AiCon
全球人工智能与机器学习技术大会

iQIYI自然语言处理和视频大数据分析应用

吴友政

Powered by 爱奇艺技术产品中心


人工智能基础课
“通俗易懂的人工智能入门课”
王天一
博士 副教授
扫一扫，免费试读


AI技术内参
你的360度人工智能信息助理
洪亮劼
Etsy 数据科学主管
扫一扫，免费试读

个人介绍

- 教育背景
  - 中科院自动化所（NLPR，CASIA）毕业
- 工作经历
  - iQIYI
  - SONY China Research Lab.
  - University of Edinburgh，UK
  - NICT，Japan
- 参与项目/研究兴趣
  - 自然语言处理、语音助手、问答系统、机器翻译
  - 语音识别（Kaldi）
  - 商业智能

TABLE OF CONTENTES

- 理解视频内容
  - 中文词法分析
  - 预测（票房和流量）
- 理解视频用户
  - 舆情监测
  - 查询理解和意图搜索
- 总结

NLP

支持业务	泡泡	头条	VR	BI	广告
	搜索	推荐	用户画像	客服中心	审核平台
应用研发	查询理解	打标签	语音助手	舆情监测	热点事件
中文词法分析	分词	词性标注	词权重	实体识别	实体链接
数据挖掘	知识图谱	领域词典	同义词	情感词	embedding

视频大数据分析

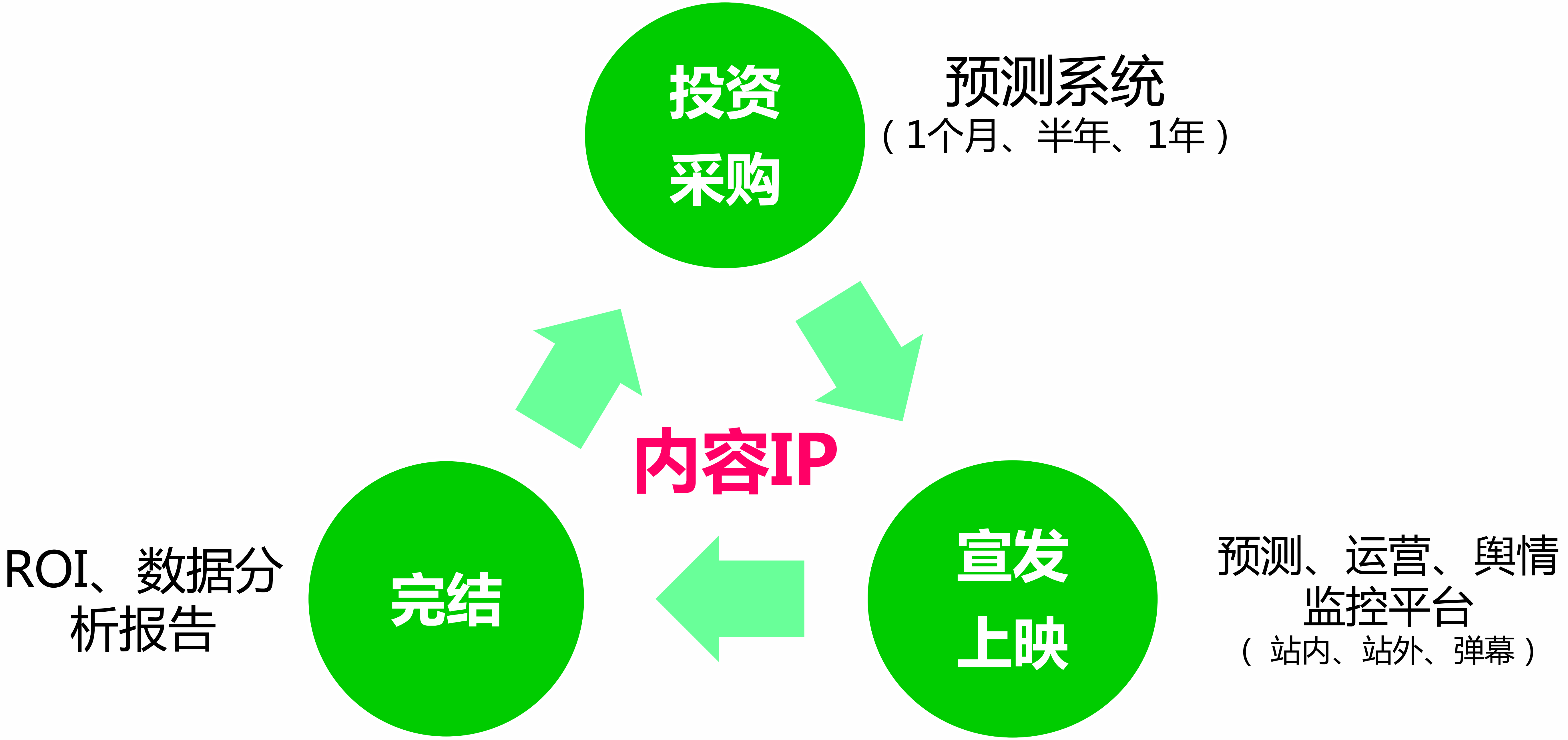

投资
采购
预测系统
（1个月、半年、1年）
内容IP
宣发
上映
预测、运营、舆情
监控平台
（ 站内、站外、弹幕 ）
完结
ROI、数据分
析报告

TABLE OF CONTENTES

- 理解视频内容
  - 中文词法分析
  - 预测（票房和流量）
- 理解视频用户
  - 舆情监测
  - 查询理解和意图搜索
- 总结

中文词法分析

鹿晗	演绎	帅气	陈长生	逆天改命	择天记	终	迎	结局
person	v	adj	character	idiom	album	d	v	n
0.758	0.284	0.402	0.577	0.476	0.756	0.0	0.0	0.370

实体识别简介

- 识别文本中具有特定意义的实体，并标注出其位置以及类型
  - <ALB>中国有嘻哈</ALB><PER>VAVA</PER>个性说唱<SONG>不想长大</SONG>
- 实体类型
  - 人名、地名、机构名、产品名、专有名词等
- 研究领域
  - 新闻、Social Network Service、Query、生物、金融等
- 现有系统
  - Stanford、哈工大、Jieba、百度云、阿里云、腾讯云
- 模型：序列标注问题

