京东-贪心 NLP 项目实验手册

项目 3：京东客服对话系统

项目设计和编写: 林培鑫 姜冰钰

后期编辑: 李文哲

单 位: 贪心科技

2020 年 8 月 27 日

目 录

1 项目描述与目标

对话系统（Dialogue Systems）又可以称之为**聊天机器人（ChatBot）**，主要是实现自动与用户进行对话的功能。帮助用户完成某些具体的任务（下单、打车、订座等）的对话系统可以称之为**任务导向型（Task-oriented）**的对话系统；解答用户的某些问题**（询问天气、股价、交通等）**的对话系统可以称之为**问答型（QA-based）**的对话系统；除此之外，还有和用户聊天的**闲聊型（Chatting）**对话系统。大多数的对话系统都是混合了几种类型的功能。

对话系统中语言的生成主要可以分为两种方式：**检索式（Retrieval）**和**生成式（Generative）**。**检索式方法**我们一般会构建一个语料库，为 FAQ 存放 query-response pairs，然后用户发起一个新的 query 时，我们去匹配为这个 query 检索最佳的 response。上述过程一般又可以分为召回（Retrieve）和排序（ranking）两个部分：召回即通过 query 找到语料库中最相似的几十个或几百个 query，大大减小候选 Response 的数量。这一部分我们一般采用轻量级的方法，如倒排索引（Inverted Index）和近似近邻搜索（Approximate Nearest Neighbor Search）等进行快速检索。排序则是对召回的结果进行进一步的筛选，可以构建更复杂的特征，使用机器学习或深度学习的方法来进行排序。**生成式方法**基本可以使用我们在上一阶段的文本生成项目所用的 Seq2seq 或 PGN 技术，但是现在更热门的是基于 Self-Attention 的预训练模型（BERT、GPT2 等）。另外，对话中回答的生成会比摘要或者翻译的难度大很多，原因在于给定 Encoder 的输入，在这一任务中 Decoder 的候选序列比起摘要或翻译要多得多。[1]

通过本项目的练习，你能掌握对话系统的技术：

- **意图识别**：这个地方我们做成一个简单的文本二分类任务，根据用户的开场白识别用户的意图是业务需求还是闲聊。
- **检索模型**：这个项目中我们用检索式的方法来做业务型对话的回答生成。我们将用 Approximate Nearest Neighbor Search(ANNS) 方法中较为常用的 Hierarchical Navigable Small World(HNSW) 来做召回的部分；然后将构建各种相似度特征（包括深度匹配网络），并利用

LightGBM 来训练一个 Learning2Rank 模型。

- **生成模型**：这个项目中我们将使用 BERT 模型来实现闲聊对话的生成。然后对于体积过大的 BERT 模型，我们还会进行压缩以便进行部署。
- **对话管理**：对于多轮的对话，我们需要对考虑上一轮的内容，以便对下一轮生成的回复进行合理性的控制。

同时，通过本项目，你将：

- 1. 熟练掌握如何使用 FastText 做文本分类。
- 2. 熟练掌握 ANNS 理论和技术。
- 3. 熟练掌握 Learning2Rank 的特征工程和 LightGBM 模型。
- 4. 熟练掌握 Transformer 和 BERT 模型。
- 5. 熟练掌握模型压缩的方法。

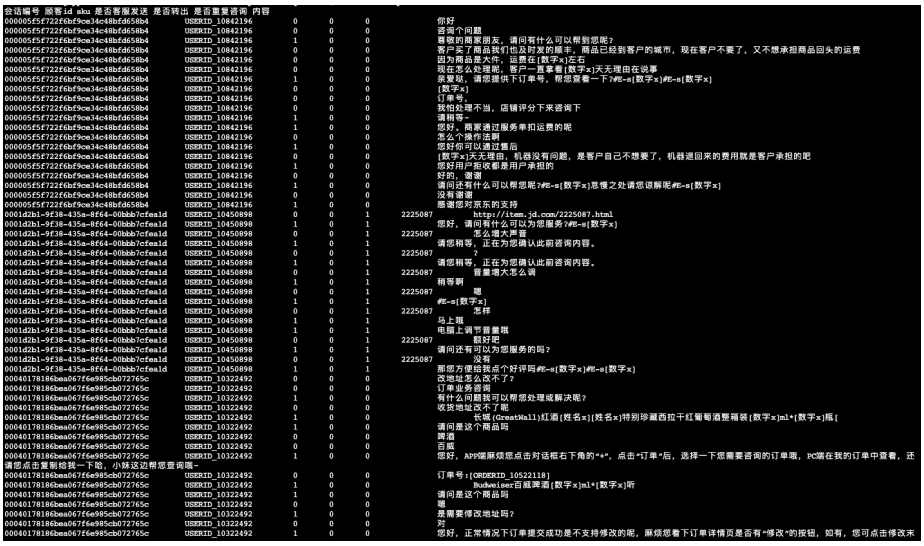
作为第三个项目，我们的目标是帮助大家完成一个客服智能对话系统，通过这个任务让大家熟悉对话系统的基本架构和常用技术。该项目的数据和部分技术用于京东的 JIMI 和京小智。

