

特征:

IT大课堂

▶ 数据预处理

随机shuffle之后按9:1的比例划分训练集和验证集

	平均长度	截断补齐长度
Word	718	2000
Article	1177	3200

截断补齐长度能覆盖95%的训练样本

达观 DATA 文本智能处理专家

模型训练:

IT大课堂

▶ 深度学习模型

- TextCNN
- GRU
- RCNN
- FastText
- Attention-GRU

19维向量

Classifier

CNN/GRU/RCN....

Embedding

文本

达观 DATA 文本智能处理专家

► 单模型结果 & 模型融合

单模型	模型输入	验证集F1值
TextCNN	word/article	0.765/0.737
GRU	word/article	0.773/0.747
RCNN	word/article	0.778/0.743
FastText	word/article	0.754/0.716
Attention-GRU	word/article	0.777/0.738

传统机器学习方法+lgb : 线上0.7875

达观数据
DATA GRAND

文本智能处理专家

训练集：验证集一开始作为调参训练使用，最终把验证集加入了训练集

融合方式：生成的模型的结果进行融合

-----2-----

模型训练&特征：

Summary of our approach



达观数据
DATA GRAND

文本智能处理专家

basemodel(5fold)

- CNN
- GRU/LSTM
- CNN+RNN
- LR
- LGB(XGB太慢懒得弄了)
- SVC
- MLP

DL models

- Embeddings : w2v/GloVe/Fasttext
- Automatic summarization: LexRank/TextRank/HDP...
- Different *MAX_WORDS/MAX_SEQUENCE_LENGTH*
- Flip text
- Different architecture of network

ML models

- BOW
- Automatic summarization:
LexRank/TextRank/HDP...
- n-Gram

模型结果: 样本数*19列矩阵喂给lgbm

投票 $\geq 50\%$

Stacking & Voting

- Stacking with models from 2018.7.18 to 2018.8.7
- Vote with different lgbl attempts

其他:

Failed attempt

- HAN 效果差
- LDA 内存爆
- Pseudo Labeling 过拟合

HAN 0.78 融合总体降低，所以扔掉了

-----3-----

特征：

一、数据处理

1. 无监督词向量训练：GloVe and Word2vec
 - 最小词频2、最大窗口5、词向量维度100
 - 使用train和test一起训练
 - Word2vec 使用gensim库中的skip-gram 模型
 - 在深度模型中，两种词向量拼接起来一起训练效果最好，Word2vec次之
2. 生成训练文件和测试文件，截断长度1200词
 - 从文章开头往后1200词
 - 文章开头取400词，中间400词，结尾400词
 - 词频为1的词被去掉

--莱姆大做平滑

传统特征的选择

- 1、计算词频，过滤掉文档频率80%以上或10文档以下的词语。词频使用对数处理，即词频

$$TF(w_i) = \log(\text{Count}(w_i) + 1)$$

- 2、idf对稀疏词敏感，并不适合用于分类，因此采用新的系数HC

$$\bullet \text{CF}(W_i, C_j) = \frac{TF(W_i, C_j) + \lambda}{\sum_{C_j} TF(W_i, C_j) + \lambda C}$$

$$\bullet \text{HC}(W_i) = -\sum_{C_j} \text{CF}(W_i, C_j) \cdot \log(\text{CF}(W_i, C_j))$$

- 3、SVM线性核：A榜得分：0.7783

卡方20w-svd800

传统特征

- 1、词袋模型的特征维度过高，需要进行降维处理。在本次比赛中，我们使用了卡方检验，通过LSVC的CV结果选出的最佳维度。

- 2、TF-IDF特征没有考虑到文档中的语义信息，因此通过LSI提取语义信息，通过TruncatedSVD实现。

- 3、我们还尝试了LDA主题模型，但是因为LDA的计算耗时大，并且结果也不好，最终没有使用LDA特征。

- 4、考虑文档中的上下文信息，使用doc2vec。

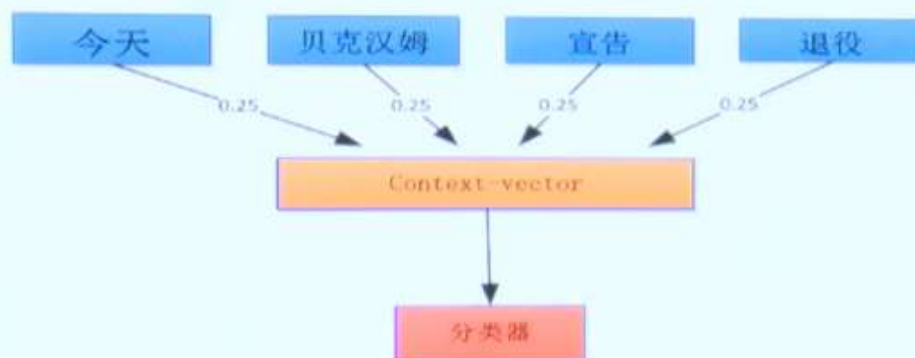
模型：

ML Model

在传统特征上使用Linear_SVM, LR, LGB, XGB模型。

ML Model	word	article
Linear_SVM	0.7803	0.7793
LR	0.7750	0.7659
LGB	0.7538	
XGB	0.7310	

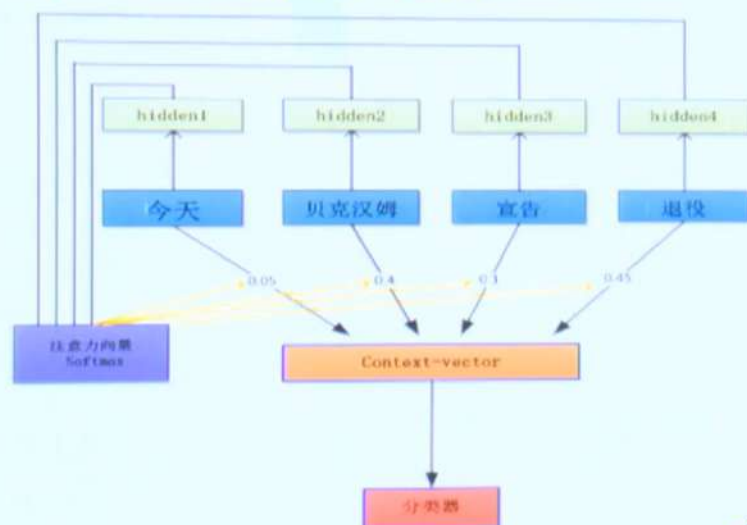
Fast-Text



对所有词语重视程度一样，不合理，F1score得分只有0.75

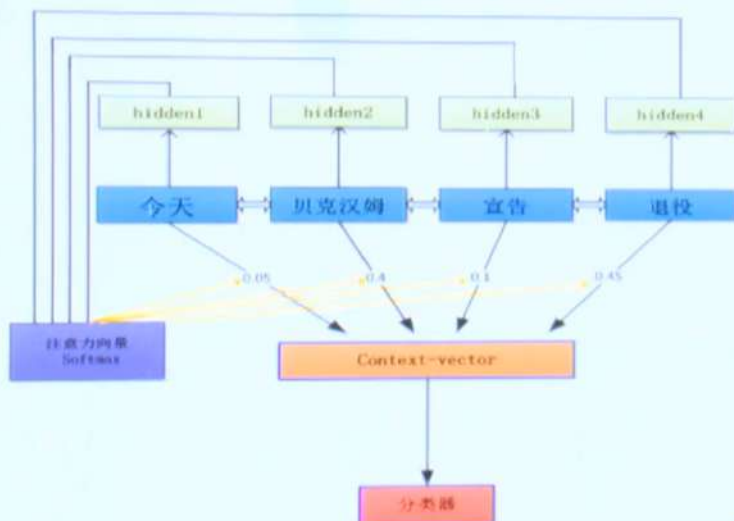
attention

Fast-attention-Text



- 先将词向量用2层MLP变换到另一个空间中，利用注意力向量计算每个词语的重要程度。
- 本次比赛中，我们采用了10个不同注意力向量用来提取不同的文本模式，产生了10个对应的 context-vector，共同输入到分类器中。

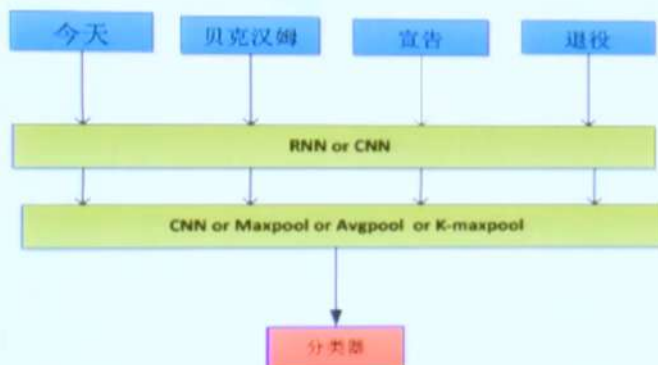
RNN-attention-Text



宣告破产----商业类
宣告退役----体育类

- 词语具有一词多意的属性，必须结合上下文语境才能确定自身的含义。Fast的不足之处在于缺乏这种考虑。
- 将两层MLP改为BiLSTM，结合上下文信息，使用隐状态来表示词语的确切含义，A榜得分0.7850。

RNN、CNN、RCNN

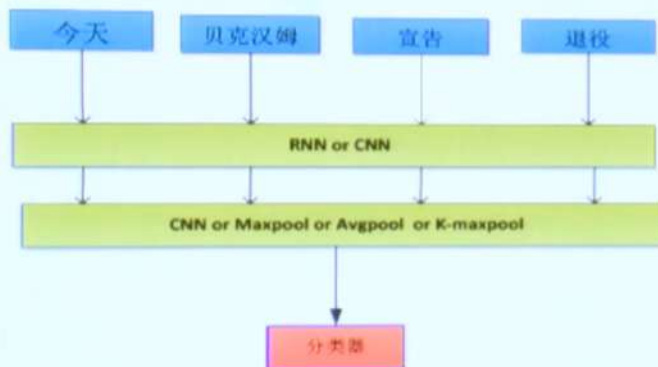


Model	score
RNN+kmaxpoo	0.7715
CNN+kmaxpoo	0.7776
RNN+CNN	0.7617

两种方式：

1. 从文章开头往后1200词
2. 文章开头中间结尾各400词，分别做卷积、RNN和pooling等操作。

RNN、CNN、RCNN



Model	score
RNN+kmaxpoo	0.7715
CNN+kmaxpoo	0.7776
RNN+CNN	0.7617

两种方式：

1. 从文章开头往后1200词
2. 文章开头中间结尾各400词，分别做卷积、RNN和pooling等操作。

易错类加权重：在损失函数上 等权softmax改为非等权 能提千分之2

关注易错类

通过改变损失函数，增加易错类的权重，使其在训练时得到更多的关注

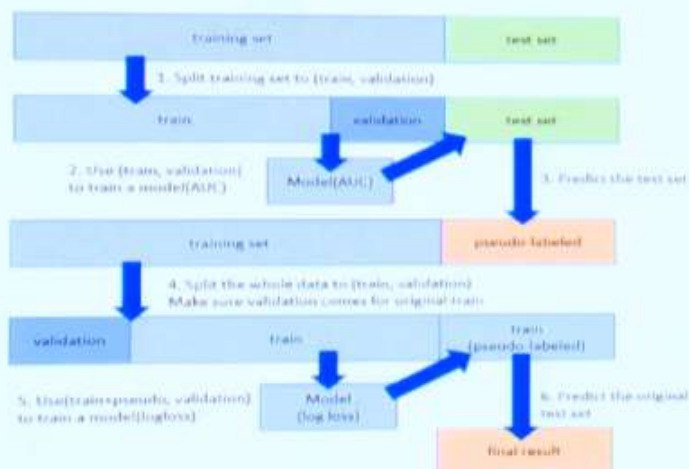
model	word	word_c w	combine	art	art_cw	combine
cnnbn	0.7247	0.7186	0.7364	0.6813	0.6727	0.6921
dpcnn	0.7346	0.7237	0.7419	0.7129	0.7012	0.7262
mhan	0.7791	0.7256	0.7824	0.7533	0.7128	0.7669

深度模型+传统特征

- 在基于word的深度模型中，dense层中加入基于char的SVD降维特征
- 在基于char的深度模型中，dense层中加入基于word的SVD降维特征

model	没加SVD (CV结果)	加入SVD (CV结果)
RNN-attention	0.7808	0.7856
Fast-attention	0.7776	0.7831

pseudo-labelling



- 模型可以容忍10%以内的标注噪声
- Test 样本的准确率大概是79%左右，因此每次训练时，取训练集8万+test中随机取5万，满足噪声容忍要求，同时扩大了训练集。
- 好处：allows your network to see a larger set of combinations of words

