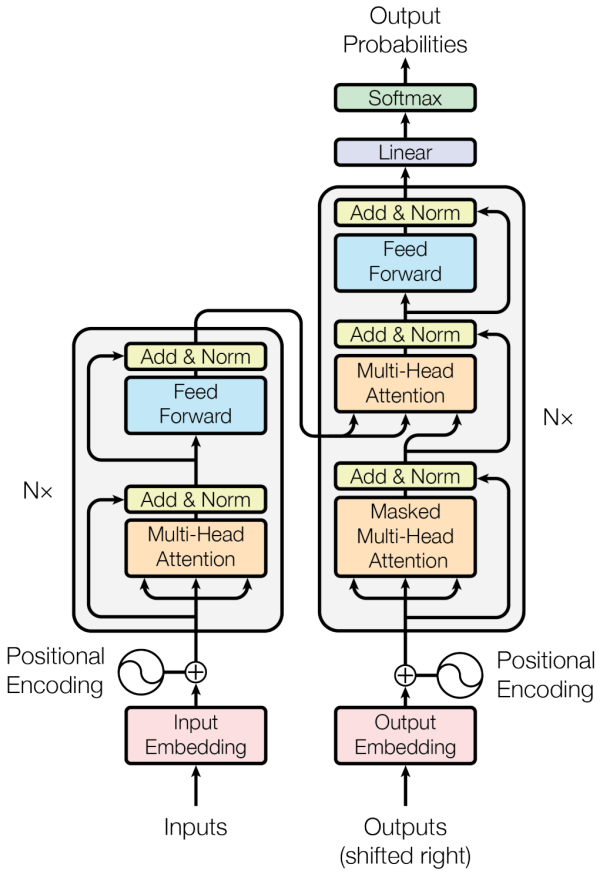
- Note! Not Pre-training yet!

> We trained on the standard WMT 2014 English-German dataset consisting of about 4.5 million sentence pairs. Sentences were encoded using byte-pair encoding [3], which has a shared source-target vocabulary of about 37000 tokens. For English-French, we used the significantly larger WMT 2014 English-French dataset consisting of 36M sentences and split tokens into a 32000 word-piece vocabulary [31]. Sentence pairs were batched together by approximate sequence length. Each training batch contained a set of sentence pairs containing approximately 25000 source tokens and 25000 target tokens.

8 NVIDIA P100 GPUs, base models for a total of 100,000 steps or 12 hours, big models were trained for 300,000 steps (3.5 days)

Transformer-Based Pre-training Models

- Encoders: BERT
- Decoders: GPT
- Encoder-Decoders: BART, T5



GPT

Improving Language Understanding by Generative Pre-Training

Alec Radford OpenAI alec@openai.com
Karthik Narasimhan OpenAI karthikn@openai.com
Tim Salimans OpenAI tim@openai.com
Ilya Sutskever OpenAI ilyasu@openai.com

Language Modeling
auto-regressive generation

Text: Second Law of Robotics: A robot must obey the orders given it by human beings

Generated training examples

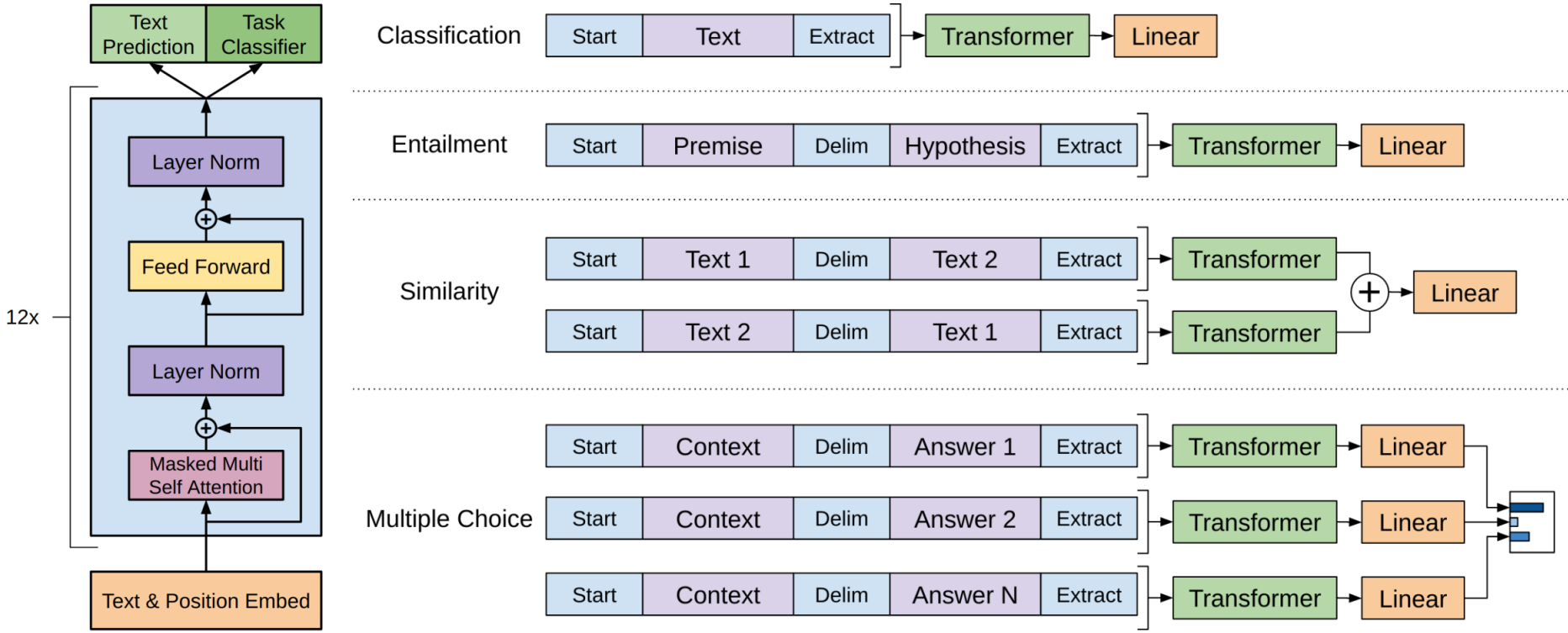
Example #	Input (features)	Correct output (labels)
1	Second law of robotics :	a
2	Second law of robotics : a	robot
3	Second law of robotics : a robot	must
...		

GPT

Improving Language Understanding by Generative Pre-Training

Alec Radford OpenAI alec@openai.com
Karthik Narasimhan OpenAI karthikn@openai.com
Tim Salimans OpenAI tim@openai.com
Ilya Sutskever OpenAI ilyasu@openai.com

Transformer “Decoder”, Language Modeling



BooksCorpus dataset (800M words)

GPT

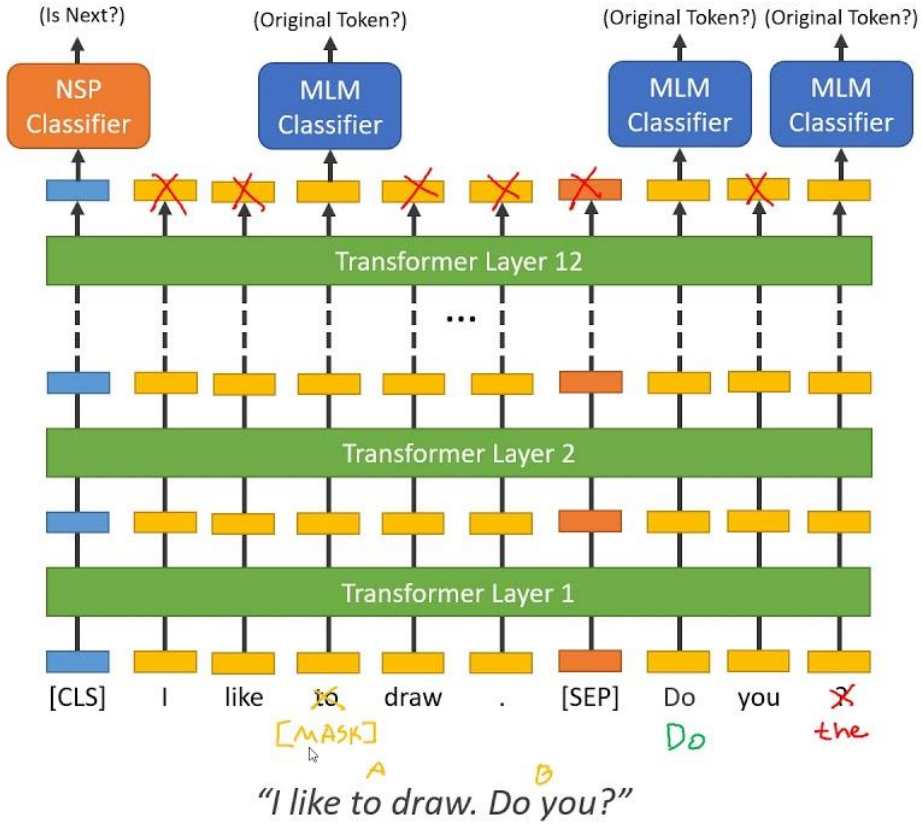
Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

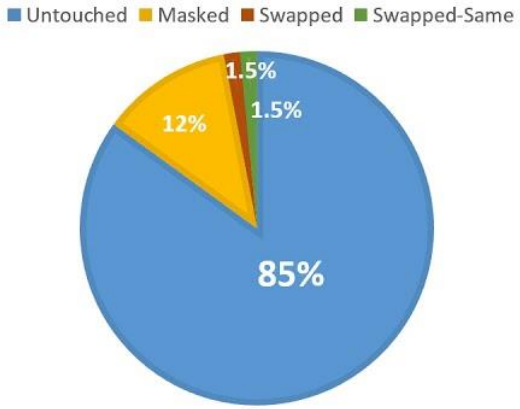
BERT

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language

Transformer Encoder





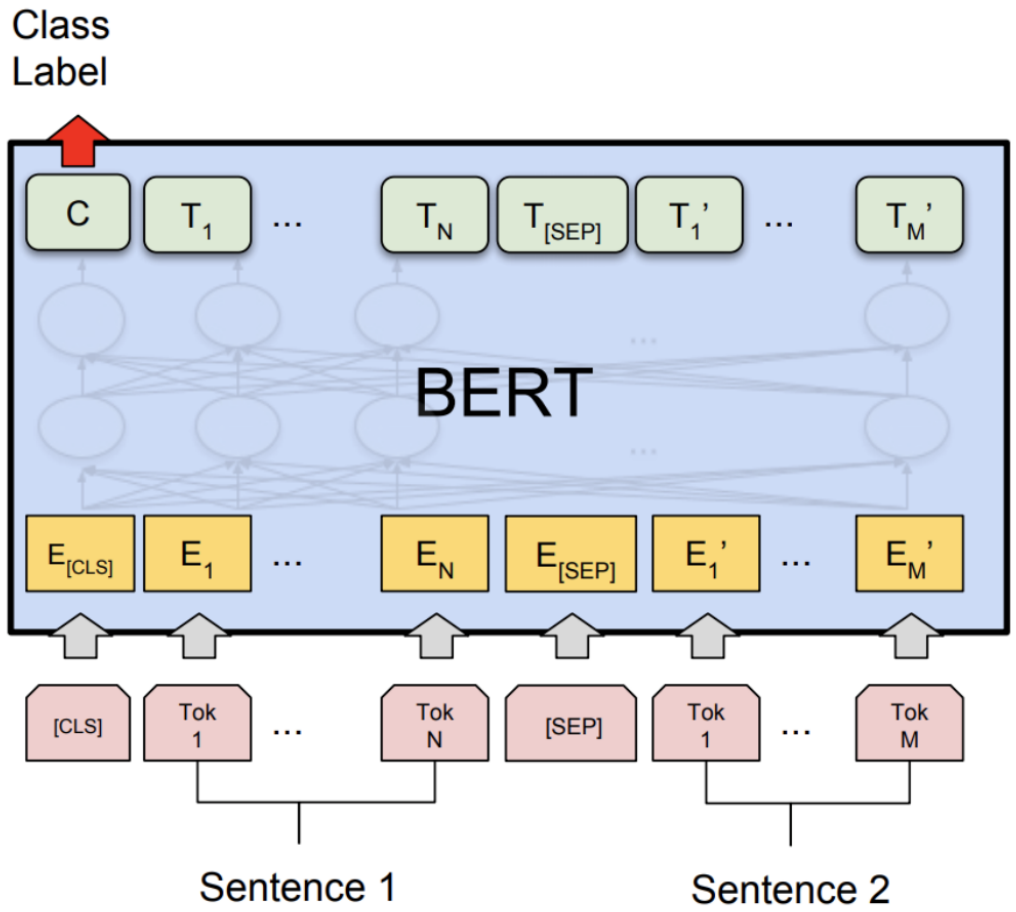
Chris McCormick

BooksCorpus (800M words) and English Wikipedia (2,500M words)

Training of $\text{BERT}_{\text{BASE}}$ was performed on 4 Cloud TPUs in Pod configuration (16 TPU chips total).[13] Training of $\text{BERT}_{\text{LARGE}}$ was performed on 16 Cloud TPUs (64 TPU chips total). Each pre-training took 4 days to complete.

