第六届泰迪杯数据挖掘挑战赛

一种基于潜在语义索引和卷积神经网络的智能阅读模型

- 系统名称：智能阅读助手EVA
- 答辩小组：83137C

2018年5月26日

目录

Contents

1.1 项目背景

数字阅读席卷社会

随着互联网的高速发展以及智能设备的普及，**数字阅读**以方便、快捷的优势,越来越被大众所接受和认可。据中国数字阅读大会上的调研数据显示，2017年全国数字阅读用户近**4**亿，人均电子书阅读量为**10.1**本，而纸质书阅读量仅**7.5**本。

新一代的阅读

借助自然语言处理技术，通过端到端的处理技术辅助快速阅读，直接对用户的问题进行处理，无需基于关键词搜索即可直接定位文档中的相关段落，并将答案直接反馈至用户。

未来的刚性需求

在传统的数字阅读中存在用户无法精准定位关键信息的问题，即无法满足用户仅需查找文档中某些片段以获取关键信息的需求。例如，当用户需要查找**法律文献**中的一些段落来解决法律疑惑时，只需要理解关键部分而无需精读整个法律文献；同样，对于**小说阅读**，如果用户仅需了解其中的特殊细节，也不需要对整部小说进行精细化阅读。

1.2 模型设计

智能阅读模型

读者

智能阅读系统
WEB/移动端

段落文本集合

问题 → 分词、字典化、序列化、填充 → 潜在语义索引(LSI) → 关键词匹配 → 候选回答 → 卷积神经网络(CNN) → 精准匹配 → 候选回答排序 → 回答

如果用户不满意，返回下一个回答

1.3 处理流程

第一部分：数据分析与预处理

对问题给出的数据集进行统计分析，提出该数据集进行处理时的关键挑战，并给出相应的预处理步骤；

第二部分：关键词匹配

首先对用户提出的问题进行分词，并将需要在其中寻找答案的文本构建成问答数据库。进而使用词频-逆向文件词频(TF-IDF)计算出问题以及段落的词频矩阵，再利用基于奇异值分解(SVD)的LSI方法将其转化为奇异矩阵，计算相似度，将相似度较大的若干个可能答案段落作为问题的粗匹配结果；

第三部分：精准匹配

我们在经典的TextCNN模型上进行优化，提出一个新的CNN模型在粗匹配结果上进行二次优化达到精确匹配的目的。

1.4 数据分析与预处理

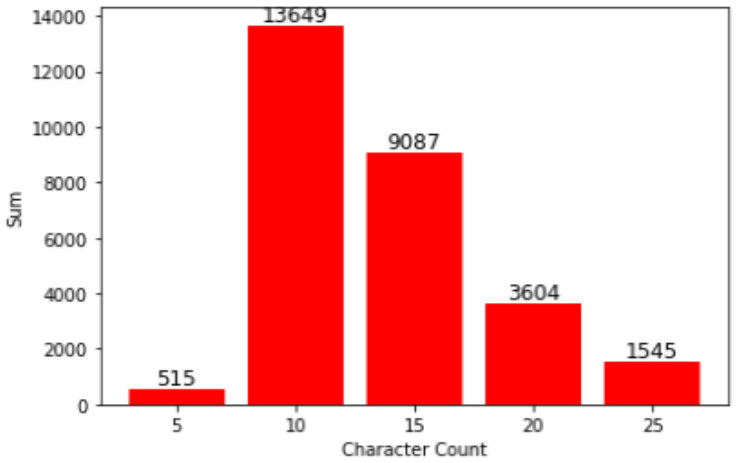
数据分析

- 高质量的数据集是模型匹配和优化的基础，对整个数据集进行分析处理可以促进对数据集的全面认知，从而更好地对数据进行特征工程编码表示，进一步提高数据集的质量。

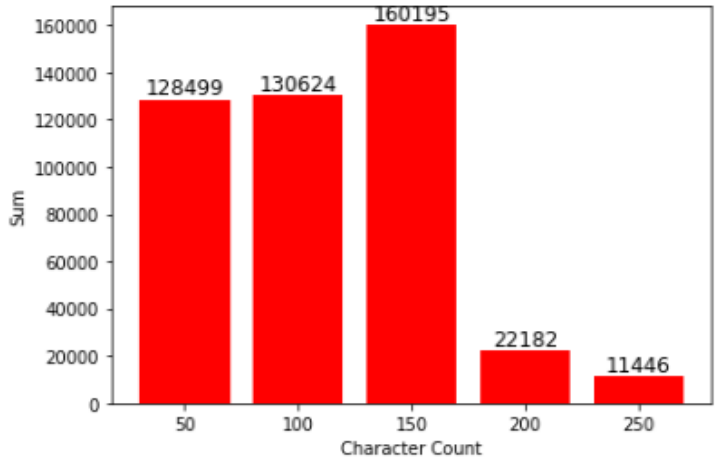
分词前	问题集	问题数量/个		最长问题/字		最短问题/字	平均长度/字
		30000		243		4	13
	答案集	回答数量/个	正确数量/个	错误数量/个	最长回答/字	最短回答/字	平均长度/字
		477019	127328	349691	6425	0	95
分词后	问题集	问题数量/个		最长问题/词		最短问题/词	平均长度/词
		30000		148		2	8
	答案集	回答数量/个	正确数量/个	错误数量/个	最长回答/词	最短回答/词	平均长度/词
		477019	127328	349691	3545	0	60

