Delta Text Analysis
机器阅读理解技术竞赛报告

Yiying Tseng 曾偳穎,
Yuting Lai 賴郁婷,
Chih-Chieh Shao 邵志杰,
Pocheng Lin 林柏誠,
Vincent Hsiao 蕭瑞辰

台达电子 台达研究院

July 28th, 2018

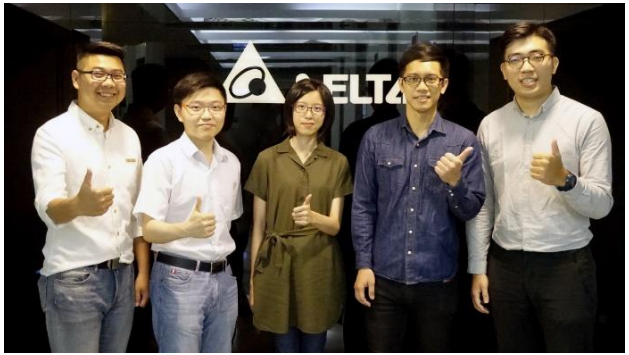
报告大纲

- 台达团队介绍
- 阅读理解系统
- 观察与讨论
- 结语

台达团队介绍

- 台达电子集团(Delta)由郑崇华先生创立于 1971 年，为全球电源管理与散热解决方案的领导厂商。深耕「**电源及零组件**」、「**自动化**」与「**基础设施**」三大业务范畴。
- 台达研究院(Delta Research Center)成立于2013年，致力于**大数据分析及物联网应用**、**前瞻技术加速企业转型**和**产官学研生态体系协作**。大数据分析包含图像处理、语音处理、文字分析及数值分析等研究主题。
- 文字分析主要研究项目: **NLP基础研究**、**智能机器人**、**知识问答系统**、**阅读理解**。
- 台达阅读理解数据集 (Delta Reading Comprehension Dataset, DRCD) 属于通用领域中文机器阅读理解数据集。

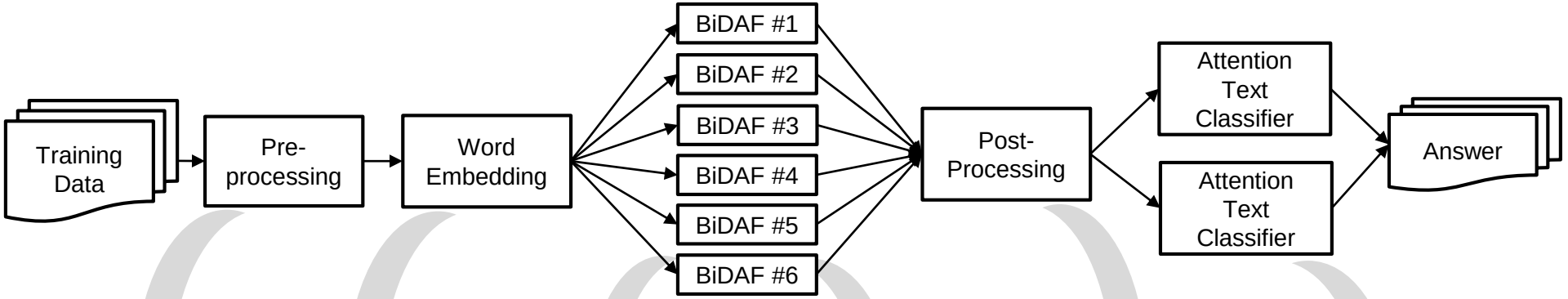
台达电子台北总部



台达电子参赛成员

台达阅读理解数据集

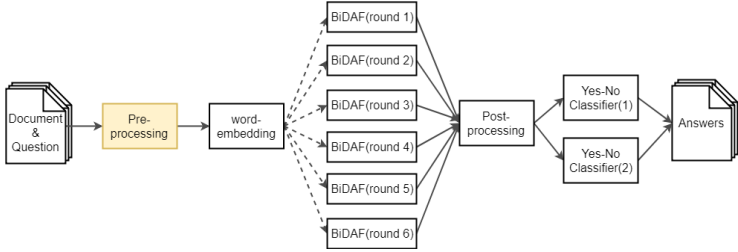
阅读理解系统



1	2	3	4	5	6
扩充数据	预训练词向量	改良预测代码	集成模型	标点正规化	是非题答案分类
+0.46% (ROUGE-L)	+3.93%	+7.23%	+0.96%	+0.16%	+0.42%

