**知识库增强的双向注意力流模型**

张诗禹1

1. 浙江大学 计算机科学与技术学院，浙江 杭州 310000）

**摘要：**机器阅读理解是自然语言处理的关键技术，是通往通用人工智能的必经之路。真实情况下的阅读理解文本通常是多篇多段落的，需要对文章进行筛选，并且对每篇文章产生的答案进行甄别，这给传统阅读理解模型造成了极大的挑战。为此，该文借助知识库，选出最可能包含答案的段落，并且在流行的双向注意力流模型中融合文章排序学习和多答案投票机制，很好的克服了上述问题。该文提出的模型在2018机器阅读理解竞赛中取得了第十名的成绩，测试集的Rouge-L分数53.59，远高于基线系统的44.57。

**关键词：**机器阅读理解；知识库；文章排序；多答案投票

**Knowledge Base Enhanced Bi-directional Attention Flow**

Shiyu Zhang1

1. College of Computer Science and Technology, Zhejiang University, Zhejiang, Hangzhou 310000, China)

**Abstract:** Machine reading comprehension is the key technology of natural language processing and the crucial step to general artificial intelligence. The reading comprehension text under real conditions is usually multi-paragraph, which requires screening of passages and answers generated by each passage. This poses a great challenge to the traditional reading comprehension model. For this reason, this paper uses the knowledge base to select the paragraph that is most likely to contain answers, and integrates passage ranking learning and multi-answer voting mechanism in the popular bi-directional attention flow model, which overcomes these problems well. The model proposed in this paper achieved the 10th place in the 2018 Machine Reading Comprehension Contest and the Rouge-L score of the test set was 53.59, which was much higher than the 44.57 of the baseline system.

**Key words:** machine reading comprehension; knowledge base; passage ranking; multi-answer voting

**0 引言**

机器阅读理解使得机器能够根据问题，分析和理解相关文本，并且生成答案。近年来，很多模型（BiDAF[1]，R-Net[2]）被提出来解决机器阅读理解问题，并且在SQuAD[3]数据集上取得优异的成绩，甚至超过了人类。但是这些模型都只针对单篇候选文章，并且每篇文章只有一个段落。

微软提出的MS MARCO[4]数据集进一步把候选文章扩大到多篇，这使得数据集更加接近真实场景，很多新的模型尝试也相继出现。其中代表性的S-Net[5]是在R-Net基础上增加了文章排序的多任务学习模块，使得模型能够利用多篇文章的信息共同学习参数，百度提出的V-Net[6]采用了多文章的交叉验证方法来做多篇文章的答案生成，猿辅导提出的MARS模型则采用了多答案投票的后处理方法来甄别最终答案。

2018机器阅读理解竞赛的数据集DuReader[7]则完全模拟真实环境，从实际搜索引擎的日志中提取数据，每个问题对应多篇文章，并且每篇文章有多段，这是至今最为复杂的机器阅读理解数据集，这也为机器阅读理解模型提出了新的挑战。

因此，本文提出知识库增强的双向注意力流模型，通过知识库抽取出最有可能包含答案的段落，然后将S-Net中的文章排序融合进BiDAF模型中，最后对答案进行MARS所采用的多答案投票，得到最终的答案。这一模型在DuReader的测试集上取得Rouge-L 得分53.59，获得了第十名的成绩。

**1 知识库增强的双向注意力流模型**

* 1. **问题描述**

对于给定问题*q*及其对应的文本形式的候选文档集合，阅读理解系统需要能够自动对问题及候选文档进行分析，输出能够满足问题的文本答案*a*，目标是*a*能够正确、完整、简洁地回答问题*q*。

**1.2 预处理**

对每一个例子的问题和文章做分词和词性标注。然后根据参考答案在文章中找出召回率最大的片段作为训练目标。对于段落的选择，首先建立知识库，知识库主要包括段落模式和问题-答案词的配对关系。段落模式例如“1..2..3..”，问题-答案词的配对关系例如，问题包含“时间”，“时候”，答案则要包括“年”“月”“日”“星期”等。这些知识库都是建立在人的知识和一些常识之上，比如段落包括“1..2..3..”这种列表模式时，则很有可能是对于答案的枚举或者步骤的说明，极有可能包含答案。然后如果问题是问时间相关的即包含“时间”“时候”类似的词时，那么包含“年”“月”“日”“星期”这些词的段落则极有可能在说明一个时间，那么就很有可能成为答案。进行上述处理后，最后输入模型的就是每篇文章最有可能包含答案的段落。