实体识别模型

Lafferty, et al.（ ICML2001 ）

Yang, et al.（AAAI2018）

MEMM、HMM

CRF

DNN
（CRF）

RL、AL
（DNN+CRF）

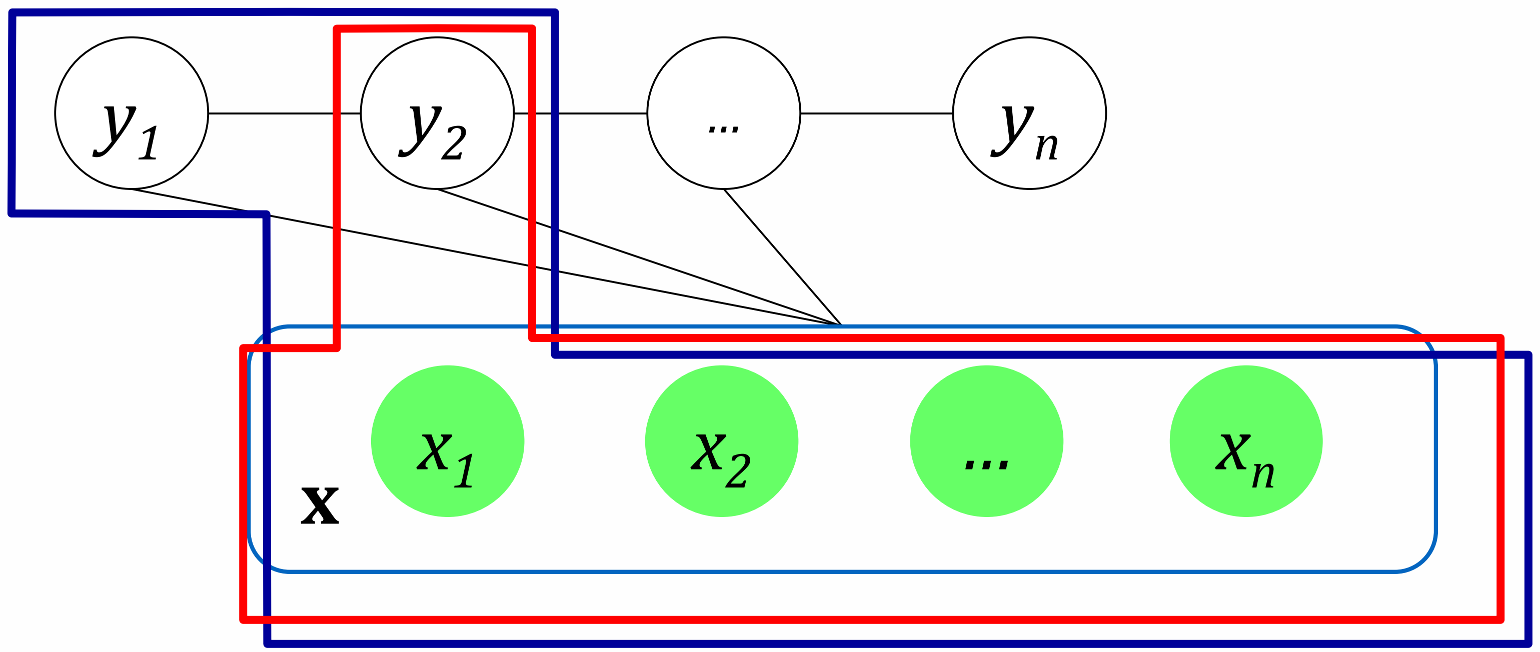
Bikel et al.（JML1999）

Collobert et al.（ JMLR2011 ）

CRF

- 基于词/字的序列标注（BIEO）

中国	有	嘻哈	VAVA	个性	说唱	不	想	长大
B-ALB	I-ALB	E-ALB	B-PER	O	O	B-SONG	I-SONG	E-SONG



$$p(\mathbf{y}|\mathbf{x}) \propto \prod_{t=1}^{n} \exp\left\{\sum_{l=1}^{L} \mu_l g_l(y_t, y_{t-1}, \mathbf{x})\right\} \cdot \exp\left\{\sum_{k=1}^{K} \theta_k f_k(y_t, \mathbf{x})\right\}$$

```
# Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-1,0]/%x[0,0]
U06:%x[0,0]/%x[1,0]
U07:%x[-1,0]/%x[1,0]

# Bigram
B
```

视频领域实体识别

- 实体类型
  - 影视剧名、游戏名、音乐名、人名、角色名
- 挑战
  - 歧义性多
    - 功夫、十二生肖、长城、非诚勿扰
  - 规律性弱
    - 西游记之孙悟空三打白骨精
  - 别名多
    - 琅琊榜之风起长林、琅琊榜2、风起长林, etc.
  - 缺少语料

解决方法（1）

- 词典
  - 实体词
    - 实时挖掘全网影视资料库（hourly update）
  - 新词（失恋哥、蓝瘦香菇）
- 语料
  - 通用和视频领域的准确人工标注语料
  - 半监督的自动标注语料学习（Liu, et al. ACL2011）
  - 启发式规则

长城鹿晗成最怂士兵被马特达蒙踢飞

Algorithm 1 NER for Tweets.

Require: Tweet stream i; output stream o.
Require: Training tweets ts; gazetteers ga.

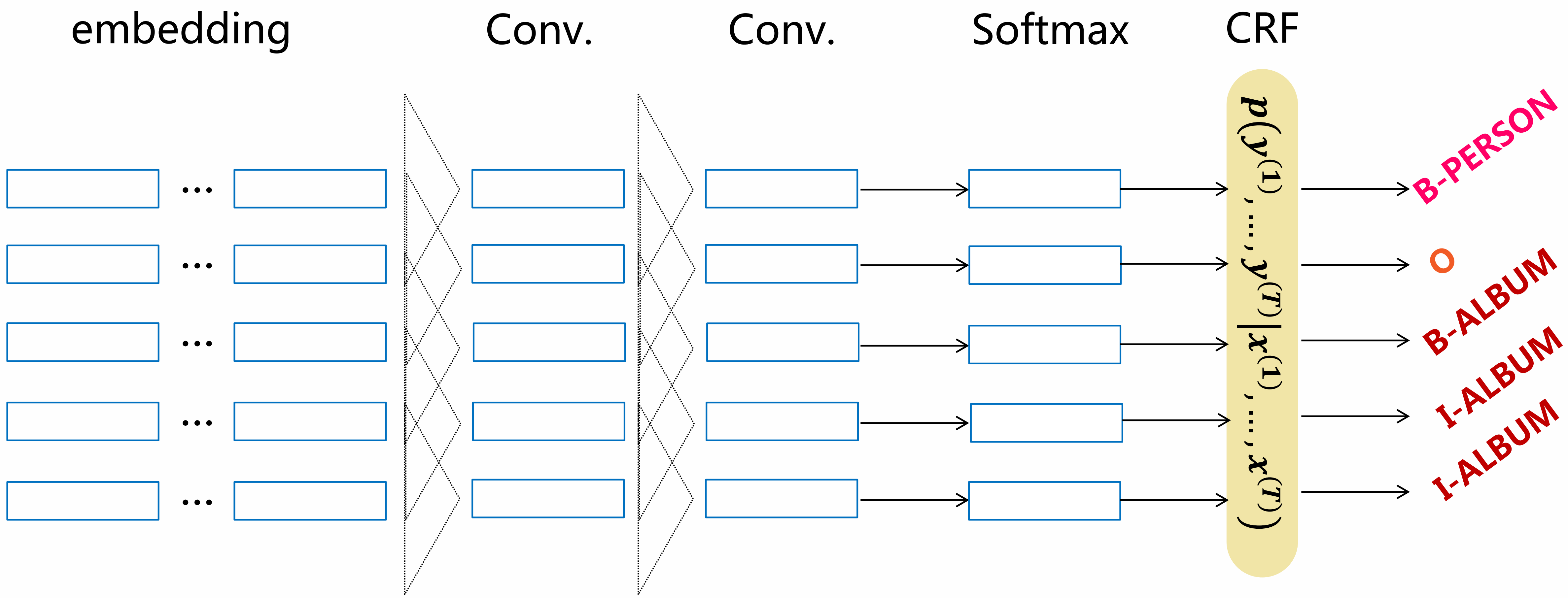
Initialize l_s, the CRF labeler: $l_s = train_s(ts)$.
Initialize l_k, the KNN classifier: $l_k = train_k(ts)$.
Initialize n, the # of new training tweets: $n = 0$.
while Pop a tweet t from i and $t \neq null$ **do**
  for Each word $w \in t$ **do**
    Get the feature vector $\vec{w}$: $\vec{w} = repr_w(w, t)$.
    Classify $\vec{w}$ with knn: $(c, cf) = knn(l_k, \vec{w})$.
    if $cf > \tau$ **then**
      Pre-label: $t = update(t, w, c)$.
    end if
  end for
  Get the feature vector $\vec{t}$: $\vec{t} = repr_t(t, ga)$.
  Label $\vec{t}$ with crf: $(t, cf) = crf(l_s, \vec{t})$.
  Put labeled result (t, cf) into o.
  if $cf > \gamma$ **then**
    Add labeled result t to ts , $n = n + 1$.
  end if
  if $n > N$ **then**
    Retrain l_s: $l_s = train_s(ts)$.
    Retrain l_k: $l_k = train_k(ts)$.
    $n = 0$.
  end if
end while
return o.