模型融合---后向选择算法，不断淘汰一个

模型筛选

- 本次比赛中一共训练了43个模型，在合队过程中必然存在一些冗余的模型。
- 使用后向选择算法，在LR分类器上交叉验证，进行模型筛选。
 1. 首先将所有模型加入，共同训练
 2. 进行43次迭代，每次剔除一个模型。最后删除CV得分提升最高的该次迭代中剔除的模型。
 3. 不断重复以上步骤，直到CV得分不再提高。
- Stacking第一层中一共使用了4个模型，分别是2层MLP，lightgbm，LR，SVC(linear)
- Stacking第二层使用了SVC(linear)
- 最后得分A榜0.80025，B榜0.79895

总结

- 使用了pseudo-labeling，降低了过拟合的程度
- 基于Char的训练结果虽然不好，但是对融合很有帮助
- 使用两种词向量的拼接进行训练，比仅使用单种词向量效果好
- 深度模型+传统模型降维过后的特征，很有帮助。Word的深度模型加char的传统特征效果更佳。
- 使用后向选择算法对模型进行筛选
- 修改了损失函数为类相关损失函数，提升了在某些类上的性能

特征:

问题描述

- 参赛者需要根据达观提供的脱敏文本数据，实现精准分类。

正文在字级别上表示(article)

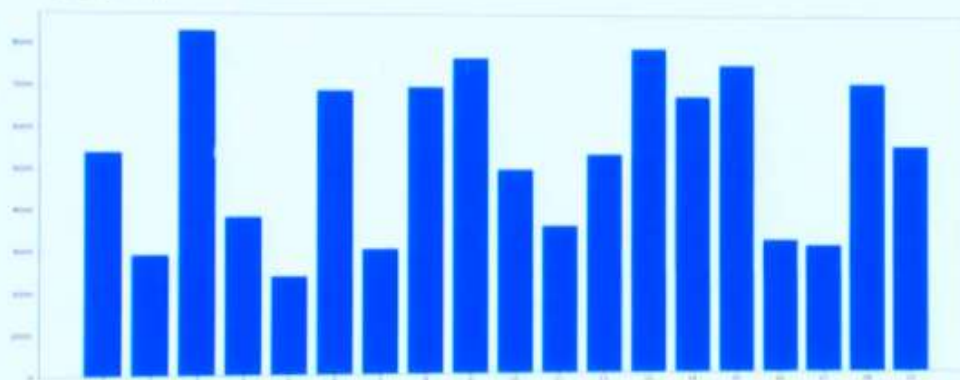
正文在词级别上表示(word_seg)



每篇文章对应的分类

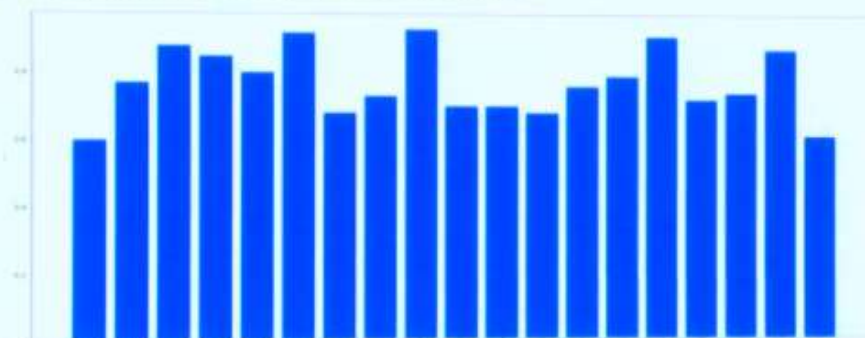
数据分布

- 各分类分布



模型预测数据情况

- 对于模型预测的数据f1score分值



- 第一类与最后一类较其他类的模型分类能力明显偏低

打乱第一类和最后一类的词序

- 删除低频词,及对模型分类能力较弱的类别进行shuffle数据增强
- 删除出现频率低于 5 次的词汇
- 对第一类和最后一类数据进行shuffle

模型:

词向量600维

RCNN

单模a榜 分数: 0.790.

04

01

RNN

单模a榜分数: 0.789.

DLMODEL

02

DPCNN

单模a榜分数0.77

Capsule

单模a榜分数0.788

03

unigram + 2gram tfidf + lda



融合（其他实验：bleeding），竞赛结果实验

DL模型的结果与
ML模型的结果使用
LightGBM模型进
行Stacking



- 过早进行模型融合导致过度拟合a榜
- 未尝试使用deepmoji
- 并未做完bagging方法

关于我们 ——CIKE实验室

CIKE实验室依托教育部大数据与机器人智能粤港澳联合实验室，在国际学术期刊和会议上发表论文上百篇，拥有多项发明专利，成员在阿里天池、Kaggle、CCF BDCI竞赛、教育部大数据挑战赛等多个国内外智能大赛中获得多个大奖，承担多个企事业单位委托的人工智能项目，有丰富的人工智能技术落地应用经验。实验室与许多企业建立长期合作关系，在数据挖掘、文本分类、知识图谱、问答与对话系统等方面都有相关合作项目。

蔡毅

教授、博导、香港中文大学博士

广东省特支计划青年拔尖人才

华南理工大学CIKE实验室负责人，指导教师
任20多个知名国际学术会议的主席和程序委员会委员
多个学术期刊编委和客座主编



达观数据

文本智能处理专家

DATA GRAND

数据&特征

分析

数据集

训练：测试 = 1 : 1

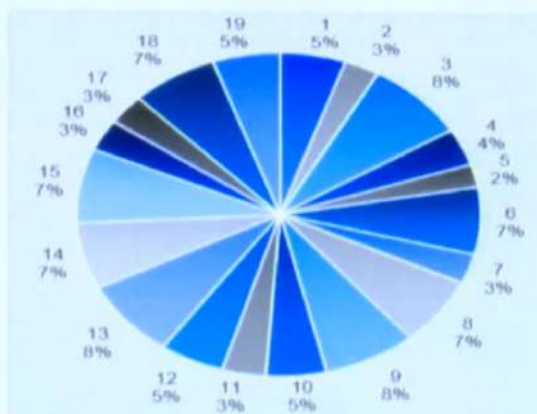
划分后的字、词的脱敏文本

训练集

- 字数量13516，平均长度1177
- 词数量875129，平均长度717

总共

- 字数量16052
- 词数量1271462



达观数据

文本智能处理专家

DATA GRAND

- Bag of words with term weighting

$$(tf_1 * w_1, tf_2 * w_2, tf_3 * w_3, \dots, tf_n * w_n)$$

- *idf* term weighting 衡量词对文档的区分度

- Entropy-based term weighting^[1]

当一个词语集中分布在少数类别中，可以视为“熵”较小，对类别的区分度较高；

反之，当一个词语较为均匀地分布在多个类别中，“熵”就较大，对类别区分度较低；

[1] Tao Wang, Yi Cai, Ho-fung Leung, Zhiwei Cai, and Huaqing Min. 2015. Entropy-based term weighting schemes for text categorization in VSM. In ICTAI 2015, pages 325–332.

tf-idf && 熵

- Bag of words with term weighting

$$(tf_1 * w_1, tf_2 * w_2, tf_3 * w_3, \dots, tf_n * w_n)$$

- *idf* term weighting 衡量词对文档的区分度

- Entropy-based term weighting^[1]

当一个词语集中分布在少数类别中，可以视为“熵”较小，对类别的区分度较高；

反之，当一个词语较为均匀地分布在多个类别中，“熵”就较大，对类别区分度较低；

[1] Tao Wang, Yi Cai, Ho-fung Leung, Zhiwei Cai, and Huaqing Min. 2015. Entropy-based term weighting schemes for text categorization in VSM. In ICTAI 2015, pages 325–332.

&类别权重平衡

特征融合linersvc分类，xgb融合

Model | Bag-of-Words

- Distributional Concentration (dc)

$$dc(t) = 1 - \frac{H(t)}{\log(|C|)} = 1 + \frac{\sum_{i=1}^{|C|} \frac{f(t, c_i)}{f(t)} \log \frac{f(t, c_i)}{f(t)}}{\log(|C|)}$$

- Balance Distributional Concentration (bdc)

$$bdc(t) = 1 - \frac{BH(t)}{\log(|C|)} = 1 + \frac{\sum_{i=1}^{|C|} \frac{p(t|c_i)}{\sum_{i=1}^{|C|} p(t|c_i)} \log \frac{p(t|c_i)}{\sum_{i=1}^{|C|} p(t|c_i)}}{\log(|C|)}$$

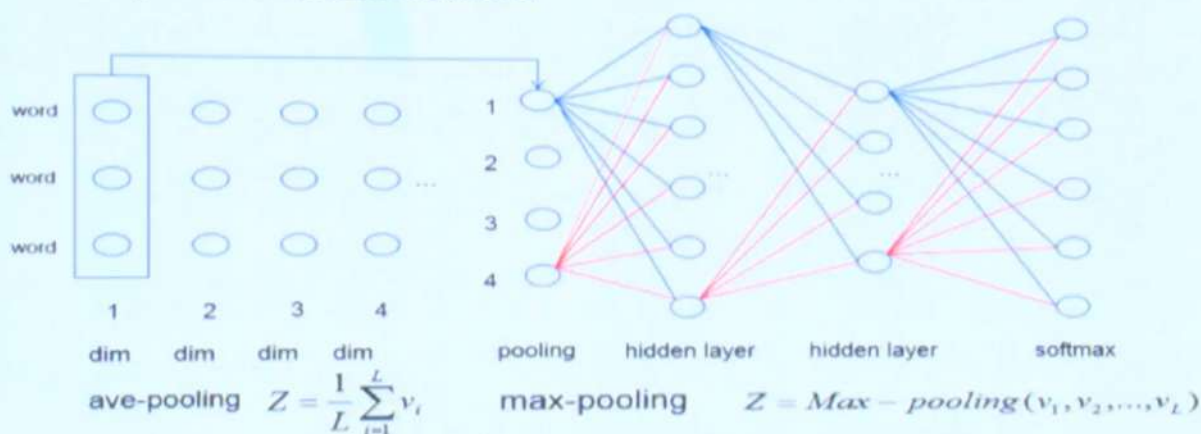
- 不同的 Term Weighting 方案以及不同的 n-gram 所得到的文本中的信息也不同，我们组合不同方案进行特征融合

[1] Tao Wang, Yi Cai, Ho-fung Leung, Zhiwei Cai, and Huaqing Min. 2015. Entropy-based term weighting schemes for text categorization in VSM. In ICTAI 2015, pages 325–332.

模型

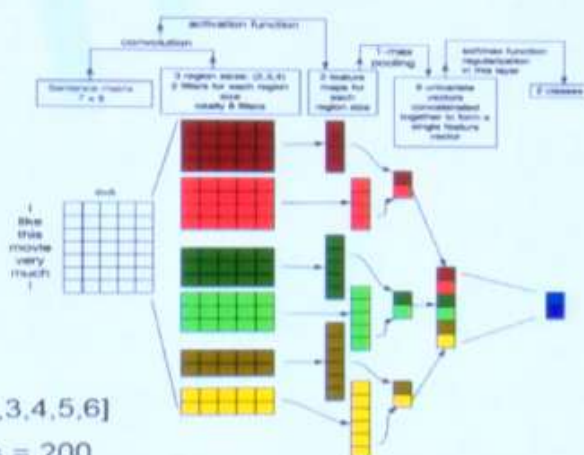
Model | SWEMS

Simple Word-Embedding Model



Ricardo Henao, Chunyuan Li, Lawrence Carin, Qiliang Su, Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Yizhe Zhang. Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms. ACL (1) 2018. 440-450.