BERT

Fine-tuning

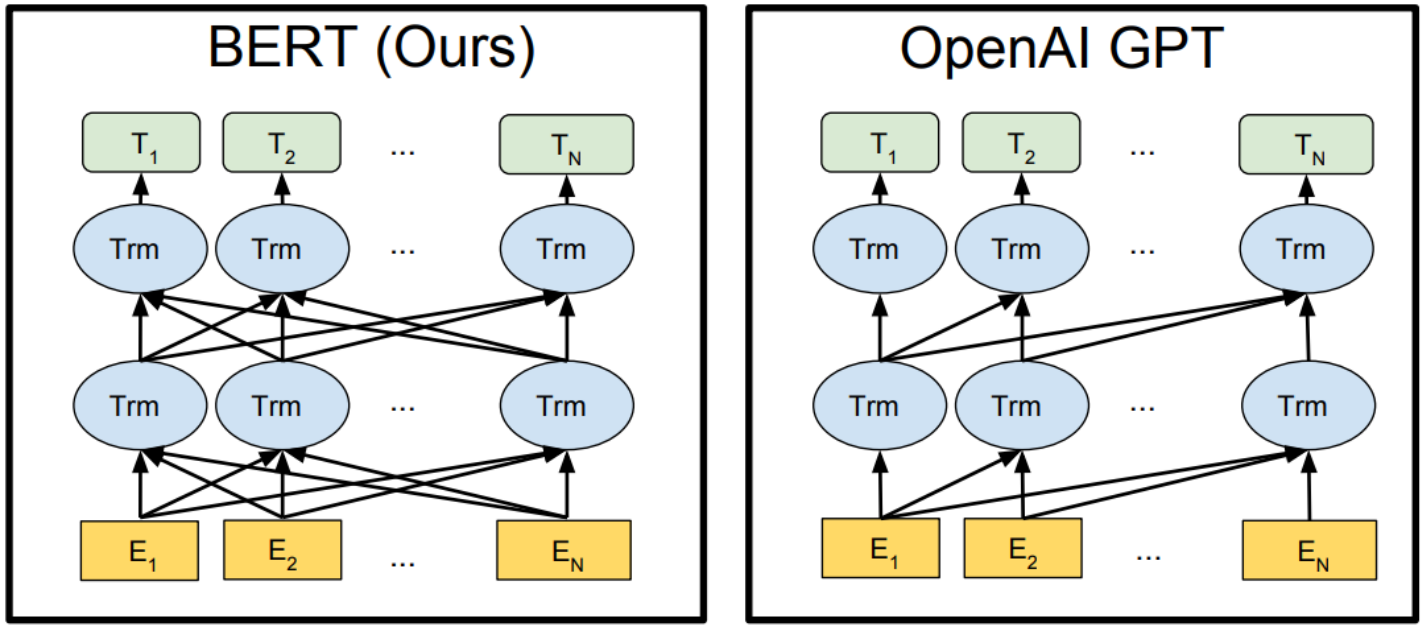


BERT

- BERT results

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

BERT VS GPT



BERT: Masked Language Modeling, Next Sentence Prediction
GPT: Language Modeling (auto-regressively predict the next word)

RoBERTa

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Yinhan Liu[*§] **Myle Ott**[*§] **Naman Goyal**[*§] **Jingfei Du**[*§] **Mandar Joshi**[†]
Danqi Chen[§] **Omer Levy**[§] **Mike Lewis**[§] **Luke Zettlemoyer**[†§] **Veselin Stoyanov**[§]

[†] Paul G. Allen School of Computer Science & Engineering,
University of Washington, Seattle, WA
{mandar90,lsz}@cs.washington.edu

[§] Facebook AI
{yinhanliu,myleott,naman,jingfeidu,
danqi,omerlevy,mikelewis,lsz,ves}@fb.com

RoBERTa

1. Training the model longer, with bigger batches, over more data (160G vs 16G);
2. Removing the next sentence prediction objective;
3. Training on longer sequences;
4. Dynamically changing the masking pattern applied to the training data

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7
$\text{XLNet}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	94.0/87.8	88.4	94.4
+ additional data	126GB	2K	500K	94.5/88.8	89.8	95.6

ELECTRA

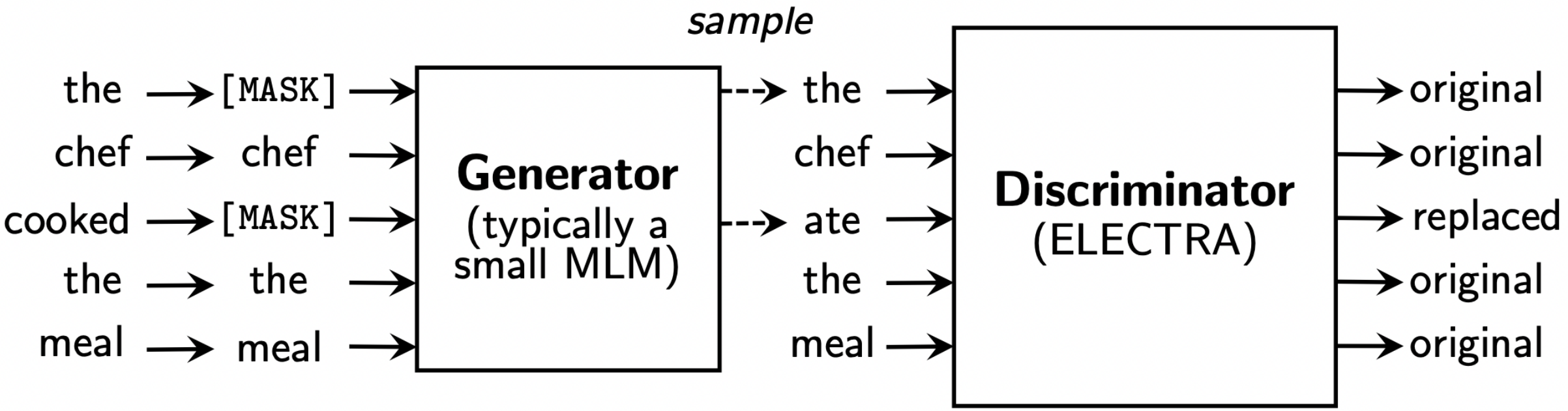
ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators

Kevin Clark
Stanford University
kevclark@cs.stanford.edu

Minh-Thang Luong
Google Brain
thangluong@google.com

Quoc V. Le
Google Brain
qvl@google.com

Christopher D. Manning
Stanford University & CIFAR Fellow
manning@cs.stanford.edu



33B tokens from ClueWeb, CommonCrawl, and Gigaword

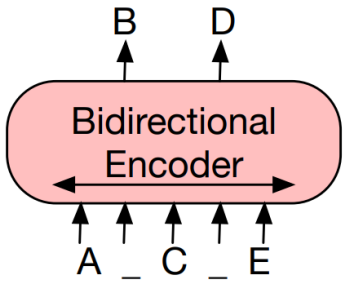
ELECTRA

Model	Train / Infer FLOPs	Speedup	Params	Train Time + Hardware	GLUE
ELMo	3.3e18 / 2.6e10	19x / 1.2x	96M	14d on 3 GTX 1080 GPUs	71.2
GPT	4.0e19 / 3.0e10	1.6x / 0.97x	117M	25d on 8 P6000 GPUs	78.8
BERT-Small	1.4e18 / 3.7e9	45x / 8x	14M	4d on 1 V100 GPU	75.1
BERT-Base	6.4e19 / 2.9e10	1x / 1x	110M	4d on 16 TPUv3s	82.2
ELECTRA-Small	1.4e18 / 3.7e9	45x / 8x	14M	4d on 1 V100 GPU	79.9
50% trained	7.1e17 / 3.7e9	90x / 8x	14M	2d on 1 V100 GPU	79.0
25% trained	3.6e17 / 3.7e9	181x / 8x	14M	1d on 1 V100 GPU	77.7
12.5% trained	1.8e17 / 3.7e9	361x / 8x	14M	12h on 1 V100 GPU	76.0
6.25% trained	8.9e16 / 3.7e9	722x / 8x	14M	6h on 1 V100 GPU	74.1
ELECTRA-Base	6.4e19 / 2.9e10	1x / 1x	110M	4d on 16 TPUv3s	85.1

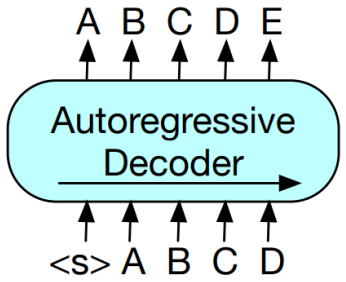
BART

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

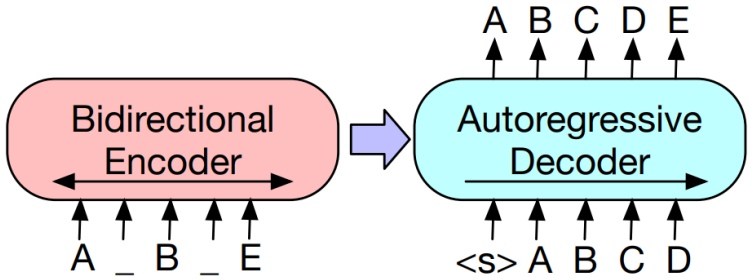
Mike Lewis*, Yinhan Liu*, Naman Goyal*, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer
Facebook AI
{mikelewis,yinhanliu,naman}@fb.com



(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.



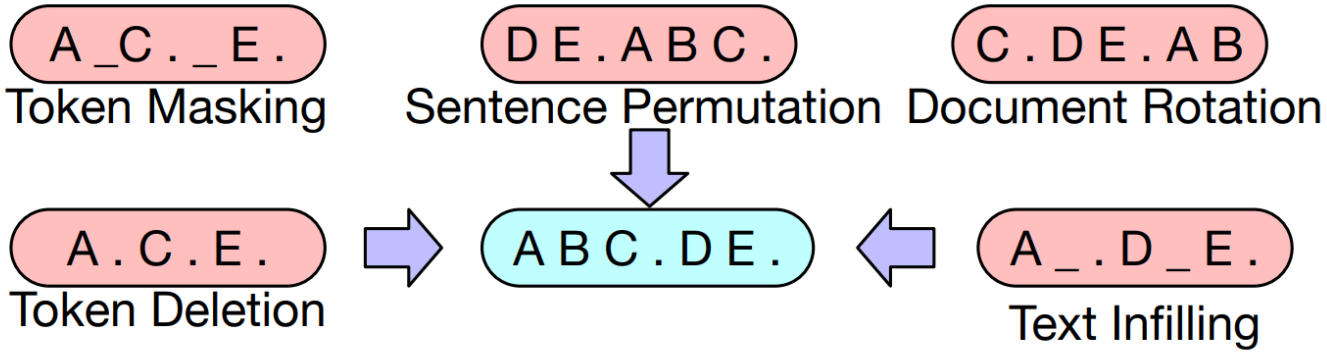
(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.



(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

BART

Same training data as RoBERTa(160G)



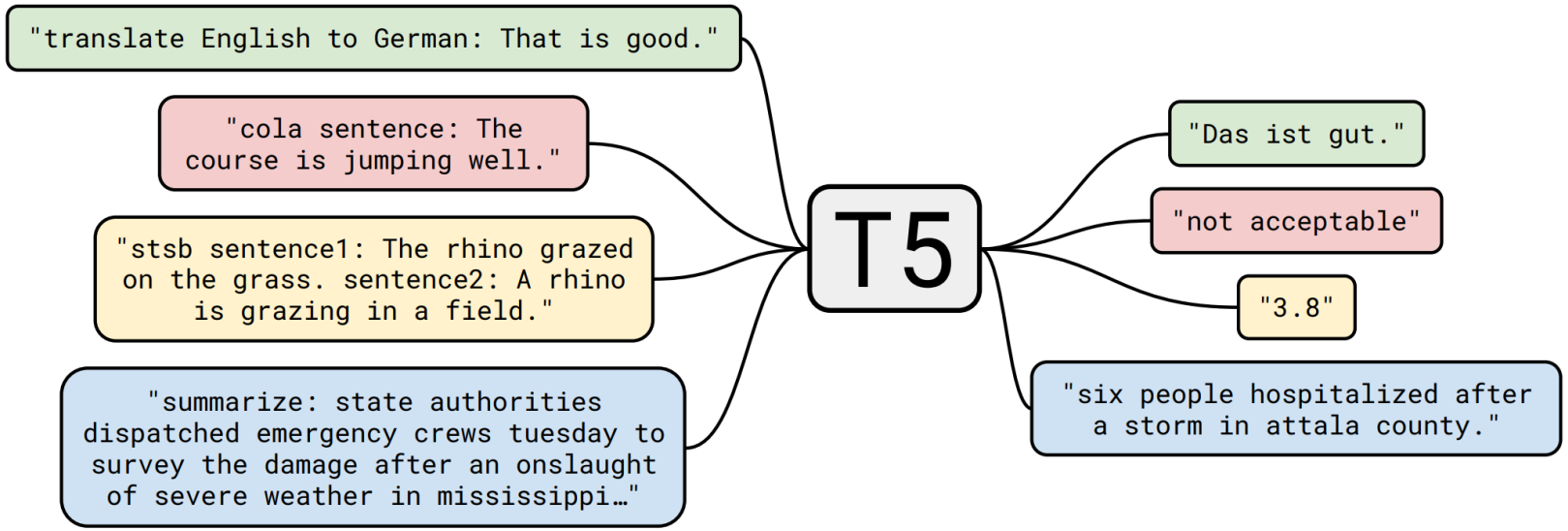
Model	SQuAD 1.1 F1	MNLI Acc	ELI5 PPL	XSum PPL	ConvAI2 PPL	CNN/DM PPL
BERT Base (Devlin et al., 2019)	88.5	**84.3**	-	-	-	-
Masked Language Model	90.0	83.5	24.77	7.87	12.59	7.06
Masked Seq2seq	87.0	82.1	23.40	6.80	11.43	6.19
Language Model	76.7	80.1	**21.40**	7.00	11.51	6.56
Permuted Language Model	89.1	83.7	24.03	7.69	12.23	6.96
Multitask Masked Language Model	89.2	82.4	23.73	7.50	12.39	6.74
BART Base						
w/ Token Masking	90.4	84.1	25.05	7.08	11.73	6.10
w/ Token Deletion	90.4	84.1	24.61	6.90	11.46	5.87
w/ Text Infilling	**90.8**	84.0	24.26	**6.61**	**11.05**	5.83
w/ Document Rotation	77.2	75.3	53.69	17.14	19.87	10.59
w/ Sentence Shuffling	85.4	81.5	41.87	10.93	16.67	7.89
w/ Text Infilling + Sentence Shuffling	**90.8**	83.8	24.17	6.62	11.12	**5.41**

