



昆明理工大学
KUNMING UNIVERSITY OF SCIENCE AND TECHNOLOGY



多语言事件关联分析及观点挖掘研究

余正涛

昆明理工大学

2020年9月6日

目 录

1

研究背景

2

多语言知识图谱构建

3

多语言文本语义表征

4

多语言事件分析及观点挖掘

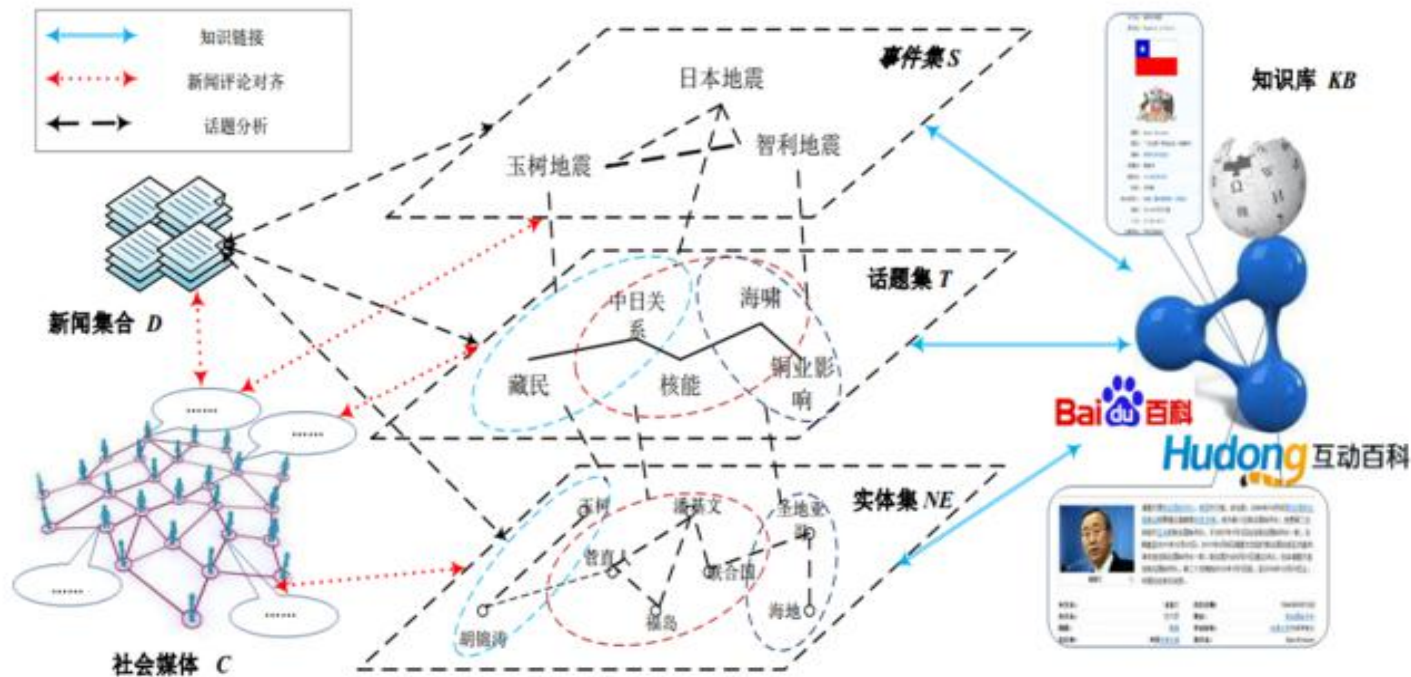
5

研究趋势



研究背景

国家之间的交流日益密切，共同关注的事件越来越多。及时获取不同国家的事件信息，把握不同国家的事件动态及事件间的关联关系，掌握不同国家对同一事件的观点，对于促进国际交流合作有着非常重要的意义。



#中越两国外长为边界界碑描红# 【中越陆地边界划界20周年 王毅同越外长行碰肘礼 亲手为界碑描红】据新华社南宁8月23日报道，国务委员兼外长王毅同越南副总理兼外长范平明23日在广西东兴出席中越陆地边界划界二十周年和勘界立碑十周年纪念活动。两国外长共同现场查看了边界界碑，并为各自的第一块界碑描 ...全文



Viet Nam Government Portal @VNGovtPortal · 8月23日
DPM, @FMPhamBinhMinh repaints the inscription on the boundary marker placed at the #MongCai Int'l Border Gate in Quang Ninh Province after he and Chinese FM Wang Yi co-chair 20th anniversary of signing of 🇻🇳🇨🇳 Land Border Treaty this morning