1.4 数据分析与预处理

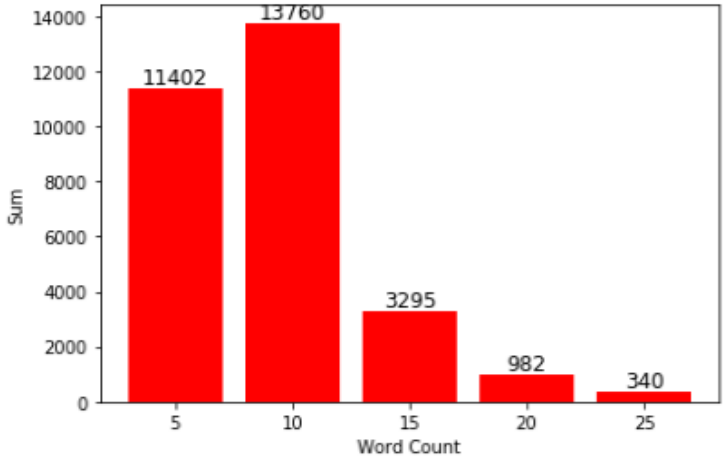
数据分析



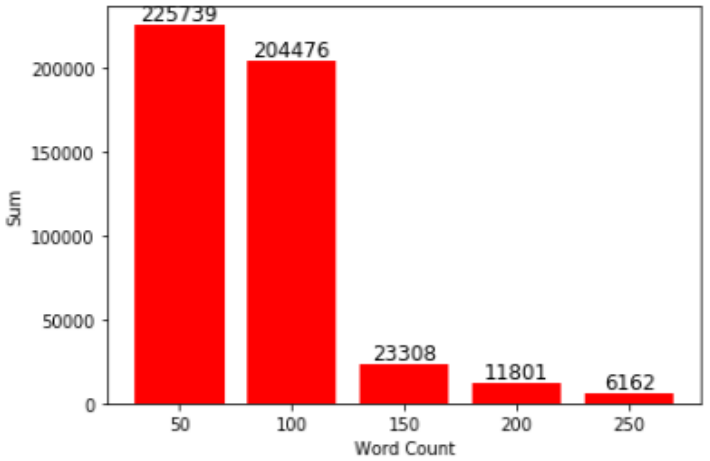
分词前统计字符频率（问题集）



分词前统计字符频率（答案集）



分词后统计字符频率（问题集）



分词后统计字符频率（答案集）

1.4 数据分析与预处理

数据预处理

例子

问题：“射雕英雄传中谁的武功天下第一”

回答：“王重阳武功天下第一”

- 分词
- 输出：['射雕 英雄传 中 谁 的 武功 天下第一', '王重阳 武功 天下第一']

- 字典化
- 输出：{'中': 5, '天下第一': 2, '射雕': 3, '武功': 1, '王重阳': 8, '的': 7, '英雄传': 4, '谁': 6}

- 序列化
- 输出： [[3, 4, 5, 6, 7, 1, 2]], [[8, 1, 2]]

- 填充
- 输出：[[0 0 0 3 4 5 6 7 1 2]], [[0 0 0 0 0 0 0 8 1 2]]

1.4 数据分析与预处理

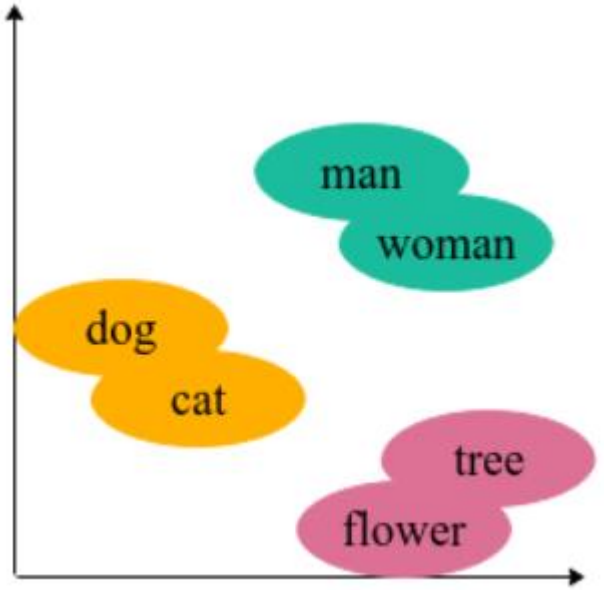
数据预处理

自然语言处理的问题转化为机器学习的问题，
需要将自然语音进行数字化表示：

独热表示(One-Hot)

man	[1, 0, 0, 0, 0, 0]
woman	[0, 1, 0, 0, 0, 0]
dog	[0, 0, 1, 0, 0, 0]
cat	[0, 0, 0, 1, 0, 0]
tree	[0, 0, 0, 0, 1, 0]
flower	[0, 0, 0, 0, 0, 1]

分布式表示



采用分布式表示进行单词嵌入，避免独热编码中的“维度灾难”和“词汇鸿沟”情况：

“维度灾难”：当维度增加时，所需存储空间呈指数增长。

“词汇鸿沟”：任意两个词之间都是孤立的，光从这两个向量看不出两个词是否存在关系。

1.5 关键词匹配

□ 词频-逆向文件频率模型(TF-IDF模型)

主要思想是指在一篇文章中，某个词语的重要性与该词语在这篇文章中出现的次数成正相关，同时与整个语料库中出现该词语的文章数成负相关。

$$\mathrm{TF}(\text{词频}) = \frac{\text{某个词在段落中的出现次数}}{\text{文章出现次数最多的词语的次数}}$$ 表示一个词语与一篇文章的相关性

$$\mathrm{IDF}(\text{逆文档频率}) = \log\left(\frac{\text{语料库中段落总数}}{\text{包含该词的段落数}}\right)$$ 表示一个词语的出现的普遍程度

□ 潜在语义索引模型(LSI模型)

LSI模型采用了基于奇异值分解（SVD）的方法，利用 SVD，将使用TF-IDF方法计算得出的词频矩阵转化为奇异矩阵，再将词语和文本映射到一个新的空间进行降维。

$$W_{m*n} = U_{m*k} * \Sigma_{k*k} * V^{T}_{k*n}$$

得到文本主题矩阵V^{T}_{k*n}之后，就可以计算文本之间的相似度，从而匹配出候选答案。

1.5 关键词匹配

问题：射雕英雄传中谁的武功天下第一？

文章段落数据库

关键词匹配

潜在语义索引(LSI)

1. 2880 0.6837087869644165 武功天下第一的王真人已经逝世，剩下我们四个大家半斤八两，各有所忌。
2. 7965 0.646489679813385 两人回到帐中，这番当真研习起《九阴真经》上的武功来，谈论之下，均觉对方一年来武功大有长进，均感欣慰。
3. 8377 0.6456590890884399 你上得华山来，妄想争那武功天下第一的荣号，莫说你武功未必能独魁群雄，纵然是当世无敌，天下英雄能服你这卖国好徒么？”
4. 2626 0.5858802199363708 丘处机道：“韩女侠，天下武学之士，肩上受了这样的一扳，若是抵挡不住，必向后跌，只有九指神丐的独家武功，却是向前俯跌。只因他的武功刚猛绝伦，遇强愈强。穆姑娘受教时日虽短，却已习得洪老前辈这派武功的要旨。她抵不住王师弟的一扳，但决不随势屈服，就算跌倒，也要跌得与敌人用力的方向相反。”
5. 5277 0.5856705904006958 武术中有言道：”未学打人，先学挨打。”初练粗浅功夫，却须由师父传授怎生挨打而不受重伤，到了武功精深之时，就得研习护身保命、解穴救伤、接骨疗毒诸般法门。须知强中更有强中手，任你武功盖世，也难保没失手的日子。这《九阴真经》中的“疗伤篇”，讲的是若为高手以气功击伤，如何以气功调理真元，治疗内伤。至于折骨、金创等外伤的治疗，研习真经之人自也不用再学。