T5

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Colin Raffel*	CRAFFEL@GMAIL.COM
Noam Shazeer*	NOAM@GOOGLE.COM
Adam Roberts*	ADAROB@GOOGLE.COM
Katherine Lee*	KATHERINELEE@GOOGLE.COM
Sharan Narang	SHARANNARANG@GOOGLE.COM
Michael Matena	MMATENA@GOOGLE.COM
Yanqi Zhou	YANQIZ@GOOGLE.COM
Wei Li	MWEILI@GOOGLE.COM
Peter J. Liu	PETERJLIU@GOOGLE.COM

Google, Mountain View, CA 94043, USA



Colossal Clean Crawled Corpus (750 GB)

T5

Model	GLUE Average	CoLA Matthew's	SST-2 Accuracy	MRPC F1	MRPC Accuracy	STS-B Pearson	STS-B Spearman
Previous best	89.4[a]	69.2[b]	97.1[a]	**93.6**[b]	**91.5**[b]	92.7[b]	92.3[b]
T5-Small	77.4	41.0	91.8	89.7	86.6	85.6	85.0
T5-Base	82.7	51.1	95.2	90.7	87.5	89.4	88.6
T5-Large	86.4	61.2	96.3	92.4	89.9	89.9	89.2
T5-3B	88.5	67.1	97.4	92.5	90.0	90.6	89.8
T5-11B	**90.3**	**71.6**	**97.5**	92.8	90.4	**93.1**	**92.8**

Model	QQP F1	QQP Accuracy	MNLI-m Accuracy	MNLI-mm Accuracy	QNLI Accuracy	RTE Accuracy	WNLI Accuracy
Previous best	74.8[c]	**90.7**[b]	91.3[a]	91.0[a]	**99.2**[a]	89.2[a]	91.8[a]
T5-Small	70.0	88.0	82.4	82.3	90.3	69.9	69.2
T5-Base	72.6	89.4	87.1	86.2	93.7	80.1	78.8
T5-Large	73.9	89.9	89.9	89.6	94.8	87.2	85.6
T5-3B	74.4	89.7	91.4	91.2	96.3	91.1	89.7
T5-11B	**75.1**	90.6	**92.2**	**91.9**	96.9	**92.8**	**94.5**

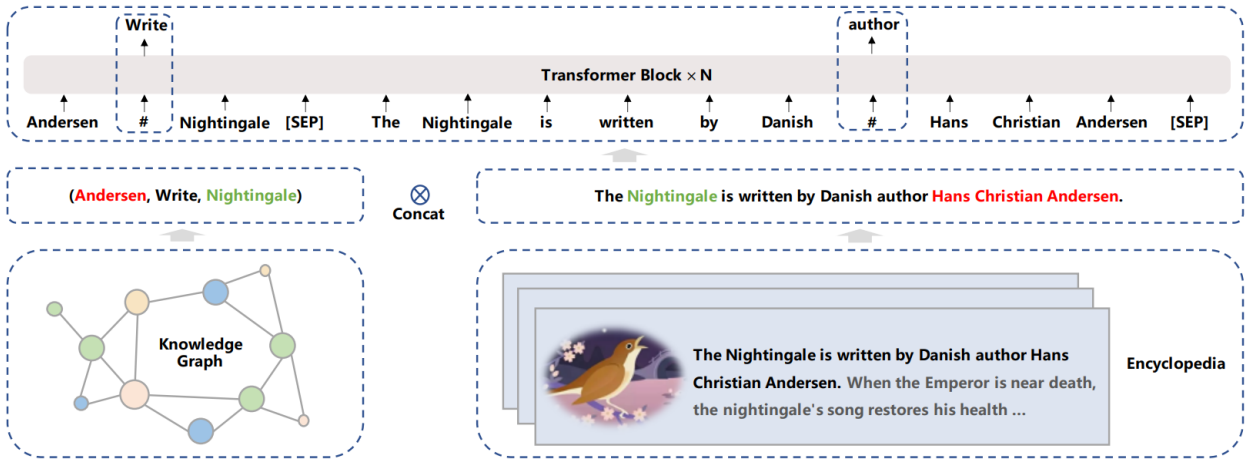
Model	SQuAD EM	SQuAD F1	SuperGLUE Average	BoolQ Accuracy	CB F1	CB Accuracy	COPA Accuracy
Previous best	90.1[a]	95.5[a]	84.6[d]	87.1[d]	90.5[d]	95.2[d]	90.6[d]
T5-Small	79.10	87.24	63.3	76.4	56.9	81.6	46.0
T5-Base	85.44	92.08	76.2	81.4	86.2	94.0	71.2
T5-Large	86.66	93.79	82.3	85.4	91.6	94.8	83.4
T5-3B	88.53	94.95	86.4	89.9	90.3	94.4	92.0
T5-11B	**91.26**	**96.22**	**88.9**	**91.2**	**93.9**	**96.8**	**94.8**

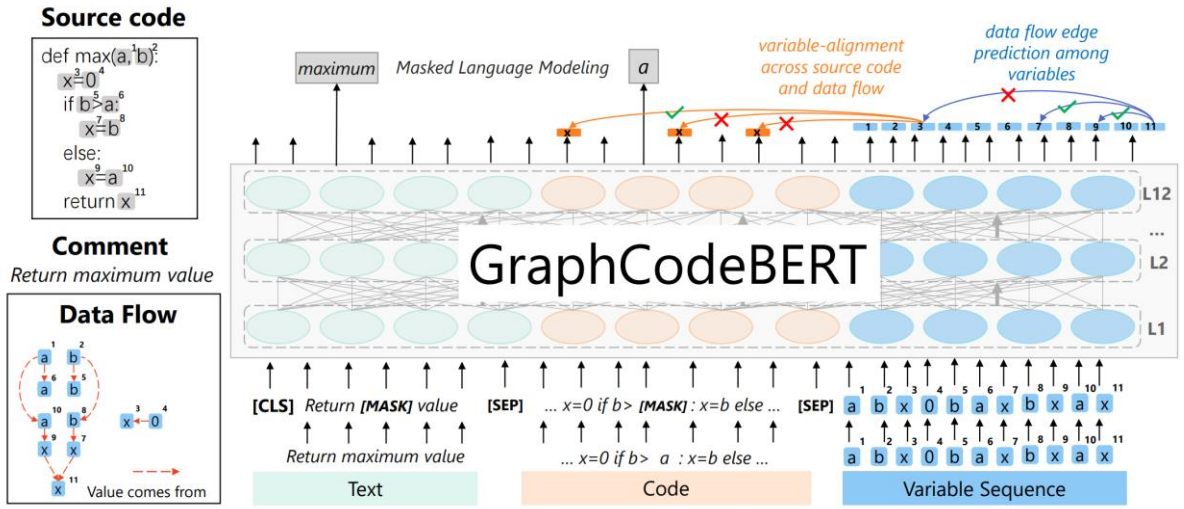
Model	MultiRC F1a	MultiRC EM	ReCoRD F1	ReCoRD Accuracy	RTE Accuracy	WiC Accuracy	WSC Accuracy
Previous best	84.4[d]	52.5[d]	90.6[d]	90.0[d]	88.2[d]	69.9[d]	89.0[d]
T5-Small	69.3	26.3	56.3	55.4	73.3	66.9	70.5
T5-Base	79.7	43.1	75.0	74.2	81.5	68.3	80.8
T5-Large	83.3	50.7	86.8	85.9	87.8	69.3	86.3
T5-3B	86.8	58.3	91.2	90.4	90.7	72.1	90.4
T5-11B	**88.1**	**63.3**	**94.1**	**93.4**	**92.5**	**76.9**	**93.8**

Model	WMT EnDe BLEU	WMT EnFr BLEU	WMT EnRo BLEU	CNN/DM ROUGE-1	CNN/DM ROUGE-2	CNN/DM ROUGE-L
Previous best	**33.8**[e]	**43.8**[e]	**38.5**[f]	43.47[g]	20.30[g]	40.63[g]
T5-Small	26.7	36.0	26.8	41.12	19.56	38.35
T5-Base	30.9	41.2	28.0	42.05	20.34	39.40
T5-Large	32.0	41.5	28.1	42.50	20.68	39.75
T5-3B	31.8	42.6	28.2	42.72	21.02	39.94
T5-11B	32.1	43.4	28.1	**43.52**	**21.55**	**40.69**

Other Pre-training Models

- Numerous other pre-training models with various objectives
- Ernie (Baidu): inject knowledge
- GraphCodeBERT: Programming language



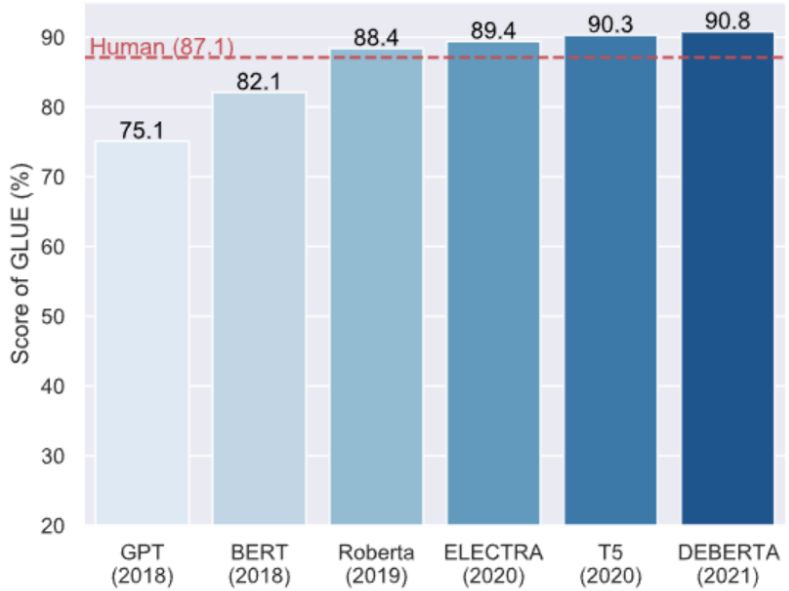


Trend

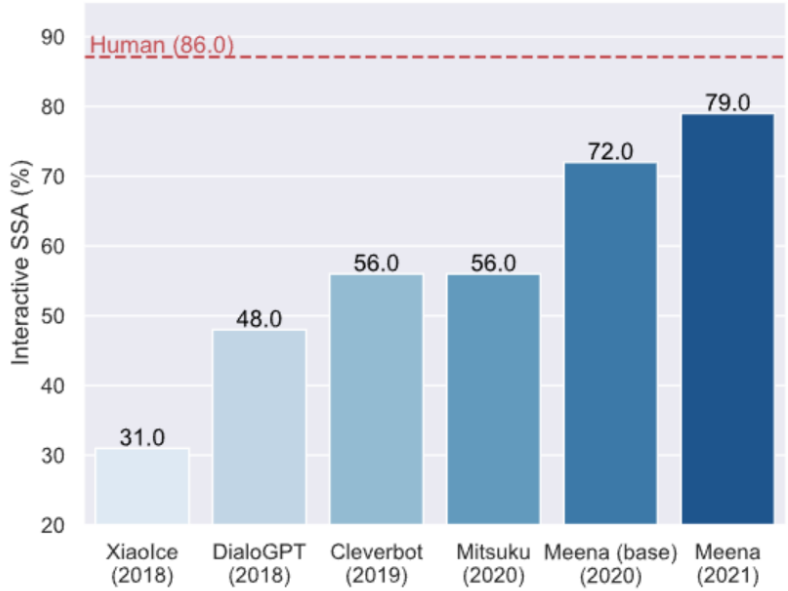
Pre-trained models: Past, present and future

Xu Han [a,1,*], Zhengyan Zhang [a,1], Ning Ding [a,1], Yuxian Gu [a,1], Xiao Liu [a,1], Yuqi Huo [b,1], Jiezhong Qiu [a], Yuan Yao [a], Ao Zhang [a], Liang Zhang [b], Wentao Han [a,2], Minlie Huang [a,2], Qin Jin [b,2], Yanyan Lan [d,2], Yang Liu [a,d,2], Zhiyuan Liu [a,2], Zhiwu Lu [c,2], Xipeng Qiu [e,2], Ruihua Song [c,2], Jie Tang [a,2], Ji-Rong Wen [c,2], Jinhui Yuan [f,2], Wayne Xin Zhao [c,2], Jun Zhu [a,2]

[a] *Department of Computer Science and Technology, Tsinghua University, Beijing, China*
[b] *School of Information, Renmin University of China, Beijing, China*
[c] *Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China*
[d] *Institute for AI Industry Research, Tsinghua University, Beijing, China*
[e] *School of Computer Science, Fudan University, Shanghai, China*
[f] *OneFlow Inc., Beijing, China*

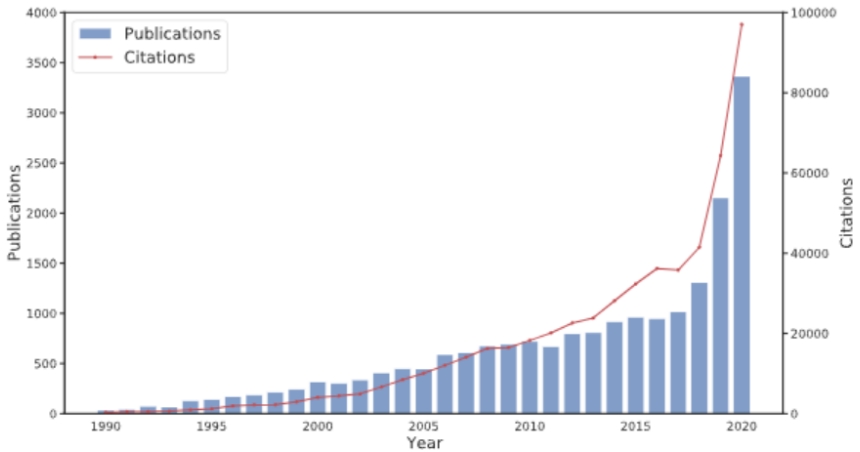


(a) Evaluation on language understanding benchmark GLUE.

