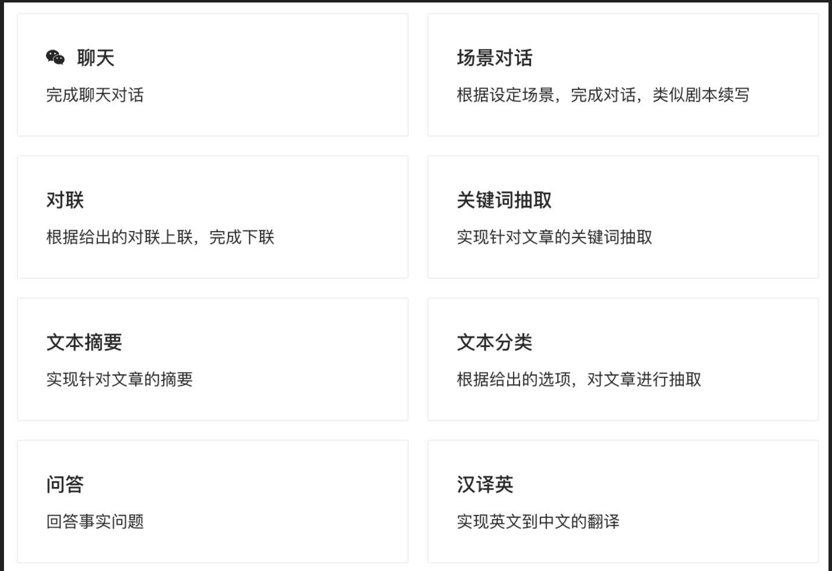
用于对话机器人的一些技术发展盘点

2021

GPT-3 is ‘Open’



