学习目标

- 了解传智大脑项目立项的背景。
- 了解市场上比较主流的AI平台形式与功能。

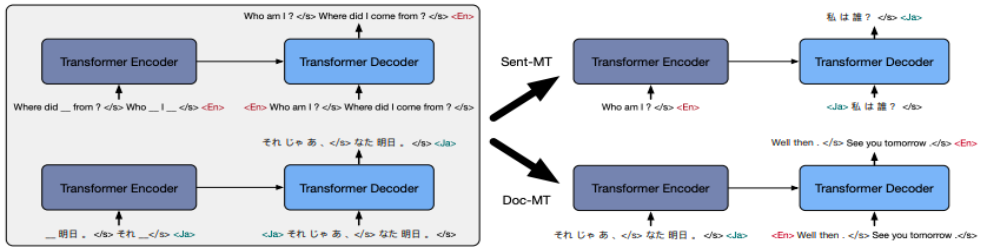
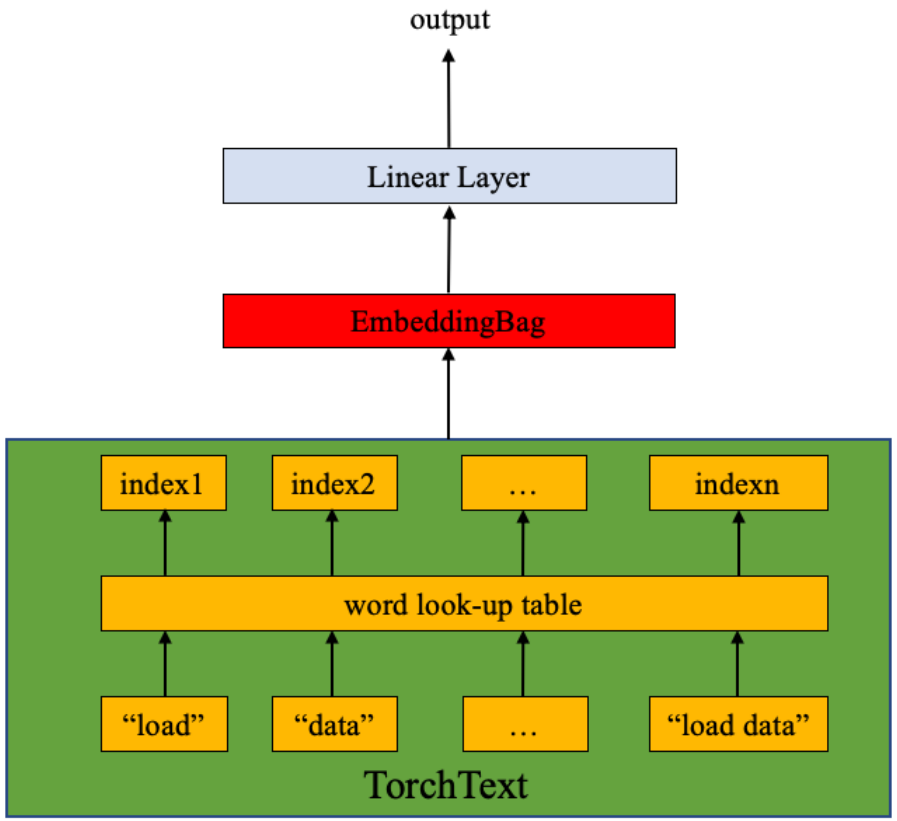


Figure 1: Framework for our Multilingual Denoising Pre-training (left) and fine-tuning on downstream MT tasks (right), where we use (1) sentence permutation (2) word-span masking as the injected noise. A special language id token is added at both the encoder and decoder. One multilingual pre-trained model is used for all tasks.


```
import requests
requests.post("url")
```

- 根据当前的硬件设备和在线需求，我们将选择具有快速推断优势的模型，简单过程如下：
  - 1，采集并标注不同类型问题的样本数据
  - 2，训练并验证fasttext进行模型多分类