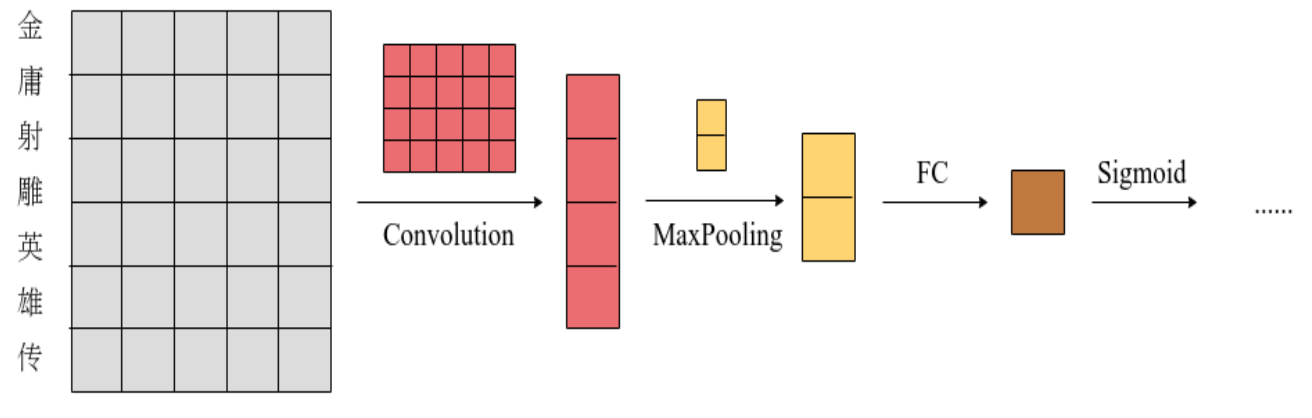
候选答案列表

1.6 精准匹配

□ 卷积神经网络(CNN)

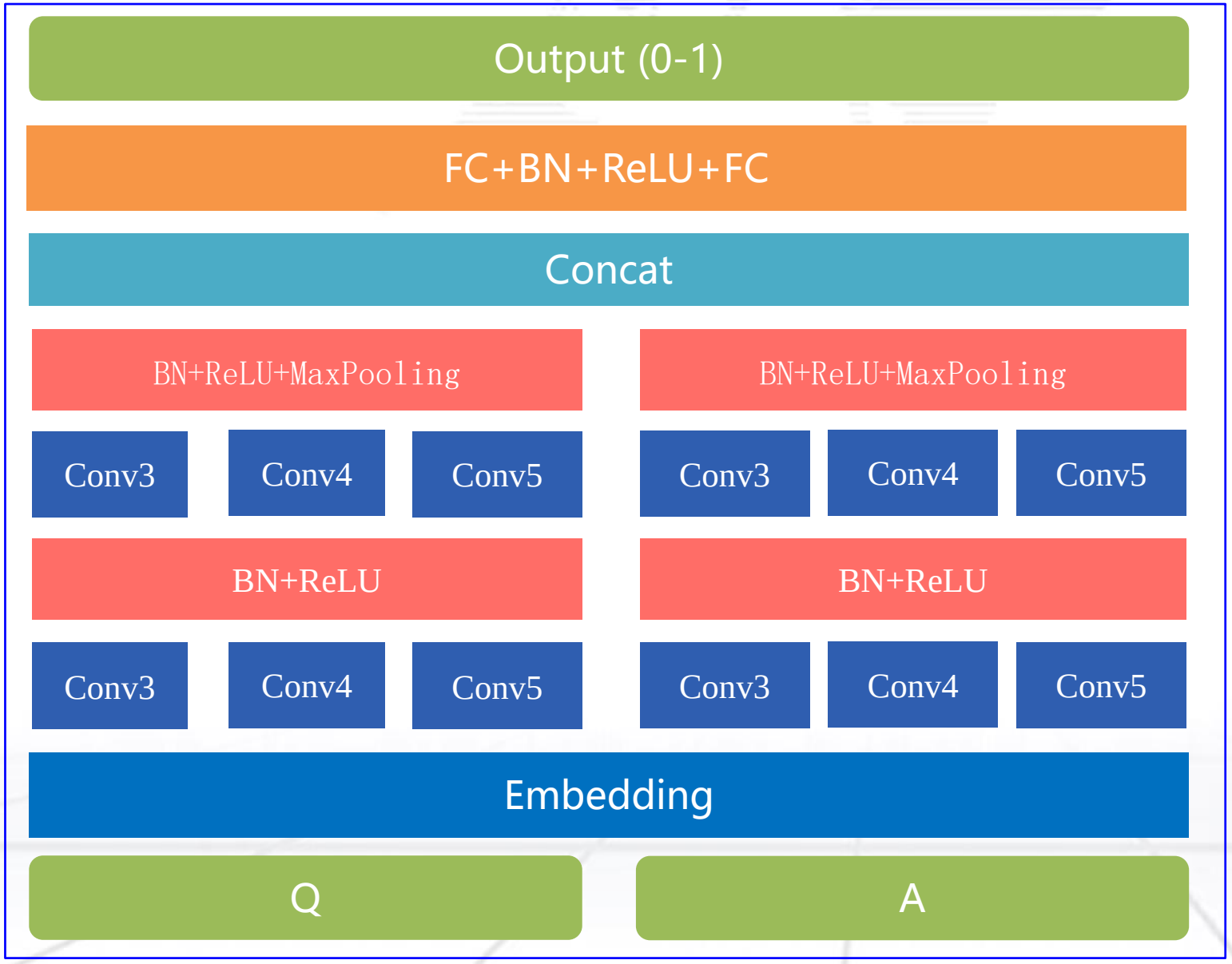
CNN不仅在图像领域表现优秀：图像识别、图像分割、图像检索等；
在自然语言处理方面也是大有用武之地：情感分析、文本分类、问答系统等。