資料來源:Delta Text Analysis阅读理解技术竞赛系统报告

Method (1/5)



1. 数据扩展

– 每一个问题有多个人工产生的参考答案，基线系统只采用第一个参考答案，产生一笔数据；模型使用文本区间位置作为表示，但参考答案不一定在段落当中，容易产生错误的训练资料。

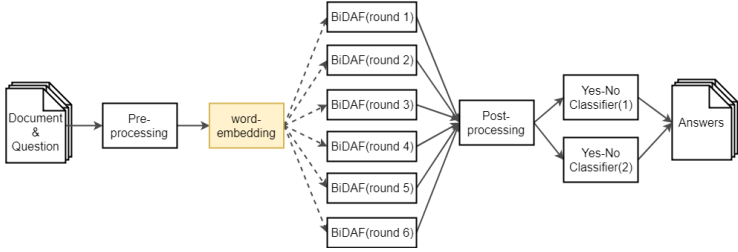
- [**扩增**]妥善利用所有资源，扩增**多个参考答案**做为训练数据。
- [**筛选**]保持训练数据的质量，滤除**相似度分数低于0.7**的训练数据。

Question	**绣眉有哪几种?**
Question Type	**实体(Entity)**
Human Answers	**绣眉的五种基本方法：1、雕润眉；2、平面绣眉；3、点状绣眉；4、立体绣眉；5、仿真立体绣眉。**
Training Answers	**。5、仿真立体绣眉仿真立体绣眉是比较流行的绣眉**

Score < 0.7
不采用

– 扩展DuReader数据集，基于时间与硬件的限制，最终使用**347,723**笔训练数据。

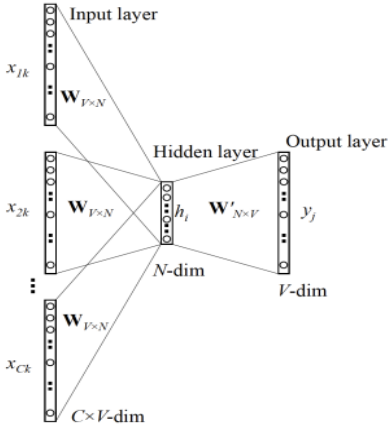
Method (2/5)



2. Word Embedding

– 使用DuReader 数据集进行词向量训练，透过FastText算法，学习词汇之间**语意关系**。

- 输入层利用**上下文信息**并同时考虑**子词信息**，强化字词的关系，解决OOV现象。
- 输出层利用**分层式 Softmax**的方式，大大降低训练时间。



Continuous Bag-of-Words Model
(T Mikolov, 2013)

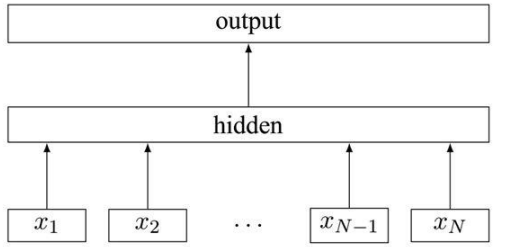
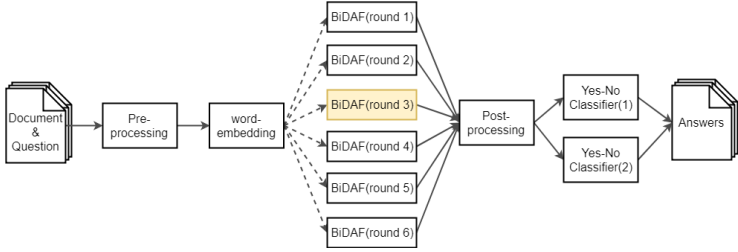


Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.