**1.3 网络模型**

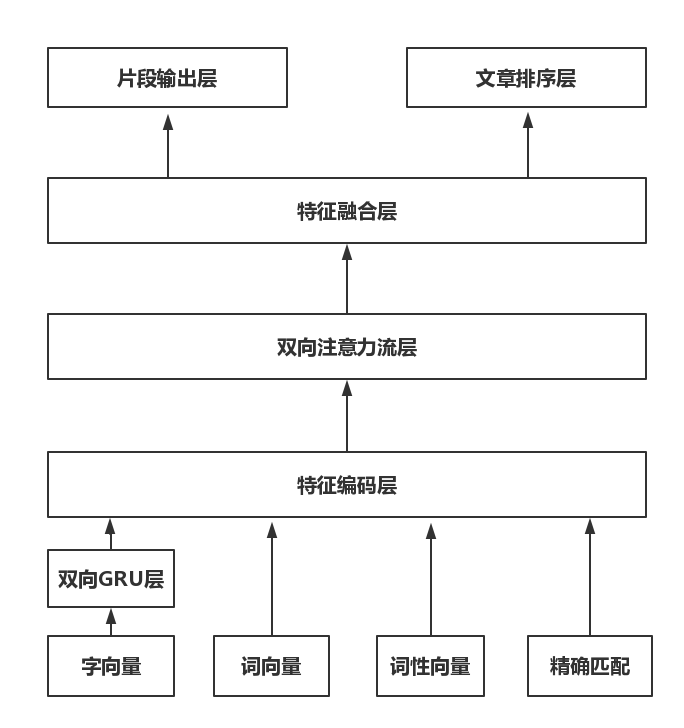


图1 模型架构

模型的顶层架构如图1所示。

1.3.1 特征编码层

模型输入包含字向量，词向量，词性向量和精确匹配位。其中字向量和词向量都用预先训练好的，字向量通过一个双向GRU得到相应词的表示，词性向量随机初始化，并且不通过模型进行训练。精确匹配位表示，如果有一个词在问题中出现过则置为1，否则为0。将以上所有特征向量连接起来作为特征向量，再经过一个双向LSTM层的编码后，文章和问题的特征向量***C***，***Q***流向网络的下一层。

1.3.2 双向注意力流层

这里采用BiDAF[1]中的设计，假设文章和问题中词数分别为*T*和*J*。我们计算两个方向上的注意力流：从文章到问题和从问题到文章。这两个注意力流都是从***C***，***Q***的相似矩阵中得到，这里表示文章中第*t*个词和问题中第*j*个词的相似度,这里相似矩阵。

从文章到问题的注意力***U***表示对于文章中的每一个词，问题中哪一个词与之最相关，表示如下：

从问题到文章的注意力***H***表示哪一个文章中的词对问题中某个词的相似度最大，表示如下：

*h*重复*T*次得到***H***。

最后本层的输出***G***表示如下：

其中[;]表示连接操作，表示逐位相乘。

1.3.3 特征融合层

该层又是一个双向LSTM层，把***G***进行再次编码，这次编码使得文章和问题中的信息得到融合，得到输出***M***。

1.3.4 片段输出层

该层采用Pointer Network预测片段起始和终止位置。起始位置通过一个softmax得到各个词的概率：

为了预测终止位置，把***M***经过一个双向LSTM层得到***M***2，然后以相同方式得到终止词的概率：

其中都是可以训练的权重向量。

1.3.5 文章排序层

这里参考S-Net[5]中的设计。我们首先把问题***Q***经过两个全连接层后再用softmax得到每个词的权重，然后计算问题中所有词的加权和，作为问题的整体表示*rQ*，计算过程如下：

把文章中每个词和*rQ*进行融合，以同样方式计算文章的整体表示*rP*：

最后再把*rP*和*rQ*结合起来经过两个全连接层得到最终每篇文章的分数：

1.3.6 损失函数

片段输出层和文章排序层都采用交叉熵作为损失函数，然后将两个损失进行加权求和作为最终模型训练的损失函数。

**1.4 后处理**

网络输出每个词作为起始和终止的概率，选择最大的词对作为最终抽取答案的起始和终止位置。抽取出答案后，首先保证所有标点都是中文标点，然后剔除答案开头的标点符号，接着由于数据中很多都是从百度经验里提取的答案，所以剔除“步骤阅读”字样，最后进行多答案投票选择最终的答案。得到最终答案后，同样建立一个知识库，来判断YES\_NO答案的态度。