在图像中，卷积核通常是对图像的一小块像素区域进行计算；
而在自然语言处理中，卷积核通常是对文本所构成的词向量进行计算。



通用的CNN-NLP模型示意图

1.6 精准匹配



Q&A

模型分别读入问题和回答。

Embedding

分别对问题和回答进行词嵌入，该层会在每次迭代中训练词向量，训练出来的词向量可以更好的适应自然语言处理任务。

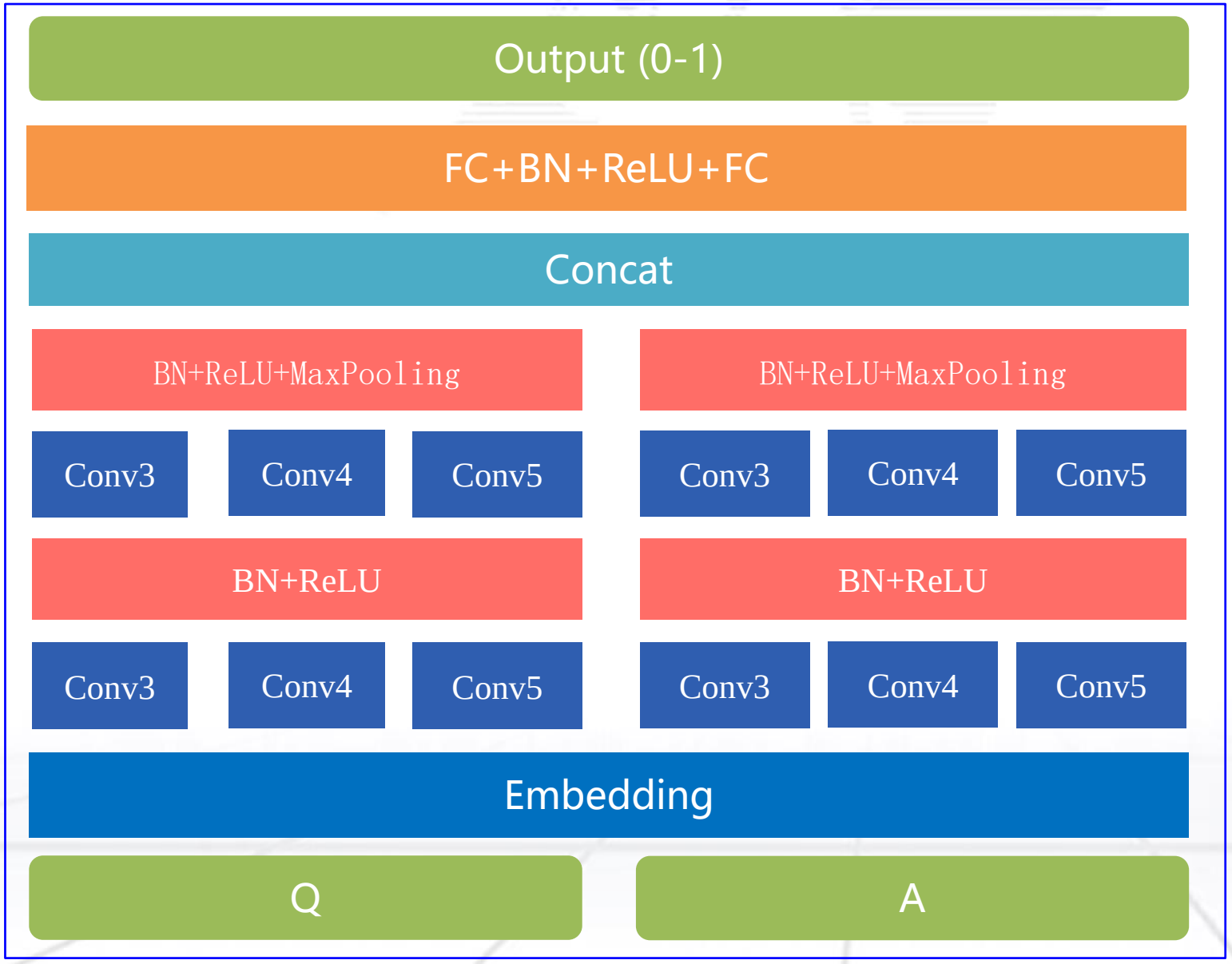
Conv3/Conv4/Conv5

分别对问题和回答进行两次卷积核大小为3、4、5的卷积操作，提取问答特征。

BN+ReLU

使用批规范化(Batch Normalization，简称BN)，加速收敛；同时使用线性整流函数(Rectified Linear Unit，简称ReLU)激活函数，防止反向传播过程中的梯度问题（梯度消失和梯度爆炸）。

1.6 精准匹配



Conv3/Conv4/Conv5

再次进行相同的卷积操作，进一步提取问答特征。

BN+ReLU+MaxPooling

再次进行同④的批规范化(BN)和激活函数(ReLU)操作，紧接着经过最大池化层(MaxPooling)，对数据进行降维，降低后续全连接层的复杂度。

Concat

将池化的向量连接起来。

FC+BN+ReLU+FC

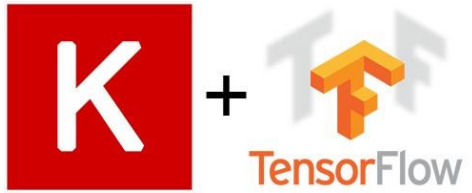
最后一步的FC起到“分类器”的作用，而第一步的FC则是起到降维的作用，如果直接进入最后分类阶段，神经元参数过多，容易导致模型过拟合。

目录

Contents

2.1 实验环境

项目	内容
操作系统	Ubuntu 16.04
内存大小	128G
硬盘容量	8T
GPU型号	GTX 1080Ti * 4
开发语言	Python3
依赖框架	Keras、Tensorflow、Matplotlib等



2.2 关键词匹配实验

在关键词匹配阶段，我们对比了TF-IDF模型 、LSI模型、Doc2Vec-DM、Doc2Vec-DOW 模型，由于LSI模型在TF-IDF模型的基础上，采用了奇异值分解，保留了一定的语义信息，进一步优化了搜索结果。而Doc2Vec-DM、Doc2Vec-DOW根据文本单词的空间向量匹配，面对全文搜索显得有些吃力。

因此，我们在这个阶段选择了LSI模型作为关键词匹配方案。

TF-IDF

LSI

Doc2Vec-DM

Doc2Vec-DOW

2.2.1 关键词匹配实验

TF-IDF匹配结果

- 测试问题：射雕英雄传中谁的武功天下第一?

```
1061 0.2566992044448525 第五回　弯弓射雕(1)
1172 0.2566992044448525 第五回　弯弓射雕(2)
3880 0.20211602747440338 郭靖涨红了脸，答道：“我想，王真人的武功既已天下第一，他再练得更强，仍也不过是天下第一。我还想，他到华山论剑，倒不是为了争天下第一的名头，而是要得这部《九阴真经》。他要得到经书，也不是为了要练其中的功夫，却是相救普天下的英雄豪杰，教他们免得互相所杀，大家不得好死。”
2880 0.18251502513885498 武功天下第一的王真人已经逝世，剩下我们四个大家半斤八两，各有所忌。
8377 0.18021109700202942 你上得华山来，妄想争那武功天下第一的荣号，莫说你武功未必能独魁群雄，纵然是当世无敌，天下英雄能服你这卖国好徒么？”
```

2.2.2 关键词匹配实验

LSI匹配结果

- 测试问题：射雕英雄传中谁的武功天下第一？

1. 2880 0.6837087869644165 武功天下第一的王真人已经逝世，剩下我们四个大家半斤八两，各有所忌。
2. 7965 0.646489679813385 两人回到帐中，这番当真研习起《九阴真经》上的武功来，谈论之下，均觉对方一年来武功大有长进，均感欣慰。
3. 8377 0.6456590890884399 你上得华山来，妄想争那武功天下第一的荣号，莫说你武功未必能独魁群雄，纵然是当世无敌，天下英雄能服你这卖国好徒么？”
4. 2626 0.5858802199363708 丘处机道：“韩女侠，天下武学之士，肩上受了这样的一扳，若是抵挡不住，必向后跌，只有九指神丐的独家武功，却是向前俯跌。只因他的武功刚猛绝伦，遇强愈强。穆姑娘受教时日虽短，却已习得洪老前辈这派武功的要旨。她抵不住王师弟的一扳，但决不随势屈服，就算跌倒，也要跌得与敌人用力的方向相反。”
5. 5277 0.5856705904006958 武术中有言道：”未学打人，先学挨打。”初练粗浅功夫，却须由师父传授怎生挨打而不受重伤，到了武功精深之时，就得研习护身保命、解穴救伤、接骨疗毒诸般法门。须知强中更有强中手，任你武功盖世，也难保没失手的日子。这《九阴真经》中的“疗伤篇”，讲的是若为高手以气功击伤，如何以气功调理真元，治疗内伤。至于折骨、金创等外伤的治疗，研习真经之人自也不用再学。

2.2.3 关键词匹配实验

Doc2Vec-DM匹配结果

- 测试问题：射雕英雄传中谁的武功天下第一?