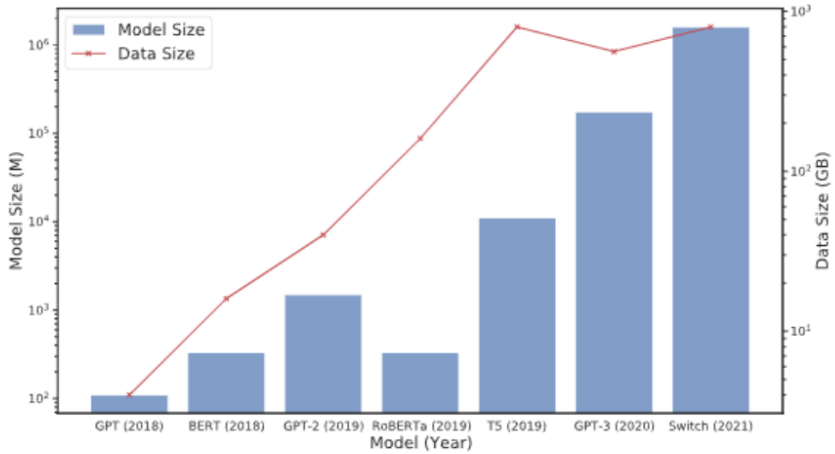


(b) Manual evaluation on dialogue systems.

Fig. 1. The two figures show the significant improvement on performance of both language understanding and language generation after using large-scale PTMs.



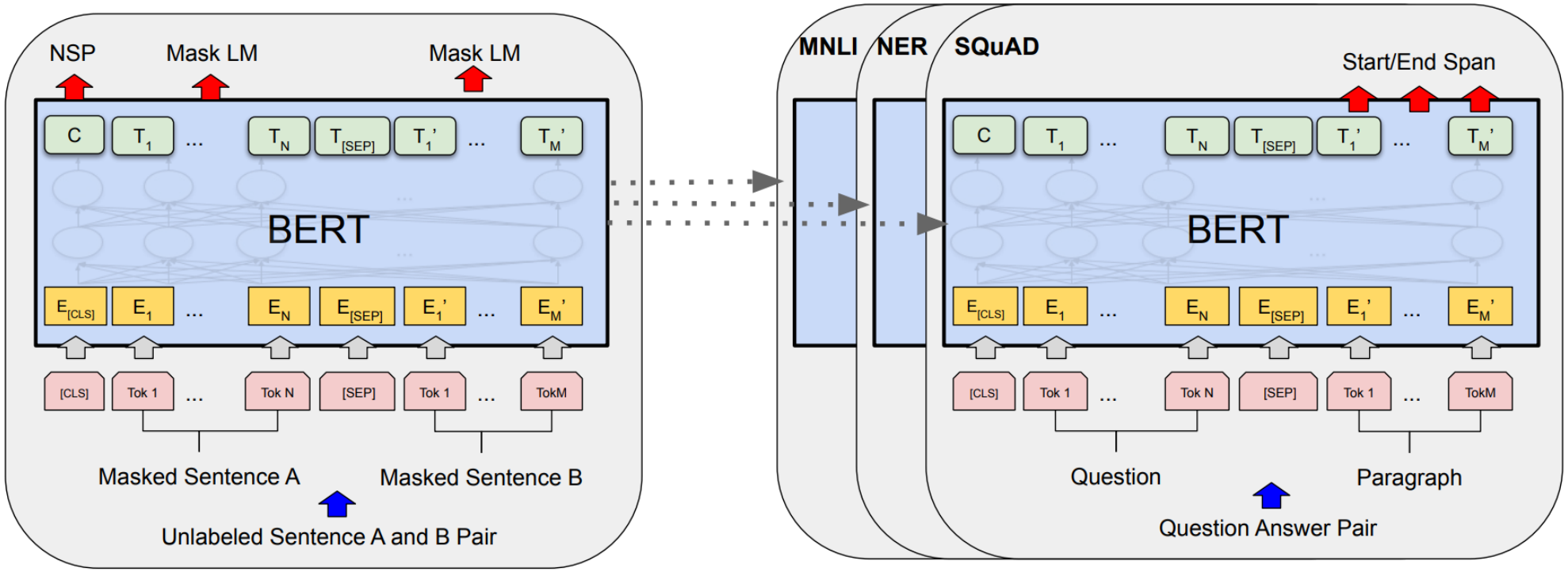
(a) The number of publications on "language models" and their citations in recent years.



(b) The model size and data size applied by recent NLP PTMs. A base-10 log scale is used for the figure.

Prompt

Recall: Pre-train and Fine-Tune



Pre-training

Large Corpus

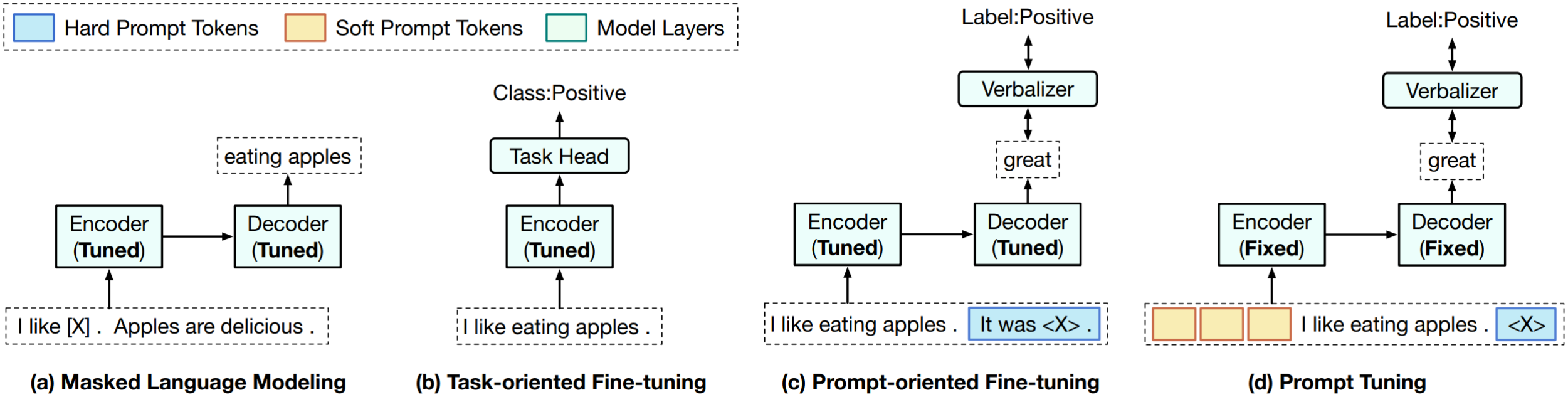
Language understanding/generation objectives

Fine-Tuning

Downstream tasks

Specific losses

Fine-tune -> Prompt



Transfer **from** Pre-train domain **to** downstream tasks

Transfer **from** downstream tasks **to** Pre-train domain

Outline

- What Language Models learn? ->**LAMA** Petroni et al. (2019)
- What is prompt? ->**PET** Schick and Schutze. (2020)
- How to automatically find template?
- -> **Autoprompt** (discrete) Shin et al. (2020)
- -> **P-tuning** (continuous) Liu et al. (2021)
- -> **Prefix-tuning** (continuous) Li and Liang. (2021)

What Language Models learn? ->**LAMA**

- GPT:

*Transformers have achieved …… multiple pieces of texts. To **summarize**:*
->auto generate summary for this text

prompt

LAMA ->dataset

- Google-RE -> “place of birth”, “date of birth” and “place of death”
- T-REx -> 41 Wikidata relations
- ConceptNet -> Open Mind Common Sense between words and/or phrases
- SQuAD -> context-insensitive questions with single token answers

	Relation	Query	Answer	Generation
T-Rex	P19	Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8] , **Florence**
	P20	Adolphe Adam died in ____.	Paris	**Paris [-0.5]** , London [-
	P279	English bulldog is a subclass of ____.	dog	dogs [-0.3] , breeds [-2.2
	P37	The official language of Mauritius is ____.	English	**English [-0.6]** , French
	P413	Patrick Oboya plays in ____ position.	midfielder	centre [-2.0] , center [-2.
	P138	Hamburg Airport is named after ____.	Hamburg	Hess [-7.0] , Hermann [-
	P364	The original language of Mon oncle Benjamin is ____.	French	**French [-0.2]** , Breton [
	P54	Dani Alves plays with ____ .	Barcelona	Santos [-2.4] , Porto [-2.
	P106	Paul Toungui is a ____ by profession .	politician	lawyer [-1.1] , journalist
	P527	Sodium sulfide consists of ____.	sodium	water [-1.2] , sulfur [-1.7
	P102	Gordon Scholes is a member of the ____ political party.	Labor	Labour [-1.3] , Conserva
	P530	Kenya maintains diplomatic relations with ____.	Uganda	India [-3.0] , **Uganda [-**
	P176	iPod Touch is produced by ____.	Apple	**Apple [-1.6]** , Nokia [-1.
	P30	Bailey Peninsula is located in ____.	Antarctica	**Antarctica [-1.4]** , Berm
	P178	JDK is developed by ____.	Oracle	IBM [-2.0] , Intel [-2.3] ,
	P1412	Carl III used to communicate in ____.	Swedish	German [-1.6] , Latin [-
	P17	Sunshine Coast, British Columbia is located in ____.	Canada	**Canada [-1.2]** , Alberta
	P39	Pope Clement VII has the position of ____ .	pope	cardinal [-2.4] , Pope [-2
	P264	Joe Cocker is represented by music label ____.	Capitol	EMI [-2.6] , BMG [-2.6]
	P276	London Jazz Festival is located in ____.	London	**London [-0.3]** , Greenw
	P127	Border TV is owned by ____.	ITV	Sky [-3.1] , **ITV [-3.3]** , (
	P103	The native language of Mammootty is ____.	Malayalam	**Malayalam [-0.2]** , Tam
	P495	The Sharon Cuneta Show was created in ____.	Philippines	Manila [-3.2] , **Philippi**

ConceptNet	AtLocation	You are likely to find a overflow in a ____.	drain	sewer [-3.1] , c
	CapableOf	Ravens can ____.	fly	**fly [-1.5]** , fight
	CausesDesire	Joke would make you want to ____.	laugh	cry [-1.7] , die
	Causes	Sometimes virus causes ____.	infection	disease [-1.2] ,
	HasA	Birds have ____.	feathers	wings [-1.8] , n
	HasPrerequisite	Typing requires ____.	speed	patience [-3.5]
	HasProperty	Time is ____.	finite	short [-1.7] , pa
	MotivatedByGoal	You would celebrate because you are ____.	alive	happy [-2.4] , h
	ReceivesAction	Skills can be ____.	taught	acquired [-2.5]
	UsedFor	A pond is for ____.	fish	swimming [-1.

Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE_n	RE_o	Fs	Txl	Eb	E5B	Bb	Bl
Google-RE	`birth-place`	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	**16.1**
	`birth-date`	1825	1	1.9	-	0.0	**1.9**	0.3	1.1	0.1	0.1	1.5	1.4
	`death-place`	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	**14.0**
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**
T-REx	1-1	937	2	1.78	-	0.6	10.0	17.0	36.5	10.1	13.1	68.0	**74.5**
	N-1	20006	23	23.85	-	5.4	**33.8**	6.1	18.0	3.6	6.5	32.4	34.2
	N-*M*	13096	16	21.95	-	7.7	**36.7**	12.0	16.5	5.7	7.4	24.7	24.3
	Total	34039	41	22.03	-	6.1	**33.8**	8.9	18.3	4.7	7.1	31.1	32.3
ConceptNet	Total	11458	16	4.8	-	-	-	3.6	5.7	6.1	6.2	15.6	**19.2**
SQuAD	Total	305	-	-	**37.5**	-	-	3.6	3.9	1.6	4.3	14.1	17.4

Table 2: Mean precision at one (P@1) for a frequency baseline (Freq), DrQA, a relation extraction with naïve entity linking (RE_n), oracle entity linking (RE_o), fairseq-fconv (Fs), Transformer-XL large (Txl), ELMo original (Eb), ELMo 5.5B (E5B), BERT-base (Bb) and BERT-large (Bl) across the set of evaluation corpora.

What is prompt? ->**PET** Schick and Schutze. (2020)

- Pattern-Exploiting Training

Name	Notation	Example	Description
Input	$\boldsymbol{x}$	I love this movie.	One or multiple texts
Output	$\boldsymbol{y}$	++ (very positive)	Output label or text
Prompting Function	$f_{\text{prompt}}(\boldsymbol{x})$	`[X]` Overall, it was a `[Z]` movie.	A function that converts the input into a specific form by inserting the input $\boldsymbol{x}$ and adding a slot `[Z]` where answer $\boldsymbol{z}$ may be filled later.
Prompt	$\boldsymbol{x}'$	I love this movie. Overall, it was a `[Z]` movie.	A text where `[X]` is instantiated by input $\boldsymbol{x}$ but answer slot `[Z]` is not.
Filled Prompt	$f_{\text{fill}}(\boldsymbol{x}', \boldsymbol{z})$	I love this movie. Overall, it was a bad movie.	A prompt where slot `[Z]` is filled with any answer.
Answered Prompt	$f_{\text{fill}}(\boldsymbol{x}', \boldsymbol{z}^*)$	I love this movie. Overall, it was a good movie.	A prompt where slot `[Z]` is filled with a true answer.
Answer	$\boldsymbol{z}$	"good", "fantastic", "boring"	A token, phrase, or sentence that fills `[Z]`

Yelp For the Yelp Reviews Full Star dataset (Zhang et al., 2015), the task is to estimate the rating that a customer gave to a restaurant on a 1- to 5-star scale based on their review's text. We define the following patterns for an input text a:

$P_1(a) =$ It was ____. a $\quad P_2(a) =$ Just ____! $\|$ a

$P_3(a) =$ a. All in all, it was ____.