解决方法（2）

- 特征与模型

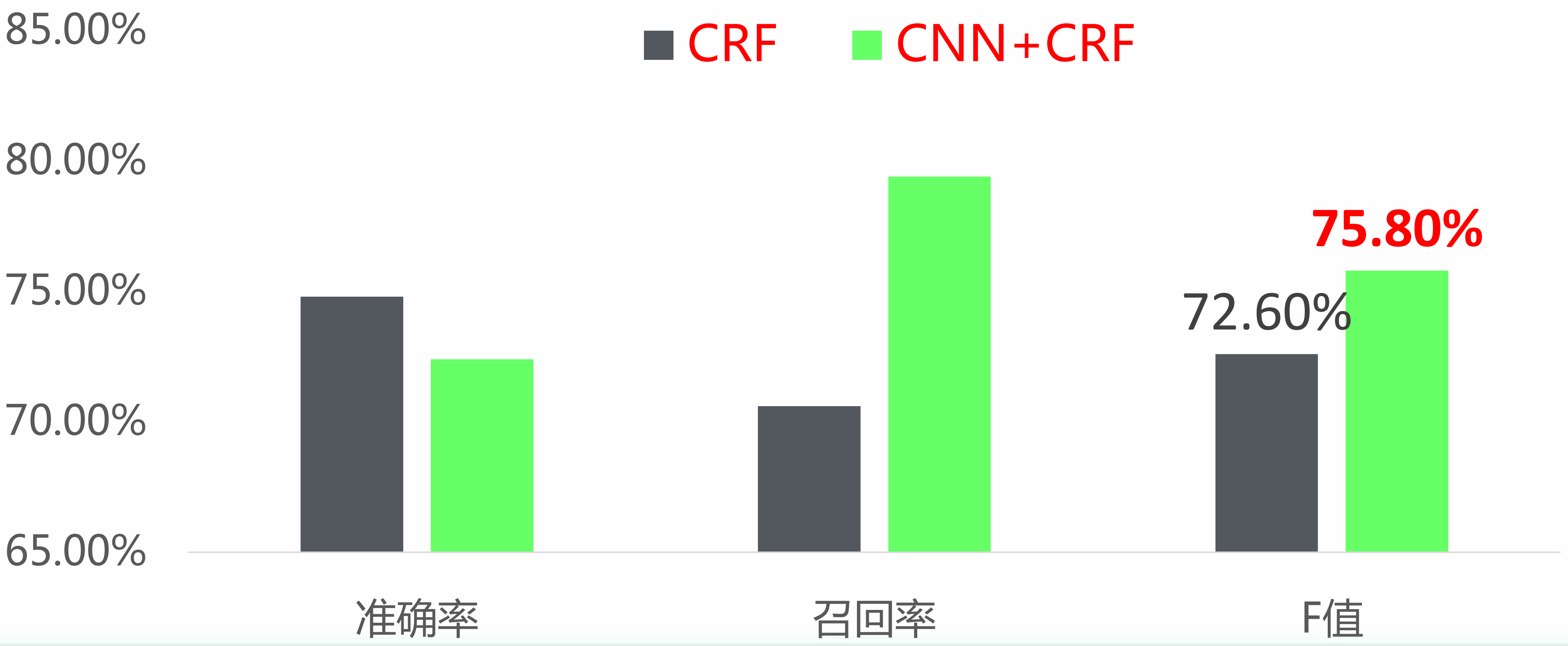
Word	POS	…	Brown Cluster
池子	$x_1^1(\mathrm{n})$	…	x_1^L
diss	$x_2^1(\mathrm{nx})$	…	x_2^L
中国	$x_3^1(\mathrm{ns})$	…	x_3^L
有	$x_4^1(\mathrm{v})$	…	x_4^L
嘻哈	$x_5^1(\mathrm{n})$	…	x_5^L

#L特征个数

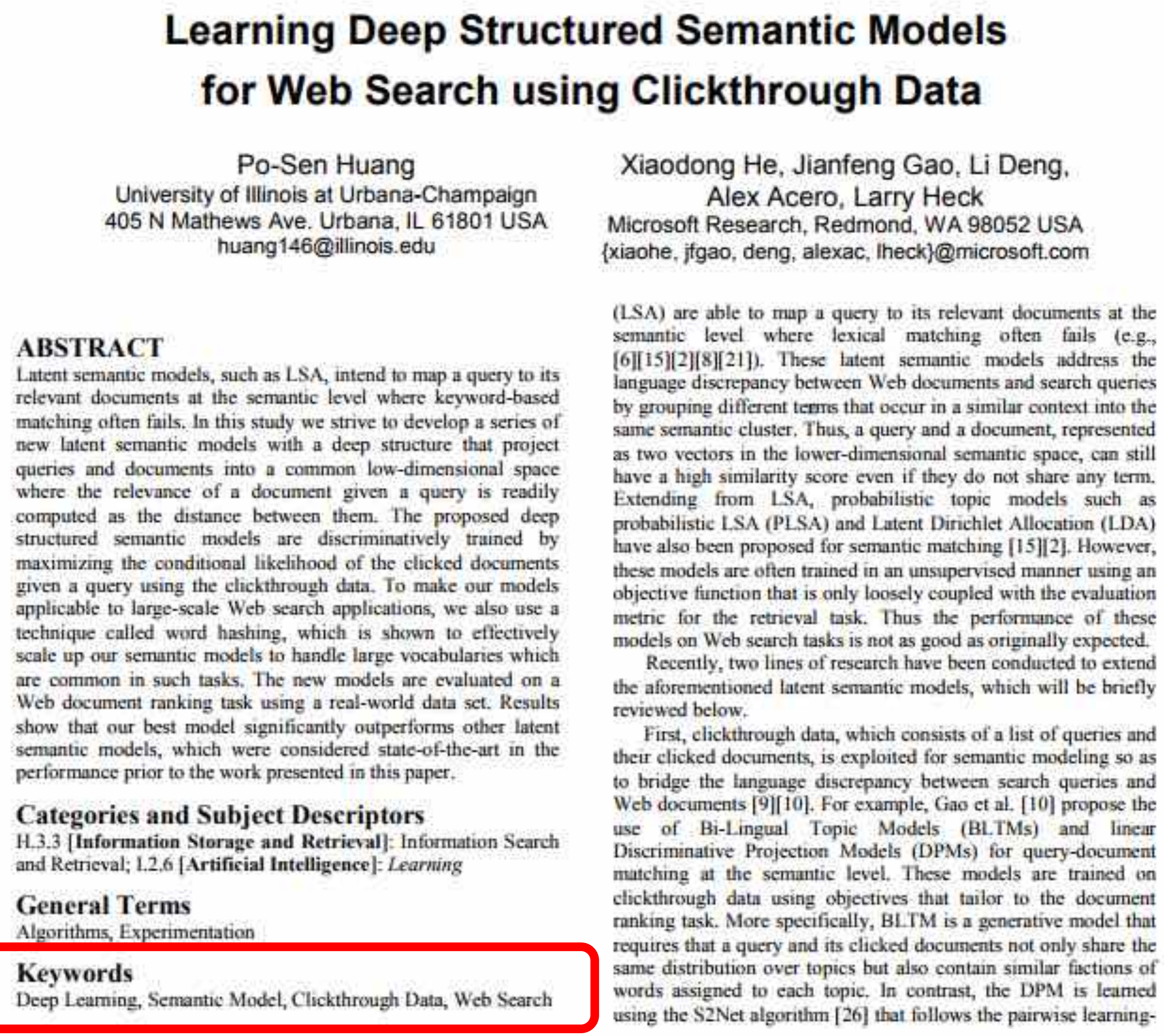


实体识别性能评测

- 分词效果
  - 视频：91.21%、 微博：94.35%
- 影视剧名识别效果：CRF vs. CNN+CRF



标签



Learning Deep Structured Semantic Models for Web Search using Clickthrough Data

Po-Sen Huang
University of Illinois at Urbana-Champaign
405 N Mathews Ave. Urbana, IL 61801 USA
huang146@illinois.edu

Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, Larry Heck
Microsoft Research, Redmond, WA 98052 USA
{xiaohe, jfgao, deng, alexac, lheck}@microsoft.com

ABSTRACT

Latent semantic models, such as LSA, intend to map a query to its relevant documents at the semantic level where keyword-based matching often fails. In this study we strive to develop a series of new latent semantic models with a deep structure that project queries and documents into a common low-dimensional space where the relevance of a document given a query is readily computed as the distance between them. The proposed deep structured semantic models are discriminatively trained by maximizing the conditional likelihood of the clicked documents given a query using the clickthrough data. To make our models applicable to large-scale Web search applications, we also use a technique called word hashing, which is shown to effectively scale up our semantic models to handle large vocabularies which are common in such tasks. The new models are evaluated on a Web document ranking task using a real-world data set. Results show that our best model significantly outperforms other latent semantic models, which were considered state-of-the-art in the performance prior to the work presented in this paper.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: *Learning*

General Terms

Algorithms, Experimentation

Keywords

Deep Learning, Semantic Model, Clickthrough Data, Web Search

(LSA) are able to map a query to its relevant documents at the semantic level where lexical matching often fails (e.g., [6][15][2][8][21]). These latent semantic models address the language discrepancy between Web documents and search queries by grouping different terms that occur in a similar context into the same semantic cluster. Thus, a query and a document, represented as two vectors in the lower-dimensional semantic space, can still have a high similarity score even if they do not share any term. Extending from LSA, probabilistic topic models such as probabilistic LSA (PLSA) and Latent Dirichlet Allocation (LDA) have also been proposed for semantic matching [15][2]. However, these models are often trained in an unsupervised manner using an objective function that is only loosely coupled with the evaluation metric for the retrieval task. Thus the performance of these models on Web search tasks is not as good as originally expected.

Recently, two lines of research have been conducted to extend the aforementioned latent semantic models, which will be briefly reviewed below.

First, clickthrough data, which consists of a list of queries and their clicked documents, is exploited for semantic modeling so as to bridge the language discrepancy between search queries and Web documents [9][10]. For example, Gao et al. [10] propose the use of Bi-Lingual Topic Models (BLTMs) and linear Discriminative Projection Models (DPMs) for query-document matching at the semantic level. These models are trained on clickthrough data using objectives that tailor to the document ranking task. More specifically, BLTM is a generative model that requires that a query and its clicked documents not only share the same distribution over topics but also contain similar factions of words assigned to each topic. In contrast, the DPM is learned using the S2Net algorithm [26] that follows the pairwise learning-

相关工作

DNN	CopyRNN（Meng, et al. ACL2017） DeepRNN（Zhang, et al. EMNLP2017）	生成文本中没有出现的 key-phrase
有监督	CRF Tagging（Gollapalli, et al. AAAI2017） Maui（Medelyan et al., 2010）, WAM（Liu, et al. CoNLL2011）	联合学习候选生成和排序
无监督	TextRank（Mihalcea and Tarau, 2004） CiteTextRank	候选生成、候选排序

相关工作

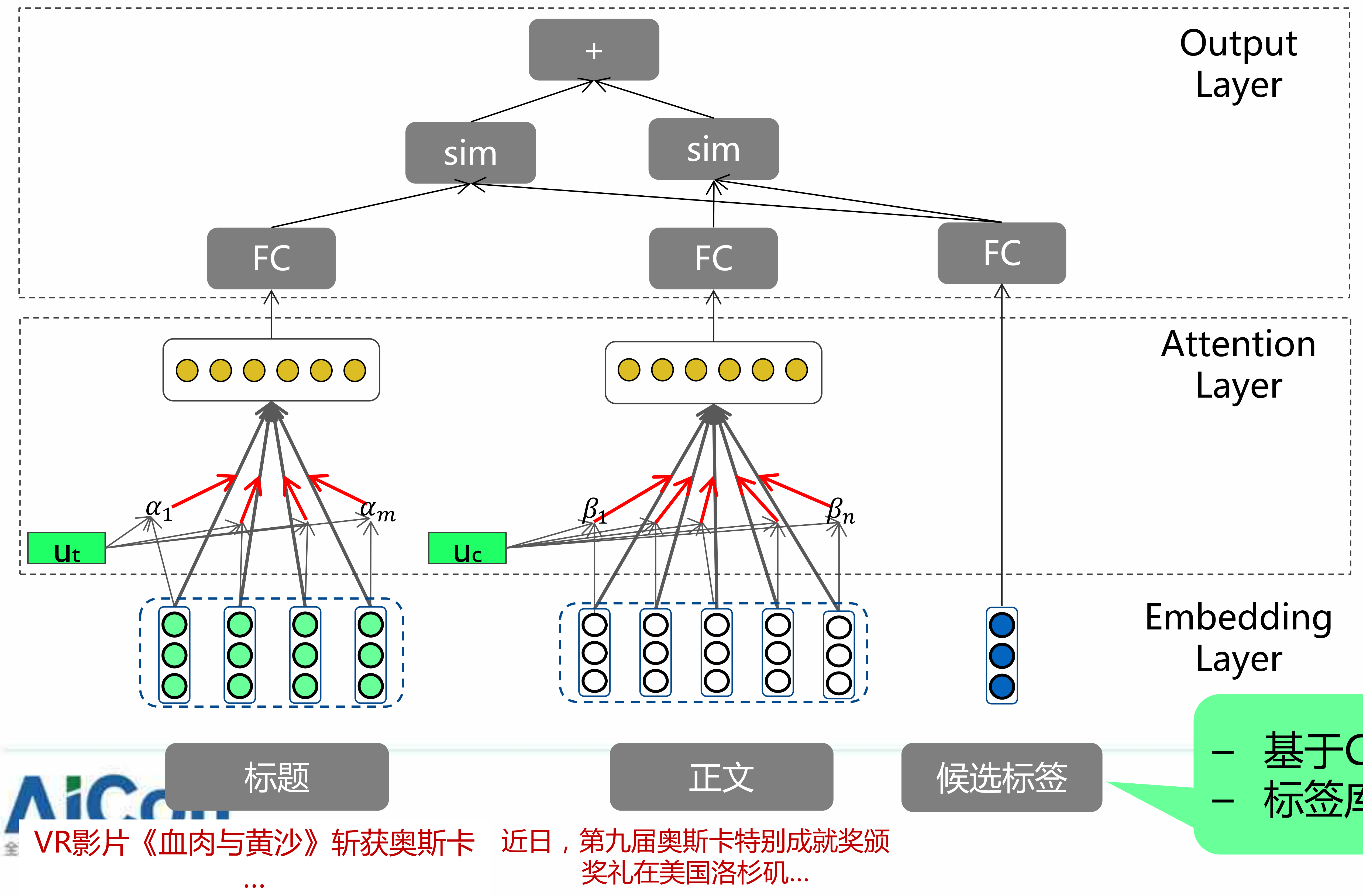
VR影片《血肉与黄沙》斩获奥斯卡特别成就奖

……第九届奥斯卡特别成就奖颁奖礼在美国洛杉矶举行，由著名导演亚利桑德罗·冈萨雷斯·伊纳里多（曾执导《鸟人》《荒野猎人》）拍摄的VR影片《血肉与黄沙》（Carne y Arena）获得了奥斯卡特别成就奖，成为首部获得奥斯卡奖的VR影片……