Model | TextCNN



Filter size = [1,2,3,4,5,6]

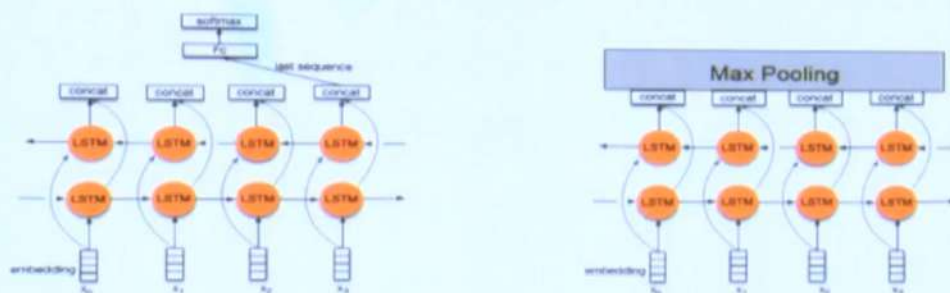
Number of filters = 200

Yoon Kim: Convolutional Neural Networks for Sentence Classification. EMNLP 2014. 1746-1751

达观数据 DATA GRAND 文本智能处理专家

按 Esc 即可退出全屏模式

Model | LSTM



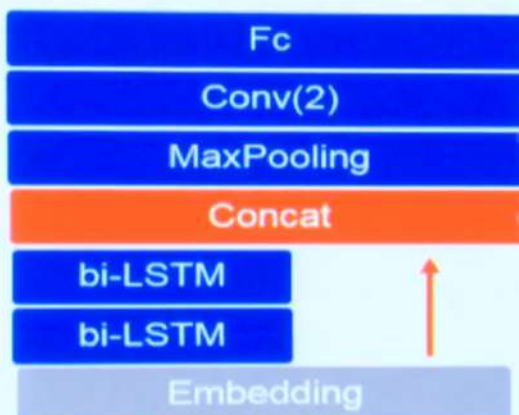
- 与通常的做法不同，这一 LSTM 模型并不是将最后一步的输出直接作为整段文本的表示，而是对所有步的输出一起做 max pooling 操作；
- 缓解定长隐向量编码过程中可能存在的信息损失问题；

参考知乎“曹山杯”夺冠记

达观数据 DATA GRAND 文本智能处理专家

Model | RCNN

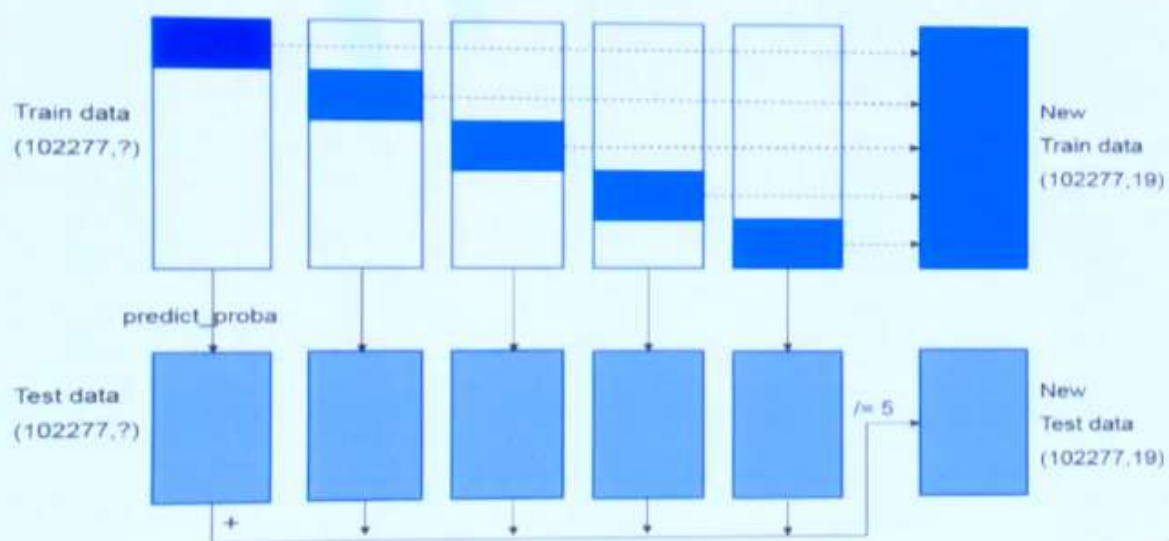
- Recurrent Convolutional Neural Network (RCNN) 将循环神经网络(RNN)和卷积神经网络(CNN)的优点结合起来
- RNN神经网络将一整段较长的文本编码进一个定长向量中，不可避免会存在信息损失，导致容易忽略句子中对分类有用的关键模式的问题
- CNN 擅长于挖掘出局部关键模式，但受限于固定大小的卷积核，不擅长挖掘词之间的长跨度依赖关系

Model |
Augmentation

删词：这个 例句 有点 萌
打乱：有点 这个 萌 例句

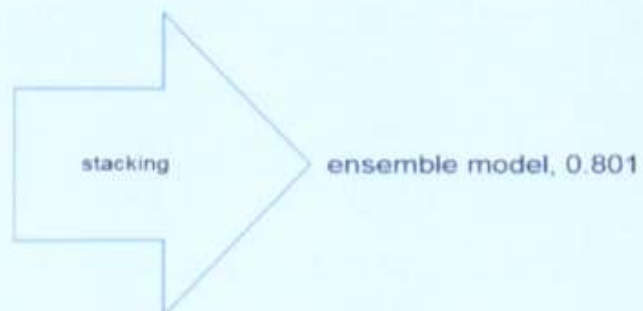
- 我们对抽取的训练样本，以一定的概率进行删词或打乱的数据增强操作；
- 当随机删词删除的恰好是那些对分类作用不大的词时，相当于新增了训练数据；
- 可能会产生干扰，但由于本数据集文本较长，最终产生有效的增强数据更多；
- 对基于LSTM的模型(LSTM, RCNN)提升了1到2个百分点；
- 为LSTM编码减少干扰，提升编码质量；

Model | Stacking



Model | Stacking

Bag-of-words model, 0.787
RCNN model, 0.792
LSTM model, 0.79
SWEMS model, 0.76
TextCNN model, 0.77
FastText model, 0.75
HAN 0.767



总结

- 当使用了较好的词权重衡量方法时，传统的向量空间模型仍然有一定竞争力；
- 基于LSTM的模型容易受噪声影响，在文本长度较长的情况下，适当使用数据增强能够提升效果；
- 模型之间的差异性越大，融合之后的结果就会越好，传统模型和深度学习模型融合之后能提高1个百分点；
- 融合模型较少的时候，直接使用等权重概率融合就会有很好的结果（优于stacking）；

Q&A

1、数据增强：

词随机打乱（python包）

词删除的概率0.4，句子被删掉0.5（训练的时候）

测试的时候：tta，数据增强

-----6-----地表最强-----

一、团队介绍

1.1 团队介绍

1.2 所获奖项

三、总结及致谢

二、解决方案

2.1 问题场景与解决思路

2.2 模型算法

2.3 其他



二 解决方案

2.1 问题场景与解决思路

队伍成员最近一年部分获奖经历

国内：

- 1、2017年CCF大数据计算智能大赛（BDCI）小超市供销存优化赛题：冠军
- 2、2018年第三届阿里云安全算法挑战赛：冠军
- 3、2018年云移杯全国旅游大数据挑战赛：冠军
- 4、2017年摩拜杯算法挑战赛：冠军
- 5、2017年移动公司4G用户流失预警赛题：一等奖
- 6、2017年CCF大数据计算智能大赛（BDCI）企业经营退出风险预测赛题：亚军
- 7、2018年拍拍贷第三届魔镜杯 语义相似度算法设计：亚军
- 8、2018年北京市校园高校大数据竞赛 校园人流量预测：亚军
- 9、2017年神州优车UAI 数据大赛：亚军
- 10、2017年智慧中国杯交通算法赛：亚军
- 11、2017年首届腾讯社交广告高校算法大赛——移动App广告转化率预估：季军

国际：

- 12、CIKM AnalytiCup 2018 跨语言短文本匹配：3rd/1027
- 13、IJCAI18 阿里国际广告算法大赛：5th/5204
- 14、KDDCup 2018 未来天气预测：7th/4170
- 15、G-Financial Forecasting Challenge Can you predict the future?：3rd/406

达观数据 文本智能处理专家
DATA GRAND

二 解决方案

2.1 问题场景与解决思路

新华社诺拉达沃托克9月11日电（记者 霍小光 李建敏）国家主席习近平11日在俄罗斯总统普京陪同下，共同出席中俄地方领导人对话会。

习近平在听取双方代表汇报本次对话会情况和中俄地方合作情况后，分别致辞。

习近平指出，中俄互为最大邻国和最重要的全面战略协作伙伴，拥有广泛共同利益。双方加强合作，深化互利交融，有利于携手化解外部风险和压力，促进共同发展繁荣。地方合作在中俄关系中所扮演重要角色。国家合作要依托地方、落脚地方、造福地方。地方合作越密切，两国互利合作基础就越牢固。在新的时代背景下，中俄地方合作面临着新形势、新任务、新要求，同时也迎来了新的历史性机遇。

习近平就未来两国地方合作提出四点建议。一要发挥地方政府作用，加强统筹协调，切实优化营商环境，鼓励更多地方结对，为两国企业相互投资营造更优质的营商环境。更便利的合作条件。二要创新合作思路，拓展合作地域，善用合作平台，发展好现有机制，深入探讨推进区域合作新模式。三要深挖互补优势，突出地方特色，实现合作精准对接，整合优质资源，激发合作内生动力，化优势为收获，打造合作亮点。四要密切人文交流，深化合作的主张民意和社会基础，推动两国地方文化、旅游、教育、媒体等领域交流机制化、常态化，增进彼此好感和认同感。

习近平强调，今年年初，我同普京总统约定2018年至2019年举办中俄地方合作交流年，将地方合作确定为今后两年双方合作主线。中俄地方合作正当其时。两国政府将支持各地方做大合作蛋糕，共享合作成果。希望中俄两国各省州代表借助中俄地方合作交流年的东风，共同开启中俄地方合作新时代，为两国关系发展添砖加瓦。

普京表示，地方合作是俄中全面战略协作伙伴关系重要组成部分。很高兴两国地方开展了密切的经济和人文交流合作。俄罗斯政府欢迎中国企业来俄罗斯投资兴业，愿继续为加深两国地方合作提供良好条件和环境。新形势下，俄中双方要以地方合作交流年为契机，提升互联互通水平，推进贸易和投资自由化便利化，增进民间友好，推动两国地方合作取得更多惠及两国人民的成果。

人类在阅读长文本时遇到的问题

问题一： 整篇文章过长。

整篇段落过长
导致很难把握全局信息，许多文章整篇都在体现自己的观点或者类型。

达观数据 文本智能处理专家
DATA GRAND

二 解决方案

2.1 问题场景与解决思路

人类在阅读长文本时候遇到的问题

新华社符拉迪沃斯托克9月11日电（记者 霍小光 李健敏）国家主席习近平11日在符拉迪沃斯托克和俄罗斯总统普京共同出席中俄地方领导人对话会。

两国元首在听取双方代表汇报本次对话会情况和中俄地方合作情况后，分别致辞。

习近平指出，中俄互为最大邻国和最重要的全面战略协作伙伴，拥有广泛共同利益。双方加强合作，深化利益交融，有利于携手化解外部风险和挑战，促进共同发展振兴。地方合作在中俄关系中的分量越来越重，国家合作要依托地方、带动地方、造福地方。地方合作越密切，两国互利合作基础就越牢固。在新的时代背景下，中俄地方合作面临着新形势、新任务、新要求，同时也迎来了新的历史机遇。

习近平就未来两国地方合作提出四点建议。一要发挥地方政府作用，加强统筹协调，切实优化营商环境，鼓励更多地方结对，为两国企业相互投资营造更优质的营商环境、更便利的合作条件。二要创新合作思路，拓展合作地域，善用合作平台，发展好现有机制，深入探讨推进区域合作新模式。三要深挖互补优势，突出地方特色，实现合作精准对接，整合优质资源，激发合作内生动力，化优势为收获，打造合作亮点。四要密切人文交流，强化合作的主流民意和社会基础，推动两国地方文化、旅游、教育、媒体等领域交流机制化、常态化，增进彼此好感 and 认同感。

习近平强调，今年年初，我同普京总统决定2018年至2019年举办中俄地方合作交流年，将地方合作确定为今后两年双方合作主线。中俄地方合作正当其时。两国政府将支持各地方做大合作蛋糕，共享合作成果。希望中俄两国各省市代表借助中俄地方合作交流年的东风，共同开启中俄地方合作新时代，为两国关系发展增添动力。

普京表示，地方合作是俄中全面战略协作伙伴关系重要组成部分。很高兴两国地方开展了密切的经贸和人文交流合作。俄罗斯政府欢迎中国企业来俄罗斯投资兴业，愿继续为加强两国地方合作提供良好条件和环境。新形势下，俄中双方要以地方合作交流年为契机，提升互联互通水平，推进贸易和投资自由化便利化，增进民间友好，推动两国地方合作取得更多惠及两国人民的成果。