1. 1158 0.8241669535636902 正自怔怔出神，突然听到华筝的声音在后叫道：”郭靖，快来，快来！”
2. 2608 0.8154236674308777 郭靖在赵王府中见过包惜弱的居所，听到这里，心下已是恍然。
3. 2441 0.8090466260910034 郭靖听得语声，心中大喜，叫道：“师父，快救弟子！”
4. 7378 0.8052639365196228 第三十五回　铁枪庙中(1)
5. 6354 0.8042024970054626 两人走入林中，郭靖将黄蓉背起，仍由她指点路径，一步步的向外走去。

2.2.4 关键词匹配实验

Doc2Vec-DOW匹配结果

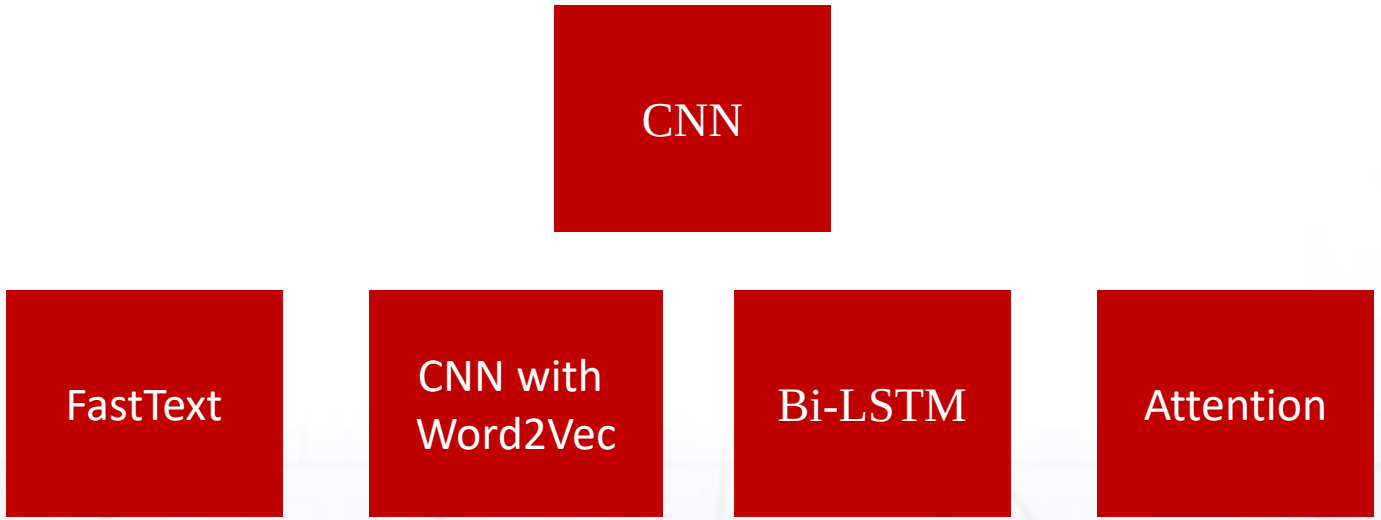
- 测试问题：射雕英雄传中谁的武功天下第一？

1. 5177 0.9147704839706421 宫内带刀护卫巡逻严紧，但周、郭、黄轻身功夫何等了得，岂能让护卫发见？洪七公识得御厨房的所在，低声指路，片刻间来到了六部山后的御厨。
2. 3014 0.9035613536834717 这词黄蓉曾由父亲教过，知道是岳飞所作的《小重山》，又见下款写着“五湖废人病中涂鸦”八字，想来这“五湖废人”必是那庄主的别号了。但见书法与图画中的笔致波磔森森，如剑如戟，岂但力透纸背，直欲破纸飞出一般。
3. 4405 0.9006547331809998 他武功既强，眼力又高，搜罗的奇珍异宝不计其数，这时都供在亡妻的圹室之中。黄蓉见那些明珠美玉、翡翠玛瑙之属在灯光下发出淡淡光芒，心想：
4. 1240 0.8996094465255737 那道人道：“睡觉之前，必须脑中空明澄澈，没一丝思虑。然后敛身侧卧，鼻息绵绵，魂不内荡，神不外游。”当下传授了呼吸运气之法、静坐敛虑之术。
5. 2428 0.8995554447174072 郭靖道：“眼不视而魂在肝、耳不闻而精在肾、舌不吟而神在心、鼻不香而魄在肺、四肢不动而意在脾，是为五气朝元。”

2.3 精匹配实验

在精准匹配阶段，我们的CNN模型与FastText模型、CNN with Word2Vec模型、Bi-LSTM模型以及Attention模型做了对比，综合训练时间、训练效率、准确率、F1-Score以及泛化能力的评估，本文提出的模型均得到优秀的表现。

因此，我们在这个阶段选择了CNN模型作为关键词匹配方案。



2.3.1 实验结果

- 测试问题：丐帮帮主是谁?