VR影片
血肉与黄沙
奥斯卡
特别成就奖
第九届奥斯卡
奖颁奖礼
亚利桑德罗·冈萨雷斯·伊纳里多
鸟人
荒野猎人

VR	**影片**	**《**	**血肉**	**与**	**黄沙**	**》**	**斩获**	**奥斯卡**	**特别**	**成就奖**	…
B-KP	I-KP	O	B-KP	I-KP	I-KP	O	O	B-KP	O	O	…

注意力模型



- 可以抽取没有文本中出现的词组作为标签
- 更好地计算语义相关性

- 基于CRF的keyphrase抽取
- 标签库中的热门标签

标签性能

	召回率	准确率	F值
CRF tagging [Gollapalli, et al. 2017]	43.3%	62.3%	51.2%
CNN [Kim, et al. 2014]	52.3%	61.5%	56.5%
Our Attention Model	60.6%	70.5%	65.2%

TABLE OF CONTENTES

- 理解视频内容
  - 中文词法分析
  - 预测（票房和流量）
- 理解视频用户
  - 舆情监测
  - 查询理解和意图搜索
- 总结

预测及应用场景

电视剧流量

版权采买，自制立项
广告售卖，内容定级
HCDN

- 长周期（**提前1年**）
- 中长周期（**半年**）
- 中短周期（**60日**）
- 播映中

+

电影票房

影业投资，版权采买
内容宣发

- 长周期（**提前1年**）
- 中长周期（**半年**）
- 中短周期（**60、30日**）

+

动漫 综艺
…

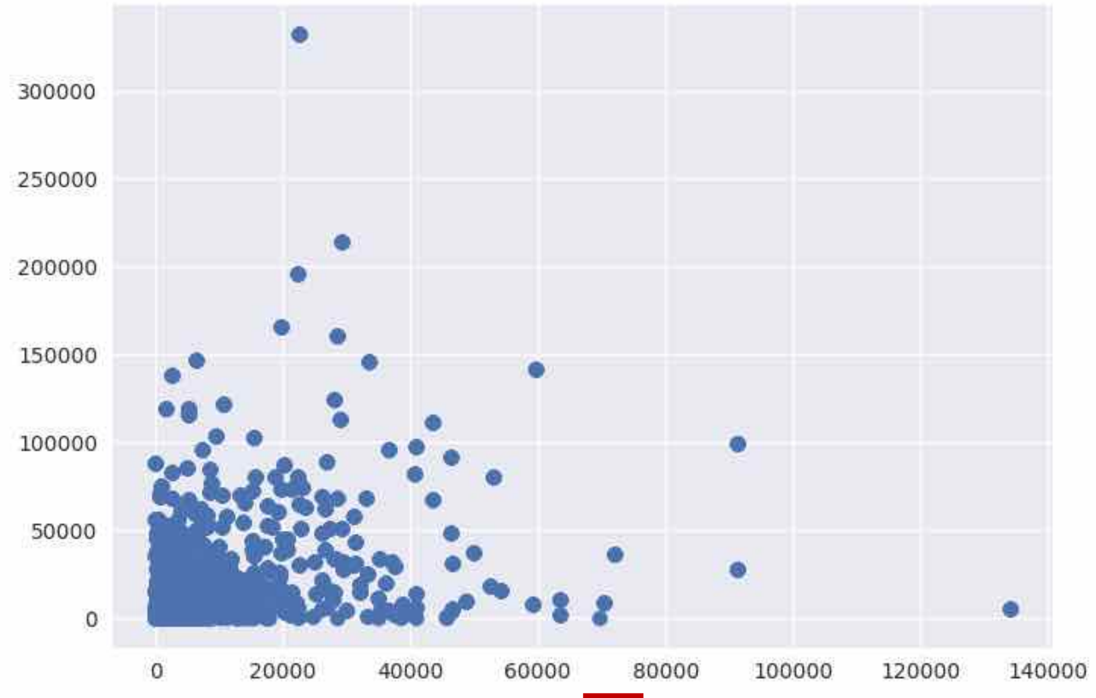
挑战与现状

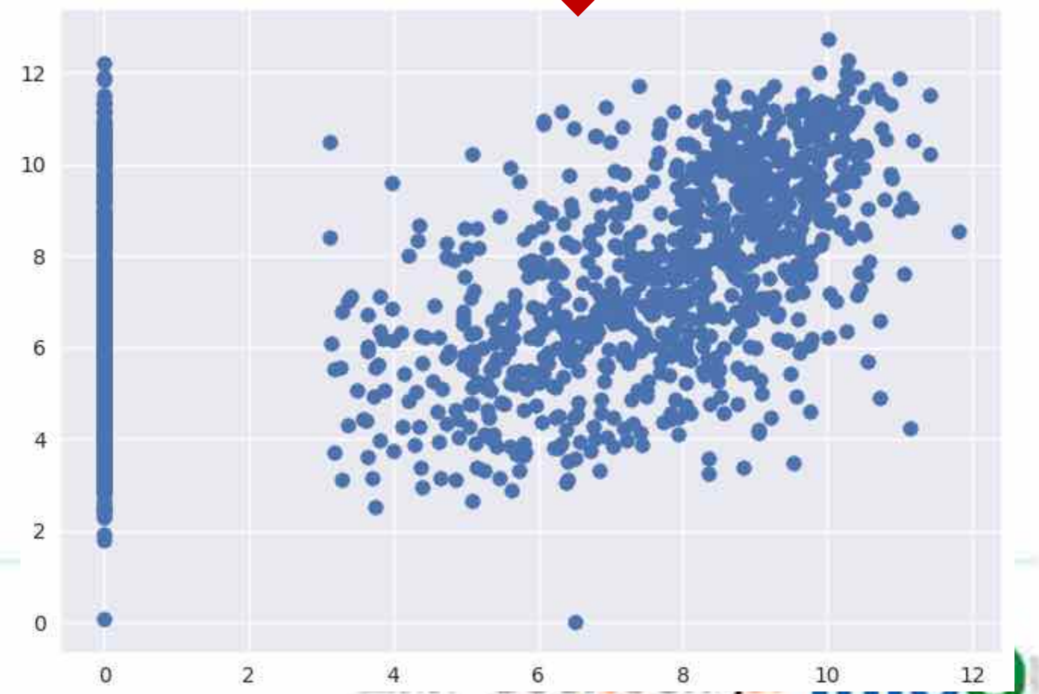
- 挑战
  - 样本少
  - 影响因素多
    - 同档期竞争的影视剧、卡司突发事件、天气、电影排片率
  - 样片/剧本理解的难度
- 电影票房
  - 国外
    - Google（上映前一周）：搜索、广告点击数据以及院线排片来预测票房（2013）
    - Epagogix（投资阶段）：分镜头剧本中提取30,073,680个特定的评分指标
  - 国内
    - 猫眼（上映影片）、百度（上映影片）、艾漫（投资）