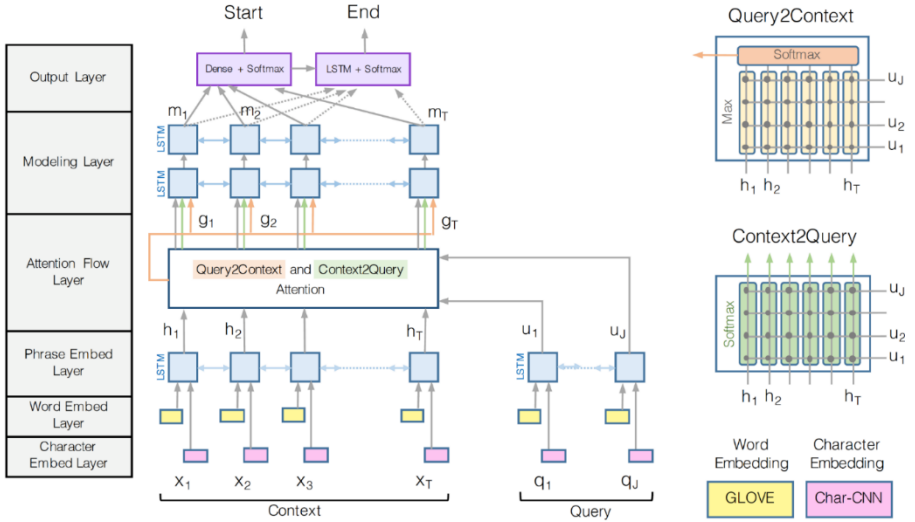
FastText Model
(A Joulin, 2016)

Method (3/5)



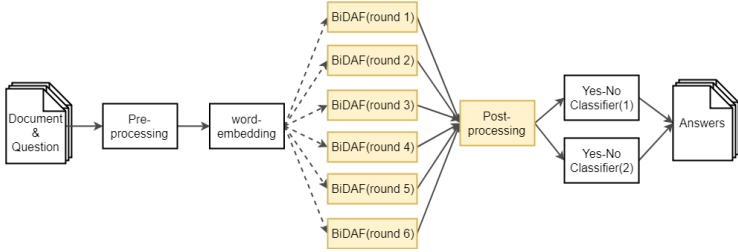
3. BiDAF 答案预测

- 改良DuReader提供的Bi-Directional Attention Flow(BiDAF)基线模型
  - [**发现**]系统多挑选短句及复述问句作为答案，预处理筛选**最相关段落**进行预测。
  - [**做法**] **全文串接**，将文章段落以**句号串接**起来，以**整篇**来预测答案。
- 改善幅度最为明显，ROUGE-L 提升7.23%



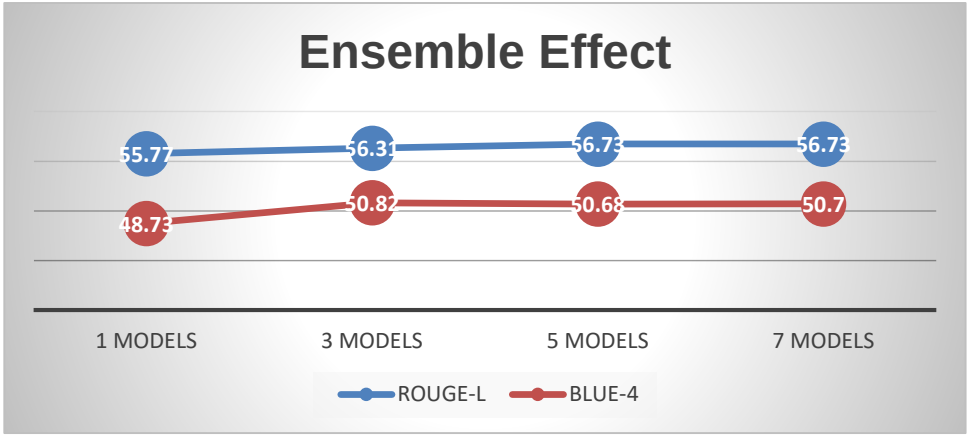
Bi-Directional Attention Flow Model (M Seo, 2016)

Method (4/5)