1. 5953 0.24029719829559326 奉立帮主是丐帮中的第一等大事，丐帮的兴衰成败，倒有一大半决定于帮主是否有德有能。当年第十七代钱帮主昏喑懦弱，武功虽高，但处事不当，净衣派与污衣派纷争不休，丐帮声势大衰。直至洪七公接任帮主，强行镇压两派不许内奸.丐帮方得在江湖上重振雄风。这些旧事此日与会群丐尽皆知晓，是以一听到要奉立帮主，人人全神贯注，屏息无声。
2. 6099 0.2300946315129598 鲁有脚道:“自来打狗棒法，非丐帮帮主不传，简长老难道不知这个规矩？”
3. 6009 0.19652195771535239 丐帮自洪七公接掌帮主以来，在江湖上从未失过半点威风，现下洪七公一死，新帮主竟如此软弱，群丐听了他这几句言语，无不愤恨难平。
4. 5961 0.1742761731147766 鲁有脚侧目斜睨杨康，心道:“凭你这小子也配作本帮帮主，统率天下各路丐帮？”伸手接过竹杖，见那杖碧绿晶莹，果是本帮帮主世代相传之物，心想，“必是洪帮主感念相救之德，是以传他。老帮主既有遗命，我辈岂敢不遵？我当赤胆忠心的辅他，莫要堕了洪帮主建下的基业。”于是双手举杖过顶，恭恭敬敬的将竹杖递还给杨康，朗声说道:“我等遵从老帮主遗命，奉杨相公为本帮第十九代帮主。”众丐齐声欢呼。
5. 6876 0.15333682298660278 “到了岳州后，丐帮大会君山。他事先悄悄对我说道：洪恩师曾有遗命，着他接任丐帮的帮主，我又惊又喜，实在难以相信，但见丐帮中连辈份最高的众长老对他也是十分敬重，却又不由得我不信。我不是丐帮的人，不能去参预大会，便在岳州城里等他，心里想着，他一旦领袖丐帮群雄，必能为国为民，做一番轰轰烈烈的大事出来，将来也必能手刃大寇，为义父义母报仇。

2.3.2 实验结果

- 测试问题：江南七怪分别是哪几位?

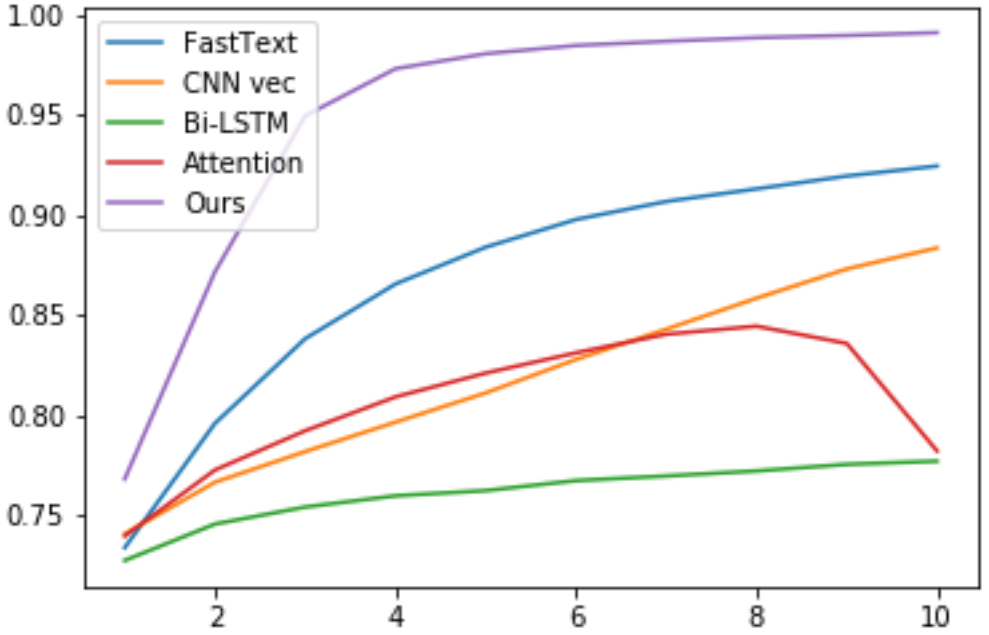
1. 944 0.4063337246576945 江南六怪这时已知那男子并非她丈夫，只是一个被她捉来喂招练功的活靶子，这女子自必是铁尸梅超风了。
2. 3373 0.3740468025207519 5 江南六怪面面相觑，都是又惊又喜：“靖儿从哪里学来这样高的武功？”
3. 7584 0.3443427880605061 7 欧阳伯伯拦在墓门，那江南六怪如何能再逃脱毒手？这是个瓮中捉鳖之计啊。”
4. 1137 0.3165022134780884 “全真教下弟子丘处机沐手稽首，谨拜上江南六侠柯公、朱公、韩公、南公、全公、韩女侠尊前：江南一别，忽忽十有六载。七侠千金一诺，间关万里，云天高义，海内同钦，识与不识，皆相顾击掌而言曰：不意古人仁侠之风，复见之于今日也。”
5. 1574 0.2901011705398559 6 江南六怪与郭靖晓行夜宿，向东南进发，在路非止一日，过了大漠草原。