$P_4(a) =$ a $\|$ In summary, the restaurant is ____.

AG's News AG's News is a news classification dataset, where given a headline a and text body b, news have to be classified as belonging to one of the categories *World* (1), *Sports* (2), *Business* (3) or *Science/Tech* (4). For $\mathbf{x} = (a, b)$, we define the following patterns:

$P_1(\mathbf{x}) =$ ____: a b $\quad P_2(\mathbf{x}) =$ a (____) b

$P_3(\mathbf{x}) =$ ____ – a b $\quad P_4(\mathbf{x}) =$ a b (____)

$P_5(\mathbf{x}) =$ ____ News: a b

$P_6(\mathbf{x}) =$ [Category: ____] a b

We use a verbalizer that maps 1–4 to "World", "Sports", "Business" and "Tech", respectively.

Yahoo Yahoo Questions (Zhang et al., 2015) is a text classification dataset. Given a question a and an answer b, one of ten possible categories has to be assigned. We use the same patterns as for AG's News, but we replace the word "News" in P_5 with the word "Question". We define a verbalizer that maps categories 1–10 to "Society", "Science", "Health", "Education", "Computer", "Sports", "Business", "Entertainment", "Relationship" and "Politics".

MNLI The MNLI dataset (Williams et al., 2018) consists of text pairs $\mathbf{x} = (a, b)$. The task is to find out whether a implies b (0), a and b contradict each other (1) or neither (2). We define

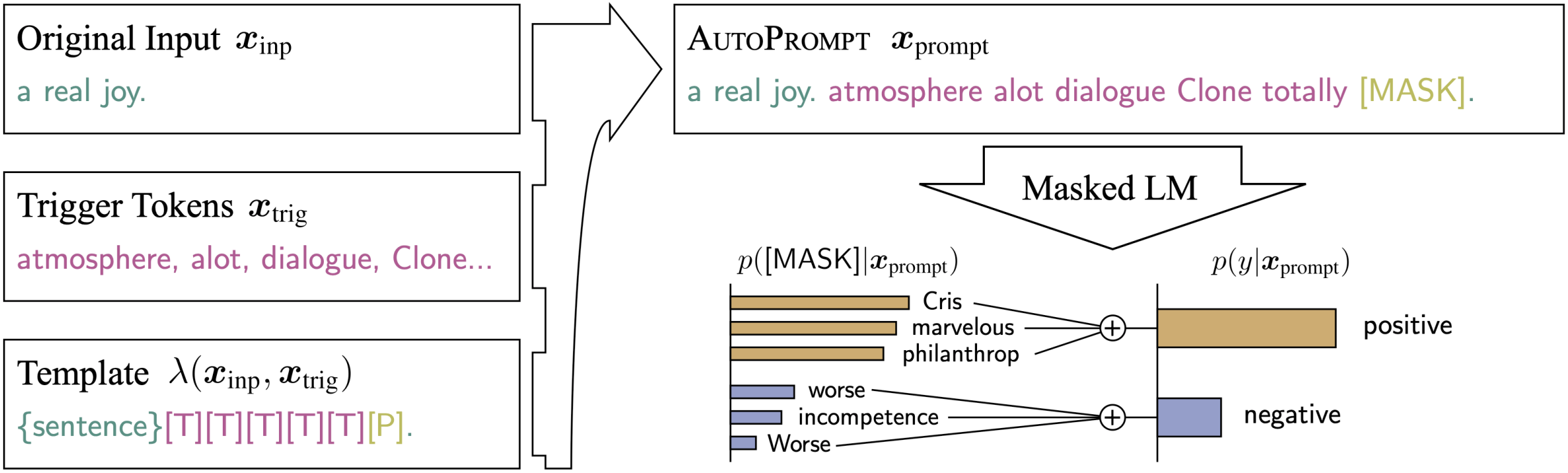
$P_1(\mathbf{x}) =$ "a"? $\|$ ____, "b" $\quad P_2(\mathbf{x}) =$ a? $\|$ ____, b

and consider two different verbalizers v_1 and v_2:

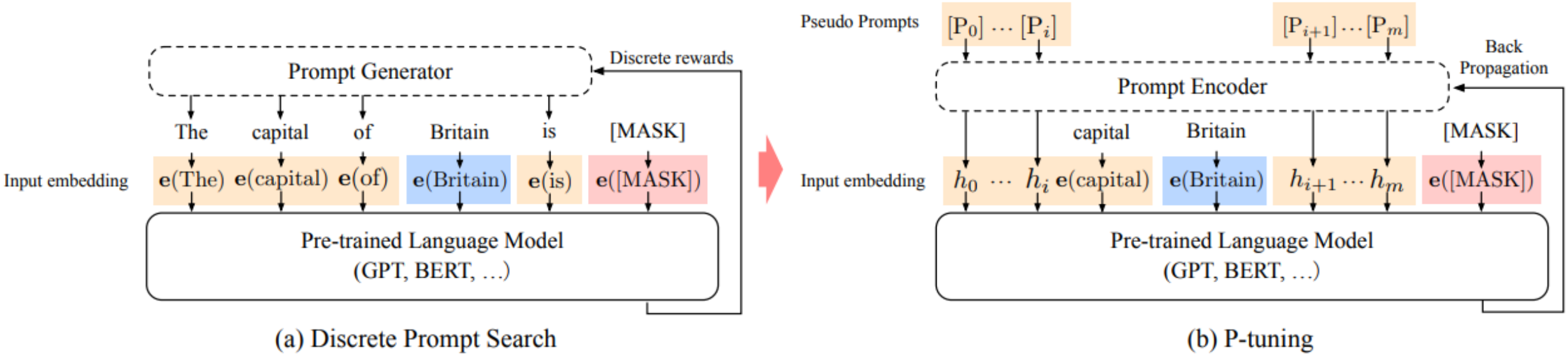
$v_1(0) =$ Wrong $\quad v_1(1) =$ Right $\quad v_1(2) =$ Maybe

$v_2(0) =$ No $\quad v_2(1) =$ Yes $\quad v_2(2) =$ Maybe

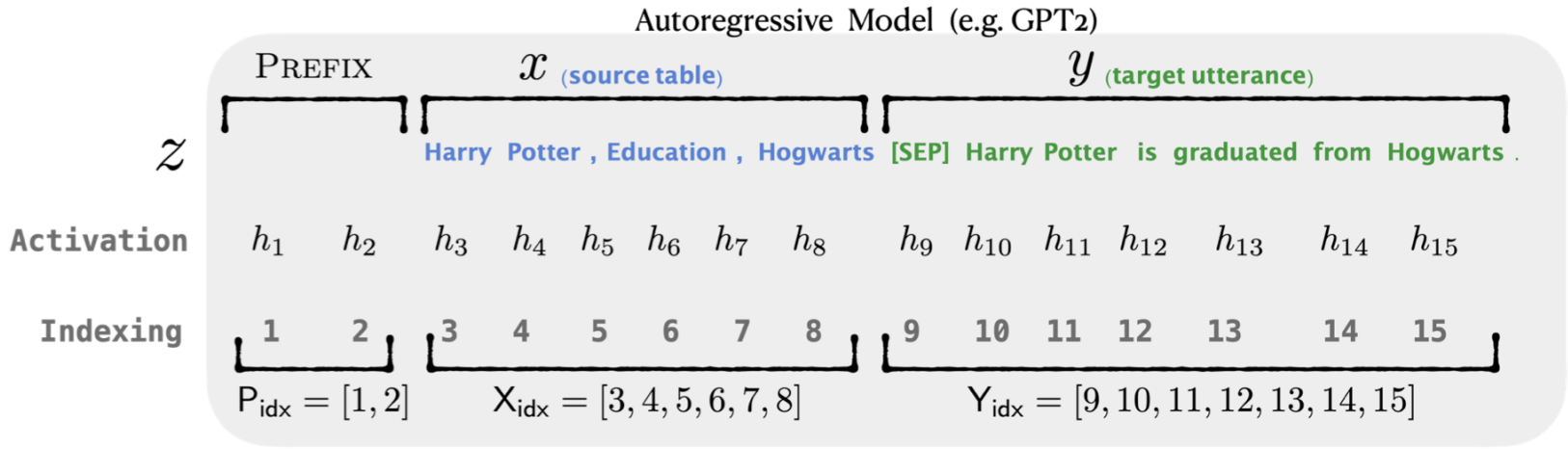
How to automatically find template?
-> **Autoprompt** (discrete) Shin et al. (2020)



How to automatically find template?
-> **P-tuning** (continuous) Liu et al. (2021)
GPT Understands, Too



How to automatically find template?
-> **Prefix-tuning** (continuous) Li and Liang. (2021)



Summarization Example

Article: Scientists at University College London discovered people tend to think that their hands are wider and their fingers are shorter than they truly are.They say the confusion may lie in the way the brain receives information from different parts of the body.Distorted perception may dominate in some people, leading to body image problems ... [ignoring 308 words] could be very motivating for people with eating disorders to know that there was a biological explanation for their experiences, rather than feeling it was their fault."

Summary: The brain naturally distorts body image - a finding which could explain eating disorders like anorexia, say experts.

Questions?

Roadmap

- Introduction to Cantonese NLP
- Background and Current Progress in Cantonese NLP
- Pre-training and the state-of-the-art NLP technology
- Preliminary Experiments on Cantonese NLP
- Challenges of Cantonese NLP and the Future Directions

Possible solutions for Cantonese NLP?

1 Tackle Cantonese NLP with models trained on **Mandarin** corpora

Pros: numerous methods including pre-trained models

Cons: different character sets, vocabulary, grammar, etc.

2 Tackle Cantonese NLP with models trained on **Traditional Chinese** corpora

Pros: traditional character set

Cons: different vocabulary, grammar, etc.

System	Training Corpus	Language	Character Type
BERT-CKIP	Wikipedia, Gigaword	SCN	Traditional
RoBERTa-HFL	Wikipedia, news websites	SCN	Both
XLNet-HK	Wikipedia, news, blogs	SCN, Cantonese	Both
ELECTRA-HFL	Wikipedia, news websites	SCN	Both
ELECTRA-HK	Wikipedia, news, blogs	SCN, Cantonese	Both

Experimental Steps:

1 Collect Cantonese **corpora**

2 Build Cantonese **benchmark**

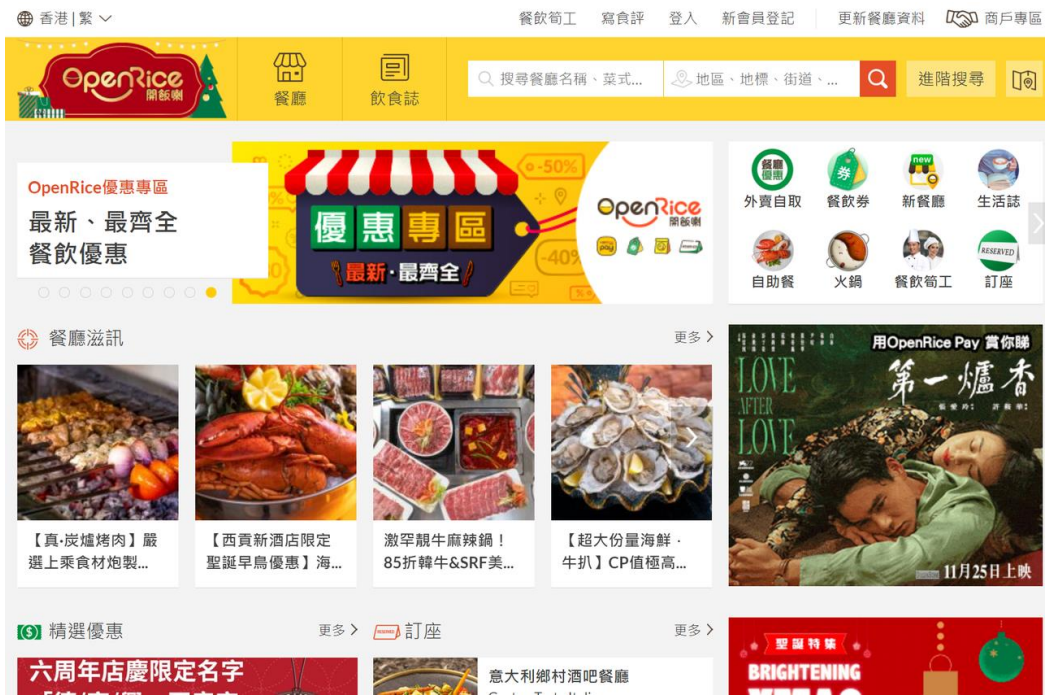
3 Performance evaluation on **existing pre-training models**

4 Pilot Study on **Cantonese pre-training**

Data Source of Corpora

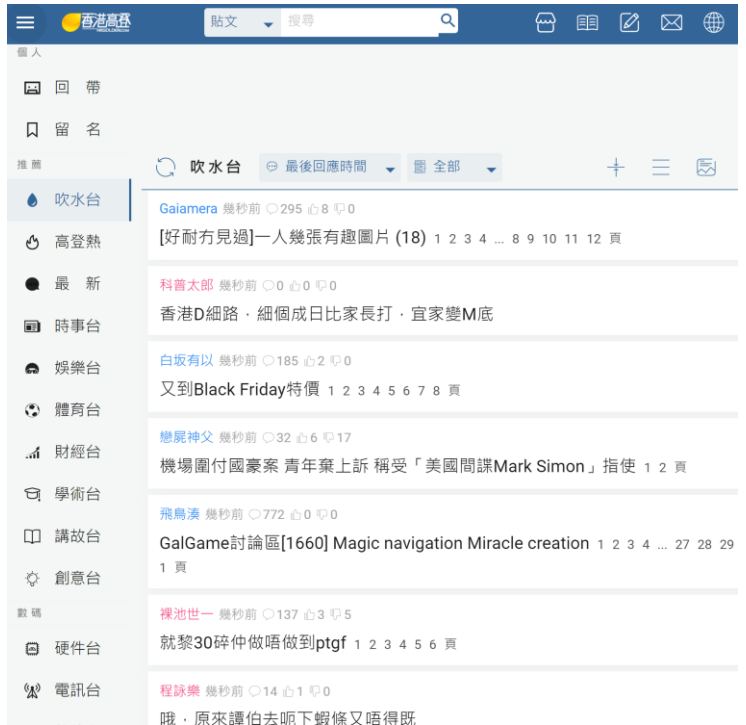
Hong Kong Discuss Forum

Openrice.com



Hong Kong LIHKG Forum

Hong Kong Golden Forum



Questionable data source?