4. 集成式模型

- 相同参数配置所训练的模型，会因为神经网络的随机特性带来差异。
- 改写答案输出格式，以**机率分布**呈现，将多个模型做**加总平均**。

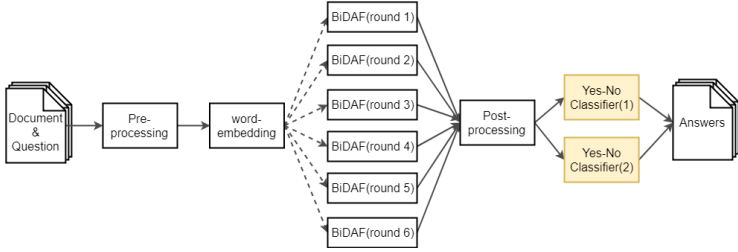


集成式模型实验数据

5. 标点符号正规化

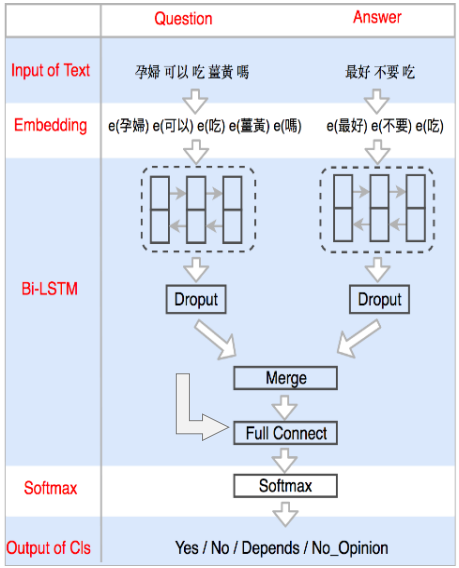
- 答案**移除**多余标点，精简答案。
- 答案**填补**缺失句号，使答案更加完整。

Method (5/5)

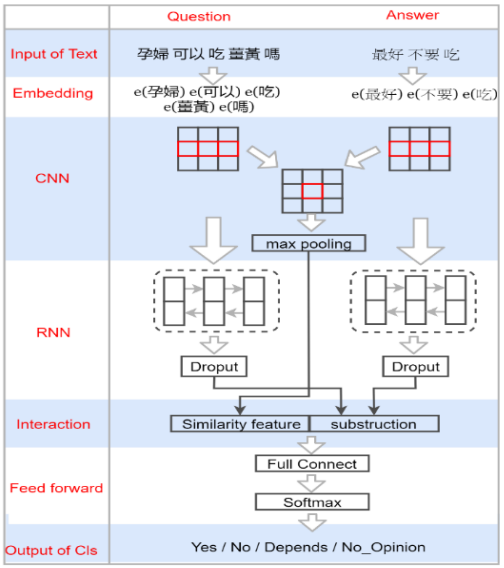


6. 是非題答案分类

– 将问题与模型预测的答案配对，产生QA Pair，基于**注意力机制**与**相似度机制**设计两套分类模型，将两个分数做加总平均，在 test1 Set 的分类结果，正确率可达**72%**。



注意力机制模型



相似度机制模型

观察与讨论 (1/2)

1. 数据集特性迥异。我们认为可以分开训练，透过**数据扩增**或**多任务学习**方式，降低数据不足的影响。

问题类型	实体	是非	描述
数据数量	76k	24k	170k

数据集的问题类型统计信息

2. 实体类型问题，答案由文本分散的字词组成，我们认为可以采用**生成答案**或**组合多个答案**。

问题	2017有什么好看的小说
段落	1.《将夜》 作者:猫腻(起点白金作家) 简介:与天斗,其乐无穷。 故事概要: 主角宁缺带.....2.《择天记》 作者:猫腻too相对将夜,这是一本正在写的新书,值得一看。.....3.《冒牌大英雄》 作者:七十二编。这本书.....
答案	1.《将夜》2.《择天记》3.《冒牌大英雄》4.《无限恐怖》5.《恐怖搞校》6.《大国医》7.《龙魔导》。