问题二： 段落之间相关性。

多个段落的内容相互关联，需要模型推断出必要的信息。

达观数据 文本智能处理专家
DATA GRAND

二 解决方案

按 Esc 即可退出全屏模式

2.1 问题场景与解决思路

人类在阅读长文本时候遇到的问题

新华社符拉迪沃斯托克9月11日电（记者 霍小光 李健敏）国家主席习近平11日在符拉迪沃斯托克和俄罗斯总统普京共同出席中俄地方领导人对话会。

两国元首在听取双方代表汇报本次对话会情况和中俄地方合作情况后，分别致辞。

习近平指出，中俄互为最大邻国和最重要的全面战略协作伙伴，拥有广泛共同利益。双方加强合作，深化利益交融，有利于携手化解外部风险和挑战，促进共同发展振兴。地方合作在中俄关系中的分量越来越重，国家合作要依托地方、带动地方、造福地方。地方合作越密切，两国互利合作基础就越牢固。在新的时代背景下，中俄地方合作面临着新形势、新任务、新要求，同时也迎来了新的历史机遇。

习近平就未来两国地方合作提出四点建议。一要发挥地方政府作用，加强统筹协调，切实优化营商环境，鼓励更多地方结对，为两国企业相互投资营造更优质的营商环境、更便利的合作条件。二要创新合作思路，拓展合作地域，善用合作平台，发展好现有机制，深入探讨推进区域合作新模式。三要深挖互补优势，突出地方特色，实现合作精准对接，整合优质资源，激发合作内生动力，化优势为收获，打造合作亮点。四要密切人文交流，强化合作的主流民意和社会基础，推动两国地方文化、旅游、教育、媒体等领域交流机制化、常态化，增进彼此好感 and 认同感。

习近平强调，今年年初，我同普京总统决定2018年至2019年举办中俄地方合作交流年，将地方合作确定为今后两年双方合作主线。中俄地方合作正当其时。两国政府将支持各地方做大合作蛋糕，共享合作成果。希望中俄两国各省市代表借助中俄地方合作交流年的东风，共同开启中俄地方合作新时代，为两国关系发展增添动力。

普京表示，地方合作是俄中全面战略协作伙伴关系重要组成部分。很高兴两国地方开展了密切的经贸和人文交流合作。俄罗斯政府欢迎中国企业来俄罗斯投资兴业，愿继续为加强两国地方合作提供良好条件和环境。新形势下，俄中双方要以地方合作交流年为契机，提升互联互通水平，推进贸易和投资自由化便利化，增进民间友好，推动两国地方合作取得更多惠及两国人民的成果。

问题三： 距离与语义。

距离的概念在本文理解中非常重要。

达观数据 文本智能处理专家
DATA GRAND

二 解决方案

2.1 问题场景与解决思路

机器学习中：有多少人工就有多少智能。深度学习亦是如此。

特征工程 → 设计合理的网络结构

问题一：整篇文章过长。

问题二：段落之间相关性。

问题三：距离与语义。

怎样设计网络结构，产生一个符合Motivation的模型，提高预测精度？

二 解决方案

按 Esc 即可退出全屏模式

2.2 网络结构——处理文本过长

paper : Sliced Recurrent Neural Networks

The standard RNN structure is shown in Figure 1, where A denotes the recurrent units.

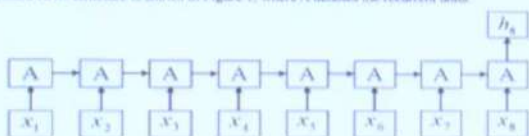


Figure 1: The standard RNN structure. Each step waits for the output of its previous step, which is computed by the recurrent unit A.

Standard RNN

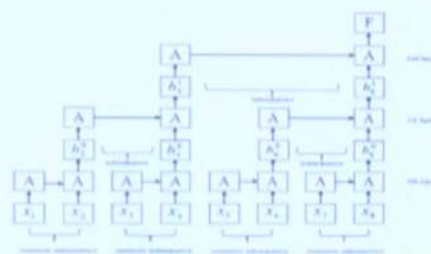


Figure 2: The SRNN structure. It is constructed by slicing the input sequence into several minimum subsequences with equal length. The recurrent units could work on each subsequence simultaneously on each layer, and the information could be transmitted through multiple layers.

SRNN

在数学上证明了在激活函数为线性函数时，SRNN = Standard RNN，前者速度快135倍

二 解决方案

2.2 网络结构——处理文本过长

paper : Sliced Recurrent Neural Networks

数据中

- 1、没有分段标志
- 2、没有标点符号

$$X = [x_1, x_2, \dots, x_T]$$

文本总长度

$$t = \frac{T}{n}$$

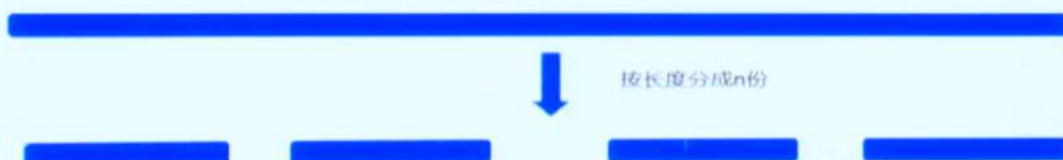
子集长度

$$X = [N_1, N_2, \dots, N_n]$$

文本子集表示

$$N_p = [x_{(p-1)t+1}, x_{(p-1)t+2}, \dots, x_{pt}]$$

子集表示



达观数据
DATA GRAND

文本智能处理专家

二 解决方案

按 Esc 即可退出全屏模式

2.2 网络结构——段落之间相关性

paper : DR-BiLSTM: Dependent Reading Bidirectional LSTM for Natural Language Inference



NLU中的依赖阅读：将推理和问题
首尾相连进行初始化

$$\begin{aligned} \bar{v}, s_v &= BiLSTM(v, 0) \\ \hat{u}, - &= BiLSTM(u, s_v) \end{aligned} \quad (1)$$

$$\begin{aligned} \bar{u}, s_u &= BiLSTM(u, 0) \\ \hat{v}, - &= BiLSTM(v, s_u) \end{aligned} \quad (2)$$

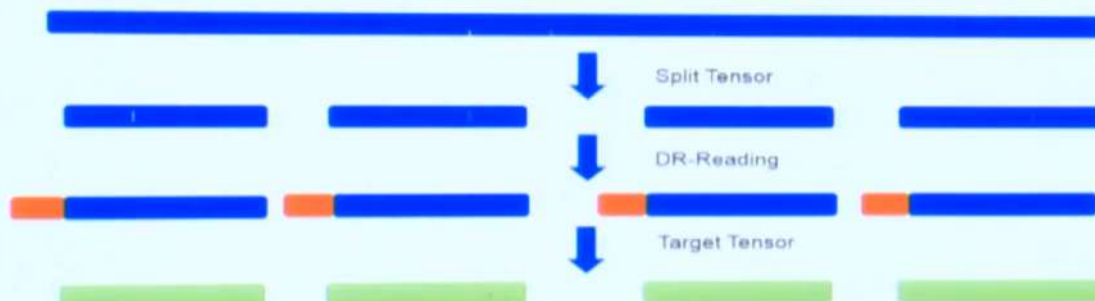
达观数据
DATA GRAND

文本智能处理专家

二 解决方案

2.2 网络结构——段落之间相关性

退而求其次，获取上下文信息



二 解决方案

2.2 网络结构——距离与语义

paper : Distance-based Self-Attention Network for Natural Language Inference

1	0	-1	∞	(0,1)
2	-1	0	∞	(0,2)
3	∞	∞	∞	∞
4	-(0,1)	-(0,2)	∞	0
	1	2	3	4

+Self-Attention

$$\text{Masked}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + \alpha M_{dis}\right)V$$

Figure 4: Distance mask

二 解决方案

2.2 网络结构

最终解决方案网络设计：

切分 (Split) + 依赖阅读 (Dependent Reading) + 距离编码 (Distance Masking)

还使用了RNN、CNN、RCNN、StackRNN、DPCNN等等模型

二 解决方案

其他

1、训练不同维度的词向量，对高频词和低频次均不作处理，有些高频词可能是标点符号或者语气词，但是对于区分文本可能有正向的作用。比如在情感打分中：

1、这个小哥哥好帅啊啊啊啊啊啊啊啊啊啊啊啊！舔屏！（5分）

这个小哥哥好帅（4分）

2、在拼多多买的东西不能用啊！！！！！！！！！！（-5分）

在拼多多买的东西不能用（-4分）

二 解决方案

按 Esc 即可退出全屏模式

其他

- 2、比赛中采用了多种传统模型
- 3、给某些类别增加类别权重
- 4、将几个类别合并为1个类别
- 5、使用全部词、前1500个词、前1000个词、前500个词、后500个词进行训练

达观数据 文本智能处理专家
DATA GRAND

三 总结及致谢

总结：

做比赛的目的是提高 从问题场景出发，理解问题并提出适合解决问题的方案的能力

致谢：

感谢主办方

感谢队友

感谢支持我们的老师、同学和朋友

感谢本次比赛的运营团队



达观数据 文本智能处理专家
DATA GRAND

Q&A

- 1、分段&分割，比如：把文章平均分为10分
- 2、不够的部分会补零，多了删除，都整理为1000个字
- 3、距离与语音
- 4、srnn速度比较快，非线性会震荡

团队介绍

数据分析

算法模型

达观 DATA 文本智能处理专家

按 **Esc** 即可退出全屏模式

比赛奖项

2017年 JDD信贷预测 冠军
2017年 UAI数据大赛 冠军
IJCAI-17 口碑商家客流量预测第二名
阿里云安全算法挑战赛线上赛第一名
【广东大赛】机场客流量的时空分布预测第二名
阿里聚安全算法挑战赛第二名
2017携程用户预订售卖房型概率预测第一名
2017“达观杯”个性化推荐算法挑战赛线上赛第一名
金融壹账通前海征信金融反欺诈创新大赛第一名
CCF-2017 小超市供销存管理优化一等奖
CCF-2017 卫星影像的AI分类与识别二等奖
国家天文数据挖掘大赛第一名
美年大健康大数据比赛第二名