基于集成学习的预测模型

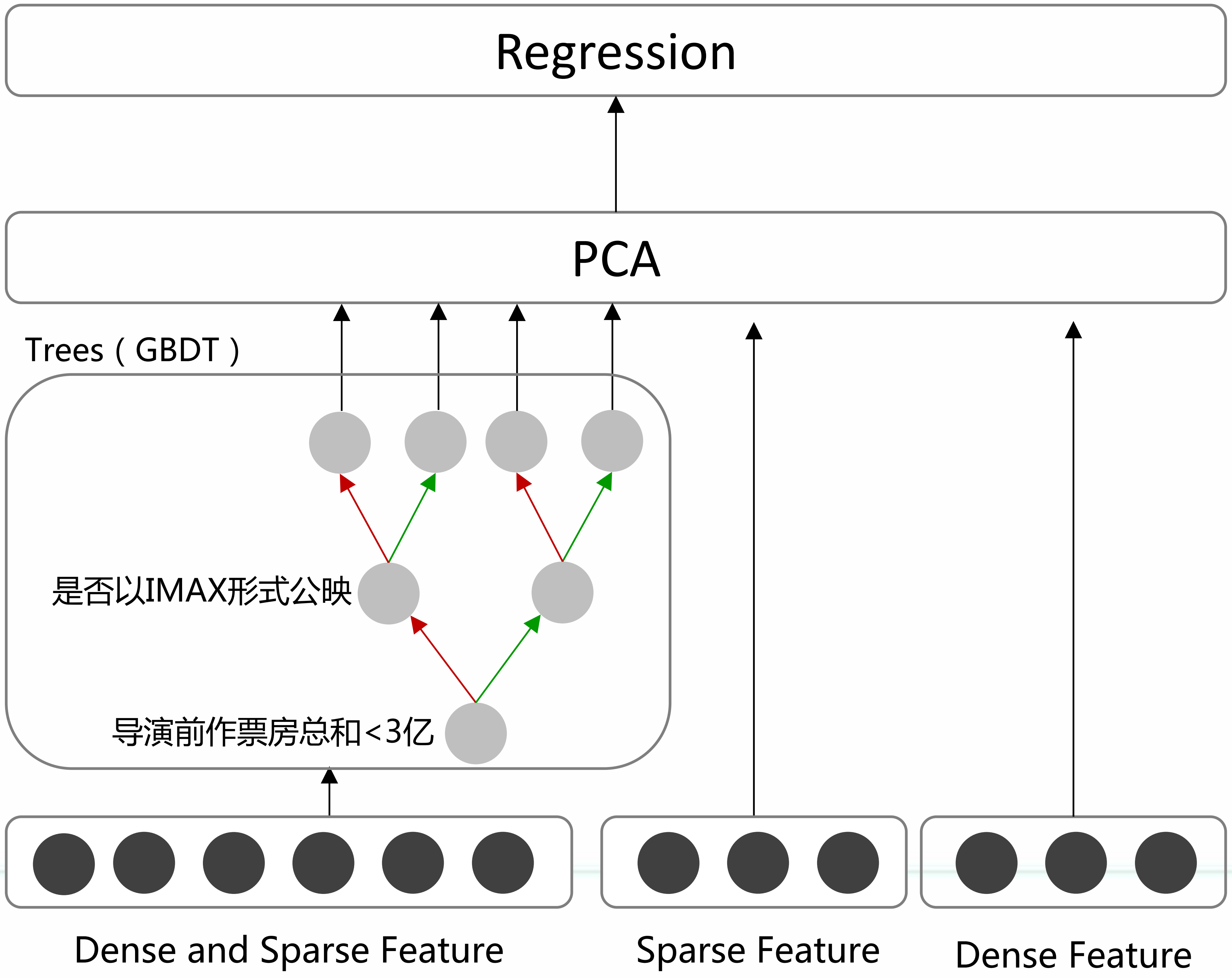
- 特征（**100+**）
  - 基本信息、历史前作、排播、搜索热度、社交热度
- 预处理
  - 缺失值处理、过采样、变换（log）
- 目标函数
  - 最小化log均方误差

$$loss = \frac{1}{2n}\sum_{i=1}^{n}(log\, y_i - log\, \hat{y}_i)^2 + \mu\|W\|_1$$