聊天

完成聊天对话

场景对话

根据设定场景，完成对话，类似剧本续写

对联

根据给出的对联上联，完成下联

关键词抽取

实现针对文章的关键词抽取

文本摘要

实现针对文章的摘要

文本分类

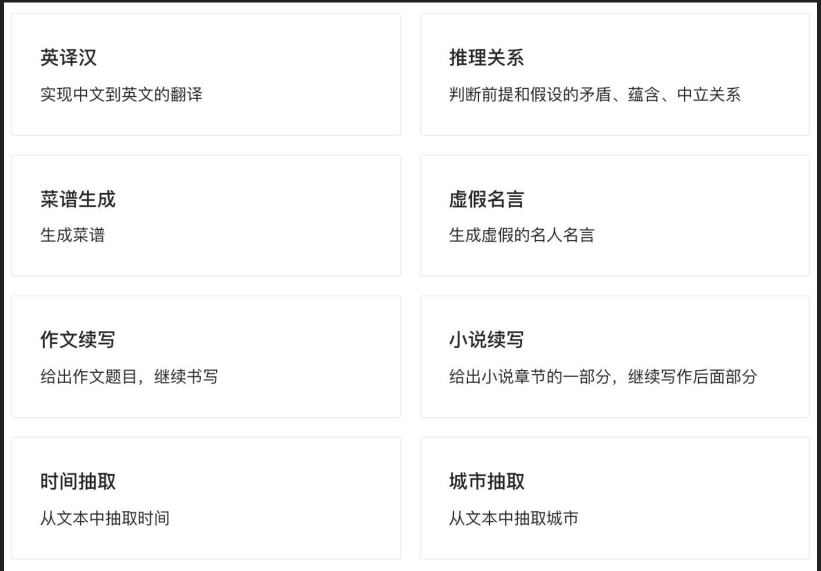
根据给出的选项，对文章进行抽取

问答

回答事实问题

汉译英

实现英文到中文的翻译



英译汉

实现中文到英文的翻译

推理关系

判断前提和假设的矛盾、蕴含、中立关系

菜谱生成

生成菜谱

虚假名言

生成虚假的名人名言

作文续写

给出作文题目，继续书写

小说续写

给出小说章节的一部分，继续写作后面部分

时间抽取

从文本中抽取时间

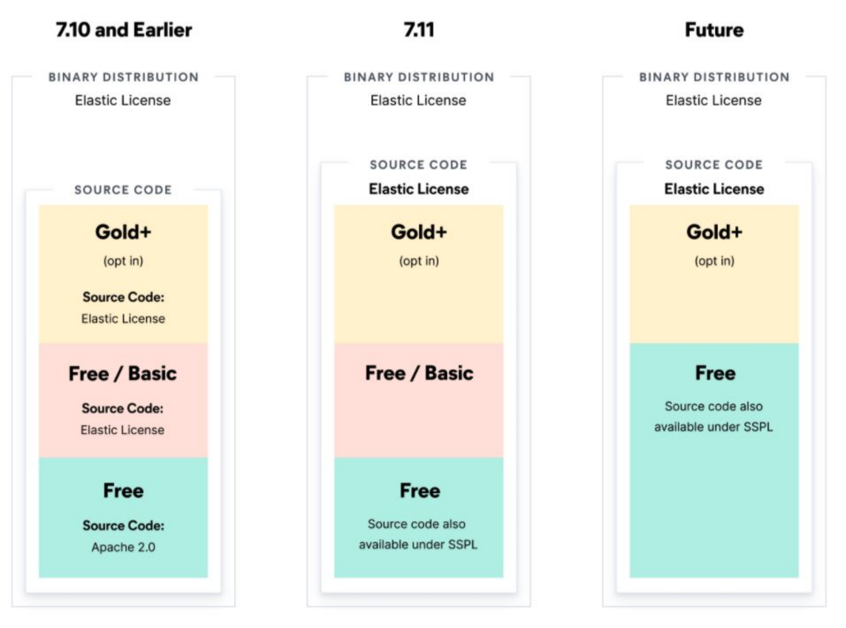
城市抽取

从文本中抽取城市

Elasticsearch
to
OpenSearch

Two things for me 1

Elaistcsearch change licence from 2021-02
https://www.elastic.co/blog/elastic-license-v2



7.10 and Earlier

BINARY DISTRIBUTION
Elastic License

SOURCE CODE

Gold+
(opt in)
Source Code:
Elastic License

Free / Basic
Source Code:
Elastic License

Free
Source Code:
Apache 2.0

7.11

BINARY DISTRIBUTION
Elastic License