2 项目数据描述

京东对话数据集在本项目中，我们使用的是京东电商的客服对话数据。原始数据有将近 200 万行，如图 1所示。
我们会只取大概 10% 的数据，对同属一个对话的句子进行合并，两两结合为 query-response pairs，如图 2所示。

3 项目事宜

本项目相对于上一个项目，将更加注重对大家动手能力的提升，让大家完整的去了解一个 NLP 项目的整体框架，并且实现各个环节的重要细节。



会话编号	顺客id aku	是否客服发送	是否转出	是否重复咨询	内容
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	0	0	0	你好
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	0	0	0	咨询个问题
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	1	0	0	尊敬的商家朋友，请问有什么可以帮到您呢？
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	0	0	0	客户买了商品我们也及时发的顺丰，商品已经到客户的城市，现在客户不要了，又不想承担商品回头的运费
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	0	0	0	因为商品金大件，运费在{数字x}左右
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	0	0	0	现在怎么处理呢，客户一直拿着{数字x}天无理由在说事
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	1	0	0	亲爱哒，请您提供下订单号，帮您查看一下?#E-s[数字x]#E-s[数字x]
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	0	0	0	{数字x}
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	0	0	0	订单号，
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	0	0	0	我怕处理不当，店铺评分下来咨询下
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	1	0	0	请稍等~
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	1	0	0	您好，商家通过服务单扣运费的呢
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	0	0	0	怎么个操作法啊
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	1	0	0	您好你可以通过售后
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	0	0	0	{数字x}天无理由，机器没有问题，是客户自己不想要了，机器退回来的费用就是客户承担的呢
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	1	0	0	您好用户拒收都是用户承担的
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	0	0	0	好的，谢谢
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	1	0	0	请问还有什么可以帮您呢?#E-s[数字x]忽懂之处请您谅解呢#E-s[数字x]
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	0	0	0	没有谢谢
000005f5f722f6bf9ce34c48bfd658b4	USERID_10842196	1	0	0	感谢您对京东的支持
0001d2b1-9f38-435a-8f64-00bbb7cfea1d	USERID_10450898	0	1	2225087	http://item.jd.com/2225087.html
0001d2b1-9f38-435a-8f64-00bbb7cfea1d	USERID_10450898	1	1	2225087	您好，请问有什么可以为您服务?#E-s[数字x]
0001d2b1-9f38-435a-8f64-00bbb7cfea1d	USERID_10450898	0	1	2225087	怎么增大声音
0001d2b1-9f38-435a-8f64-00bbb7cfea1d	USERID_10450898	1	1	2225087	请您稍等，正在为您确认此前咨询内容。