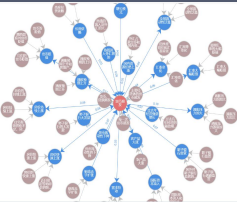




研究难点

研究现状

金融事理图谱



单语事件分析: 面向单语言的热点事件发现、跟踪、分析及预测等任务, 已形成较为完善的事件分析系统。

新浪舆情通
yuqing.sina.com

人民网
舆情监测室
yuying.people.com.cn

www.news.cn
新华网
www.xinhuanet.com

跨语言事件分析: 面向高资源语言的跨语言事件分析已有相关研究。

面向资源稀缺语言的跨语言分析

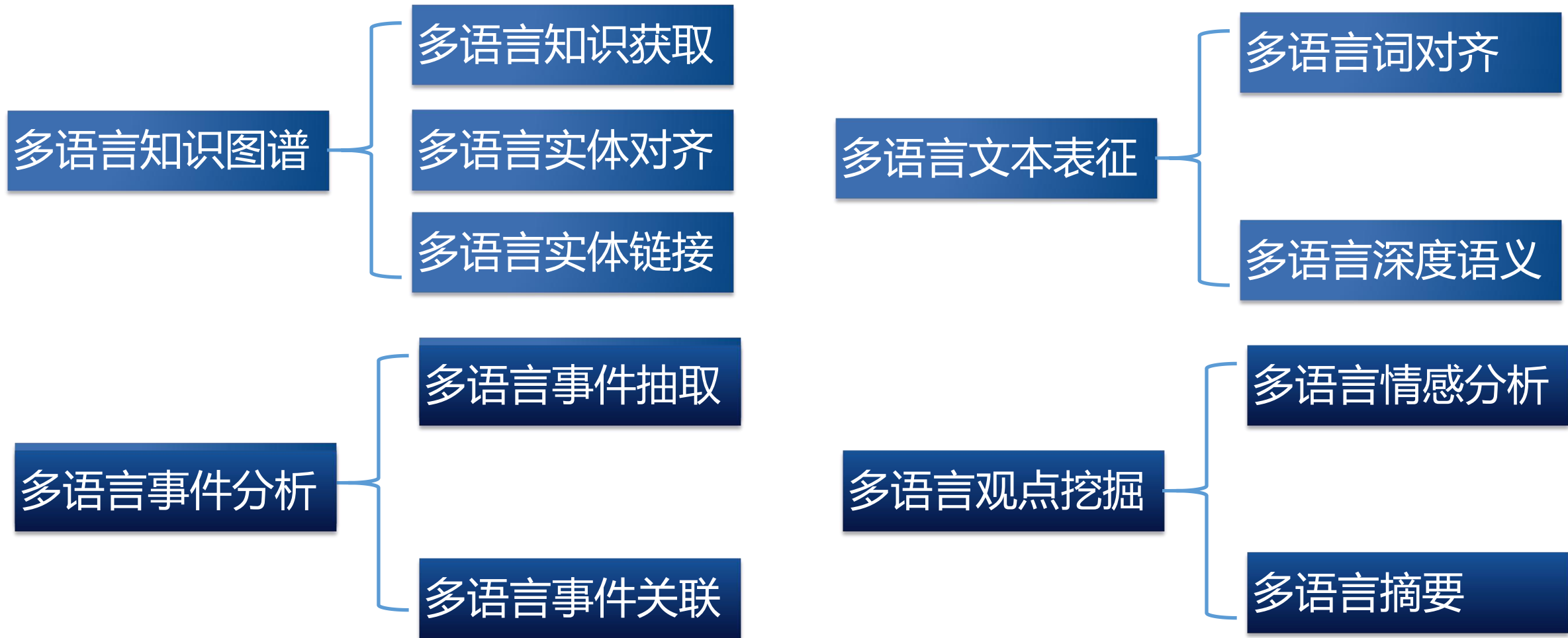


存在问题

- 多语言知识稀缺**: 如何构建**多语言知识图谱**, 为多语言事件分析等任务提供知识支撑。
- 多语言文本语义理解困难**: 低资源语言**翻译技术不成熟**, 如何实现**语义空间对齐**是难点。
- 多语言语料稀缺**: 面向东南亚等资源稀缺语言开展事件分析和观点挖掘面临**标注语料稀缺**, 有效实现**跨语言事件**是难点。