- 从特定sql数据库中获得指定数据:

```
# 请安装mysql工具: pip install pymysql
import pandas as pd
import pymysql

# 内部数据库配置, 这是由数仓工程师提供的
sql_config = {
    "host": "47.241.24.21",
    "user": "ai_test",
    "password": "**********",
    "database": "ai",
}


# 打开数据库连接
db = pymysql.connect(**sql_config)

# 使用 cursor() 方法创建一个游标对象 cursor
cursor = db.cursor()

# 使用 execute()  方法执行 SQL 查询
# 编写sql语句, 根据id取出指定范围的内容查看, 这里取前1000条
cursor.execute("SELECT * from `ai_test` WHERE 0 < id < 1000")

# 使用 fetchall() 方法获取全部数据.
data = cursor.fetchall()

# 输出成csv或excel文件, 这里是输出csv
# 输出excel使用df.to_excel API即可
df = pd.DataFrame(data)
df.to_csv("./corpus.csv")

# 关闭数据库连接
db.close()
```

- 输出效果:
  - 在当前目录下生成corpus.csv或corpus.xlsx文件
  - 文件内容如下图所示（这里为了清晰显示和说明使用excel打开）
  - 内网机器IP：172.17.0.228

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	会话ID	时间	时间戳	内容	消息发送人	姓名	手机号	QQ	微信	意向校区	意向学科	专题页标题	地区
2	2324393	2020/4/29 0:08	1588090087	您好，我们是传智播客	咨询师	132*****676	132*****676		wan*****_chn@163.com	未知	C++	黑马程序员C/C++与网络攻防培训官网-	中国 江苏
3	2324393	2020/4/29 0:08	1588090090	你好，你是想了解哪个	咨询师	132*****676	132*****676		wan*****_chn@163.com	未知	C++	黑马程序员C/C++与网络攻防培训官网-	中国 江苏
4	2324393	2020/4/29 0:08	1588090104	还在么同学？	咨询师	132*****676	132*****676		wan*****_chn@163.com	未知	C++	黑马程序员C/C++与网络攻防培训官网-	中国 江苏
5	2324393	2020/4/29 0:08	1588090117	价格？	客户	132*****676	132*****676		wan*****_chn@163.com	未知	C++	黑马程序员C/C++与网络攻防培训官网-	中国 江苏
6	2324393	2020/4/29 0:08	1588090128	您好	咨询师	132*****676	132*****676		wan*****_chn@163.com	未知	C++	黑马程序员C/C++与网络攻防培训官网-	中国 江苏
7	2324393	2020/4/29 0:08	1588090136	您想了解哪个专业的学	咨询师	132*****676	132*****676		wan*****_chn@163.com	未知	C++	黑马程序员C/C++与网络攻防培训官网-	中国 江苏
8	2324393	2020/4/29 0:09	1588090143	专业不同，学时学费也	咨询师	132*****676	132*****676		wan*****_chn@163.com	未知	C++	黑马程序员C/C++与网络攻防培训官网-	中国 江苏
9	2324393	2020/4/29 0:09	1588090149	C++	客户	132*****676	132*****676		wan*****_chn@163.com	未知	C++	黑马程序员C/C++与网络攻防培训官网-	中国 江苏
10	2324393	2020/4/29 0:09	1588090158	好的，可以给您发送一	咨询师	132*****676	132*****676		wan*****_chn@163.com	未知	C++	黑马程序员C/C++与网络攻防培训官网-	中国 江苏
11	2324393	2020/4/29 0:09	1588090166	好	客户	132*****676	132*****676		wan*****_chn@163.com	未知	C++	黑马程序员C/C++与网络攻防培训官网-	中国 江苏
12	2324393	2020/4/29 0:09	1588090172	好的，方便QQ或者微	咨询师	132*****676	132*****676		wan*****_chn@163.com	未知	C++	黑马程序员C/C++与网络攻防培训官网-	中国 江苏
13	2324393	2020/4/29 0:09	1588090192	wangtong_chn@022.c	客户	132*****676	132*****676		wan*****_chn@163.com	未知	C++	黑马程序员C/C++与网络攻防培训官网-	中国 江苏
14	2324393	2020/4/29 0:10	1588090201	好的，您备注个有效电	咨询师	132*****676	132*****676		wan*****_chn@163.com	未知	C++	黑马程序员C/C++与网络攻防培训官网-	中国 江苏