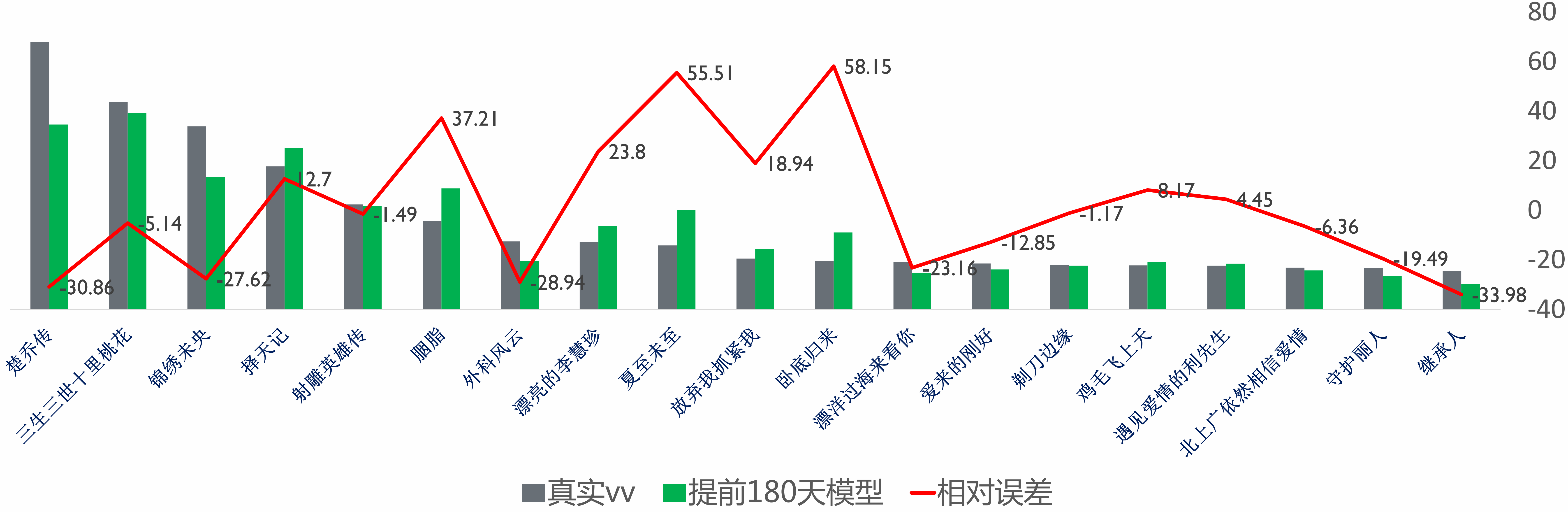


基于集成学习的预测模型



预测性能

- 电视剧流量预测R2准确率88%



预测性能

- 电影票房预测R2准确率81%

真实	4.34亿
预测	4.06亿
相对误差	6%

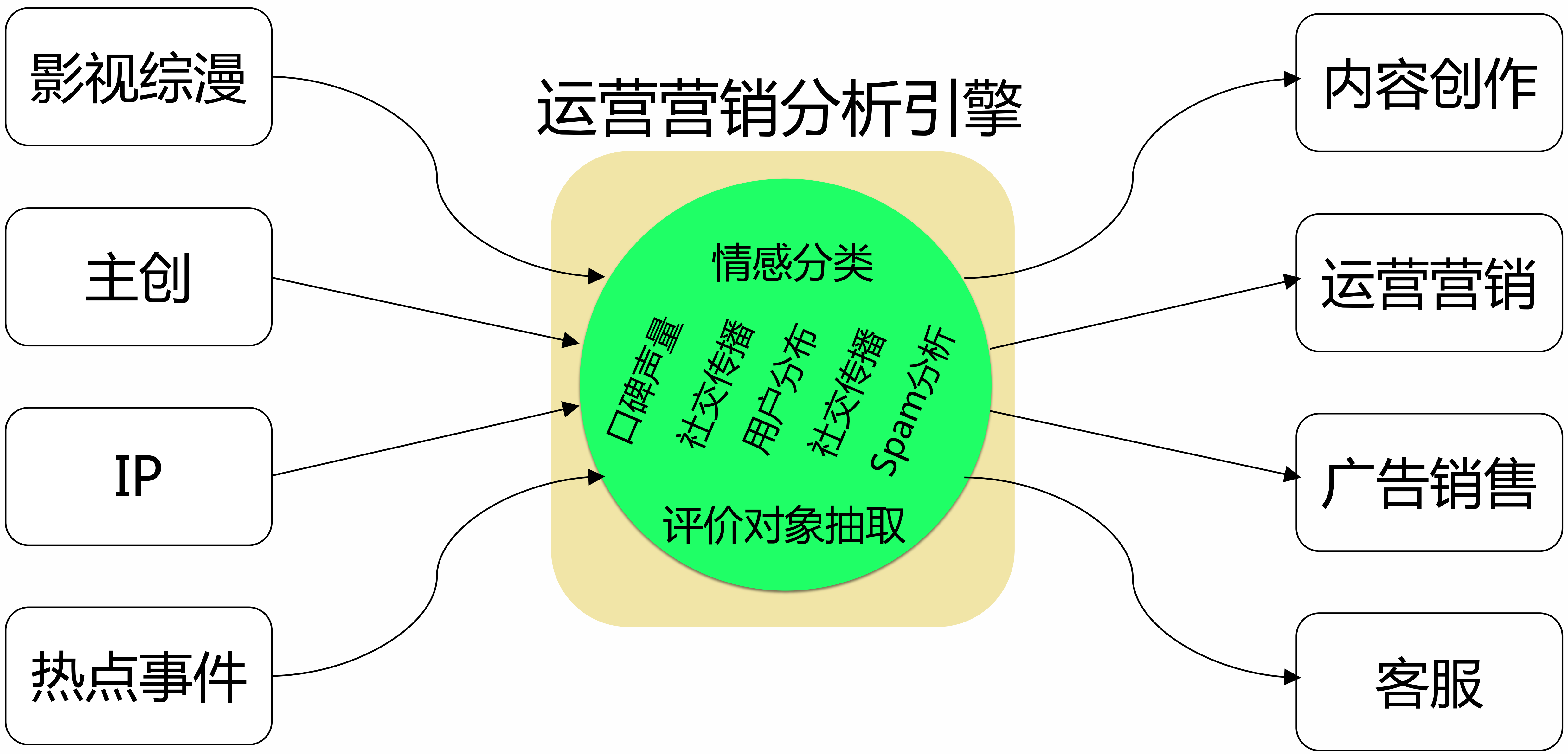
爆款

真实	13.27亿
预测	8.25亿
相对误差	37.8%

TABLE OF CONTENTES

- 理解视频内容
  - 中文词法分析
  - 预测（票房和流量）
- 理解视频用户
  - 舆情监测
  - 查询理解和意图搜索
- 总结

运营营销


影视综漫
主创
IP
热点事件
运营营销分析引擎
情感分类
口碑声量
社交传播
用户分布
社交传播
Spam分析
评价对象抽取
内容创作
运营营销
广告销售
客服

相关工作

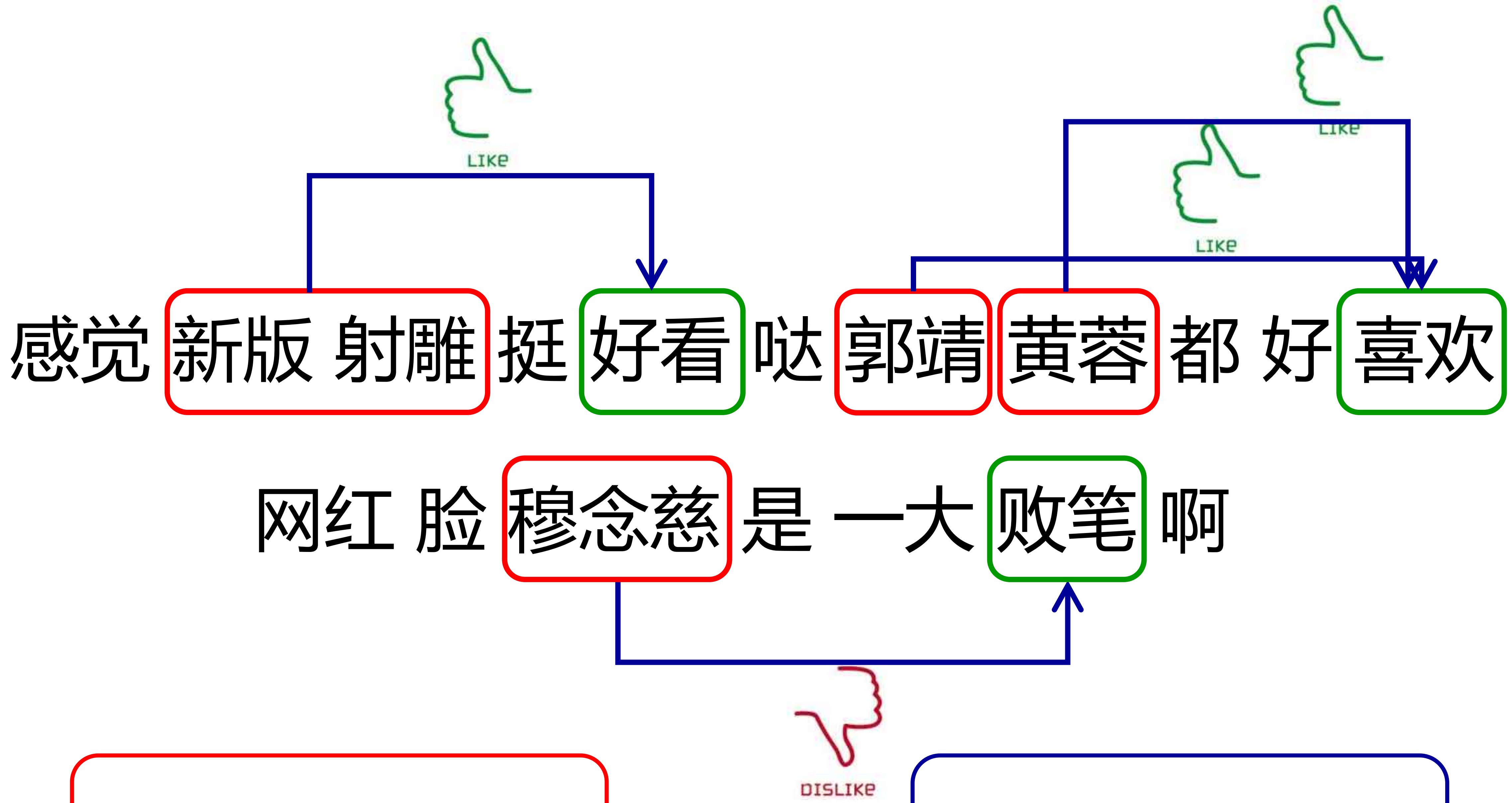
- COAE2016（中文倾向性分析评测）
  - Task A：微博观点摘要
  - Task B：影视评论的篇章级 - 句子级 - 词语级情感极性
- SemEval （2013~2017）： Sentiment Analysis in Twitter
  - Task A: Tweet Polarity Classification
  - Task B: Topic-based Tweet Polarity Classification
  - Task C: Tweet quantification

Tweet	Overall Sentiment	Topic-level Sentiment
Saturday without Leeds United is like Sunday dinner it doesnot feel normal at all (Ryan)	Weakly-Negative	Leeds United: Highly-Positive