实体类型中答案分散的例子

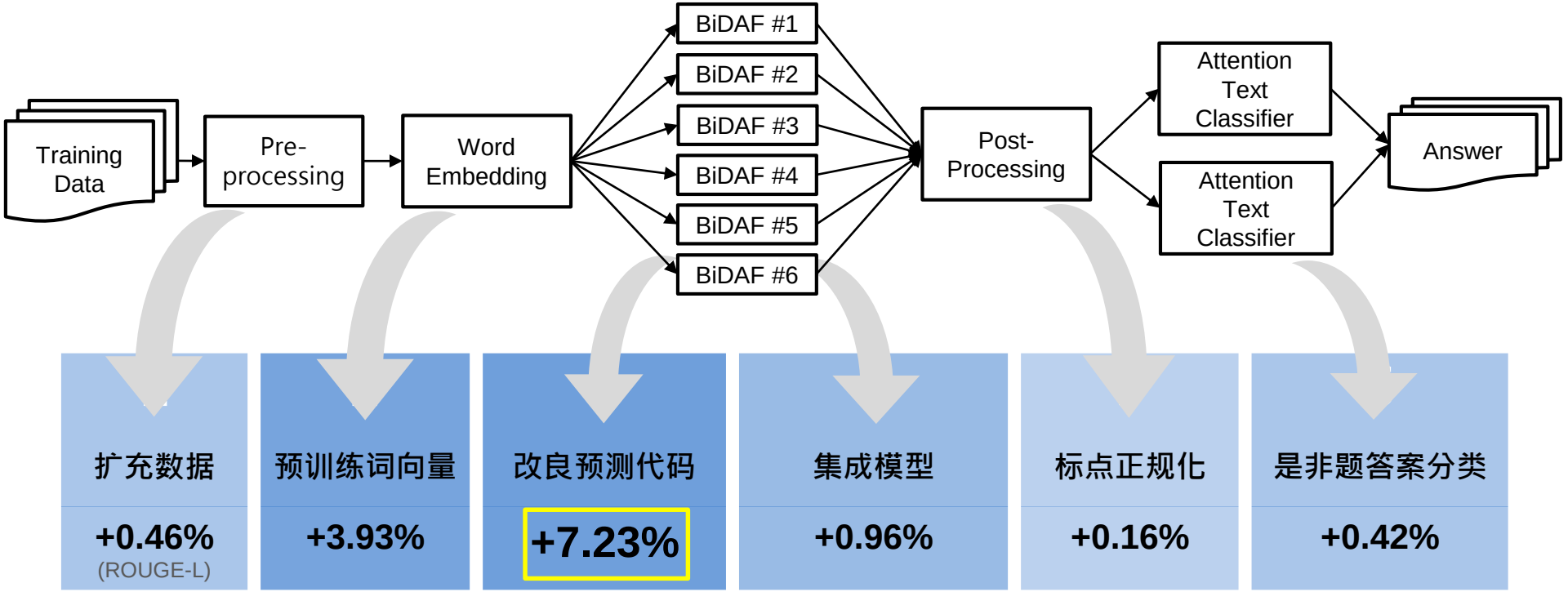
观察与讨论 (2/2)

3. 集成方法采用**统一融合(Uniform blending)**，我们认为**线性融合(Linear blending)**找出更佳的模型权重。
4. 使用**句号串接**并预测全部文章，虽然带来显着提升，也造成一些问题。
   我们认为可以改用**保留符加上后处理**以避免这种情况，或是再设计别的**段落预筛选方法**。

问题1	**怎么清除dnf缓存**
预测答案1	1、打开电脑，进入桌面，鼠标双击“这台电脑”并进入。2、选择要清理的磁盘，如C盘，鼠标右击C盘，点击“属性”。3、进入“属性”窗口后，点击右下角的“磁盘清理”。4、点击“磁盘清理”后，会弹出一个窗口扫描此磁盘下的各种垃圾文件和缓存，如果磁盘文件较多，扫描的时间会有点长一些。
问题2	**遣倦**
预测答案2	是“缱绻”，形容情投意合，难舍难分；缠绵。缱绻。应该这样写。形容情投意合，难舍难分，缠绵。

正/反面例子

結語



本系统在 MRC 2018 的评比中得到 ROUGE-L **56.57%** 与 BLEU-4 **48.03%**

資料來源:Delta Text Analysis阅读理解技术竞赛系统报告

Smarter. Greener. Together.

Q & A

台达阅读理解数据集

To learn more about Delta, please visit www.deltaww.com or scan the QR code

English

Traditional Chinese

Simplified Chinese