1.4.1 多答案投票

这里参考MARS模型里的做法。一般正确的答案会是大多数文章所共同认可的，所以可以通过投票的方法来选出每篇文章所认为的其他文章生成答案正确的分数，最后选取最高得分的答案作为最终答案。

这里的投票机制描述如下：

其中f()和g()表示平滑函数，这里分别采用四次方根和对数函数。

最终分数由两部分组成，其中span\_score就是答案起始词概率和终止词概率的乘积，vote\_score表示互相投票的得分，具体指将每个文章所生成答案和其他文章的答案计算rouge相似度，然后乘以那篇文章的span\_score，再求和得到文章答案的投票得分，最终得分是平滑后span\_score和vote\_score的乘积。

**2 实验**

**2.1 实验设置**

本文在DuReader数据集上进行测试，字向量和词向量用FastText[[1]](#footnote-1)在中文维基语料[[2]](#footnote-2)上训练得到，分词使用pyltp[[3]](#footnote-3)工具包。实验环境为debian，使用tensorflow框架，显卡为Titan XP。最终的模型为single model。

**2.2 实验参数**

字向量和词向量维度设为300，词性向量维度设为10，字向量通过的LSTM隐藏层长度设为100，其他LSTM的隐藏层长度设为150，损失函数中文章排序的比重为20%。采用adam作为优化方法，初始学习率为0.01，每经过两个epoch，学习率缩小为1/10，一般训练4-6个epoch即可。输入文章长度固定为500词，问题为60词，答案最长为500词，每个词包含4个字，不足的全部填充0，Batch大小设为32。

**2.3 实验结果与分析**

模型在DuReader的测试集上结果为BLEU-4得分47.21，ROUGE-L得分53.59，在验证集上不用多答案投票，直接选span\_score最高者为答案的结果为BLEU-4得分42.46,ROUGE-L得分50.17。

模型加上字向量，词向量，词性向量和精确匹配ROUGE-L能够提高2-3个点，加上文章排序可以提高0.5个点左右，加上基于知识库的段落筛选可以提高6-7个点，加上多答案投票可以提高1个点左右，加上基于知识库的YES\_NO态度提取可以提高1个点左右，加上中文标点处理，去除开头标点和去除“步骤阅读”字样可以提高1个点左右。

**3 结论**

本文的主要贡献是建立了知识库来筛选文章段落，并且借鉴文章排序和多答案投票等模块，加入BiDAF使得最终的模型能够很好的处理多篇多段落文章的机器阅读理解问题，并且取得了2018机器阅读理解竞赛第十名的成绩。

未来首先可以采用CoVe[8]和ELMo[9]等迁移学习方法，使得模型输入的信息更加丰富，从而能够捕捉到更多有用的信息去提取答案。然后可以思考更加通用的算法去提取文本的特征模式，自动构建知识库。最后可以考虑使用监督学习的方法，构建网络模型训练段落的提取器。

**参考文献**

[1] Seo M, Kembhavi A, Farhadi A, et al. Bidirectional Attention Flow for Machine Comprehension[J]. 2016.

[2] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017c. Gated self-matching networks for reading comprehension and question answering. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 – August4, Volume 1: Long Papers.

[3] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In Proceedings ofthe 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016.

[4] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In Proceedingsof the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 colocated with the 30th Annual Conference on NeuralInformation Processing Systems (NIPS 2016).

[5]Chuanqi Tan, Furu Wei, Nan Yang, Weifeng Lv, and Ming Zhou. 2017. S-net: From answer extraction to answer generation for machine reading comprehension. arXiv preprint arXiv:1706.04815 .

[6] Yizhong Wang, Kai Liu, Jing Liu, Wei He, Yajuan Lyu, Hua Wu, Sujian Li, Haifeng Wang. Multi-Passage Machine Reading Comprehension with Cross-Passage Answer Verification. arXiv:1805.02220.

[7] Wei He, Kai Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2017. Dureader: a chinese machine reading comprehension dataset from real-world applications. arXivpreprint arXiv:1711.05073 .

[8] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In NIPS 2017.

[9] Peters, Matthew E. and Neumann, Mark and Iyyer, Mohit and Gardner, Matt and Clark, Christopher and Lee, Kenton and Zettlemoyer, Luke. 2018. Deep contextualized word representations. Proc. of NAACL 2018.

1. https://fasttext.cc/ [↑](#footnote-ref-1)
2. https://dumps.wikimedia.org/zhwiki/latest/zhwiki-latest-pages-articles.xml.bz2 [↑](#footnote-ref-2)
3. http://pyltp.readthedocs.io/zh\_CN/develop/api.html [↑](#footnote-ref-3)