0001d2b1-9f38-435a-8f64-00bbb7cfea1d	USERID_10450898	0	1	2225087	7
0001d2b1-9f38-435a-8f64-00bbb7cfea1d	USERID_10450898	1	1	2225087	请您稍等，正在为您确认此前咨询内容。
0001d2b1-9f38-435a-8f64-00bbb7cfea1d	USERID_10450898	0	1	2225087	音量增大怎么调
0001d2b1-9f38-435a-8f64-00bbb7cfea1d	USERID_10450898	1	1	2225087	稍等啊
0001d2b1-9f38-435a-8f64-00bbb7cfea1d	USERID_10450898	0	1	2225087	嗯
0001d2b1-9f38-435a-8f64-00bbb7cfea1d	USERID_10450898	1	1		#E-s[数字x]
0001d2b1-9f38-435a-8f64-00bbb7cfea1d	USERID_10450898	0	1	2225087	怎样
0001d2b1-9f38-435a-8f64-00bbb7cfea1d	USERID_10450898	1	1		马上哦
0001d2b1-9f38-435a-8f64-00bbb7cfea1d	USERID_10450898	1	1		电脑上调节音量哦
0001d2b1-9f38-435a-8f64-00bbb7cfea1d	USERID_10450898	0	1	2225087	额好吧
0001d2b1-9f38-435a-8f64-00bbb7cfea1d	USERID_10450898	1	1		请问还有可以为您服务的吗？
0001d2b1-9f38-435a-8f64-00bbb7cfea1d	USERID_10450898	0	1	2225087	没有
0001d2b1-9f38-435a-8f64-00bbb7cfea1d	USERID_10450898	1	1		那您方便给我点个好评吗#E-s[数字x]#E-s[数字x]
00041078186bea067f6e985cb072765c	USERID_10322492	0	0		改地址怎么改不了？
00041078186bea067f6e985cb072765c	USERID_10322492	0	0		订单业务咨询
00041078186bea067f6e985cb072765c	USERID_10322492	1	0		有什么问题我可以帮您处理或解决呢？
00041078186bea067f6e985cb072765c	USERID_10322492	0	0		收货地址改不了呢
00041078186bea067f6e985cb072765c	USERID_10322492	1	0		长城(GreatWall)红酒{姓名x}{姓名x}特别珍藏西拉干红葡萄酒整箱装{数字x}ml*{数字x}瓶{
00041078186bea067f6e985cb072765c	USERID_10322492	1	0		请问是这个商品吗
00041078186bea067f6e985cb072765c	USERID_10322492	0	0		跨道
00041078186bea067f6e985cb072765c	USERID_10322492	0	0		百威
00041078186bea067f6e985cb072765c	USERID_10322492	1	0		您好，APP端麻烦您点击对话框右下角的“+”，点击“订单”后，选择一下您需要咨询的订单哦，PC端在我的订单中查看，还请您点击复制给我一下哈，小妹这边帮您查询哦~
请您点击复制给我一下哈，小妹这边帮您查询哦~					
00041078186bea067f6e985cb072765c	USERID_10322492	0	0		订单号:[ORDERID_10522118]
00041078186bea067f6e985cb072765c	USERID_10322492	1	0		Budweiser百威啤酒{数字x}ml*{数字x}听
00041078186bea067f6e985cb072765c	USERID_10322492	1	0		请问是这个商品吗
00041078186bea067f6e985cb072765c	USERID_10322492	0	0		嗯
00041078186bea067f6e985cb072765c	USERID_10322492	1	0		是需要修改地址吗？
00041078186bea067f6e985cb072765c	USERID_10322492	0	0		对
00041078186bea067f6e985cb072765c	USERID_10322492	1	0		您好，正常情况下订单提交成功是不支持修改的呢，麻烦您看下订单详情页是否有“修改”的按钮，如有，您可点击修改来