Hong Kong 01 News

Hong Kong on.cc news

Minpao News

Yahoo Hong Kong news

Statistics

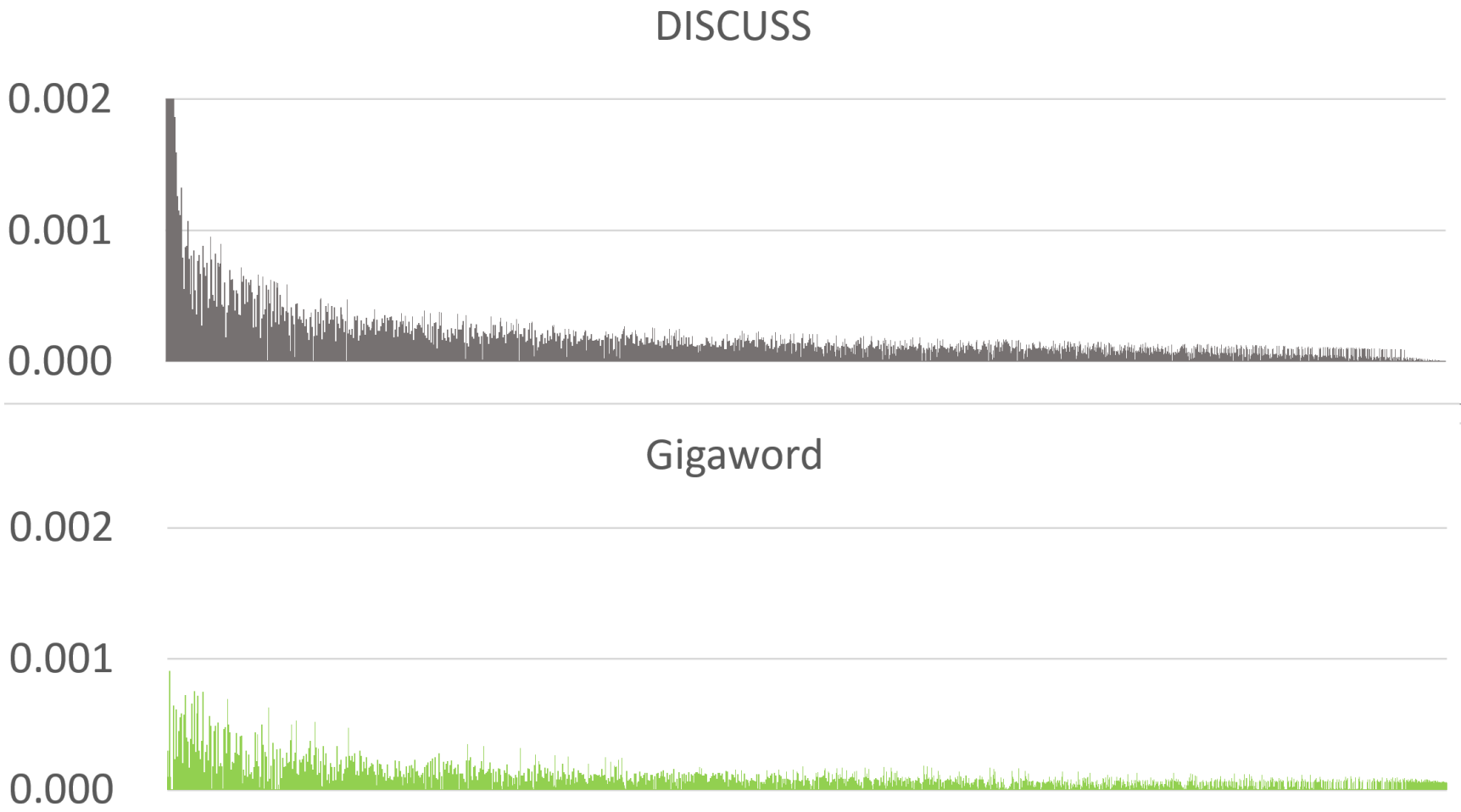
Data Source	Token Count	File Size	
DISCUSS	1,330.5 M	2,447.4 MB	
LIHKG	1,228.7 M	36,032.2 MB	Raw Html
HKGolden	50.6 M	123.1 MB	
OpenRice	584.9 M	1,259.5 MB	

Scales of text from 4 Cantonese forums
with 3.2 billion tokens.

Quantitative Analysis

——Cantonese v.s. Traditional Chinese (Taiwan)

- X-axis shows the union of the top 1,000 bigrams
- Ordered by the average frequency on two datasets.
- The top curve refers to DISCUSS while bottom Gigaword.



Distribution of bigrams from DISCUSS and Gigaword datasets.

Qualitative Analysis (Colloquialism)

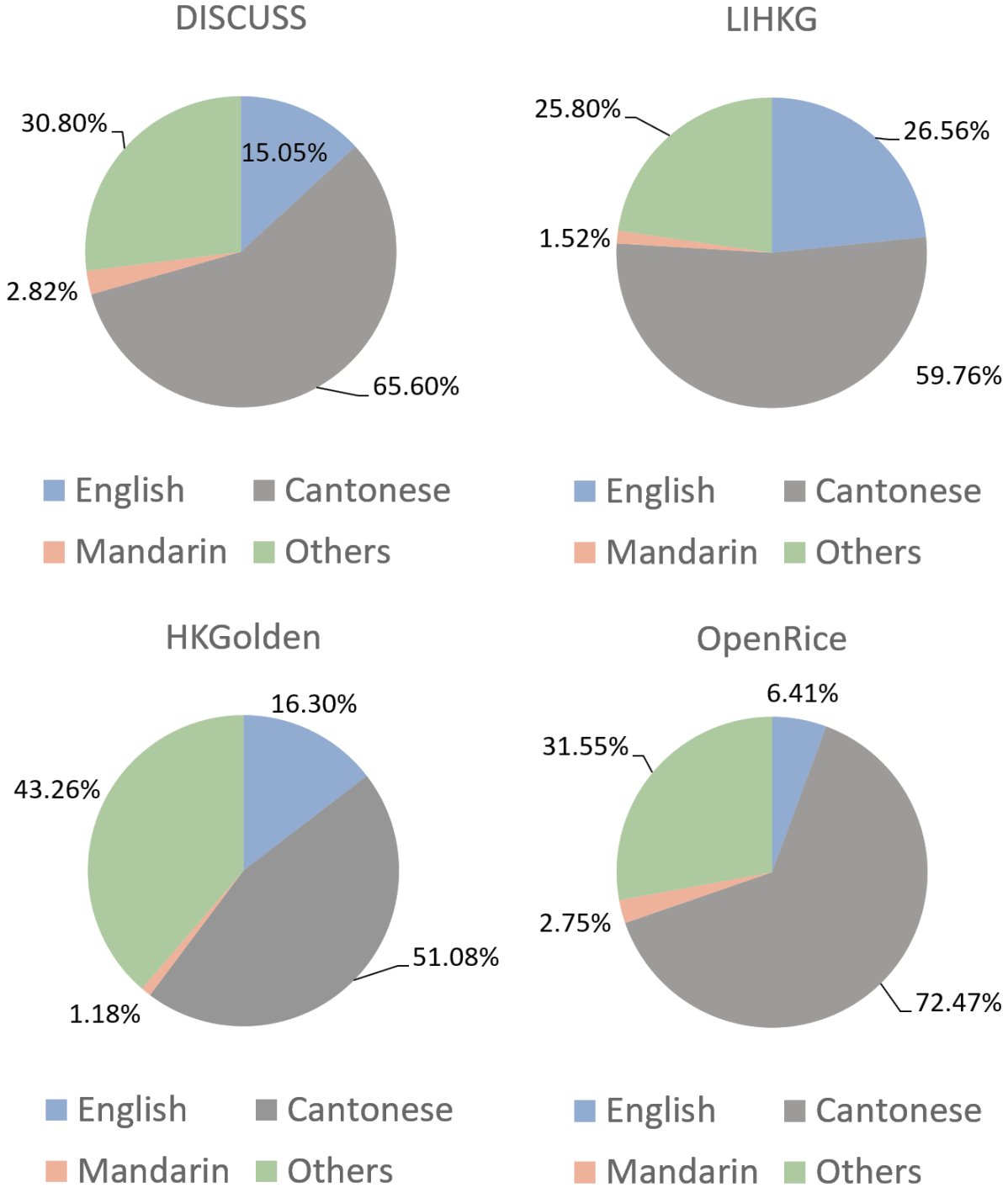
Free-style writing & spelling mistake toleration
“訓覺” other than “瞓覺” (‘sleep’)

Slangs and idioms are quite common in colloquial context
“今次演唱會好難買到飛，佢都系执死鸡先至有得睇咋。”
(‘It’s extremely hard to buy tickets for the concert. He can’t go to the concert unless he collects a lucky coin’)

Neologisms and special tokens
“海鮮義大利飯味道唔錯:)！隻蝦好有驚喜因為非常新鮮, LOL！！ ”
(‘The tastes of seafood paella are not bad! The shrimp is a surprise because it’s very fresh, LOL!! ’)

Multilingualism

- An opensource toolkit fastlangid is employed to analyze the language usage ratio.
- The codeswitching behavior across Cantonese and English is frequent
- Codemixing with languages other than Chinese and English. Cantonese - speaking areas happen to integrate speakers of multiple nationalities.



Benchmark

Two Cantonese datasets and two standard Chinese datasets

Datasets	N_train	N_test	C	NLP Task
OpenRice (Can)	54,000	7,610	5	Sentiment
LIHKG (Can)	10,000	1,000	19	Topic
TNews (SCN)	53,360	10,000	14	Topic
Ciron (SCN)	7,014	876	5	Irony

Benchmark statistics. N_train: training instance number; N_test: test instance number; C: class number; Can: Cantonese; SCN: standard Chinese.

Baseline Results

	Original language		Original language		Original language		Original language	
	OpenRice		LIHKG		TNews		Ciron	
	Can	SCN	Can	SCN	Can	SCN	Can	SCN
BERT-CKIP[1](*SCN* in TRAD_CH)	64.9	63.9	71.7	71.1	56.3	56.5	58.2	58.8
BERT-HFL (*SCN* in SIMP_CH (Cui et al., 2021))	63.8	64.1	**72.5**	**72.8**	56.0	57.4	58.1	58.6
RoBERTa-UER (*SCN* in SIMP_CH (Zhao et al., 2019))	65.3	66.1	70.3	71.5	56.1	57.6	57.8	58.0
RoBERTa-HFL (*SCN* in SIMP_CH (Cui et al., 2021))	**65.6**	**65.9**	72.4	72.6	**57.1**	**58.1**	**58.6**	59.2
ELECTRA-HFL (*SCN* in SIMP_CH (Cui et al., 2021))	64.9	65.1	71.9	72.1	56.0	57.9	57.4	**59.3**
ELECTRA-HK[2](*SCN and Cantonese* in TRAD_CH)	65.2	64.3	70.0	66.7	53.8	53.4	**58.6**	58.3
XLNet-HK[3](*SCN and Cantonese* in TRAD_CH)	65.1	64.1	72.1	66.1	53.1	52.9	58.0	57.3

1. Use Baidu translation API to generate SCN dataest for Openrice and LIHKG; generate Cantonese dataset for Tnews and Ciron.
2. Cantonese Pre-training models needs to be improved.
3. RoBERTa-HFL is generally the best performed model. Still performance drop going from Mandarin to Cantonese.
4. BERT-CKIP, the only other model trained on Traditional characters, consistently performs better than BERT-HFL

A Pilot Study on Cantonese Pre-training

	Sentiment Analysis	Topic Classification	Semantic Analysis	Semantic Analysis
Pre-train steps	Openrice (Can)	LIHKG (Can)	cmnli (SCN)	Afqmc (SCN)
0	53.1	71.7	**58.7**	70.8
6w	53.9	72.8	56.1	**71.0**
12w	54.6	72.8	56	70.4
18w	55.4	73.1	56.4	70.8
24w	55.1	73.3	57.3	70.6
30w	**55.5**	**73.6**	55.7	70.4
36w	55.2	72.9	55.4	70.8

- Cantonese pre-training based on RoBERTa-HFL
- Cantonese Pre-training works!

Questions

107

Roadmap

- Introduction to Cantonese NLP
- Background and Current Progress in Cantonese NLP
- Pre-training and the state-of-the-art NLP technology
- Preliminary Experiments on Cantonese NLP
- Challenges of Cantonese NLP and the Future Directions

Preliminary Findings from Experiments

Language Resource. Despite the potential to explore Cantonese on *social media*, the acquired resource exhibits *data quantity and quality concerns*, requiring model robustness against sparse context.

Language Representations. The Cantonese context is ubiquitously multilingual and presents the need for NLP models to gain knowledge from numerous languages and understand how they work together to form semantic coherence.