SOURCE CODE
Elastic License

Gold+
(opt in)

Free / Basic

Free
Source code also available under SSPL

Future

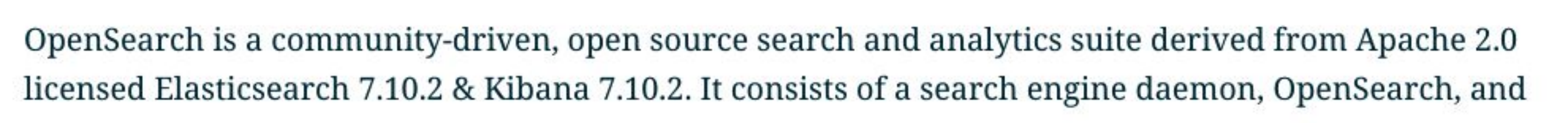
BINARY DISTRIBUTION
Elastic License

SOURCE CODE
Elastic License

Gold+
(opt in)

Free
Source code also available under SSPL

Amazon fork Elastcisearch 7.1 to OpenSearch

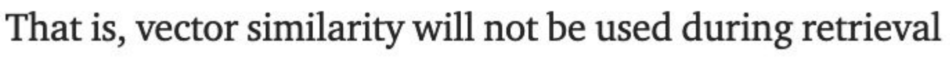


OpenSearch is a community-driven, open source search and analytics suite derived from Apache 2.0 licensed Elasticsearch 7.10.2 & Kibana 7.10.2. It consists of a search engine daemon, OpenSearch, and

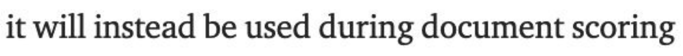
Two things for me 2

Elasticsearch not contain KNN in official build

https://towardsdatascience.com/speeding-up-bert-search-in-elasticsearch-750f1f34f455