图 1: 原始对话数据

```
session_id,custom,assistance
000005f5f722f6bf9ce34c48bfd658b4,你好{SEP}咨询个问题,尊敬的商家朋友，请问有什么可以帮到您呢？
000005f5f722f6bf9ce34c48bfd658b4,客户买了商品我们也及时发的顺丰，商品已经到客户的城市，现在客户不要了，又不想承担商品回头的运费{SEP}因为商品是大件，运费在{数字x}左右{SEP}现在怎么处理呢，客户一直拿着{数字x}天无理由在说事，是亲哒，请您提供下订单号，帮您查看一下？？
000005f5f722f6bf9ce34c48bfd658b4,{数字x}{SEP}订单号，{SEP}我怕处理不当，店铺评分下来咨询下,请稍等~{SEP}您好，商家通过服务单扣运费的呢
000005f5f722f6bf9ce34c48bfd658b4,怎么个操作法啊,您好你可以通过售后
000005f5f722f6bf9ce34c48bfd658b4,{数字x}天无理由，机器没有问题，是客户自己不想要了，机器退回来的费用就是客户承担的呢,您好用户拒收都是用户承担的
000005f5f722f6bf9ce34c48bfd658b4,好的，谢谢,请问还有什么可以帮您呢？感谢之处请您谅解呢
000005f5f722f6bf9ce34c48bfd658b4,没有谢谢,感谢您对京东的支持
0001d2b1-9f38-435a-8f64-00bbb7cfea1d,怎么增大声音,您好，请问有什么可以为您服务？
0001d2b1-9f38-435a-8f64-00bbb7cfea1d,音量增大怎么调,稍等啊
0001d2b1-9f38-435a-8f64-00bbb7cfea1d,怎样,马上哦{SEP}电脑上调节音量哦
0001d2b1-9f38-435a-8f64-00bbb7cfea1d,额好吧,请问还有可以为您服务的吗？
0001d2b1-9f38-435a-8f64-00bbb7cfea1d,没有,那您方便给我点个好评吗
00041078186bea067f6e985cb072765c,订单业务咨询,有什么问题我可以帮您处理或解决呢？
00041078186bea067f6e985cb072765c,收货地址改不了呢,长城(GreatWall)红酒{姓名x}{姓名x}特别珍藏西拉干红葡萄酒整箱装{数字x}ml*{数字x}瓶{{SEP}请问是这个商品吗
00041078186bea067f6e985cb072765c,跨道{SEP}百威,您好，APP端麻烦您点击对话框右下角的“+”，点击“订单”后，选择一下您需要咨询的订单哦，PC端在我的订单中查看，还请您点击复制给我一下哈，小妹这边帮您查询哦~
00041078186bea067f6e985cb072765c,订单号:{订单x},Budweiser百威啤酒{数字x}ml*{数字x}听{SEP}请问是这个商品吗
00041078186bea067f6e985cb072765c,跨道太大没法报{SEP}只能申请退款吗？,请问与原地址距离远吗？
00041078186bea067f6e985cb072765c,不远，对面小区{SEP}包送货的直接上门{SEP}又麻烦,配送员配送前会提前联系您的，建议您到时和配送员协商，您看可以吗？
00041078186bea067f6e985cb072765c,能备注吗？,这个配送会提前给您联系的呢，您可以和配送协商哟
00041078186bea067f6e985cb072765c,发票是纸质吗？,{SEP}请问还有其他还可以帮到您的吗？
00041078186bea067f6e985cb072765c,那发票是随货一起发还是单独发收货人地址,难度较发票的是的呢
00041078186bea067f6e985cb072765c,我收货地址跟收票地址不同{SEP}都寄到收票人地址吗？是在您完成后一个工作日会单独邮寄给您呢
00041078186bea067f6e985cb072765c,没有了,姨子态慢的地方，您多多包含哦，辛苦您点击聊天框中的{站点x}给我一个评价哦，妹子期待再次为您服务
00041078186bea067f6e985cb072765c,谢谢,不客气呢，亲爱
0004e921-0f2f-4490-abdb-75544a2481fe,vE吗,您好在的，请问有什么可以帮您呢？
0004e921-0f2f-4490-abdb-75544a2481fe,这个可以剪辑视频，ps吗{SEP}这个笔记本,可以的哈
0004e921-0f2f-4490-abdb-75544a2481fe,还有玩装网{数字x}{SEP}{数字x}u28观,可以玩的{SEP}请问还有什么能帮到您的呢？