15	2324393	2020/4/29 0:10	1588090216	别误会，资料是发您邮	咨询师	132*****676	132*****676		wan*****_chn@163.com	未知	C++	黑马程序员C/C++与网络攻防培训官网-	中国 江苏
16	2324393	2020/4/29 0:10	1588090259	02101021202	客户	132*****676	132*****676		wan*****_chn@163.com	未知	C++	黑马程序员C/C++与网络攻防培训官网-	中国 江苏
17	2324393	2020/4/29 0:11	1588090261	好的	咨询师	132*****676	132*****676		wan*****_chn@163.com	未知	C++	黑马程序员C/C++与网络攻防培训官网-	中国 江苏
18	2324393	2020/4/29 0:11	1588090281	之后留意一下邮箱哈，	咨询师	132*****676	132*****676		wan*****_chn@163.com	未知	C++	黑马程序员C/C++与网络攻防培训官网-	中国 江苏
19	2324398	2020/4/29 1:07	1588093631	来源：<a target='_bla	客户	同学	150*****899		150*****899	未知	未知	黑马程序员大数据+机器学习培训-专业	中国 河北 秦皇岛
20	2324398	2020/4/29 7:01	1588114865	感谢您的咨询以及对	咨询师	同学	150*****899		150*****899	未知	未知	黑马程序员大数据+机器学习培训-专业	中国 河北 秦皇岛
21	2324405	2020/4/29 7:54	1588118086	您好，我们是传智播客	咨询师	同学	133*****958	499*****8		未知	软件测试	软件测试培训-专业的软件测试培训机构	中国 天津 天津
22	2324405	2020/4/29 7:54	1588118089	你好，你是想了解哪个	咨询师	同学	133*****958	499*****8		未知	软件测试	软件测试培训-专业的软件测试培训机构	中国 天津 天津
23	2324405	2020/4/29 7:55	1588118104	还在么同学？	咨询师	同学	133*****958	499*****8		未知	软件测试	软件测试培训-专业的软件测试培训机构	中国 天津 天津
24	2324405	2020/4/29 7:55	1588118121	你是从零基础学习还是	咨询师	同学	133*****958	499*****8		未知	软件测试	软件测试培训-专业的软件测试培训机构	中国 天津 天津
25	2324405	2020/4/29 7:55	1588118141	同学，还在浏览网页吗	咨询师	同学	133*****958	499*****8		未知	软件测试	软件测试培训-专业的软件测试培训机构	中国 天津 天津
26	2324405	2020/4/29 7:55	1588118154	我这边给你发送一份详	咨询师	同学	133*****958	499*****8		未知	软件测试	软件测试培训-专业的软件测试培训机构	中国 天津 天津
27	2324405	2020/4/29 7:56	1588118173	022221221	客户	同学	133*****958	499*****8		未知	软件测试	软件测试培训-专业的软件测试培训机构	中国 天津 天津
28	2324405	2020/4/29 7:56	1588118182	您QQ是吗	咨询师	同学	133*****958	499*****8		未知	软件测试	软件测试培训-专业的软件测试培训机构	中国 天津 天津
29	2324405	2020/4/29 7:56	1588118193	您的电话是？这边备注	咨询师	同学	133*****958	499*****8		未知	软件测试	软件测试培训-专业的软件测试培训机构	中国 天津 天津
30	2324405	2020/4/29 7:56	1588118199	别误会，资料是发您Q	咨询师	同学	133*****958	499*****8		未知	软件测试	软件测试培训-专业的软件测试培训机构	中国 天津 天津
31	2324405	2020/4/29 7:56	1588118204	https://fs-im-kefu.0mc	咨询师	同学	133*****958	499*****8		未知	软件测试	软件测试培训-专业的软件测试培训机构	中国 天津 天津
32	2324405	2020/4/29 7:56	1588118209	在吗亲	咨询师	同学	133*****958	499*****8		未知	软件测试	软件测试培训-专业的软件测试培训机构	中国 天津 天津
33	2324405	2020/4/29 7:56	1588118214	我是在线客服老师	咨询师	同学	133*****958	499*****8		未知	软件测试	软件测试培训-专业的软件测试培训机构	中国 天津 天津
34	2324405	2020/4/29 7:56	1588118219	需要备注下的	咨询师	同学	133*****958	499*****8		未知	软件测试	软件测试培训-专业的软件测试培训机构	中国 天津 天津
35	2324405	2020/4/29 7:57	1588118224	只是申请资料备注一下	咨询师	同学	133*****958	499*****8		未知	软件测试	软件测试培训-专业的软件测试培训机构	中国 天津 天津