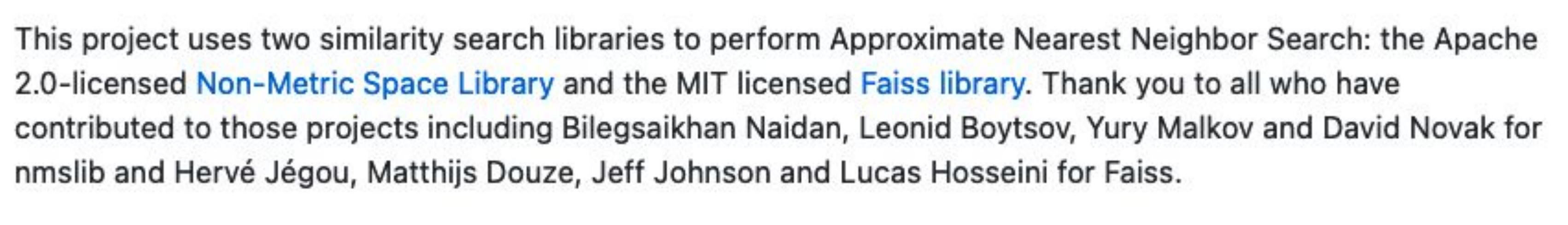


That is, vector similarity will not be used during retrieval



it will instead be used during document scoring

The KNN plugin under control by AWS and in OpenSearch official build
https://github.com/opensearch-project/k-NN



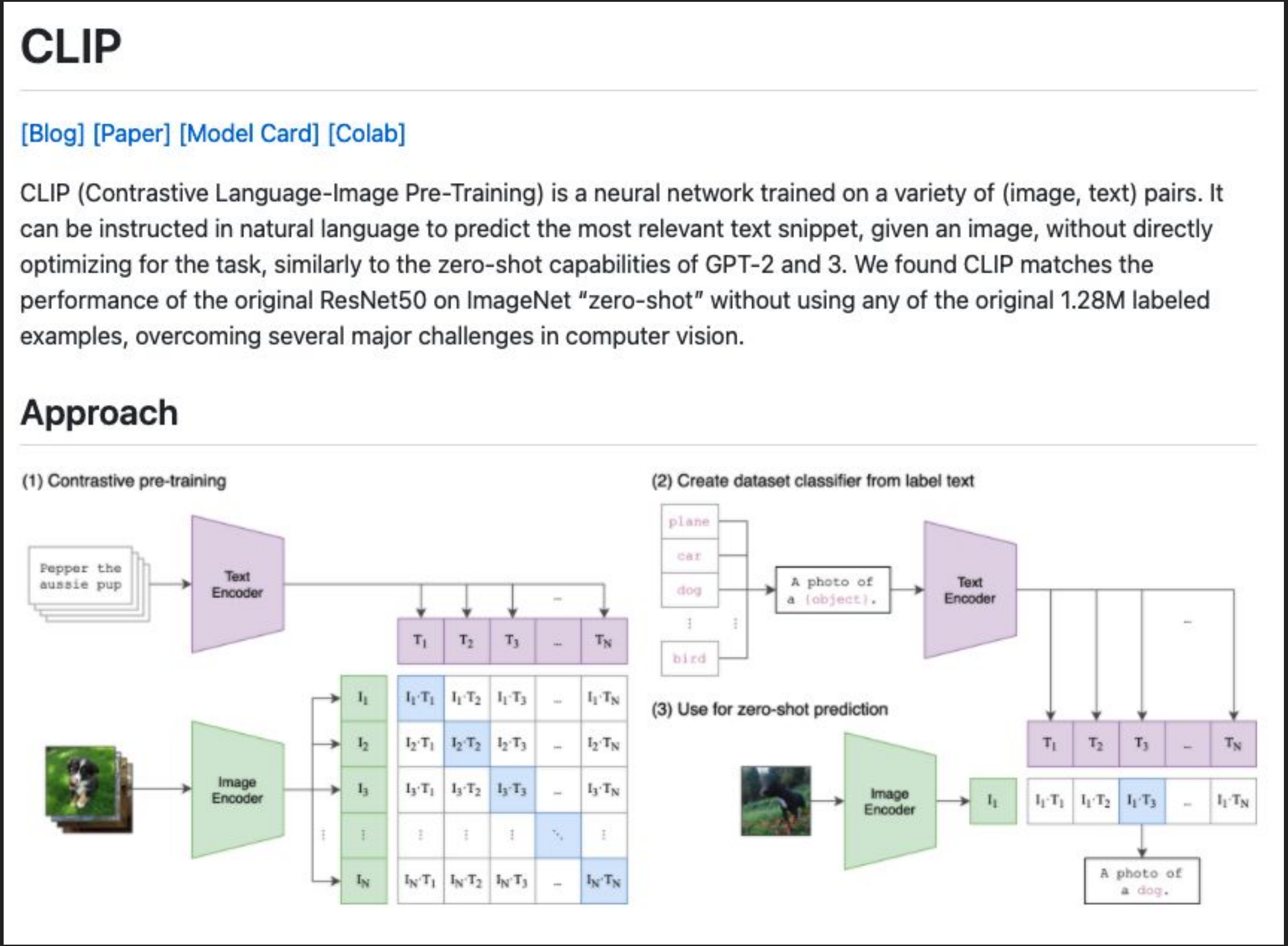
This project uses two similarity search libraries to perform Approximate Nearest Neighbor Search: the Apache 2.0-licensed Non-Metric Space Library and the MIT licensed Faiss library. Thank you to all who have contributed to those projects including Bilegsaikhan Naidan, Leonid Boytsov, Yury Malkov and David Novak for nmslib and Hervé Jégou, Matthijs Douze, Jeff Johnson and Lucas Hosseini for Faiss.