0004e921-0f2f-4490-abdb-75544a2481fe,外观只有黑色的吗{SEP}想知道那个类型的cpu和显卡适合我用才说的功能,是的哦{SEP}因特尔{数字x}代cpu和{数字x}ti显卡可以满足您的需求了哈{SEP}请问还有什么能帮到您的呢？
00059cdd9920d7a5c89ff7688ed74ee,请问京东自营的商品怎么还要等待厂商处理呢？,请问您是咨询之前的问题还是有其他的问题需要处理呢？{SEP}有什么问题我可以帮您处理或解决呢？
00059cdd9920d7a5c89ff7688ed74ee,好吧,嗯呢{SEP}请您等待下了哦{SEP}{SEP}请问还有其他还可以帮到您的吗？
00059cdd9920d7a5c89ff7688ed74ee,没有了,好哒~{SEP}感谢您对京东的支持，祝您生活愉快，再见！
00054fc7ff74ab412a9c52176b97c486,咨询订单号:{订单x}订单金额:{金额x}下单时间:{日期x},有什么问题我可以帮您处理或解决呢?{SEP}您好，您能提供下订单号，方便给小妹查询下吗？
00054fc7ff74ab412a9c52176b97c486,物流太长时间没进展了,这边看您自己已经申请了呢
00054fc7ff74ab412a9c52176b97c486,什么时候能有退货的结果呢？,给您选择吧了
00054fc7ff74ab412a9c52176b97c486,好的 @对的
00054fc7ff74ab412a9c52176b97c486,好的,要是有，您能收就行{SEP}
00054fc7ff74ab412a9c52176b97c486,没有了，谢谢!,请问还有其他还可以帮到您的吗？
00056f2a4ddbe19ecee325873c0eb21c,外观这个订单是下周准送到{数字x}号送货安装?、正在为您查询，还请您稍微等待一小会 。{姓名x}。 !{SEP}{数字x}美的(Midea){数字x}升对开门冰箱变频无霜一级能效智能APP大容量电冰箱米兰变BCD-{数字x}WKPZM(E){SEP}您看是这件商品吗
00056f2a4ddbe19ecee325873c0eb21c,这个订单有两件商品{SEP}还有一个洗衣机,好的哈{SEP}请问您是想在{数字x}月{数字x}号配送商品吗
00056f2a4ddbe19ecee325873c0eb21c,是的，您不是{数字x}号安装?安装是需要您联系厂家那边预约哦{SEP}小天鹅售后服务电话:{数字x}-{数字x}-{数字x}[SEP]冰箱是不需要安装的
00056f2a4ddbe19ecee325873c0eb21c,洗衣机不包安装吗?,是有安装服务的{SEP}需要联系厂家那边预约哈
00056f2a4ddbe19ecee325873c0eb21c,我单独联系?,是的哦
00056f2a4ddbe19ecee325873c0eb21c,不是你们联系么?{SEP}那这安装是收费还是免费的?,您稍等这面咨询下{SEP}厂家师傅免费上门给您安装调试好的{SEP}那我这面帮您预约到{数字x}号送货了{SEP}{电话x}{SEP}厂东{地址x}{SEP}电话地址是否需要更改呢
00056f2a4ddbe19ecee325873c0eb21c,不需要修改{SEP}是当天送货安装对吧?,已安排{数字x}号送货
00056f2a4ddbe19ecee325873c0eb21c,{订单x}{SEP}这个也帮我改下{SEP}{数字x}号送货,已预约过啦{SEP}配送师傅送货前会联系收货人，请保持电话畅通
00056f2a4ddbe19ecee325873c0eb21c,这个空调包改成了{数字x}号吗?{SEP}看到了，感谢。{SEP}您客气了哈，这是我们应该为您做的 。{SEP}给您带来的困扰再次真诚的给您致歉，请问还有其他可以帮助您的吗？
00056f2a4ddbe19ecee325873c0eb21c,没有了，谢谢,感客气了哈，这是我们应该为您做的 {SEP}感谢您对京东的支持，祝您生活愉快，再见{SEP}好的，后期有任何问题-请您联系我们京东为您处理哟
00075f14885579d360c76cf6a64c38a0,客户无法提供商品，,止为您查询，还请稍微等待一小会儿～{SEP}Adidas阿迪达斯基袋女{数字x}夏季新款NEO休闲运动短袖T恤阔腿裤七分裤BR{数字x}短袖+T恤BR{数字x}M{SEP}请核对下商品哦，谢谢
00075f14885579d360c76cf6a64c38a0,好的,您好，此单请终止取作，已经帮您备注服务单留言
```