视频场景下舆情分析任务定义

- Given (u_i, a_i, c_i)：u_i=user, a_i=album, c_i=评论
  - Spam
    - 水军账号识别
    - Spam帖子识别
  - Paragraph/Sentence
    - 识别段落/句子的观点倾向性：CNN、Bi-LSTM（**83%**）
    - **挑战：郭靖 黄蓉 都 好 喜欢 网红 脸 穆念慈 是 一大 败笔 啊**
  - Phrase（Aspect）
    - 评价对象、评价词抽取；关系抽取；评价对象聚合

Phrase级舆情抽取



BiLSTM+CRF

（ Lample, et al. NAACL2016 ）

BiLSTM+Attention

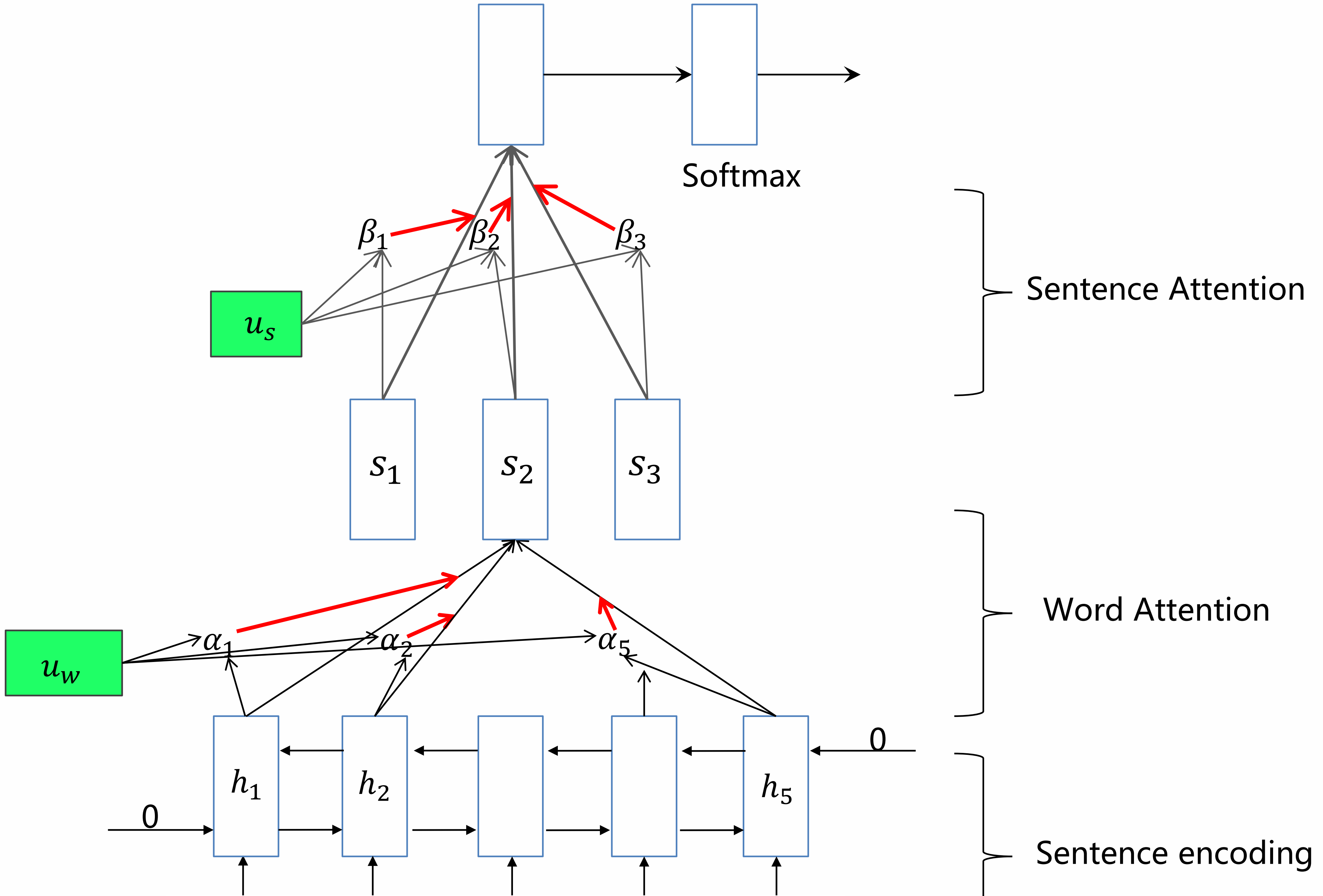
（ Zhou,et al. ACL2016；Lin, et al. ACL2016 ）

BiLSTM+Hierarchical Attention

$$s_i = \sum_i \alpha_i h_i$$

$$\alpha_i = \frac{exp(u_w^T u_i)}{\sum_j exp(u_w^T u_j)}$$

$$u_i = tanh(W_s h_i + b_s)$$



Distance to Opinion Word	4	3	2	1	0
Distance to Opinion Target	1	0	1	2	3
	郭靖	黄蓉	都	好	喜欢

DEMO

新浪微博
全量占比
战狼2(正向)
战狼2(负向)
战狼2(中性)
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
2017-07-16
2017-07-21
2017-07-26
2017-07-31
2017-08-05
2017-08-10
2017-08-15
2017-08-20
2017-08-25
2017-08-30
2017-09-04
2017-09-09
2017-08-11
战狼2(正向)：2686
战狼2(负向)：697
战狼2(中性)：5803

DEMO

视觉音效
评价词:515
uv:462
镜头好看
画面精致
画面好看
画面美
镜头恐怖
画面不错

剧情
评价词:1858
uv:1663
剧情不错
剧情精彩
剧情好看
剧情好
剧情紧凑
情节紧凑
结局紧

演员
评价词:3237
uv:2652
演技好
演技不错
喜欢李现
演技棒
李现帅
演技尴尬
喜欢顾影

TABLE OF CONTENTES

查询理解


中国有嘻哈第2期
推荐
旧时光
热点
电视剧
电影
生逢灿烂的日子
灿烂的日子：张嘉译苦恋车晓
更新至15集
文学
漫画
游戏中心
游戏直播
奇秀直播
电影票
商城
斗地主
麻将
应用商店
活动中心
没WiFi也尽情追剧，免流量限时抢！
推荐
热点
VIP会员
我的
泡泡


全网
李
搜索
李小璐
李连杰电影
李小璐出轨
李小璐出轨视频
李易峰
李连杰
李兰迪
李沁
李小龙
李小璐贾乃亮


全网
空天猎
取消
相关
最新
最热
高级筛选
已为您搜索“空天猎”，仍然搜索“天空猎”
空天猎
7.2
导演: 李晨
上映时间: 2017
播放至 33分59秒
播放源
继续观看
缓存
吴京曾被邀请出演《空天猎》 吴京用四个字回绝李晨
发布时间: 2017-10-08
上传者: 网游传奇
李晨范冰冰小两口哭了！《空天猎》票房惨淡竟是这些因素造成
发布时间: 2017-10-06
上传者: 空心城_1842


全网
中国有嘻哈·王者之路
取消
未检测到语音
试试这样说
我要买会员
播放楚乔传第十集
下载王者荣耀
按住 说出你想搜的

意图搜索

王菲的女儿	胡歌古装电视剧大全	演过黄蓉的演员有哪些	邓超老婆演过的电视剧

TABLE OF CONTENTES

- 理解视频内容
  - 中文词法分析
  - 预测（票房和流量）
- 理解视频用户
  - 意图搜索
  - 舆情监测
- 总结

总结

- 理解视频内容
  - 中文词法分析
  - 智能标签
  - 票房和流量预测
  - 热点事件发现和摘要
  - 智能审核（审核、标题党、软色情等）
- 理解视频用户
  - 舆情监测
  - Query Understanding
  - 助手与客服

Thanks!