Text, Image, Text & Image Semantic Similarity is AWESOME

Technique 1

OpenAI's CLIP

is awesome



CLIP

[Blog] [Paper] [Model Card] [Colab]

CLIP (Contrastive Language-Image Pre-Training) is a neural network trained on a variety of (image, text) pairs. It can be instructed in natural language to predict the most relevant text snippet, given an image, without directly optimizing for the task, similarly to the zero-shot capabilities of GPT-2 and 3. We found CLIP matches the performance of the original ResNet50 on ImageNet "zero-shot" without using any of the original 1.28M labeled examples, overcoming several major challenges in computer vision.

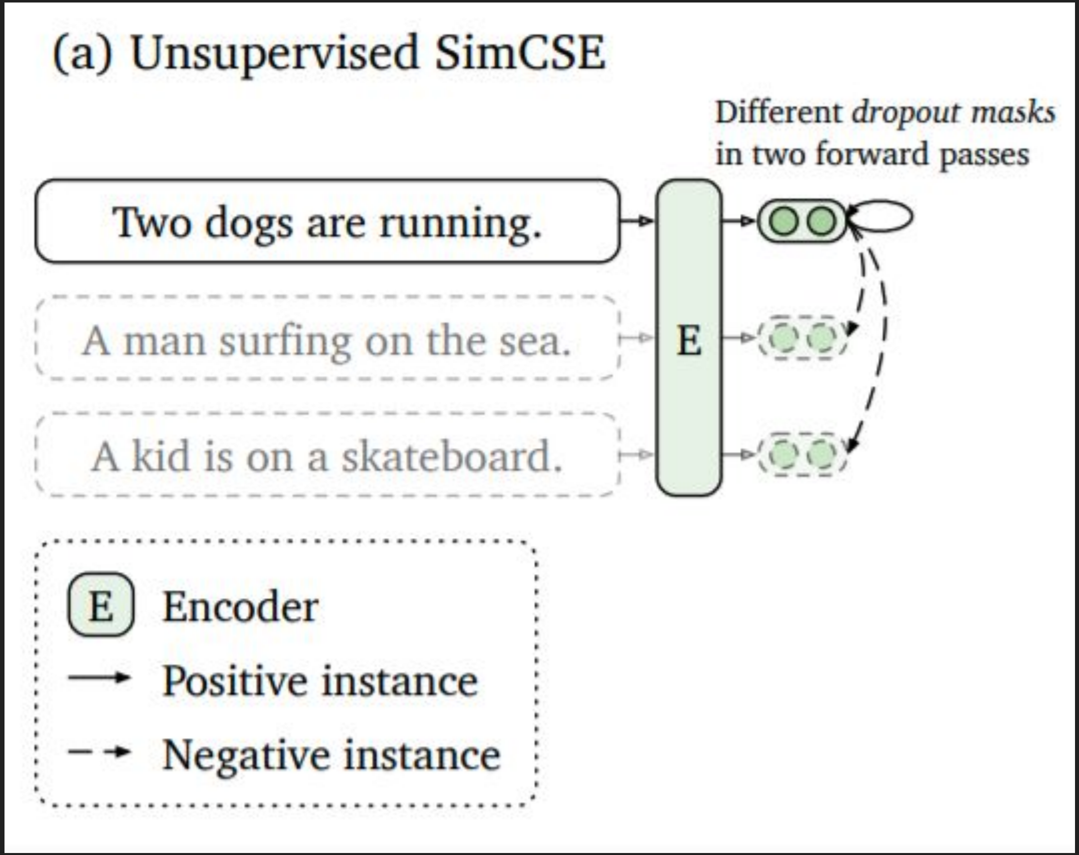
Approach

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 ... IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
IN·T1 IN·T2 IN·T3 ... IN·TN
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I1
I1·T1 I1·T2 I1·T3 ... I1·TN
A photo of a dog.

Technique 2

Text Similarity

is awesome


(a) Unsupervised SimCSE
Different *dropout masks* in two forward passes
Two dogs are running.
A man surfing on the sea.
A kid is on a skateboard.
E
E Encoder
Positive instance
Negative instance