图 2: 处理后的对话数据

3.1 项目的整个框架

整个项目框架如图 3所示，项目流程如图 4所示。下面对于图中每个模块做简要的描述，具体的细节请参考本文章后续的内容。

- **意图识别**：训练一个 FastText 模型对用户的输入进行意图识别，对业

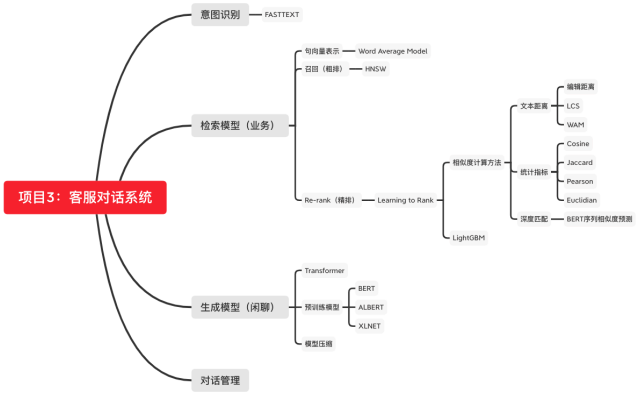


图 3: 京东客服对话系统项目模块架构图

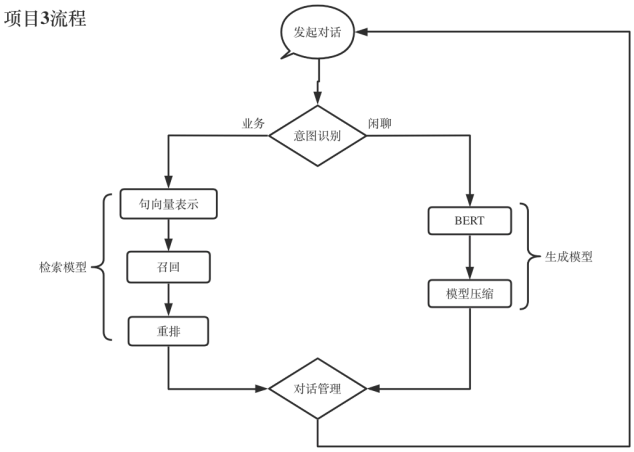


图 4: 京东客服对话系统项目流程图

务型需求使用检索模型来生成回复，对闲聊需求则使用生成模型来生成回复。这一部分的训练数据我们使用关键词来自动标注，每个样本是一句用户的输入，整理一系列业务相关的词汇，将包含业务关键词的样本标注为 1，否则为 0。

- **召回**：我们使用 Word Average Model (WAM) 来表示句向量，之后用 HNSW 模型来实现 ANNS。HNSW 我们可以使用 hnswlib 或者 faiss 来实现。

- **排序**：这一步我们需要构建多种相似度特征，主要可以分为几类：1. 基于字符串距离的（编辑距离、列文斯坦距离、LCS）；2. 基于向量距离的（cosine、Euclidian、Jaccard、WMD）；3. 基于统计量的（BM25、Pearson Correlation）；4. 基于深度匹配模型的。构建完特征后，我们使用 LightGBM 来训练一个 LTR 模型。
- **生成模型**：这个地方我们使用 BERT（也可以是 GPT2 等预训练模型）来做 seq2seq 的生成任务。之后对 BERT 模型进行压缩（蒸馏或剪裁）。
- **对话管理**：对轮对话管理需要实现 Dialogue State Tracing (DST) 和 Policy Selection。

4 项目主要技术介绍

4.1 Average word vectors

为了通过计算距离来实现近似搜索，我们需要将一个句子或者一篇文档表示为一个向量，这种方式叫做 Sentence Embedding 或者 Doc Embedding。其中最常用的一种模型是 Average word vectors 或者叫 Word Averaging Model(WAM)。其方法是得到将一个句子中每一个词向量在同一个 embedding 维度的值取平均值得到该句子在该嵌入维度的值。如图 5所示。