2.4 训练时间

模型	FastText	CNN vec	Bi-LSTM	Attetion	Ours
时间/s	94	455	2888	3889	300

训练时间示意图

2.5 训练效果

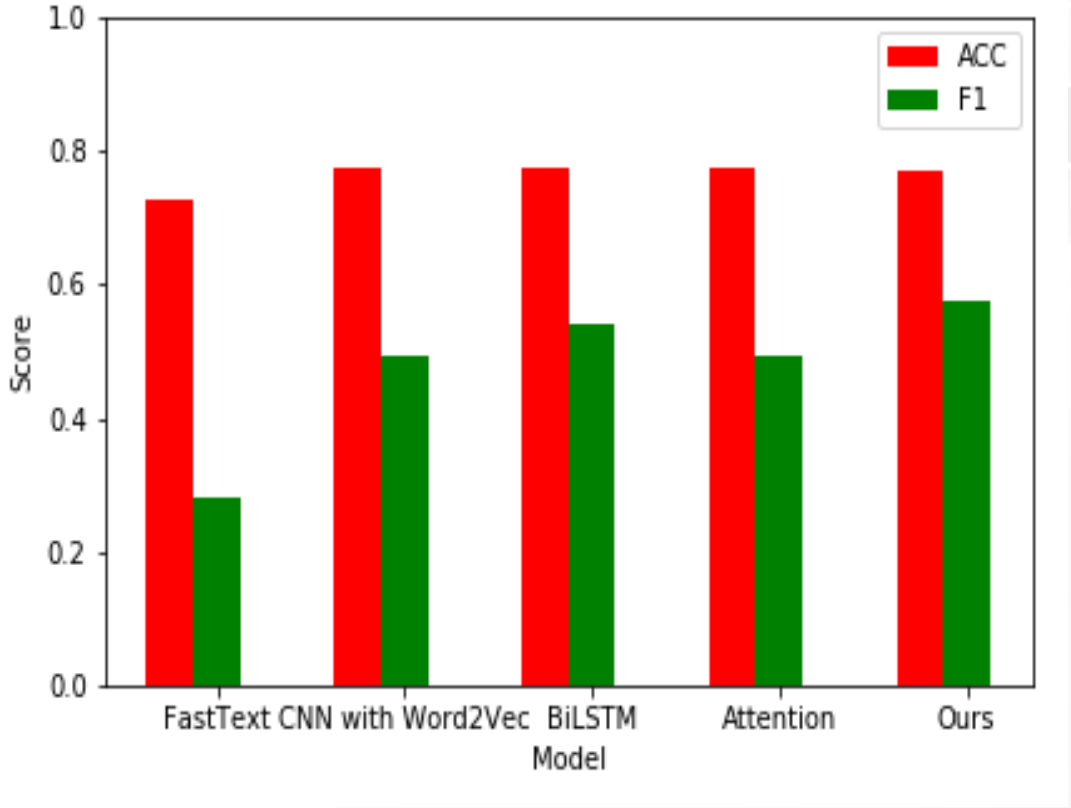


训练效果示意图

2.6 准确率、F1-Score

$$\boldsymbol{Accuracy} = \frac{\text{“预测正确”的样本数}}{\text{总样本数}}$$

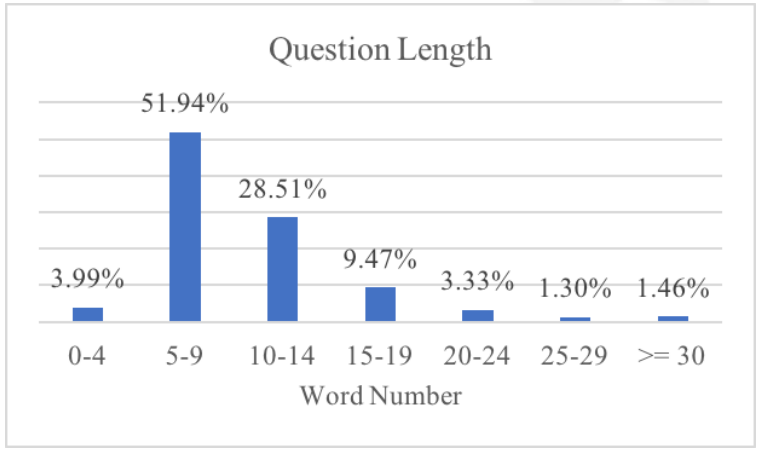
$$\boldsymbol{F1-score} = \frac{\text{“预测标签为1且真实标签也为1”的样本数}}{\text{“标签为1”的样本数} + \text{“真实标签为1”的样本数}}$$

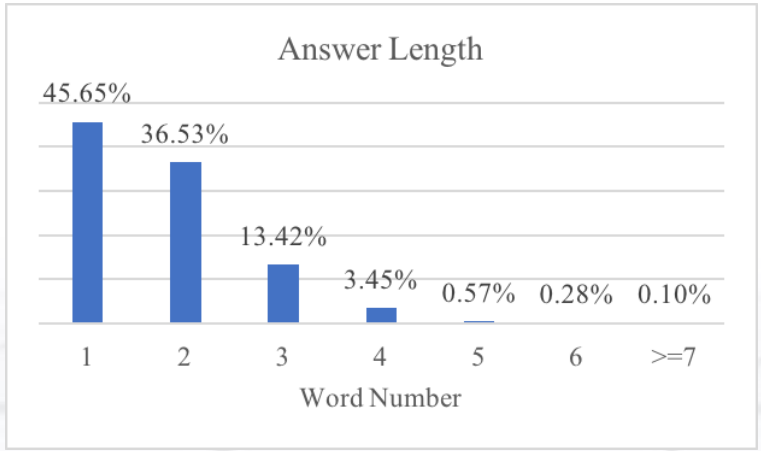


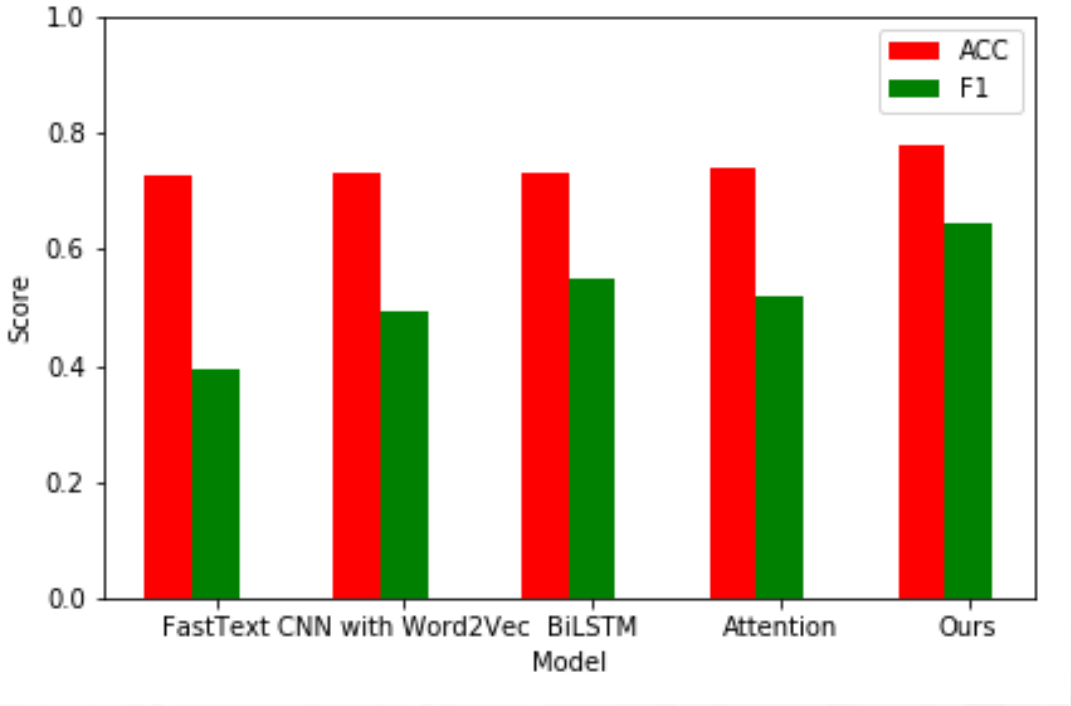
模型评估效果示意图

2.7 泛化能力

泛化实验使用百度开源数据集 WebQA(https://spaces.ac.cn/archives/4338)
该数据集含有 44万条问答记录，主要数据来源于百度知道等问答社区。