数据说明:
由于数据涉及用户隐私，学生无法请求内网数据库，但是可以直接使用给定的部分脱敏数据（如上图excel形式）。
通过上图表格，我们可以看到共有13列（13个字段），分别是"会话ID"，"时间"，"时间戳"，"内容"，"消息发送人"，"姓名"，"手机号"，"QQ"，"微信"，"意向校区"，"意向学科"，"专题页标题"，"地区"。下面对每个字段分别解释。
会话ID：整个对话块的唯一标识，用于区分消息内容是否为同一次对话。
时间：该条消息发送出去的具体结构化时间。
时间戳：该条消息发送出去的时间戳。
内容：该条消息的具体内容。

消息发送人：该条消息是由谁发送的，这里只有"客户"和"咨询师"两种。
姓名：咨询师通过对话获得的学员姓名。从"姓名"字段开始为非对话块本身的字段，而是咨询师通过对话块填写的学员信息或其他辅助信息。
手机号：咨询师通过对话获得的学员手机号。
QQ：咨询师通过对话获得的学员QQ号。
微信：咨询师通过对话获得的学员微信号。
意向校区：咨询师通过对话获得的学员意向校区。
意向学科：咨询师通过对话获得的学员意向学科。
专题页标题：这是一个咨询师辅助信息，代表学员通过官网中哪个页面进行咨询的。这个信息咨询师使用的系统是可以获得的。
地区：这也是一个辅助信息，代表学员使用的终端所在IP地区。这个信息咨询师使用的系统也是可以获得的。

```
# 导入必备工具包
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# 设置显示风格
plt.style.use("fivethirtyeight")

# 这里以给定的excel表格为输入
# 该数据可以在给定的原始代码中找到
# 可以将该段代码和数据拷贝到本地运行，查看可视化效果
path = "/data/coItcastBrain/原始数据已脱敏.xlsx"

# 读取excel表格
original_data = pd.read_excel(path, "Sheet4")

# 我们会将"客户"的消息内容和"咨询师"的消息内容分开统计
# 分别获得对应的内容
user_original_data = original_data[original_data["消息发送人"] == "客户"]
employee_original_data = original_data[original_data["消息发送人"] == "咨询师"]


# 分别在数据中添加新的句子长度列
user_original_data["sentence_length"] = list(
    map(lambda x: len(str(x)), user_original_data["内容"])
)

employee_original_data["sentence_length"] = list(
    map(lambda x: len(str(x)), employee_original_data["内容"])
)

print("绘制学员对话句子长度分布图:")
# 绘制学员对话句子长度的数量分布图
sns.countplot("sentence_length", data=user_original_data)
```