达观 DATA 文本智能处理专家

词1800, 字都保留, 置信度95, 词2000, 词向量50个词



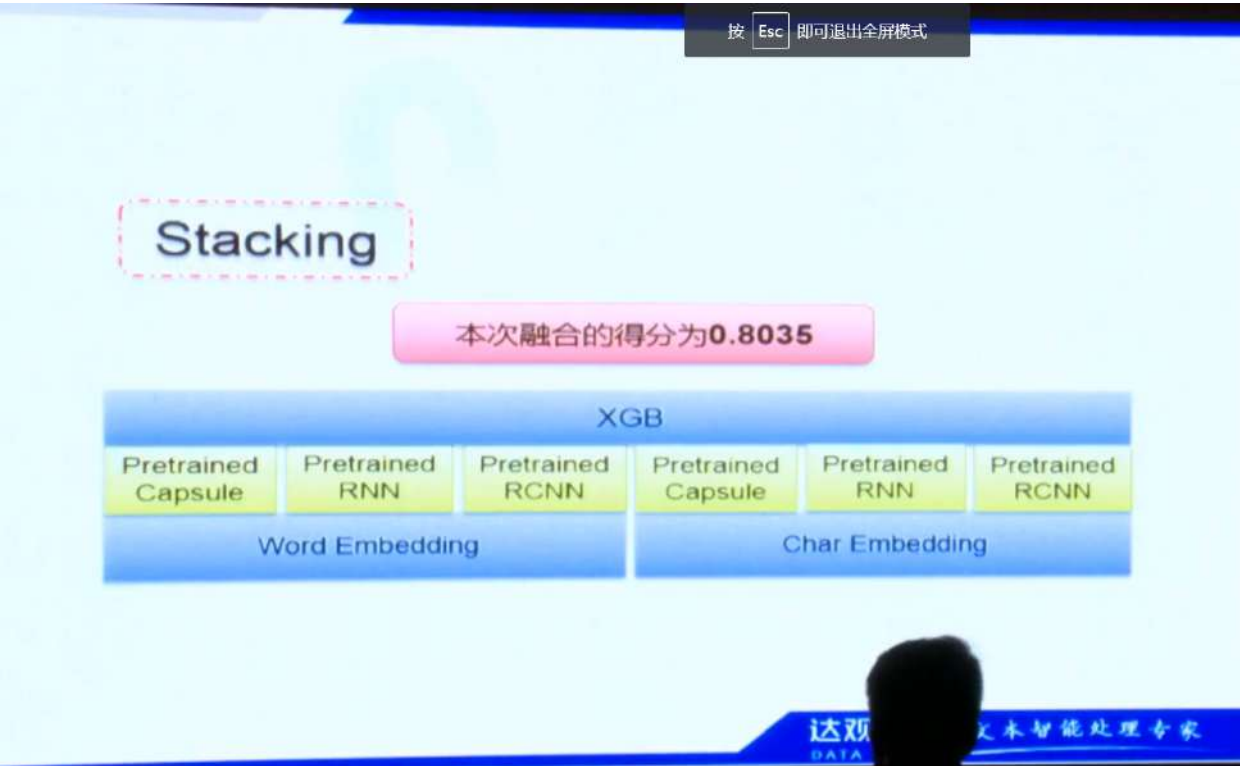
类别不均衡，用传统的损失函数

模型：

rnn已经开源，两层



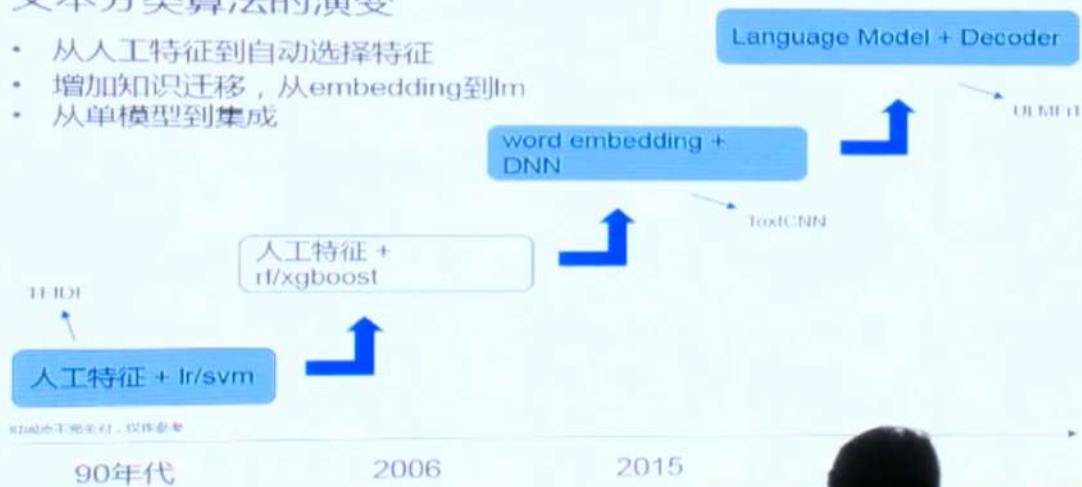
胶囊0.8,其他两个能到0.798



算法思路

文本分类算法的演变

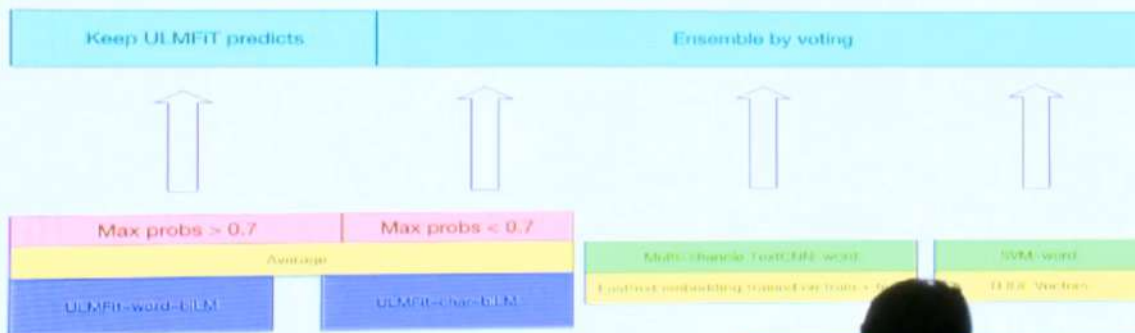
- 从人工特征到自动选择特征
- 增加知识迁移，从embedding到lm
- 从单模型到集成



算法思路

解决方案总体框架

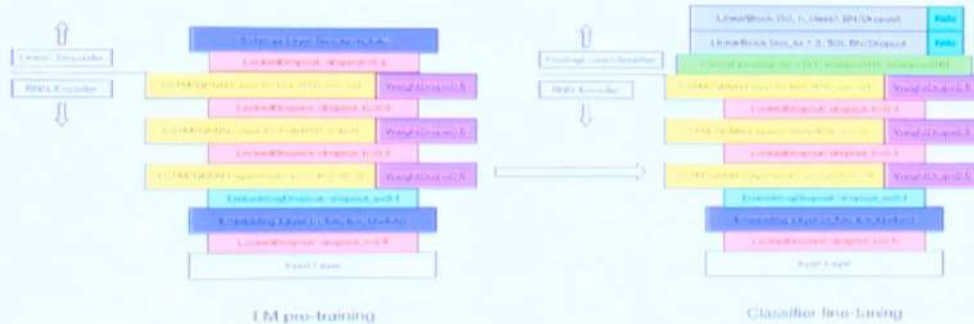
ULMFIT-biLM-word-char + TextCNN-word + SVM-word
ULMFIT-biLM-word-char: 0.80250
Ensemble: 0.805878



特征: LM 词性(前两层)+词义 (消歧) vs word2vec

ULMFiT模型

Universal language model fine-tuning for text classification



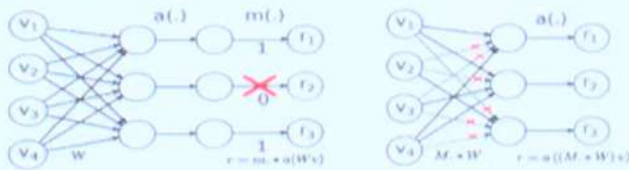
- Howard, Jeremy, and Sebastian Ruder. "Universal language model fine-tuning for text classification." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.
- Mentz, Stephen, Nitesh Shirish Keskar, and Richard Socher. "Regularizing and optimizing language models." arXiv preprint arXiv:1708.02182 (2017).

达观数据 智能处理专家
DATA GRA

AWD-LSTM LM (SOTA)

用了很多正则优化方法的语言模型结构

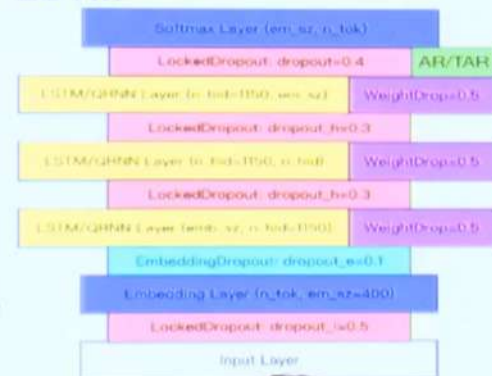
- DropConnect
 - Dropouts作用于LSTM效果不好，破坏长依赖
 - 随机置零activations改为随机置零weights
 - 不破坏cuDNN LSTM实现更高效



DropOut Network

DropConnect Network

ASGD Weight-Dropped LSTM



- Mentz, Stephen, Nitesh Shirish Keskar, and Richard Socher. "Regularizing and optimizing language models." arXiv preprint arXiv:1708.02182 (2017).

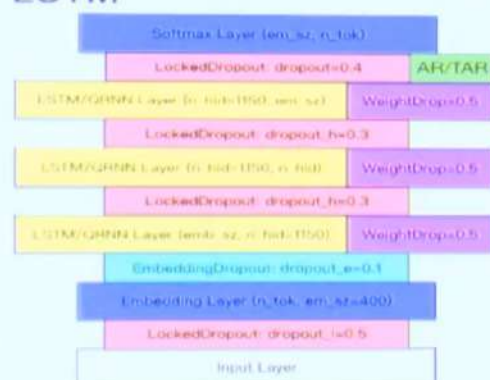
达观数据 智能处理专家
DATA GRA

AWD-LSTM LM (SOTA)

用了很多正则优化方法的语言模型结构

- DropConnect
 - Dropouts作用于LSTM效果不好，破坏长依赖
 - 随机置零activations改为随机置零weights
 - 不破坏cuDNN LSTM实现更高效
- Variational Dropout
 - 传统Dropout，mask每次都会采样
 - 只采样一次，锁定复用
 - 用于所有的dropout操作
- Embedding Dropout
- Activation Regularization(AR)/Temporal AR
- Weight Tying

ASGD Weight-Dropped LSTM



达观数据集 perplexity = $\exp(\text{val_loss}) = 71.9$

1. Moritz, Stephen, Nitish Shirish Keskar, and Richard Socher. "Regularizing and optimizing LSTM models." arXiv preprint arXiv:1708.02182 (2017).

达观数据 智能处理专家
DATA GRAB

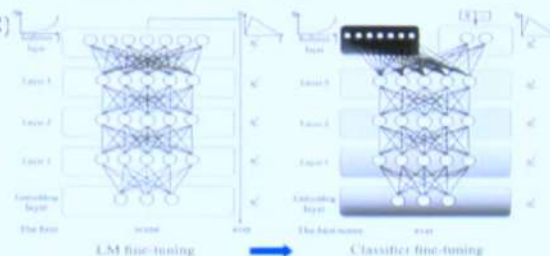
ULMFiT训练-Im

训练用到的主要方法

- Slanted Triangular Learning Rates (STLR)
 - 先线性增长后线性减少，增短减长
 - 先快速收敛，后微调



ULMFiT训练分两个阶段



$$\text{cut} = \lceil T \cdot \text{cut_frac} \rceil$$

$$p = \begin{cases} t/\text{cut}, & \text{if } t < \text{cut} \\ 1 - \frac{t - \text{cut}}{\text{cut} \cdot (1/\text{cut_frac} - 1)}, & \text{otherwise} \end{cases}$$

$$\eta_t = \eta_{\text{max}} \cdot \frac{1 + p \cdot (\text{ratio} - 1)}{\text{ratio}}$$

T: 总的迭代次数;
cut_frac: 快速收敛迭代次数占比
cut: 第cut个迭代, t从递增改为递减
t: 当前第t个迭代
p: 当前递增或递减的占比, 用于后面计算当前值
ratio: 最小值和最大值的比例
cut_frac=0.1, ratio=32, max_iter=3

1. Smith, Leslie N. "Cyclical learning rates for training neural networks." Applications of Computer Vision (WACV), 2017. IEEE Winter Conference on. IEEE, 2017.

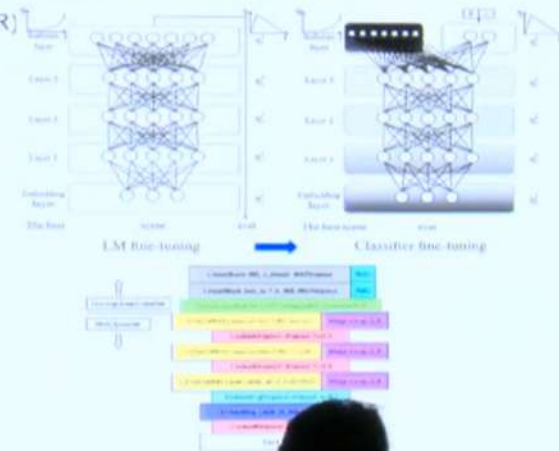
达观数据 智能处理专家
DATA GRAB

ULMFiT训练-classifier

训练用到的主要方法

- Slanted Triangular Learning Rates (STLR)
 - 先线性增长后线性减少，增短减长
 - 先快速收敛，后微调
- Discriminative Fine-Tuning
 - DL不同层提取不同的信息，从一般到特定
 - 确定最后层 h 后，其它层为 $\eta^{i-1} = \eta^i / 2.6$
- Gradual Unfreezing
 - 从上往下逐层放开训练，防止灾难性遗忘
- Variable Length BPTT
 - BPTT长度上增加随机性，类似cv shuffling
- Concat Pooling

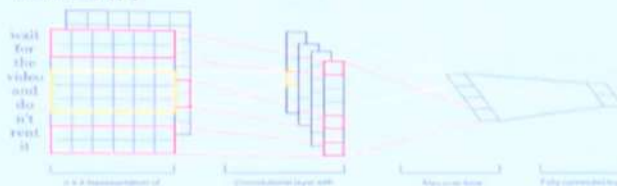
ULMFiT训练分两个阶段



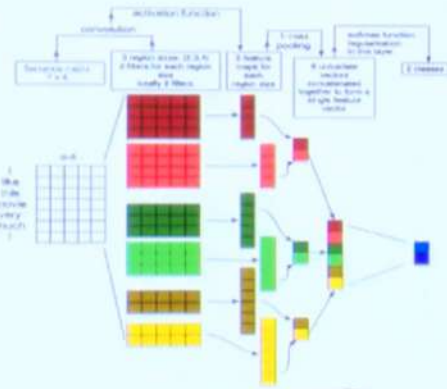
1. Howard, Jeremy, and Sebastian Ruder. "Universal language model fine-tuning for text classification." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 310-315. 2018.