Technique 3

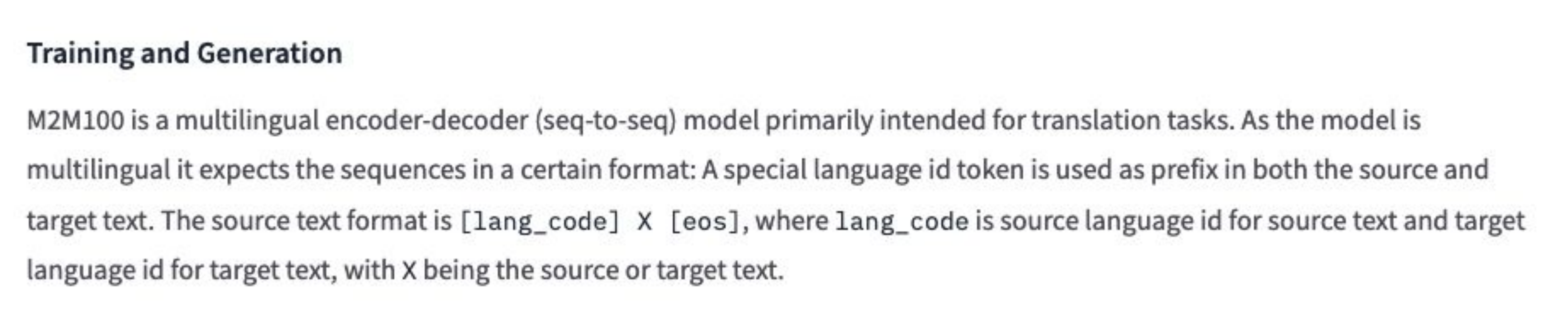
Sentence-Transformers

is awesome

Beyond Machine Translation

M2M100

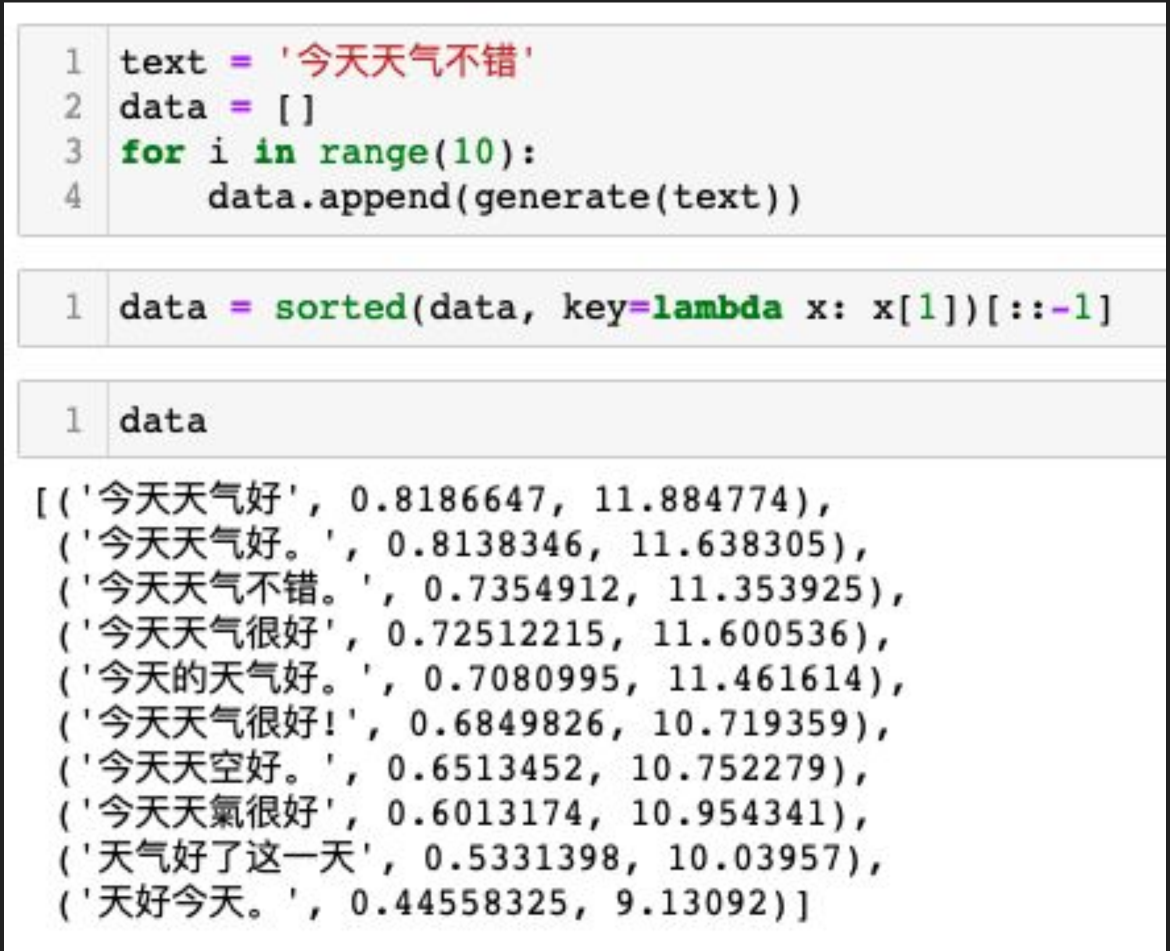
100种语言到100种语言的翻译



Training and Generation

M2M100 is a multilingual encoder-decoder (seq-to-seq) model primarily intended for translation tasks. As the model is multilingual it expects the sequences in a certain format: A special language id token is used as prefix in both the source and target text. The source text format is `[lang_code] X [eos]`, where `lang_code` is source language id for source text and target language id for target text, with `X` being the source or target text.

中文到中文的翻译会怎么样？



```python
text = '今天天气不错'
data = []
for i in range(10):
    data.append(generate(text))
```

```python
data = sorted(data, key=lambda x: x[1])[::-1]
```

```python
data
```

