阅读理解组项目报告

组员：李冬，韩玮光，郭天翼，田文雨

2018.7.27

大纲

机器阅读理解

机器阅读理解技术作为一项能够让机器阅读文章并回答与阅读内容相关问题的自然语言理解方面的基础技术，能够有效地支撑移动智能助手（如小爱同学）、智能问答系统等人工智能应用。

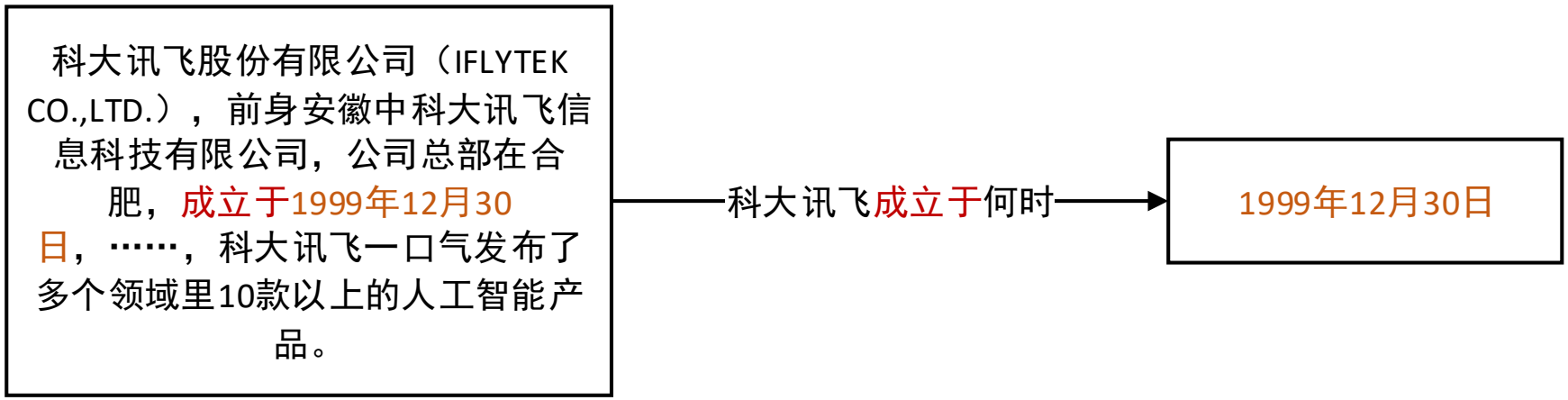


图: 机器阅读理解的一个示例

解决方案

采用解决方案：基于循环神经网络与注意力机制的答案边界预测模型。
通过门控循环单元与指针网络对候选答案的起始与结束边界进行预测。

大纲

整体架构

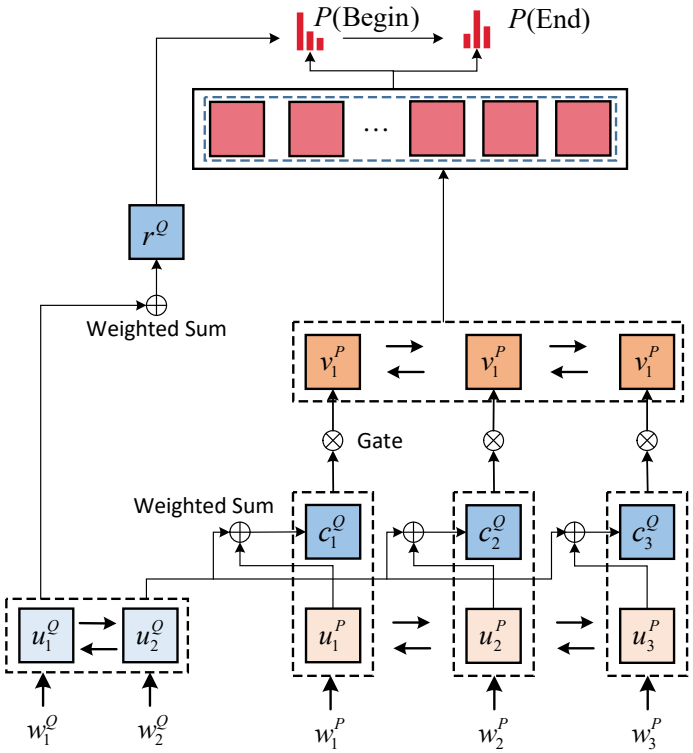


图: 模型的整体架构图

具体方法

生成基本词嵌入

$$u_t^Q = \text{BiGRU}_Q(u_{t-1}^Q, [e_t^Q, \text{char}_t^Q])$$
$$u_t^P = \text{BiGRU}_P(u_{t-1}^P, [e_t^P, \text{char}_t^P])$$