TextCNN模型

使用了传统的Kim的TextCNN



参数名	参数值
embedding_size	100
vocab_size	6000
MAX_SEQUENCE_LENGTH	4000
filter_sizes	[2, 3, 4]
num_filters	1000
drop	0.7
batch_size	64
epochs	25



multi-channel TextCNN + fasttext embedding

1. Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1406.2661. 2014.

一些失败尝试

- 传统机器学习，应该改进到boosting的方式，比如xgboost，我直接用xgboost替换了svm，效果不理想，也没时间太多调参，所以放弃了，但我相信xgboost应该可以战胜svm。
- 关于embedding我使用了fasttext，但是今年ACL最佳paper的ELMO应该能有更好的性能，可惜ELMO训练速度极慢，放弃了，之后可以尝试一下这个sota的embedding。
- text领域的test time argument是我一直想要尝试的方法，但至今没有看到有效的方法，有人尝试google翻译再返翻译的方式，在这个数据集上不适用，我简单尝试了把文章句子随机打乱，加权但是效果还是不理想，期待之后业界能有好的text TTA。

达观数据

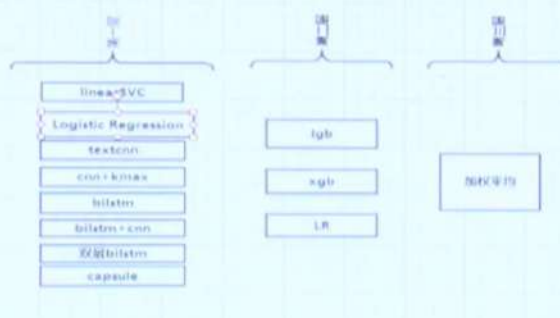
智能处理专家

9

1.1 总体架构

比赛采用三层stacking

- 第一层：
为传统的SVM,LR，以及深度学习模型
CNN,LSTM,ATTENTION,CAPSULE等模型
- 第二层：
LGB, XGB, LR
- 第三层：
加权平均



达观数据

智能处理专家

第四层，几个人的结果加权平均

1.3 第二层

按 **Esc** 即可退出全屏模式

1, 利用第一层提取的特征, 十折训练LGB, 每一折都对测试数据进行预测, 最后取平均作为一个结果

2, 利用第一层提取的特征, 十折训练XGB, 每一折都对测试数据进行预测, 最后取平均作为另一个结果

3, 利用第一层提取的特征, 十折训练LR, 每一折都对测试数据进行预测, 最后取平均作为另一个结果

达观数据
DATA GRAND

智能处理专家

1.3 第三层

1, 对第一层的XGB,LGB,LR结果进行加权融合,

线上最优分数为

XGB_0.80393,

LGB_0.80338,

LR_0.80213

取权重 (5,3,2) 线上最优分数为0.8052

2, 根据训练数据的类别分布对结果进行再平衡

A榜最优得分为0.80598

达观数据
DATA GRAND

智能处理专家

2.3 模型的得分

比赛中记录的一些模型得分

模型	线下十折平均F值	线上得分
LinearSVC	0.7766	0.7782
LR	0.7704	
textcnn	0.7715	0.7854
Conv+kmx	0.7637	
bilstm	0.7765	
bilstm+cnne	0.7788	0.7901
Bilstm+att	0.7717	0.7868
双层的 bilstm	0.7773	
capsule	0.7741	
capsule	0.7714	0.7853

2.4 后处理

class=argmax([prob1,prob2,prob3,...
,prob19])



class=argmax([k1*prob1,k2*prob2,k
3*prob3,...,k19*prob19])



如何确定k1-kn? 参数再多也得一个
个来!

训练集	argmax	加权argmax
3:8313,	3 8213	3 8239
13:7907,	13 8206	13 7963
9: 7675,	9 7685	9 7670
15:7511,	15 7418	15 7448
18:7066,	18 7093	18 7076
8: 6972,	8 6992	8 6939
6: 6888,	6 6935	6 6911
14:6740,	14 6896	14 6803
19:5524,	19 5554	19 5517
1: 5375,	12 5232	1 5348
12:5326,	10 4878	12 5294
10:4963,	1 4845	10 4950
4: 3824,	4 3980	4 3911
11:3571,	11 3553	11 3553
16:3220,	16 3351	16 3245
17:3094,	17 3227	17 3129
7: 3038,	2 2951	7 3001
2: 2901,	7 2882	2 2914
5: 2369,	5 2366	5 2366

2.4 后处理

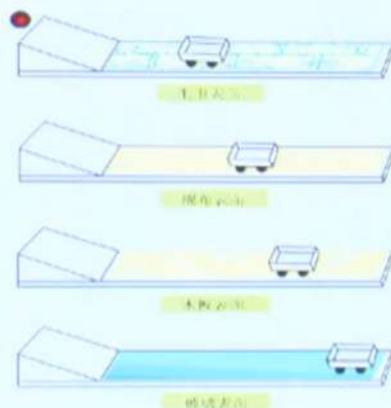
train_rt_n:训练集中某类别占比.

pred_rt_n:最终预测出来的某类别占比.

kn:赋予的每个类别的权重。

s:参数(当 $s > 1$ 时表现为阻尼, 当 $s < 1$ 时表现为惯性)

```
for i in range(5):  
    for n in range(19):  
        kn=(train_rt_n/pred_rt_n - 1.0)/s+kn
```