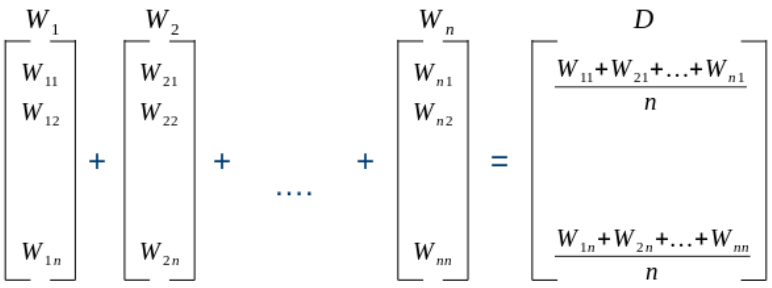


图 5: Average word vectors

4.2 Hierarchical Navigable Small World

HNSW[4] 的前身是 NSW (Navigable-Small-World-Graph)。NSW 通过设计出一个具有导航性的图来解决近邻图发散搜索的问题，但其搜索复杂度仍然过高，达到多重对数的级别，并且整体性能极易被图的大小所影响。HNSW 则是着力于解决这个问题。作者借鉴了 SkipList 的思想，提出了 Hierarchical-NSW 的设想。简单来说，按照一定的规则把一张的图分成多张，越接近上层的图，平均度数越低，节点之间的距离越远；越接近下层的图平均度数越高，节点之间的距离也就越近。搜索从最上层开始，找到本层距离最近的节点之后进入下一层。下一层搜索的起始节点即是上一层的最近节点，往复循环，直至找到结果（见图 6）。由于越是上层的图，节点越是稀少，平均度数也低，距离也远，所以可以通过非常小的代价提供了良好的搜索方向，通过这种方式减少大量没有价值的计算，减少了搜索算法复杂度。更进一步，如果把 HNSW 中节点的最大度数设为常数，这样可以获得一张搜索复杂度仅为 log(n) 的图。

实现 HNSW 主要有以下两个 package 可以选用：

hnswlib：一个轻量级的 library，专门用来实现 HNSW 算法，底层用 C++ 实现，有实现 Python 的 wrapper。

Facebook AI Similarity Search (Faiss)：Facebook 开发的专门用于做 Proximate Search 的 library，支持包括 KNN 和 HNSW 等多种搜索算法，也支持 PCA 等降维算法以提升性能。而且支持 GPU。Faiss 的底层也是用 C++ 开发，并实现 Python wrapper。Github repo: https://github.com/facebookresearch/faiss。

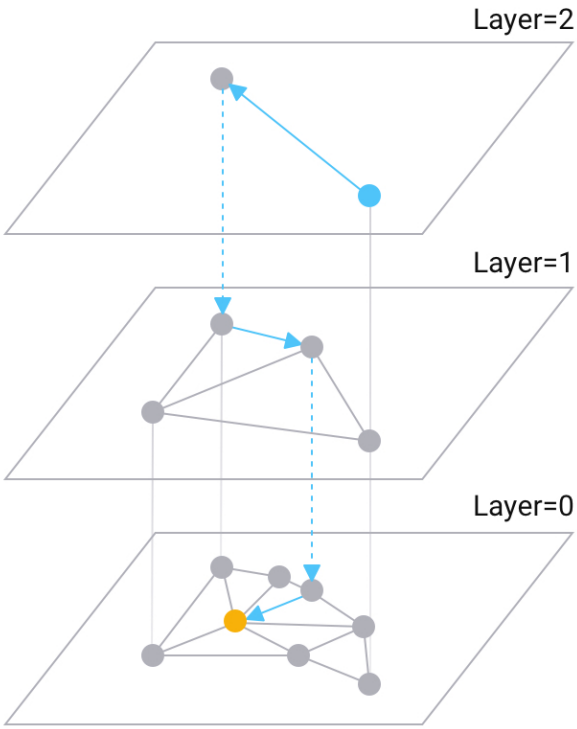


图 6: Hierarchical Navigable Small World

4.3 Learning-to-Rank

排序学习 [3] 是一个有监督的机器学习过程，对每一个给定的查询 - 文档对，抽取特征，通过日志挖掘或者人工标注的方法获得真实数据标注。然后通过排序模型，使得输入能够和实际的数据相似。常用的排序学习分为三种类型：PointWise，PairWise 和 ListWise。如图 5所示。

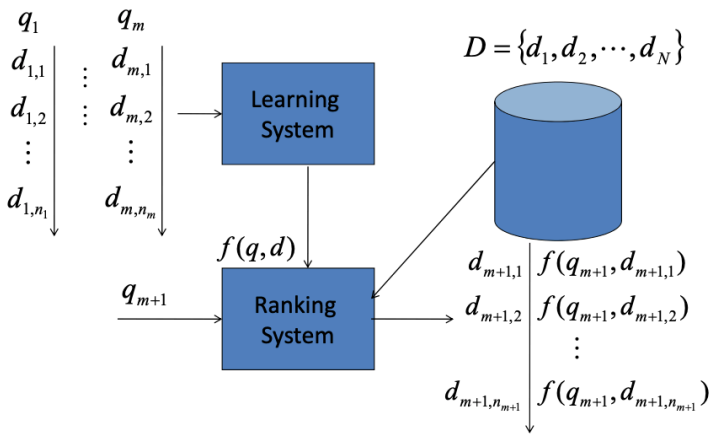


图 7: Learning-to-Rank

4.4 BERT

BERT[2] 是 Google 于 2018 年提出的语言模型，在多种 NLP 任务上有出色的表现，因此近几年成为 NLP 领域的热门模型。BERT 的全称是 Bidirectional Encoder Representation from Transformers，模型的核心有两个：1 是使用了基于 Self-Attention 的 Transformer[5]；2 是使用了 pre-train 的方法，即用了 Masked LM 和 Next Sentence Prediction 两种方法分别捕捉词语和句子级别的 representation。如图 8， 9所示。