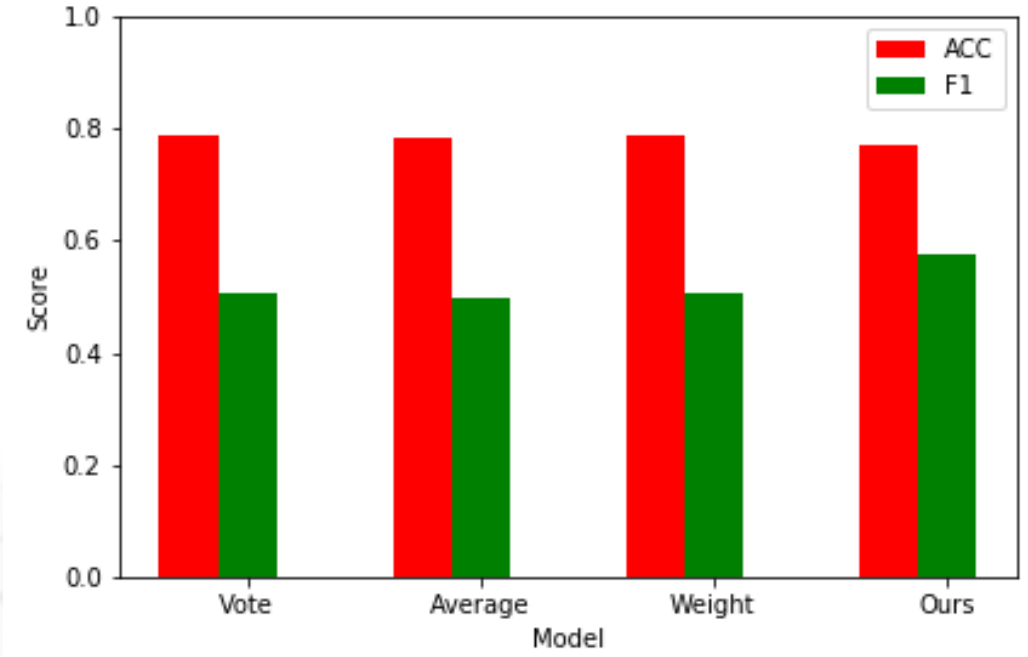




模型泛化效果示意图

2.8 模型融合

投票	平均 (1：1：1：1：1)	权重 (1：2：2：2：3)
即少数服从多数原则，分类得票数超过一半的作为预测结果。	将所有预测结果相加取平均值。	将所有预测结果按照权重计算。



模型融合效果示意图

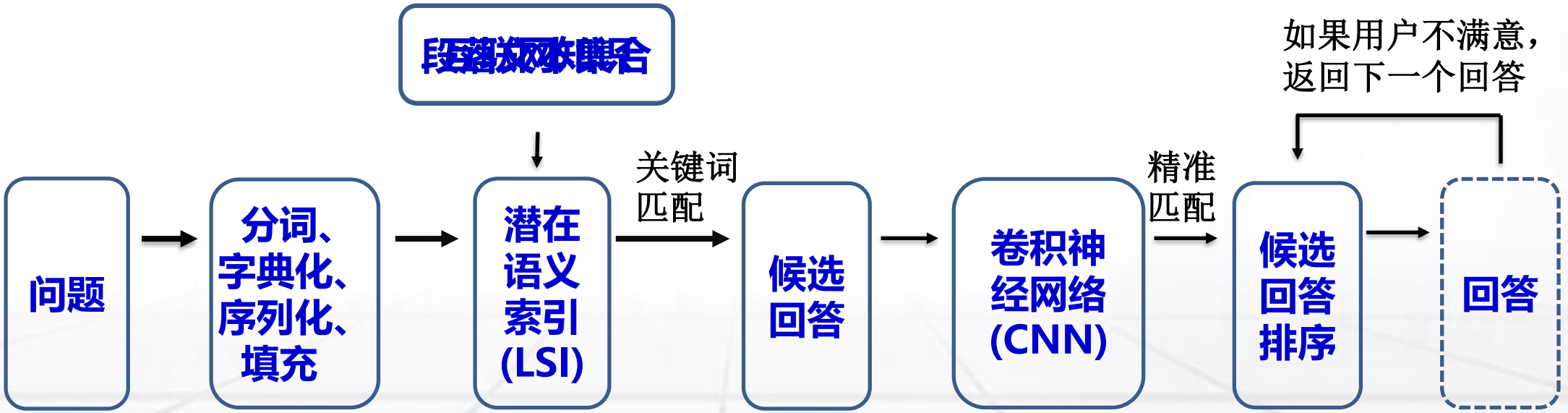
目录

Contents

3.1 总结与展望

未来构建基于Web的深度智能阅读系统

基于潜在语义索引和卷积神经网络的智能阅读模型，以开放的互联网上的 Web 知识作为问答系统的知识来源，从搜索引擎返回的相关网页片段从中抽取出用户所提问题的答案。



3.1 总结与展望


hello~~~我是你的专属宝宝！你想和我说什么？
郭靖的爱人
黄蓉
江南七怪分别是谁
1.飞天蝙蝠——柯镇恶 2.秒手书生----朱聪 3.马王神——韩宝驹 4.闹市侠隐——全金发 5.笑秘陀——张阿生 6.南山樵子——南希仁 7.越女剑——韩小莹
射雕英雄传的作者
金庸，原名查良镛，1924年3月10日生于浙江省海宁市，1948年移居香港。当代知名武侠小说作家、新闻学家、企业家、政治评论家、社会活动家，“香港四大才子”之一。1944年考入重庆中央政治大学外交系。1946年秋，金庸进入上海《大公报》任国际电讯翻译。1948年，毕业于上海东吴大学法学院。1952年调入《新晚报》编辑副刊，并写出《绝代佳人》、《兰花花》等电影剧本。1959年，金庸等人于香港创办《明报》。1985年起，历任香港特别行政区基本法起草委员会委员、政治体制小组负责人之一，基本法咨询委员会执行委员会委员，以及香港特别行政区筹备委员会委员。2000年，获得大紫荆勋章。2009年9月，被聘为中国作协第七
说点什么吧~~~
发送


Bai du 百度
郭靖的爱人
百度一下
网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»
百度为您找到相关结果约822,000个
搜索工具
郭靖配偶：
黄蓉
黄蓉，是《射雕英雄传》的女主角和《神雕侠侣》的重要角色。桃花岛主“东邪”黄药师与冯衡的独生女，精通父亲传授的桃花岛武功、五行八卦阵和奇门遁甲之术。黄蓉集天地... 详情>>
来自百度百科 | 报错


Bai du 百度
江南七怪分别是谁
百度一下
网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»
百度为您找到相关结果约81,900个
搜索工具
问 江南七怪分别是谁：
江南七怪都是谁_百度知道
江南七怪飞天蝙蝠:柯镇恶 妙手书生:朱聪 马王神:韩宝驹 南山樵子:南希仁 笑弥陀:张阿生 闹市侠隐...
来自百度知道 | 1个回答 | 2013-08-25
江南七怪分别是谁_百度知道
柯镇恶,朱聪,韩宝驹,南希仁,张阿生,全金发,韩小莹
来自百度知道 | 1个回答 | 2012-06-01
江南七怪的武功分别是什么?_百度知道
飞天蝙蝠:柯镇恶兵器为降魔杖,暗器为毒菱,有"降魔杖法" 妙手书生:朱聪使铁扇,"妙手空空"的绝技百无一失,在大漠自己研究了一套"分筋错骨手"的功夫 马王神:韩宝...
来自百度知道 | 1个回答 | 2016-02-07

非常感谢各位评委、专家！

任何问题将会是对我们的帮助和改进

答辩小组：83137C