03 比赛总结

数据挖掘比赛玩什么？

玩样本，玩模型，玩特征，玩融合，玩评分...常玩常新。

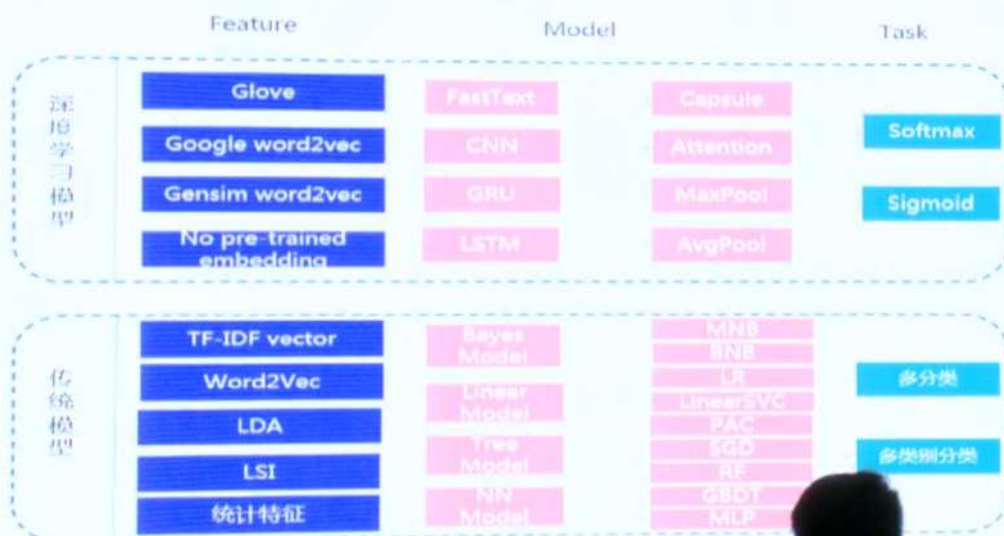
相似的题，有时候解法不同。

不同的题，有时候方法却相同。

但有一种方法却是通用的：多练习，多思考，多尝试，多总结。

融合: bagging, stacking

模型架构



特征

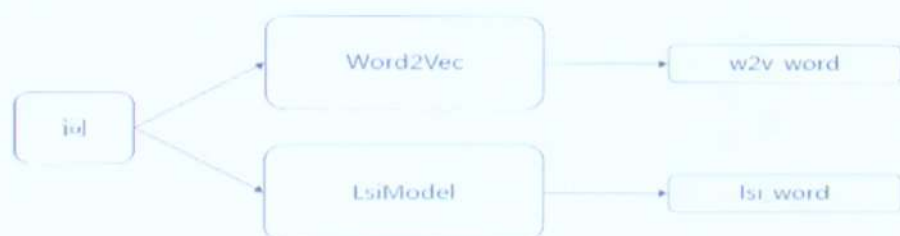
TF-IDF特征



LDA特征



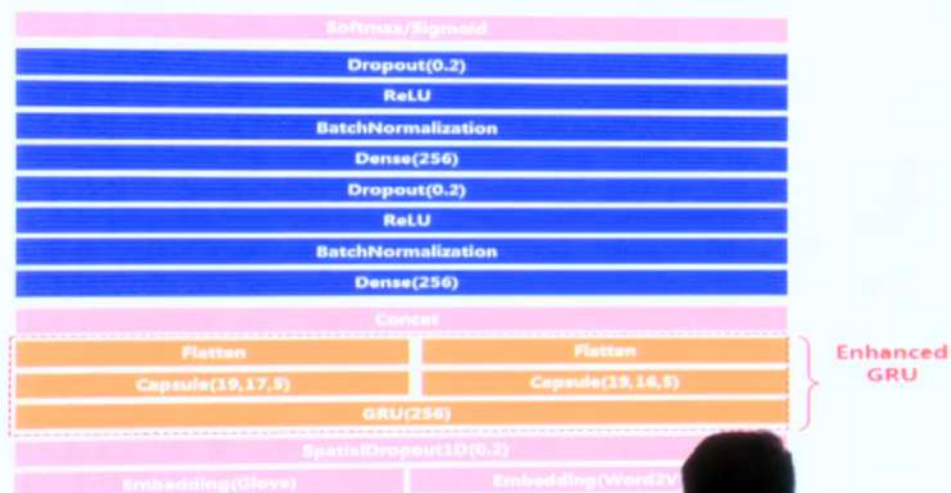
LSI特征+Word2Vec特征



模型

胶囊并行，特征多样性，学到更多的信息

Hybrid NN-1

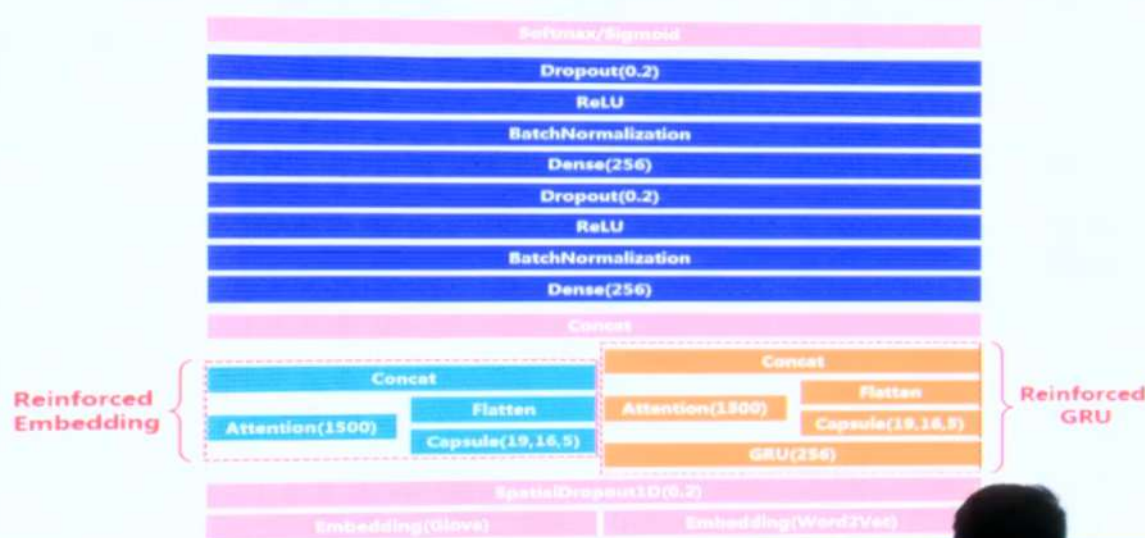


达观数据
DATA GRAND

智能处理专家

按 Esc 即可退出全屏模式

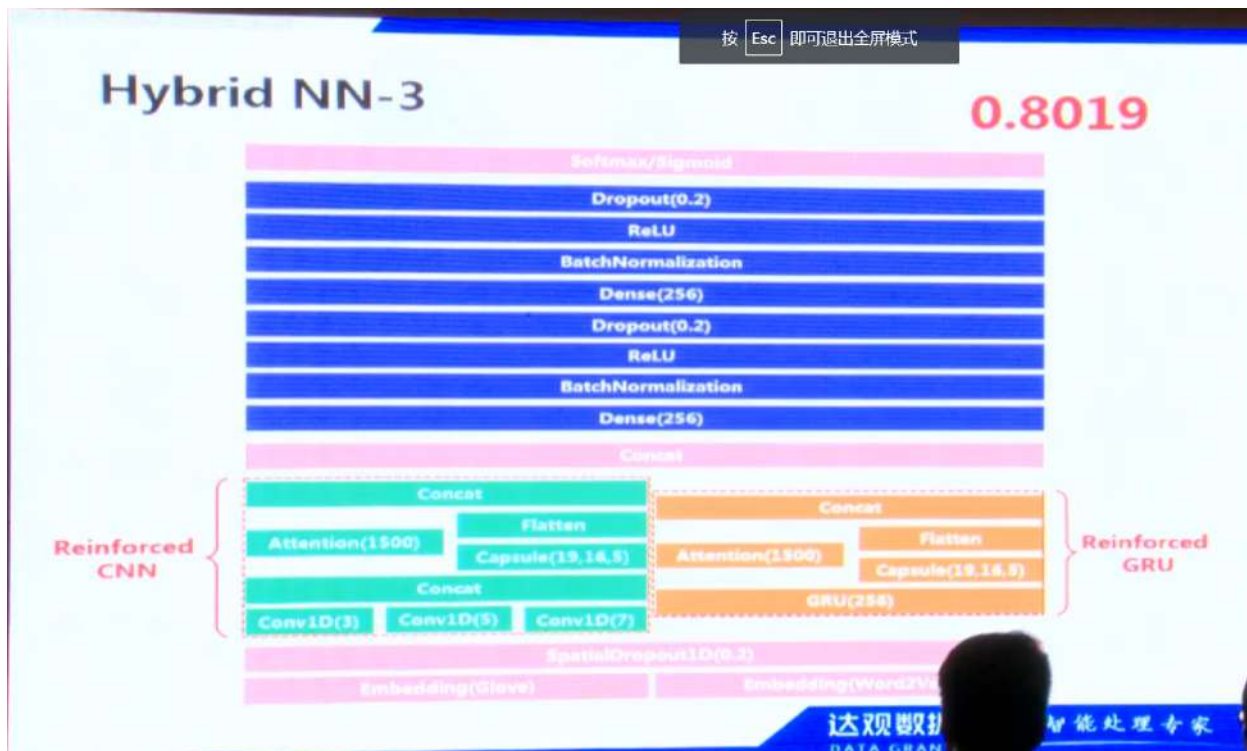
Hybrid NN-2



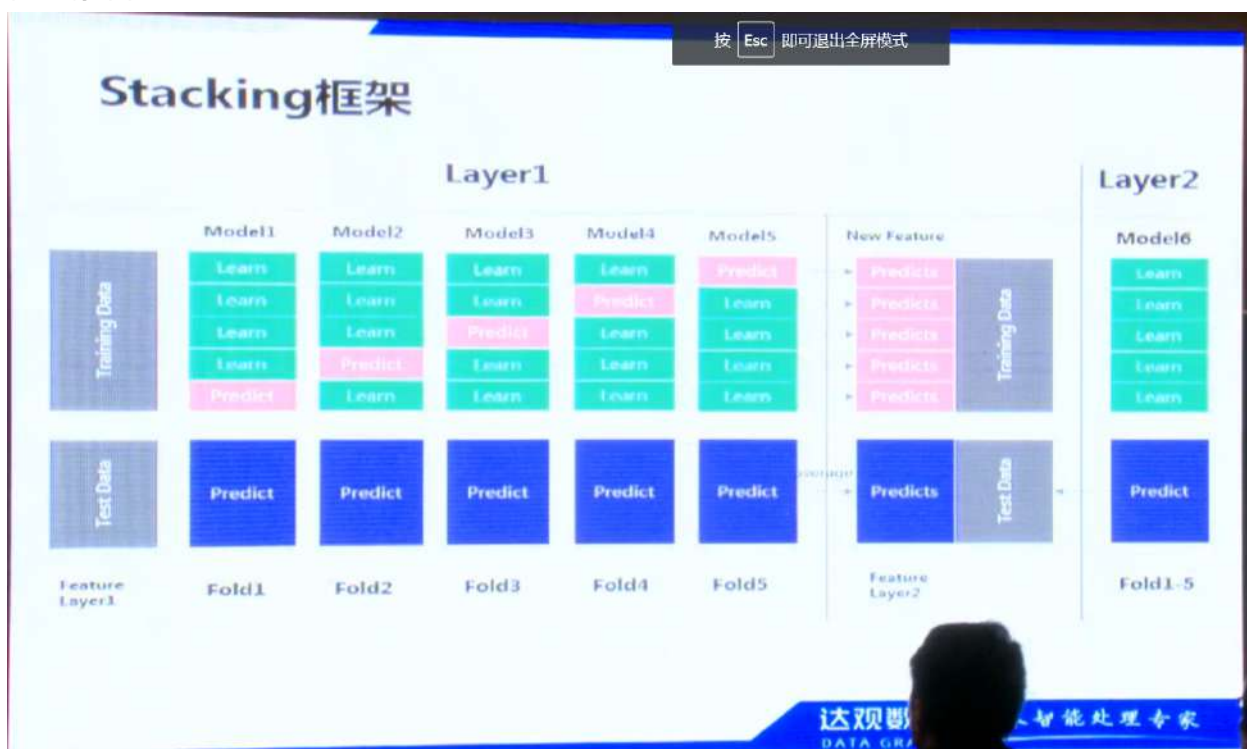
达观数据
DATA GRAND

智能处理专家

加了3个 (3,5,7) 过滤器大小的卷积层



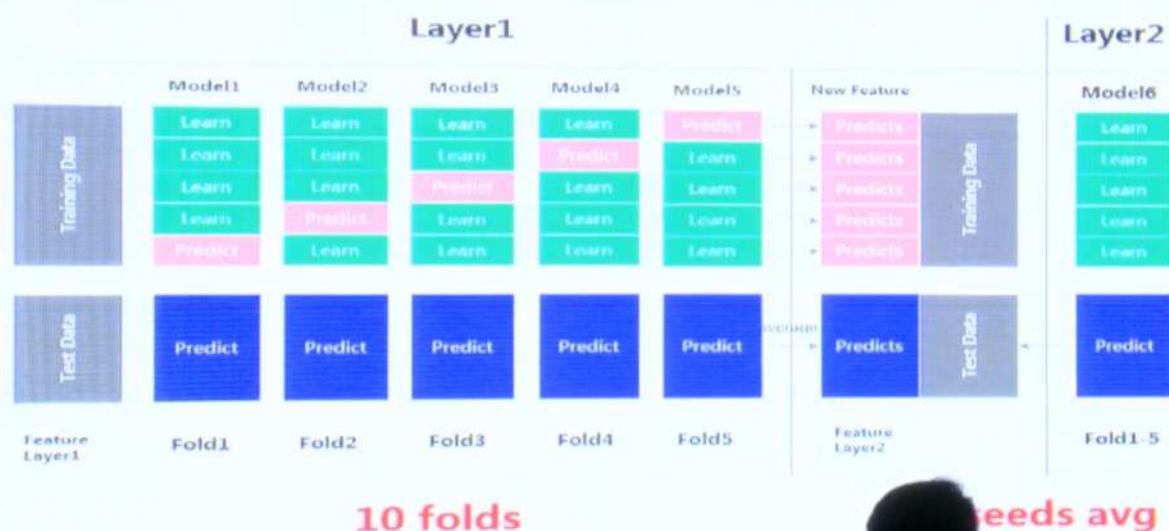
模型融合



20个seed

trick: 预训练词向量

Stacking框架



达观数据 数据智能处理专家
DATA GRA

总结

- 网络结构的创新和改进对本赛题的效果是明显的
- 预训练的Embedding能加快网络的训练，并且效果俱佳
- 传统模型对于融合的提升是巨大的

达观数据 数据智能处理专家
DATA GRA

展望

- 尝试其他的网络结构，例如：DPCNN、DenseNet、HAN等
- 混合网络的参数调优
- 融合系数的优化

达观数据 文本智能处理专家

-----达观公司介绍-----

文本智能处理的需求遍及各行各业

金融

- 银行
- 券商
- 保险
- 基金
- 信托

7,800+

国内金融机构数量

媒体

- 新闻机构
- 出版社
- 网络媒体
- 监管部门

互联网

- 社交
- 电商
- 视频
- 阅读
- 文化娱乐

政府与公共机构

- 公安
- 海关
- 机关单位
- 团体组织

法律

- 企业法务部
- 律师事务所
- 司法机构

24,000+

国内律所数量

更多行业

大量的企业和政府机构目前仍依赖人工对海量文档资料进行手工处理，缺乏技术手段将工作自动化，提升业务的运行效率，降低成本，提升可靠性

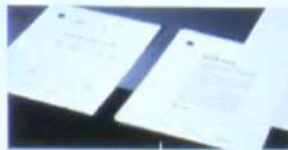
大型企业

- 科技
- 通信
- 制造

让计算机代替人类来进行文本自动化处理有广阔的应用面



法律法规资料



招股说明书 债券说明书



合同 公文



银行业务单据



资讯文章



档案 问答库

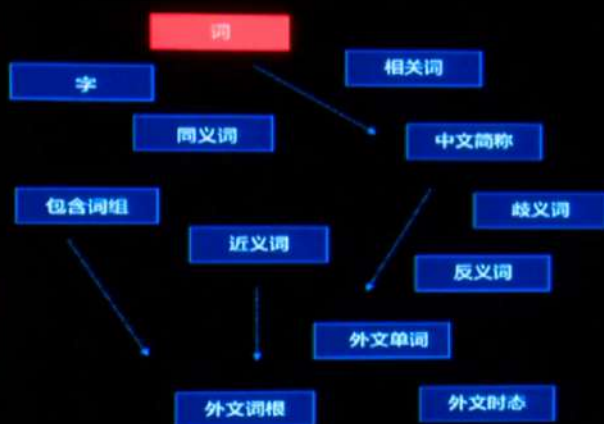
常见需求：

- 提炼文章核心观点
- 抽取文章关键信息
- 发现文章错漏内容
- 找出文档间关联关系
- 审核材料的内容
- 起草报告文书
- 核对文字和数字信息
- 文章润色修改

达观数
DATA GR

本智能处理专家

文本挖掘应用难点一：处理各类复杂的字词语义关系



- **【词汇粒度问题】** 中华人民共和国 | 共和国 | 中国
- **【指代归属问题】** 本次活动 该条款 我校 他
- **【同义近义问题】** 路易威登 LV | 球鞋 运动鞋 跑步鞋
- **【局部转义问题】** 巧克力囊肿 | 鸡翅U盘
- **【一词多义问题】** 方便 意思 杯具
- **【上下文依赖问题】** 101空降师 101路 101号 101次

文本挖掘应用难点二：多样化的句法结构的解析

网络问题的处理方法
如何解决网络故障
网络连不上怎么办
处理网络问题的经验

【解决办法】语义的归一化

你上班了吗？
班你上了吗？
上班了吗你？
你班上了吗？

句子中【主】【谓】【宾】
元素的定位和结构调整

- 通过语义归一化来把握句子含义和关联各种表达方式
- 通过定位和调整（主谓宾 定状补）等句子元素，生成句法依存树来理解句子结构



文本挖掘技术难点三：种类繁多的歧义语义的处理

咬死了猎人的狗

省略省略了【咬死】动作的主语

做手术的是他的父亲

【做】的具体含义有歧义

五个公司的工程师

【五个】修饰的对象

他的仪表很不好

【仪表】是指器材还是外貌

乒乓球拍卖了

【球拍】存在歧义切分问题

小张欺负了小王，老师
喊了他家长

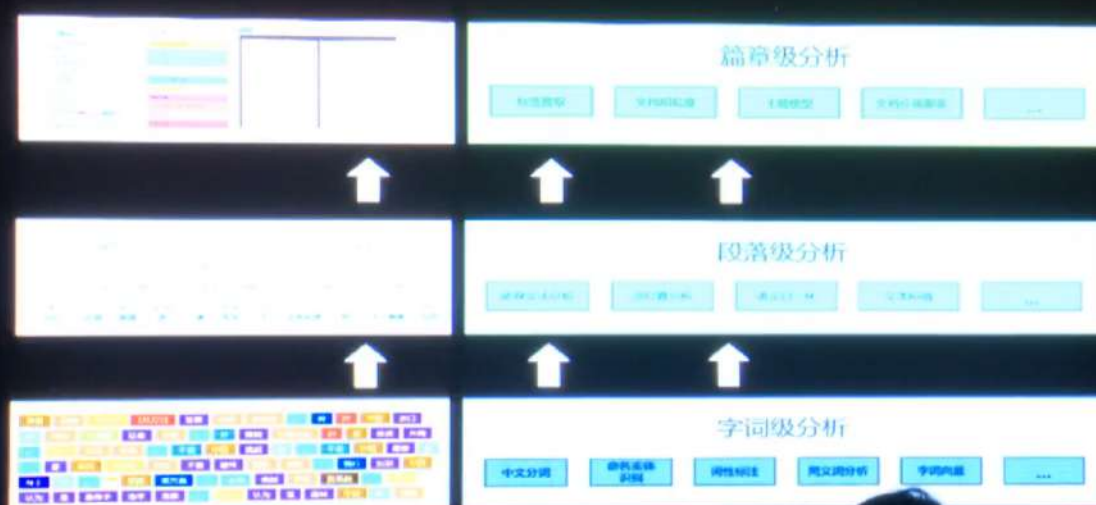
【他】的指代存在歧义

- 歧义种类非常多，在词语、句子、指代、修饰范围等各种情况下均会发生
- 常见的“省略”、“双关”、“反讽”、“假借”、“暗喻”等说法，加大了正确理解文字的难度
- 必须要从词法、句法、语义、上下文、以及领域知识等方面共同处理来消除歧义问题