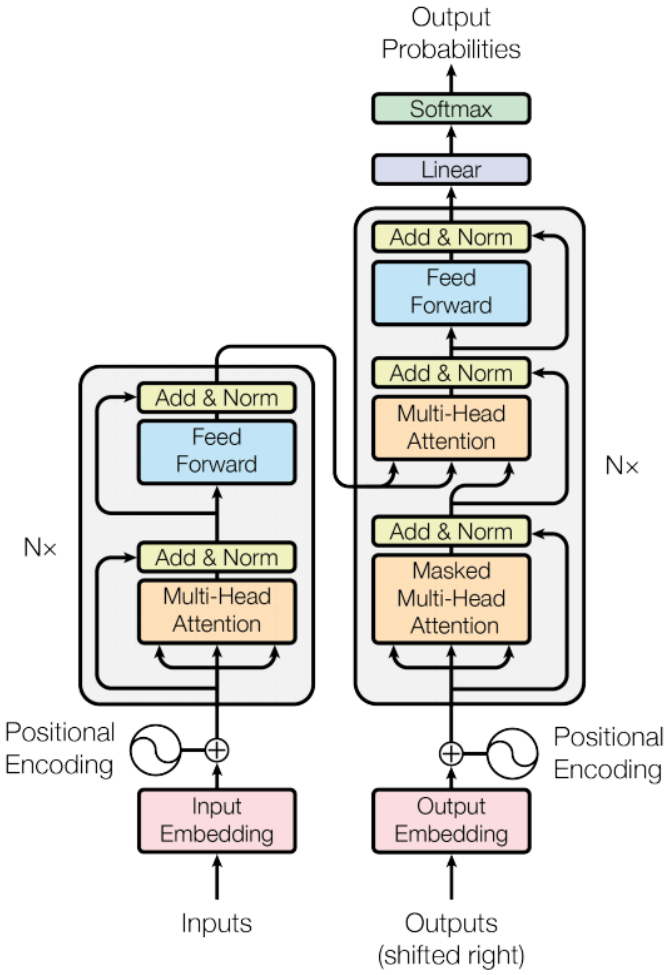


图 8: Transformer

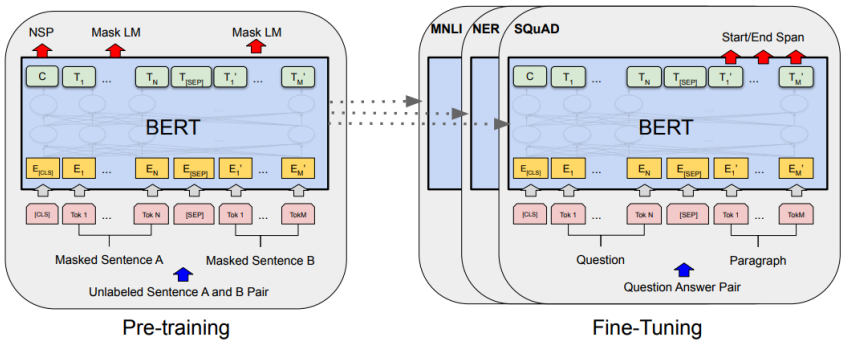


图 9: Bert

参考文献

[1] Chen, H., Liu, X., Yin, D., and Tang, J. A survey on dialogue systems: Recent advances and new frontiers. *CoRR abs/1711.01731* (2017).

[2] Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805* (2018).

[3] Li, H. A short introduction to learning to rank. *IEICE Trans. Inf. Syst. 94-D* (2011), 1854–1862.

[4] Malkov, Y. A., and Yashunin, D. A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *CoRR abs/1603.09320* (2016).

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *CoRR abs/1706.03762* (2017).