注意力机制

$$s_j^t = v^\top \tanh(W_u^Q u_j^Q + W_u^P u_t^P)$$
$$a_i^t = \exp(s_i^t) / \sum_{j=1}^{m} \exp(s_j^t)$$
$$c_t^Q = \sum_{i=1}^{m} a_i^t u_i^Q$$

生成问题-文本相关嵌入：

$$g_t = \text{sigmoid}(W_g[u_t^P, c_t^Q])$$
$$[u_t^P, c_t^Q]* = g_t \odot [u_t^P, c_t^Q]$$
$$v_t^P = \text{GRU}(v_{t-1}^P, [u_t^P, c_t^Q]*)$$

具体方法

初始化答案预测网络：

$$s_j = v^\top \tanh(W_u^Q u_j^Q + W_v^Q V_r^Q)$$
$$a_i = \exp(s_i) / \sum_{j=1}^{m} \exp(s_j)$$
$$r^Q = \sum_{i=1}^{m} a_i u_i^Q$$

预测起始与结束位置：

$$s_j^t = v^\top \tanh(W_h^P v_j^Q + W_h^a h_{t-1}^a)$$
$$a^t = \exp(s_i^t) / \sum_{j=1}^{m} \exp(s_j^t)$$
$$c_t = \sum_{i=1}^{N} a_i^t v_i^P$$
$$h_t^a = \mathrm{GRU}(h_{t-1}^a, c_t s)$$
$$p^t = \arg\max_N(a_1^t, \ldots, a_N^t)$$

大纲

模型训练

DuReader数据集的数据来源于网络。训练集包含201574个样例，每个样例包含一个问题，多篇短文。开发集包含10000个样例。

使用讯飞数据集进行训练，讯飞数据集的所有数据均来源于中文维基百科。训练集包含2403个样例，每个样例包含一篇短文，多个问题，总共包含10142个问题，每个问题只有一个答案。开发集包含848篇短文，3219个问题，每个问题有三个答案。