计算机与人脑有类似的文本阅读处理过程



经典的文本分析应用的三个层次



以NLP技术为核心，构建服务于多场景的产品矩阵

文本挖掘引擎

让计算机具备文字阅读能力，帮助
客户自动化处理海量文本数据

垂直搜索引擎

利用知识图谱和语义分析技术
为客户搭建出高效精准的智能
搜索系统

智能推荐引擎

为用户提供千人千面的个性
化推荐内容有效提升点击率、
转化率及用户粘性



达观数据 文本智能处理专家
DATA GROUP

文本挖掘引擎：应用场景及价值

常见文本类型



常见操作方式



+



提升处理效率



提高准确率



节约人力资源

自反馈+自学习

达观文本挖掘引擎的优势



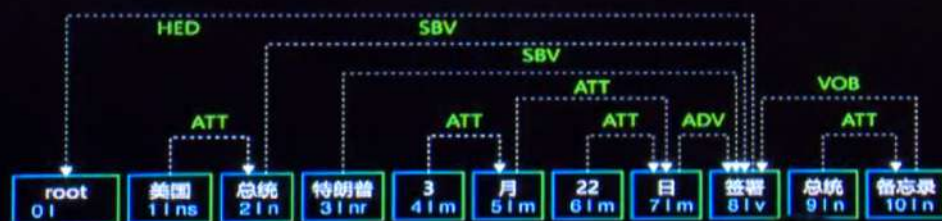
达观领先的
文本算法



丰富的
文本语料积累



自学习机制
进化模型



识别合同的风险

文本挖掘引擎的核心功能：文档智能审阅

为企业自动化抽取文档的关键信息、
对比不同版本的文档差异、智能纠正
错误文字内容，以及发现文书中潜在
的法律风险



智能审阅
内容一致性
文字规范性
引用相关性
合规性检查
关键内容比对

风险类型
条款冲突
条款不一致
无法律依据
条款冲突

风险等级
高风险
中风险
低风险
低风险

判断依据
《合同法》：双方经协商一致，
最终确定合同总价为【币种】
【大写】【小写金额】。
《合同法》：第一百二十七条
因不可抗力不能履行合同的，
根据不可抗力的影响，部分
或者全部免除责任，但法律
另有规定的除外。

达观审核算法

智能纠错

智能识别

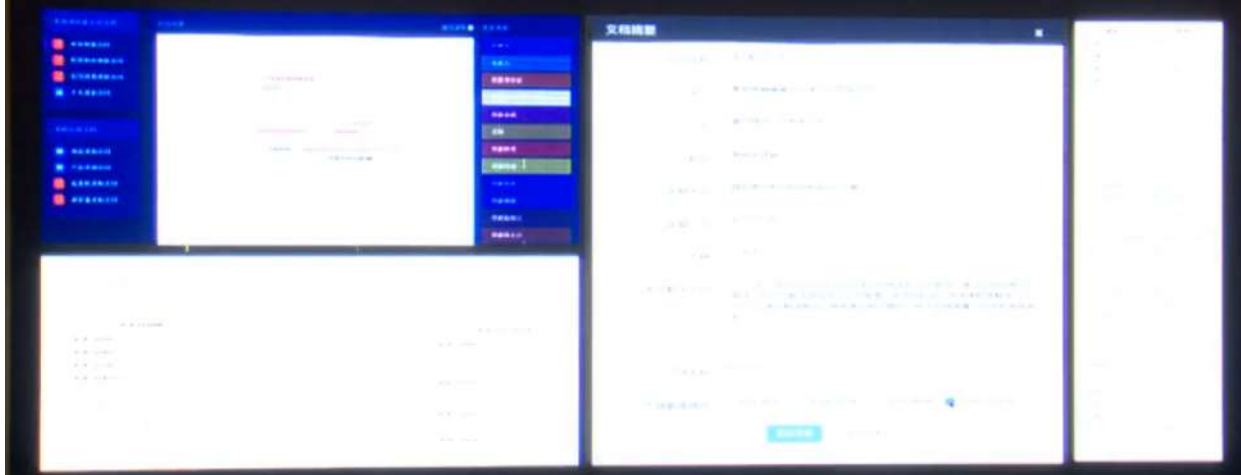
智能比对

智能审阅

文本挖掘引擎案例：合同智能审阅

国内某知名科技金融集团 合同智能审阅	业务需求 自动提取财务合同、单据等财务相关合同文档的关键信息，代替人工进行信息录入、核对、提高效率 此后，集团信托也采购了类似系统	项目周期 OCT - DEC 2017 - 2017	Deloitte. 德勤智能文档审阅
项目目标 完成40多个租赁合同字段的自动提取，同时提供纸质与电子版的多文档对比	竞标情况 在短短一周内给出的POC效果水平，远超其余2家供应商在一个月达到的效果，得到集团上下的一致认可	实现效果 95.1% 94.5% 准确率Precision 召回率Recall	华泰证券 智能信息抽取平台
项目收益 单份文档审阅操作节省90%时间 直接减少200人的文档用工人员成本	客户评价 达观在对接过程中，响应速度很快，服务态度好，技术扎实，提供精准高质量的内容处理服务，系统结合实际业务场景，上线时间之快大大超出预期。功能全面，还配套提供了文档标注系统，为以后应对其他新类型的文档做好了准备		法大大 Fadada.com 合同智能处理合作

文本挖掘引擎案例



搜索关键词-->到搜索文档

垂直搜索引擎：应用场景及价值





达观垂直搜索引擎的优势



分析邮件+word+pdf

垂直搜索引擎案例：MRO工业品行业搜索

MRO工业品行业 智能搜索引擎系统	业务需求 期望能自动提取客户订单中的商品 相关信息，并同商品库中类似商品 进行自动匹配，快速输出报价结果	项目周期 APR - MAY 2018 - 2018	 HUAWEI CSO搜索平台
项目目标 实现Excel、PDF、Msg、传真等 多种格式的客户订单商品识别和检 索，实现客户询价商品的快速报价 响应	竞标情况 根据客户提供的样例数据，在一周 内给出了测试DEMO，准确率超过 80%，效率和准确率得到了客户 的高度赞赏，双方迅速进入商务阶段	实现效果 支持Excel、PDF、Msg和传真等 多种格式的询价单商品识别和检索， 识别检索准确率均超过 90%	Haler HR简历搜索
项目收益 秒级将客户的询价单转化为报价单 极大提升了客服人员工作效率 缩短了客户的采购周期，提高了客 户品牌价值	客户评价 工业品的查找一直是客户的痛点，商品名称、型号、规格五花八门，客服人员 查找起来费时费力，有幸接触达观数据后，达观数据强大的技术实力和高效 的工作作风，快速实现了我们智能询价的功能，后续我们希望同达观在 更多AI应用中展开合作		 专利搜索

智能推荐引擎：应用场景及价值

推荐内容及场景

 商品 面向电商领域	 资讯 面向媒体领域	 简历 面向HR领域
 文件 面向法律领域	 服务 面向运营商	...

+

推荐方式

 个性化推荐	 相关推荐	 热门推荐
---	--	--



点击率提升



人均PV提升



用户停留时间延长



转化率增加

智能推荐引擎案例：招商银行



招商银行

掌上生活个性化推荐

业务需求

- 实现内容采集、个性化推荐和后台管理一体化
- 挖掘长尾内容价值，将推荐系统构建的用户画像、标签系统等用于更多管理运营场景
- 整个系统规模较大，技术架构层次和环节较多

招商银行

重大事件智能推荐

实施过程

通过用户对于金融产品和服务的相关反馈，依靠先进的自然语言处理技术进行正负面判断，深度挖掘用户反馈优化运营

CHANGHONG 长虹

液晶电视推荐平台

实现效果

提供千人千面的个性化推荐并应用于首页发现头条中
上线效果明显，核心的用户平均点击阅读数指标提升了3倍
为工具类app成功引流，提高用户留存率

澎湃

新闻推荐系统

达观拥有业界知名的文本技术工程和研究团队



陈运文
CEO

ACM、IEEE、CCF 高级会员
斯坦福大学商学院教授
腾讯文学高级总监
百度的核心技术工程师
复旦大学计算机博士



纪达麒
CIO

中国计算机学会会员
斯坦福大学博士
盛大文学技术总监
百度的工程师
北京邮电大学硕士



冯佳妮
COO

搜狗互联网COO
盛大云计算事业部总监
爱奇艺技术副总裁
品优购大/数据经理
中国政法大学教授



魏洁
副总裁

原华为AI产品部负责人
中国移动研究院AI项目主任
知乎首席技术专家
复旦大学计算机博士
4V100高级工程师



桂洪冠
副总裁

工程架构专家
中国计算机学会会员
搜狗高级架构师
搜狗AI产品部技术专家
中国计算机博士



姚学锋
副总裁

AI产品业务专家
SAI中国总经理
搜狗AI产品部技术负责人
复旦大学计算机博士

达观数据现为复旦大学计算机学院校外硕士培养基地，陈运文等兼任复旦大学校外研究生导师

聘任中文信息学会理事、知名文本挖掘教授、复旦大学博导黄孟芳教授担任首席技术顾问

聘任知识图谱知名学者、复旦大学教授、博导肖仰华教授担任高级技术顾问

达观数据
DATA GROUP

文本智能处理专家

达观数据蓬勃发展 位于中国文本智能处理领域的最前列

达观数据是一家专注于 文本自动化处理的创业企业

达观专注于为广大企业客户提供文本自动阅读、内容抽取、自动纠错、知识关联、搜索和推荐等网络文本智能技术服务，让计算机代替人工来处理企业内的各类文书资料，帮助企业提升自动化水平。

权威认证的人工智能服务， 可充分保障客户业务实践与业务安全

国家级高新技术企业，并拥有CMMI3软件成熟度认证、ISO9001质量管理体系认证、国家双软认证等全面的行业资质，并获中国人工智能创业企业30强等殊荣。

为企业提供一整套 可私有化部署的文本自动处理系统

为金融、法律、电商、传媒、制造等企业的合同、公文、报告、新闻提供自动化阅读、解析、核心信息抽取等系统，提升企业运行效率。



达观数据 文本智能处理专家

达观数据已与众多知名投资机构 跨国企业 研究院所达成战略合作

受到中国顶级投资机构投资和关注

获得软银赛富、真格基金、方广资本等国际最著名投资机构多轮投资



与多家国际知名科技企业建立合作关系

成为微软加速器二期成员，SAP创新合作伙伴，联想之星第九期成员，普华永道创新企业加速营成员



与中国知名高校、科研机构保持良好的产学研关系

成立复旦大学校外实习基地，中国计算机学会(CCF)会员单位，中文开放知识图谱 (OpenKG.CN) 发起者之一，浦东软件园加速器成员



达观数据 文本智能处理专家

高度重视核心技术研发 取得累累硕果



在文本智能处理领域，达观已获得45项国家发明专利与软件著作权

- 自主研发的核心算法已申请45项国家发明专利，文本智能处理技术处于行业领先地位
- 核心技术已成功服务近50个细分行业，涵盖百家企业



翻译出版多本著名人工智能著作，并引起业内强烈反响

- 翻译人工智能经典著作《智能Web算法》，把机器学习技术应用到工业界的先行者
- 参与撰写《数据实践之美》，与百度、腾讯、IBM、埃森哲等企业分享技术经验与心得



在国际学术期刊和CSDN、Qcon、51CTO等国内知名技术社区发表高质量学术文章

- 定期发布技术干货文章，多次被发布到网站首页置顶，受到网友广泛好评
- 在十几家技术媒体和社区开设专栏，累计发文上百篇，阅读人次