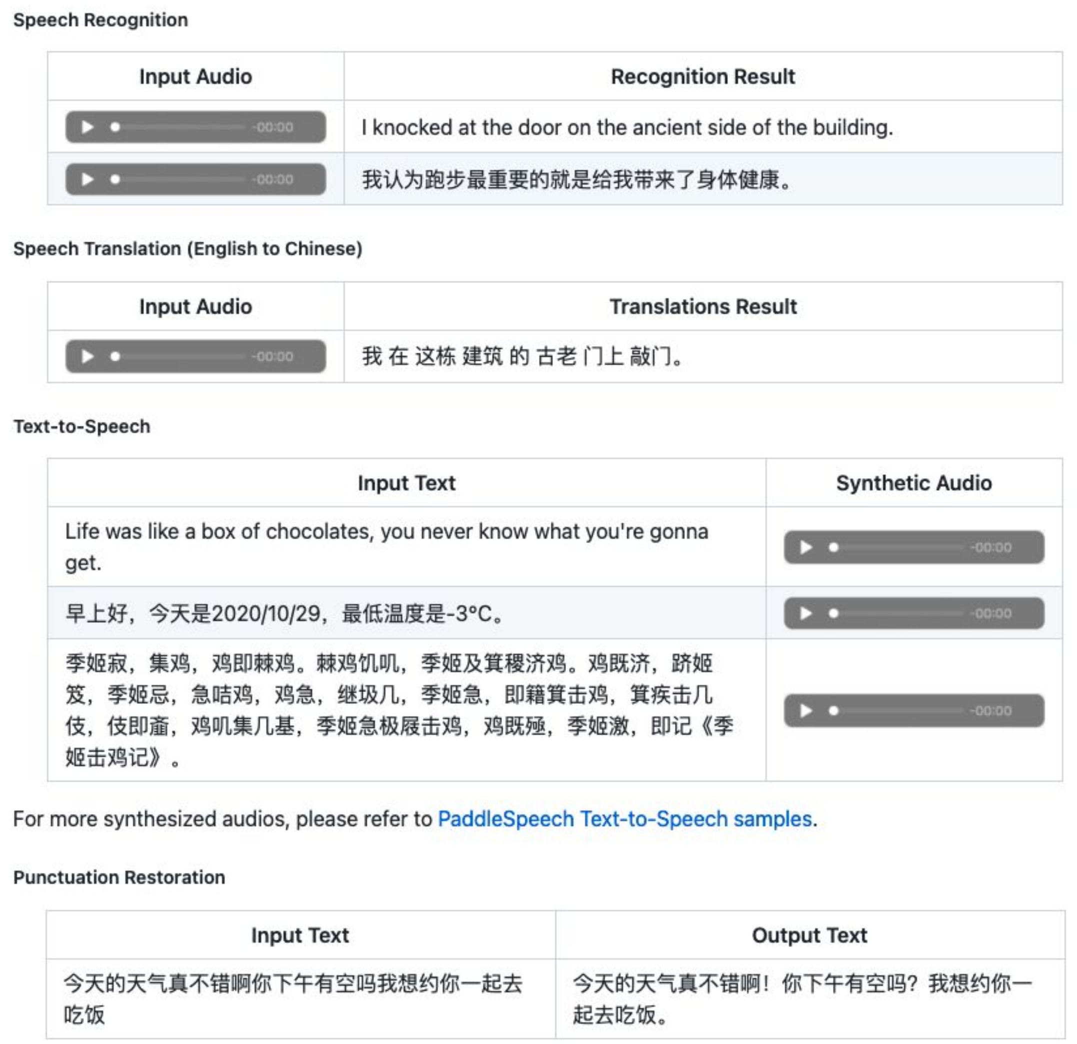
```
[('今天天气好', 0.8186647, 11.884774),
 ('今天天气好。', 0.8138346, 11.638305),
 ('今天天气不错。', 0.7354912, 11.353925),
 ('今天天气很好', 0.72512215, 11.600536),
 ('今天的天气好。', 0.7080995, 11.461614),
 ('今天天气很好!', 0.6849826, 10.719359),
 ('今天天空好。', 0.6513452, 10.752279),
 ('今天天氣很好', 0.6013174, 10.954341),
 ('天气好了这一天', 0.5331398, 10.03957),
 ('天好今天。', 0.44558325, 9.13092)]
```

TTS & ASR

PaddleSpeech

https://github.com

/PaddlePaddle/PaddleSpeech/



Speech Recognition

Input Audio	Recognition Result
▶ -00:00	I knocked at the door on the ancient side of the building.
▶ -00:00	我认为跑步最重要的就是给我带来了身体健康。

Speech Translation (English to Chinese)

Input Audio	Translations Result
▶ -00:00	我 在 这栋 建筑 的 古老 门上 敲门。

Text-to-Speech

Input Text	Synthetic Audio
Life was like a box of chocolates, you never know what you're gonna get.	▶ -00:00
早上好，今天是2020/10/29，最低温度是-3°C。	▶ -00:00
季姬寂，集鸡，鸡即棘鸡。棘鸡饥叽，季姬及箕稷济鸡。鸡既济，跻姬笈，季姬忌，急咭鸡，鸡急，继圾几，季姬急，即籍箕击鸡，箕疾击几伎，伎即齑，鸡叽集几基，季姬急极屐击鸡，鸡既殛，季姬激，即记《季姬击鸡记》。	▶ -00:00

For more synthesized audios, please refer to PaddleSpeech Text-to-Speech samples.

Punctuation Restoration

Input Text	Output Text
今天的天气真不错啊你下午有空吗我想约你一起去吃饭	今天的天气真不错啊！你下午有空吗？我想约你一起去吃饭。

Plato XL 11B

https://github.com

/PaddlePaddle/Knover/tree/develop/projects/PLATO-XL