模型基础

DuReader数据集的baseline：

- https://github.com/baidu/DuReader

S-Net：

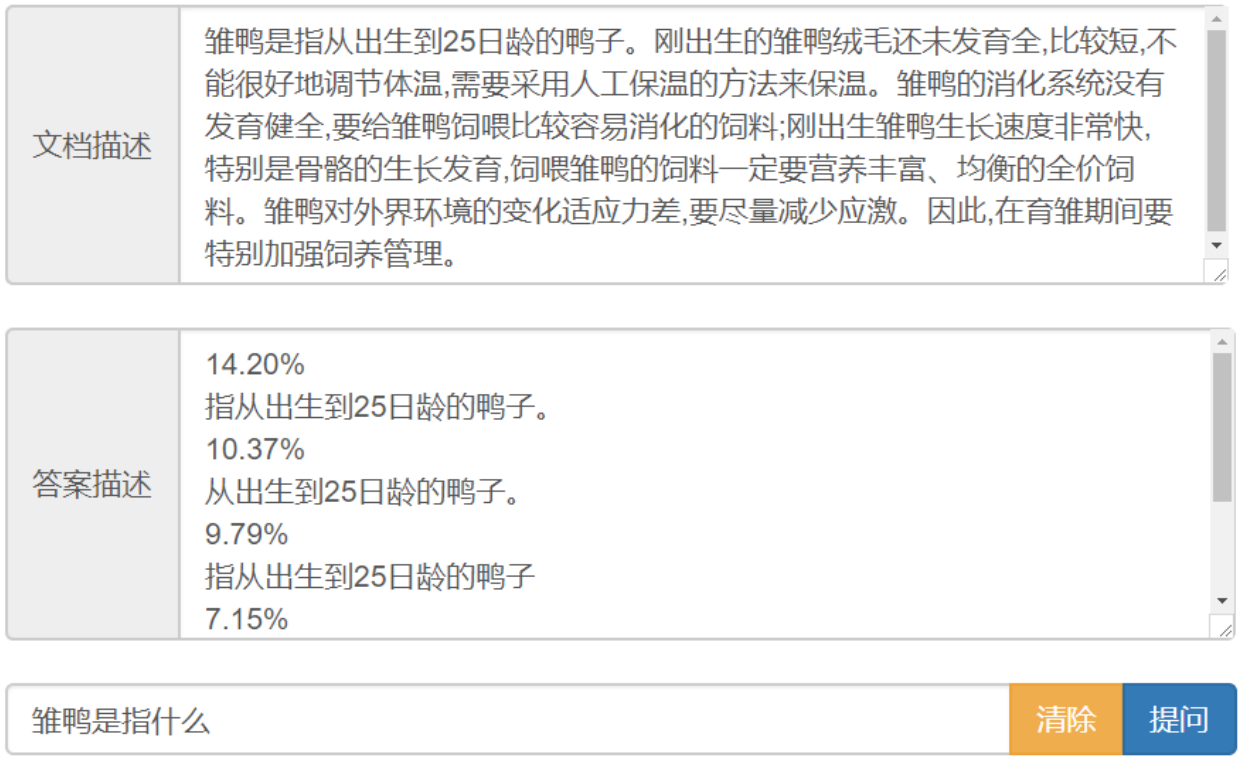
- https://arxiv.org/abs/1706.04815

2017年讯飞杯比赛代码

大纲

项目进展

模型现在基本可以回答大部分基本问题，对某些没有直接对应的答案的问题也有较好的回答效果：



效果

文档描述	科大讯飞股份有限公司（IFLYTEK CO.,LTD.），前身安徽中科大讯飞信息科技有限公司，成立于1999年12月30日，2014年4月18日变更为科大讯飞股份有限公司 [1-2] ，专业从事智能语音及语言技术研究、软件及芯片产品开发、语音信息服务及电子政务系统集成 [3] 。拥有灵犀语音助手 [4-5] ，讯飞输入法 [6] 等优秀产品。 科大讯飞信息科技股份有限公司现任董事长兼总裁为刘庆峰先生。科大讯飞信息科技股份有限公司是一家专业从事智能语音及语音技术研究、软件及芯片产品开发、语音信息服务的国家级骨干软件企业，主要股东包括：中国移动、中科大资产经营有限公司、上海广信、联想投资、盈富泰克等。语音技术实现了人机语音交互，使人与机器之间沟通变得像人与人沟通一样简单。语音技术主要包括语音合成和语音识别两项关键技术。让机器说话，用的是语音合成技术；让机器听懂人说话，用的是语音识别技术。此外，语音技术还包括语音编码、音色转换、口语评测、语音消噪和增强等技术，有着广阔应用空间。2017年6月，入选《麻省理工科技评论》2017 年度全球 50 大最聪明公司"榜单。 [7] 2017年11月9日，科大讯飞年度发布会在北京召开。从教育到医疗，从客服到智能家居，再到移动手机端和车载环境，在2017年的年度发布会上，科大讯飞一口气发布了多个领域里10款以上的人工智能产品。
答案描述	15.37% 2017年6月，入选《麻省理工科技评论》2017 年度全球 50 大最聪明公司"榜单。 11.93% 2017年6月，入选《麻省理工科技评论》2017 年度全球 50 大最聪明公司" 10.94% 2017年6月，入选《麻省理工科技评论》2017 年度全球 50 大最聪明公司"榜单 7.86% 入选《麻省理工科技评论》2017 年度全球 50 大最聪明公司"榜单。 6.10% 入选《麻省理工科技评论》2017 年度全球 50 大最聪明公司"

科大讯飞取得了哪些成就 清除 提问

效果

对于复杂的问题与答案，模型仍然存在一些问题，难以抽取到正确的答案

文档描述	七大洲指地球陆地分成的七大陆地板块，包括亚洲（全称亚细亚洲）（Asia）、欧洲（全称欧罗巴洲）（Europe）、北美洲（全称北亚美利加洲）（North America）、南美洲（全称南亚美利加洲）（South America）非洲（全称阿非利加洲）（Africa）、大洋洲（Oceania）、南极洲（Antarctica）。
答案描述	4.63% 亚洲（全称亚细亚洲） 3.83% 亚洲（全称亚细亚洲）（Asia）、欧洲（全称欧罗巴洲）（Europe）、北美洲（全称北亚美利加洲）（North America）、南美洲（全称南亚美利加洲）（South America）非洲（全称阿非利加洲）（Africa）、大洋洲（Oceania）、南极洲（Antarctica）。 3.40% 亚洲（全称亚细亚洲）（Asia）、欧洲（全称欧罗巴洲） 3.29% 亚洲（全称亚细亚洲）（Asia）、欧洲（全称欧罗巴洲）（Europe）、北美洲（全称北亚美利加洲）（North America）、南美洲（全称南亚美利加洲）（South America）非洲（全称阿非利加洲）（Africa）、大洋洲（Oceania）、南极洲（Antarctica）

七大洲包括什么 清除 提问

大纲

未来工作

阅读理解组未来的工作是：

- 尝试解决答案较长的情况下的抽取问题；
- 尝试使用生成式模型来生成答案