NLP Methods. The SOTA pretraining in standard Chinese shows compromised performance in Cantonese, while Cantonese counterparts cannot learn effective language representations on the tested Cantonese benchmarks.

Cantonese NLP Challenges

State-of-the-art NLP is built upon language resources in *large-scale* and *high-quality*

Low-Resource

Data Sparsity

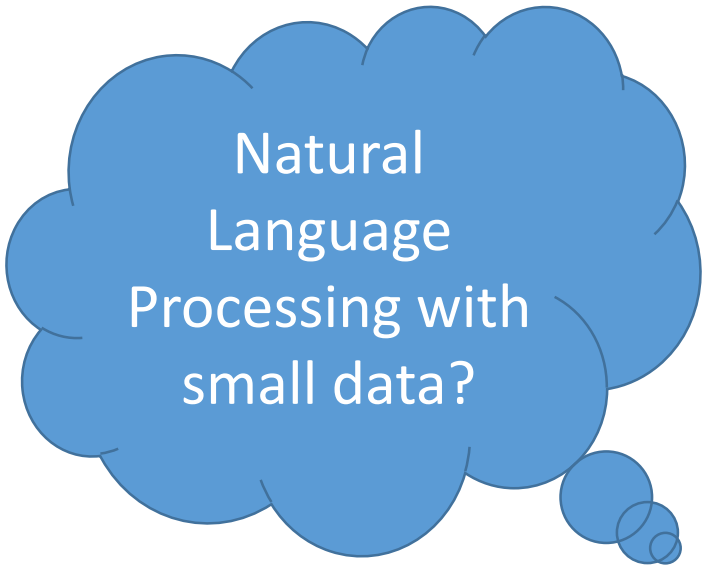
Limited Progress in existing studies and shortage of language resources.

Data scarcity and *domain/task constrain* are top issues to address in benefiting deep semantics and general NLP tasks.

Linguistic Features resulted from Cantonese history.

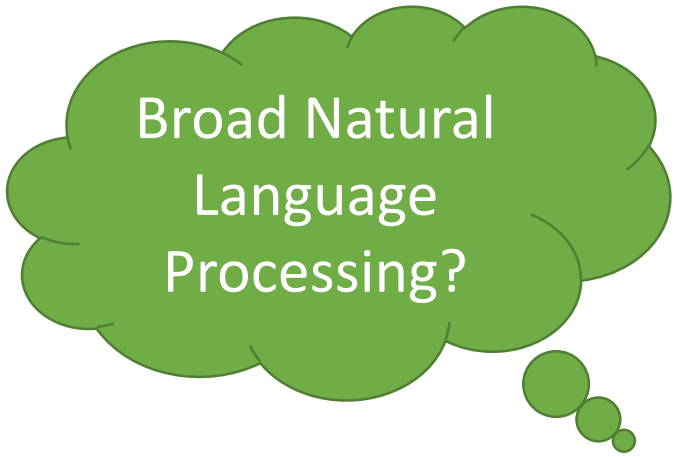
The inherent *colloquialism* and *multilingualism* in the Cantonese language would present challenges.

Low-resource Language Challenges (Data Scarcity)



- **Data scarcity** (limited data quantities for large-scale *pre-training*).
  - The modern NLP largely benefits from pre-training, enabling models to navigate large-scale text to gain generic language use capabilities. However, pre-training requires large-scale unlabeled data.
  - Take Chinese pre-training as an example, RoBERTa-WWM is trained on billions of words, and ERNIE is trained on millions of sentences.

Low-resource Language Challenges (Domain/task Constraints)



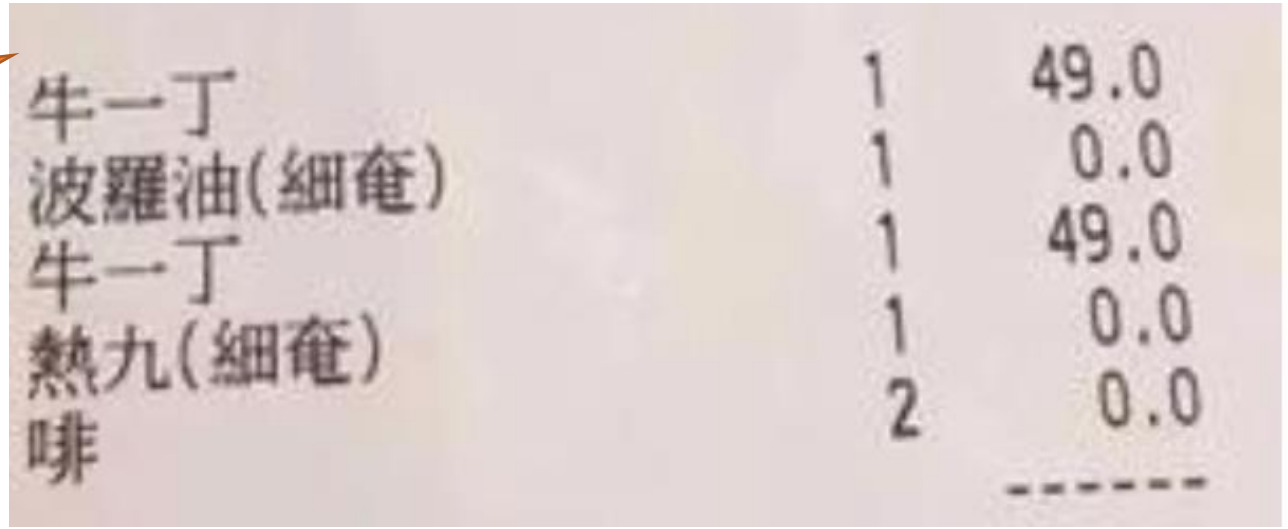
- **Domain/task constraints** (limited labeled data for *fine-tuning* in various domains/tasks).
  - Domain- and task-specific fine-tuning requires labeled datasets to narrow the generic language use skills down to handle specific tasks in certain domains.
  - The labeled datasets exhibit very few domains/tasks, and annotating datasets for a broad range of domains/tasks may not be easy.

Data Sparsity Challenges (Colloquialism)

- 熱九 == 熱狗 (hotdog)
- 細奄 == 小份奄列(small-sized omelette)
- 啡 == 咖啡 (coffee)

- Colloquialism (Cantonese is mainly derived from pronunciation and mainly used for informal communications)
  - Noisy text issues in spoken languages include (informal) abbreviations, typos, grammatical errors, fragmented syntax, etc.

An example receipt in local HK restaurants

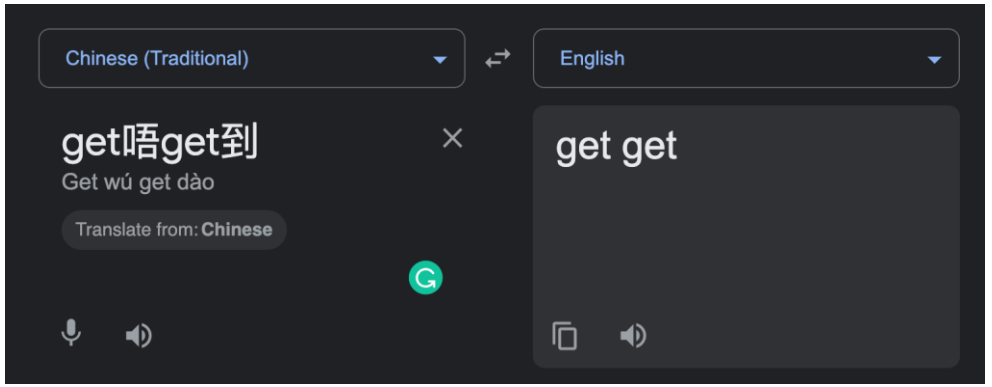


牛一丁	1	49.0
波羅油(細奄)	1	0.0
牛一丁	1	49.0
熱九(細奄)	1	0.0
啡	2	0.0

Data Sparsity Challenges (Multilingualism)

- Multilingualism (The Cantonese language historically evolved in *multilingual* environments).
  - The language tends to exhibit Chinese syntax, whereas some characteristics may be coded in other languages.
  - It results in a larger lexicon and sparsifies the word occurrence in the context for semantic learning.

It means, “Check it for me.” 😂



😂 It means, “Did you get it?”

Future Directions (a low-resource solution from the perspective of data)

- **Data Augmentation**. Generating new data by *modifying existing data* through *transformations* designed based on prior knowledge of the problem's structure (Chen et al. 2021).
  - ***Heuristic rules***, e.g., adding, replacing, editing, and removing text pieces.
  - ***Automatic translations from standard Chinese*** with the pre-trained translation model (relies on a good standard Chinese → Cantonese translator).
  - ***Generating the text with the pre-trained language generation*** model (relies on a good Cantonese generation model).
  - ***Perturbation-based augmentation***, e.g., augmenting data by adding adversarial perturbations or manipulating the hidden representations through perturbations.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, Diyi Yang: An Empirical Survey of Data Augmentation for Limited Data Learning in NLP.

Future Directions (a low-resource solution from the perspective of data!)

- **Information Retrieval**. Viewing the colloquial nature of Cantonese, it might be possible to *retrieve the data from social media* platforms.
  - ***Focusing on Cantonese sites***. The Cantonese social media platforms are not very popular (may not contain sufficient data to support large-scale pre-training), and many Cantonese speakers prefer global social media platforms, e.g., Twitter and Instagram.
  - ***Filtering Cantonese text from global social media***. We may also retrieve the data from global social media platforms, whereas it may be challenging to separate Cantonese from standard Chinese (with very similar characteristics) in an automatic manner.

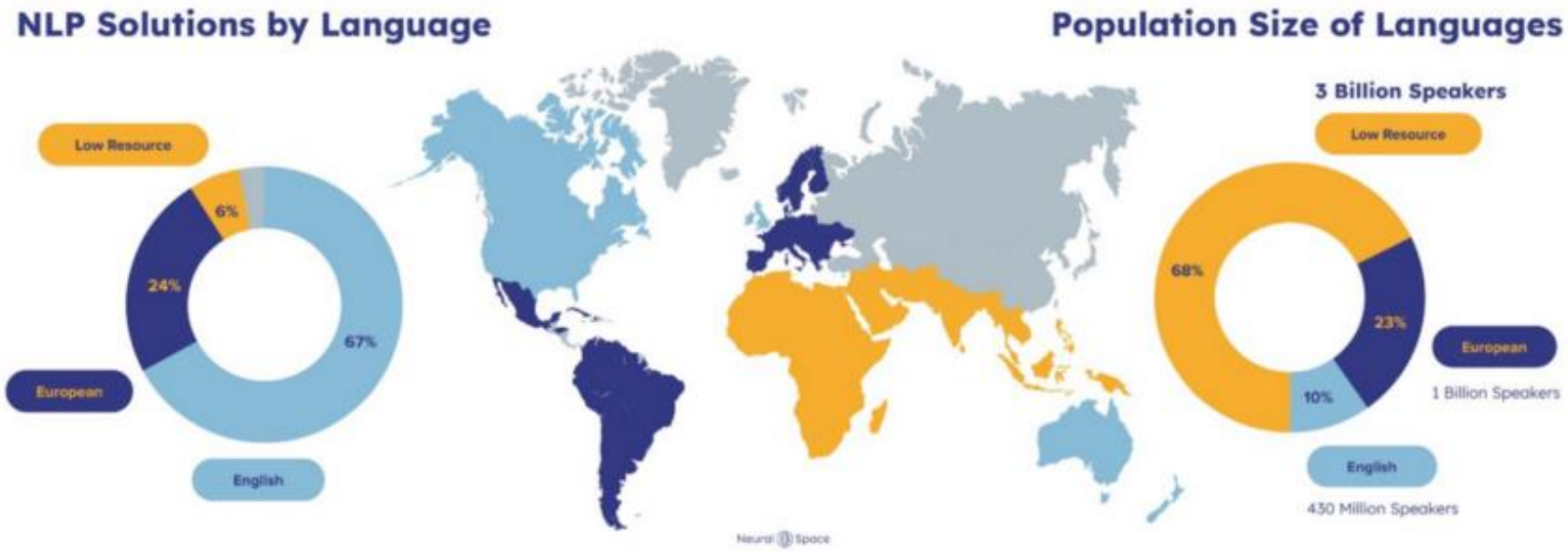
Future Direction (a low-resource solution from the perspective of methods)

- **Cross-lingual Learning.** Cantonese is a Chinese variant, so we may *borrow the knowledge from standard Chinese* (well-developed).
  - We may employ state-of-the-art *pre-trained transformers to capture the general and specific language features for transfer learnin*g, e.g., continuous pre-training with standard Chinese transformers and conduct self-supervised learning tasks to focus on Cantonese features.
- **Cross-modal Learning.** Because Cantonese is derived from pronunciations, it is strictly related to speech features.
  - *Phonetic knowledge* may play an essential role in Cantonese learning, e.g., we may include the *speech features into the pre-training* or employ *multi-modal pre-training to learn Cantonese use from both speech and text modalities*.

Low-Resource Language: beyond Cantonese

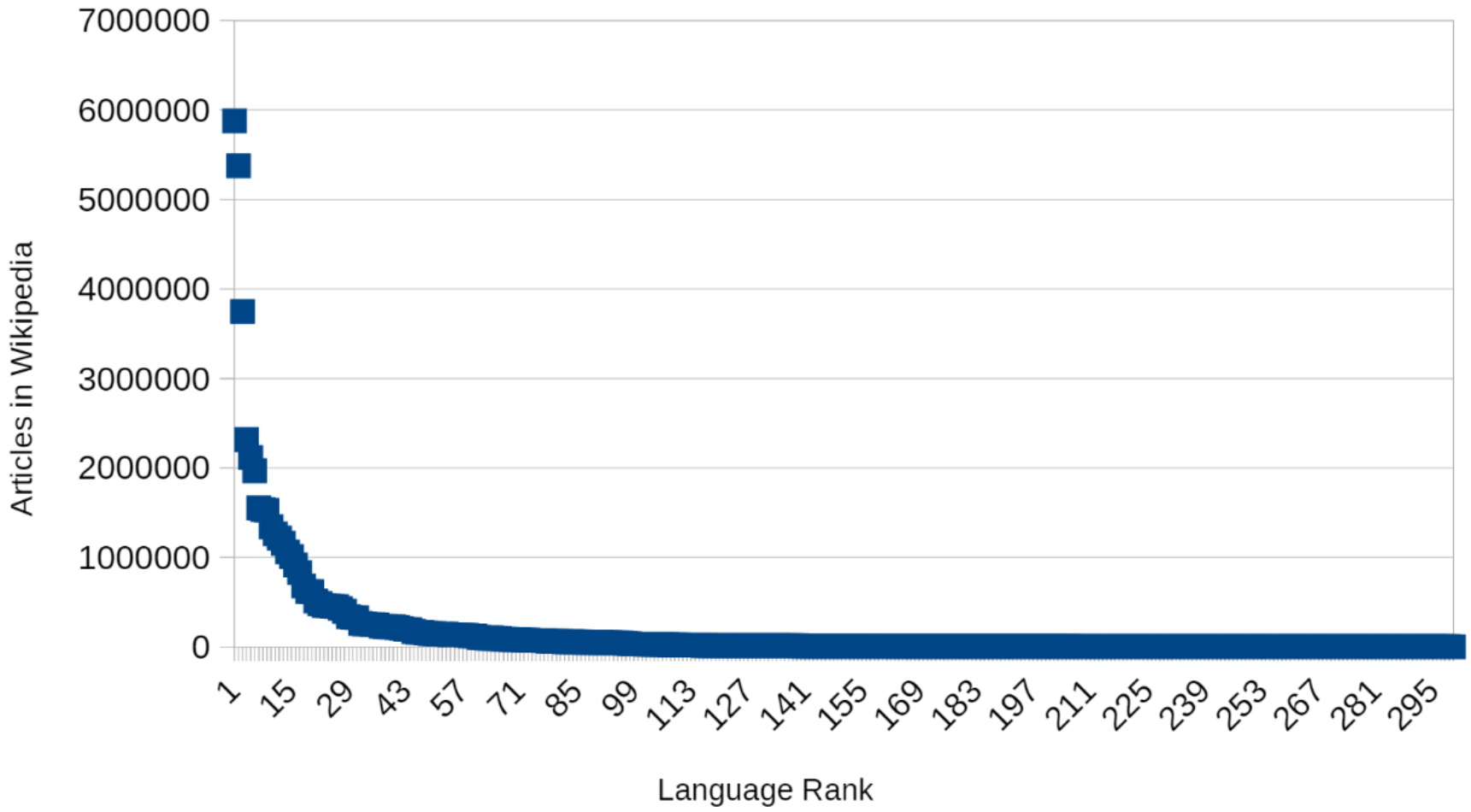
There are above 7000 languages spoken by people across the world

- Only about 20 have text corpora of hundreds of millions of words.
- 3 billion low-resource language speakers (mainly in Asia and Africa)



https://medium.com/neuralspace/low-resource-language-what-does-it-mean-d067ec85dea5

Low-Resource Language: beyond Cantonese



https://github.com/neubig/lowresource-nlp-bootcamp-2020/blob/master/slides/01-neubig-nlptasks.pdf

Top 10 languages by speaker numbers

1	English	1.5 B
2	Mandarin Chinese	1.1 B
3	Hindi	602.2 M
4	Spanish	548.3 M
5	French	274.1 M
6	Standard Arabic	274.0 M
7	Bengali	272.7 M
8	Russian	258.2 M
9	Portuguese	257.7 M
10	Urdu	231.3 M

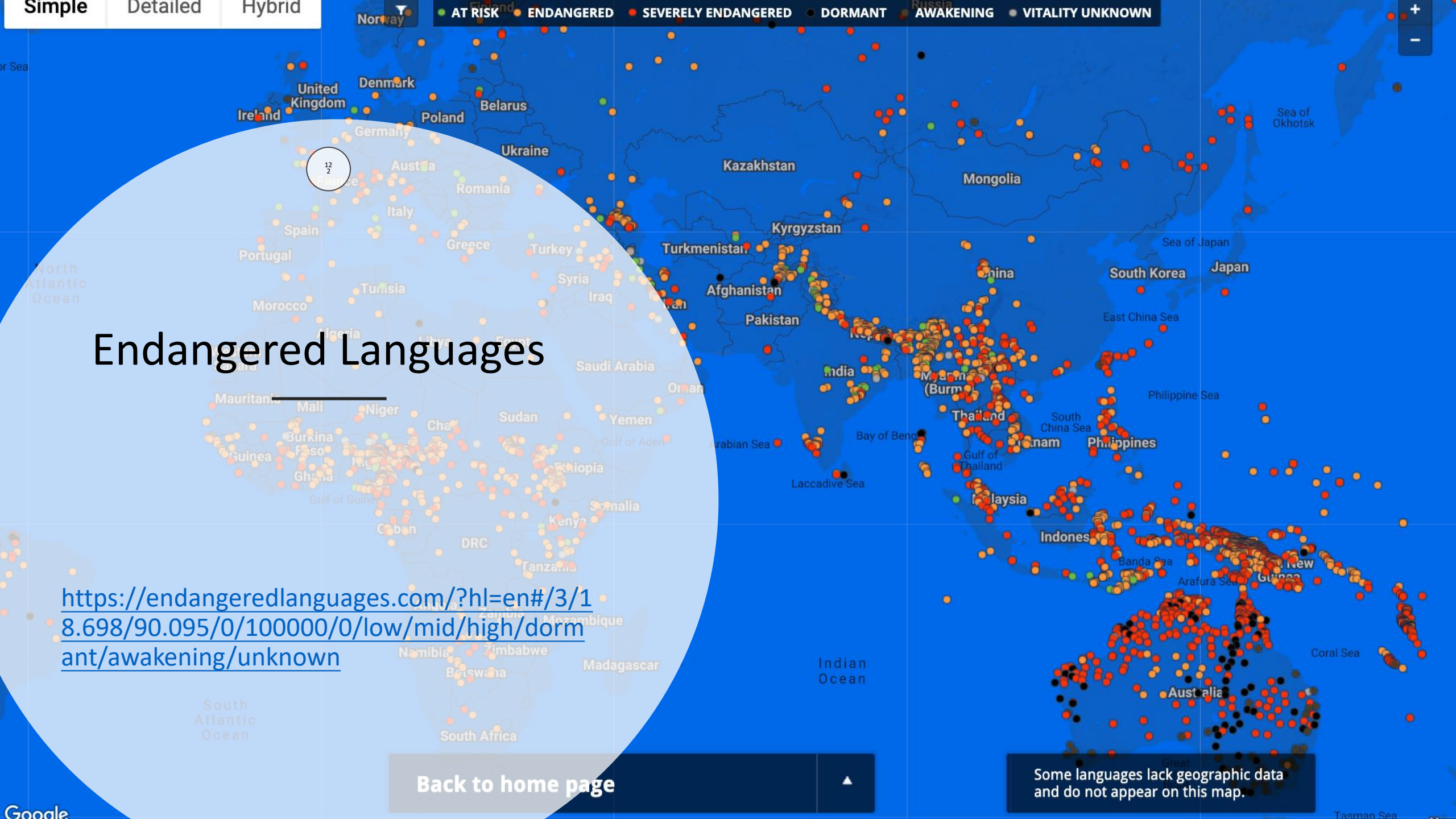
https://www.ethnologue.com/guides/ethnologue200

Top 11-20 languages by speaker numbers

11	Indonesian	199.0 M
12	Standard German	134.6 M
13	Japanese	125.4 M
14	Nigerian Pidgin	120.7 M
15	Marathi	99.1 M
16	Telugu	95.7 M
17	Turkish	88.1 M
18	Tamil	86.4 M
19	Yue Chinese	85.6 M
20	Vietnamese	85.3 M

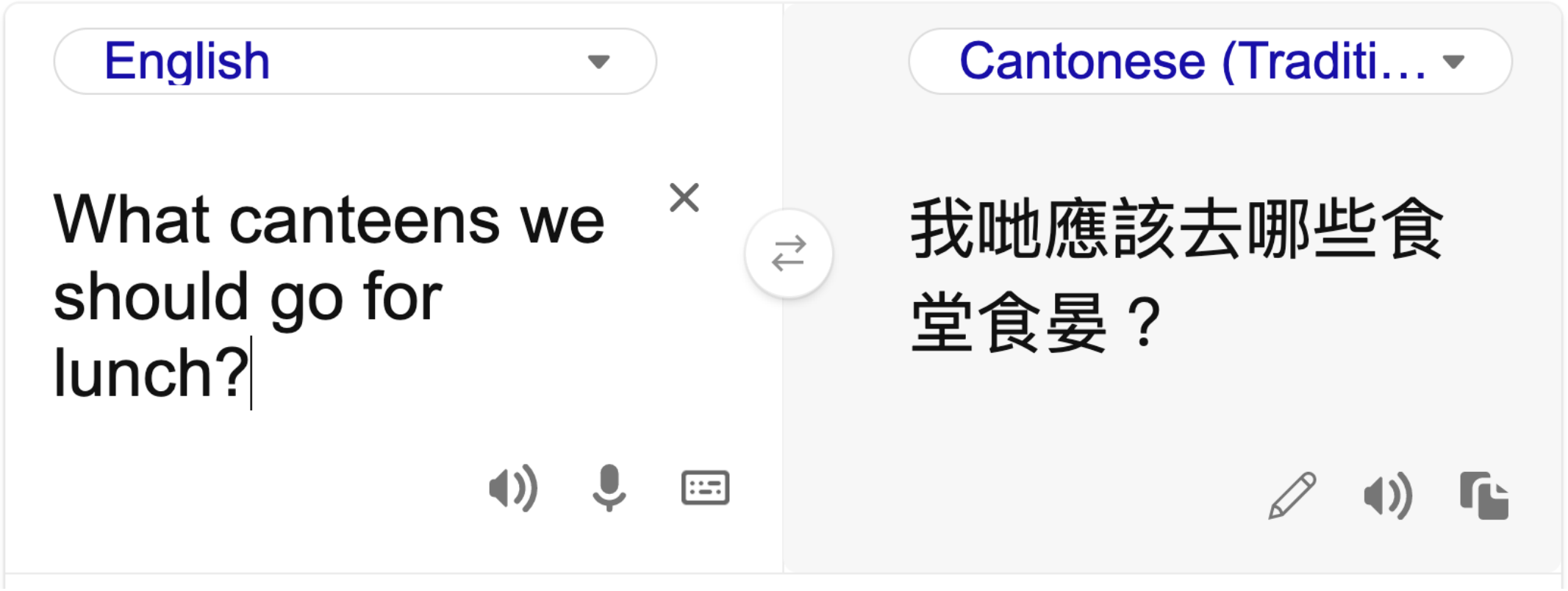
Cantonese

https://www.ethnologue.com/guides/ethnologue200


Simple
Detailed
Hybrid
AT RISK
ENDANGERED
SEVERELY ENDANGERED
DORMANT
AWAKENING
VITALITY UNKNOWN
Endangered Languages
https://endangeredlanguages.com/?hl=en#/3/18.698/90.095/0/100000/0/low/mid/high/dormant/awakening/unknown
Back to home page
Some languages lack geographic data and do not appear on this map.
Google

Applications (Human-human communications)

Machine Translation



Applications (human-machine communications)

Cantonese Dialogue Systems

Applications (Language Analysis)

Social Media Analysis

香港失業率跌至3.8%
連跌六個月至疫情初
政府統計處公布最新失業率，今年8至10月失業率為3.8%，較7月至9月失業率的3.9%，下跌0.1百分點...

- 屋企人要我夾錢畀阿哥供首期... 😦
- 好朋友移民都唔同我講... 😔
- 今屆世盃邊個小組實力陣容比較強勁? ⚽
- 嚟緊世盃大家會去邊度睇? ⚽

https://discuss.com.hk

賞心。悅目

2012-04-12 30551 瀏覽

曾幾何時，在這裡吃過一頓匆忙卻令人滿足的午餐；

數個月前，在Robuchon a Galera的晚餐亦教人如癡如醉；

今天，又回到這裡，跟男朋友度過了一個悠閒的下午。

還未夠 2 :30 pm，我倆已到達餐廳，除了成功避過長長的人龍，更可隨意選擇餐廳內的座位，但由於感覺較擠迫，最終還是選擇了餐廳外走廊的座位，那裡枱與枱之間以牆壁相間，既寬敞，又有私人空間。

想起還要呆坐半小時才開始供應tea set，我的肚子便立即發出抗議的聲音，幸好男朋友「批准」我先點一份三文治邊吃邊等。

熱烘烘的**Crispy Pocket – Tuna Confit with Fresh Tomato and Taleggio Cheese** ($72)，以兩片鬆軟的白麵包夾著吞拿魚、蕃茄、芝士和雞蛋，入口外脆內軟，滋味無窮，是超水準之作！

https://openrice.com

A Slide to Takeaway

- **Cantonese** is a *Chinese variant* (dialect) with *more than 85 million speakers* worldwide (ranked 19 in speaker number).
  - Many advances focus on the “standard” language and ignore the dialects.
- Cantonese origins from pronunciations and is historically developed in a multilingual environment, resulting in the characteristics of *colloquialism* and *multilingualism*.
- The state-of-the-art NLP is based on *pre-training and fine-tuning*, largely relying on *large-scale* and *high-quality* data, both unavailable for Cantonese learning (and many other *low-resource languages*).
- Future directions should advance NLP in low resources for the solutions in *data* (e.g., to generate more data) and *methods* (e.g., cross-lingual and cross-modal learning).