主要研究工作



目 录

1

研究背景

2

多语言知识图谱构建

3

多语言文本语义表征

4

多语言事件分析及观点挖掘

5

研究趋势



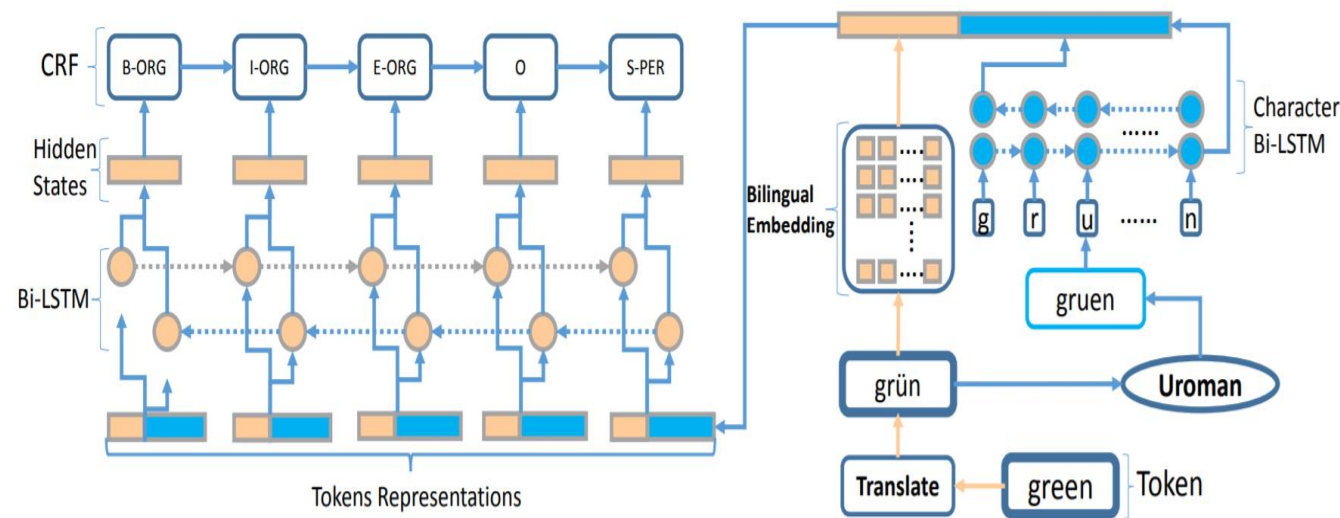
基于迁移学习的多语言实体识别

- **任务**：实体识别是构建跨语言知识图谱的基础。
- **问题**：低资源场景实体识别标注数据稀缺
- **方法**：利用高资源语言标注数据，迁移知识到低资源语言。首先基于**双语词嵌入**实现不同语言的语义对齐，然后利用**双语词典**增强不同语言的嵌入表征，最后实现跨语言实体识别。

Models	Spanish	Dutch	German
WikiNER*	60.55	61.60	48.10
CTNER*	51.82	53.94	50.96
CTNER*+	65.95	66.50	59.11
NCNER	72.37	71.25	57.76
Our Model	64.48	73.44	62.26

迈克尔·乔丹在1984年NBA选秀中于第1轮第3位被芝加哥公牛队选中，职业生涯曾效力于芝加哥公牛队以及华盛顿奇才队。新秀赛季当选NBA年度最佳新秀。

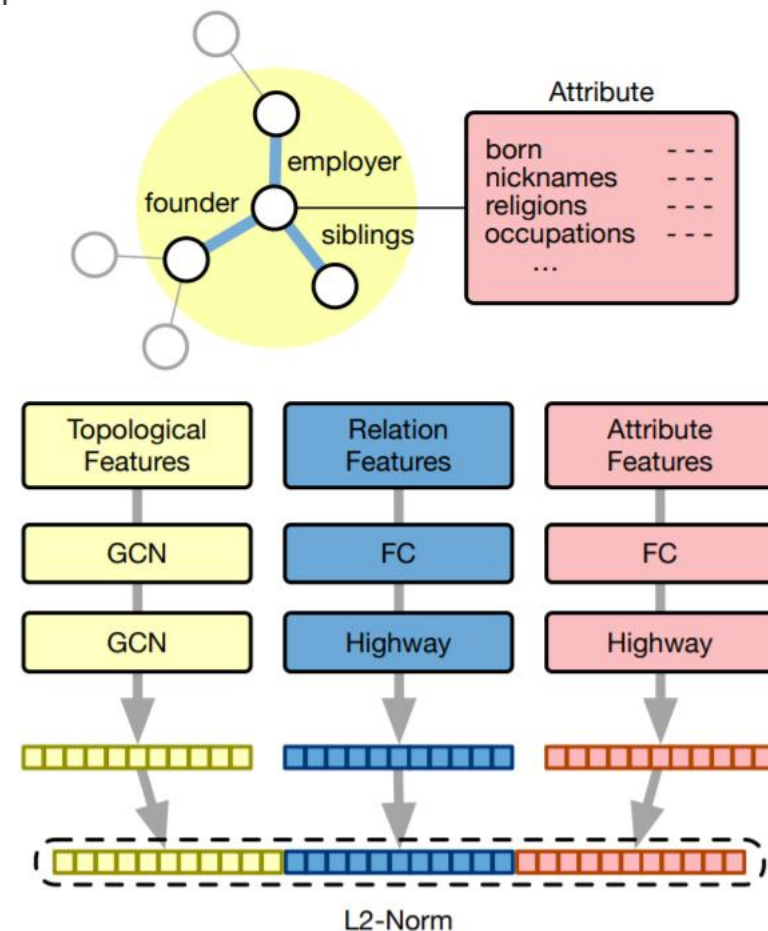
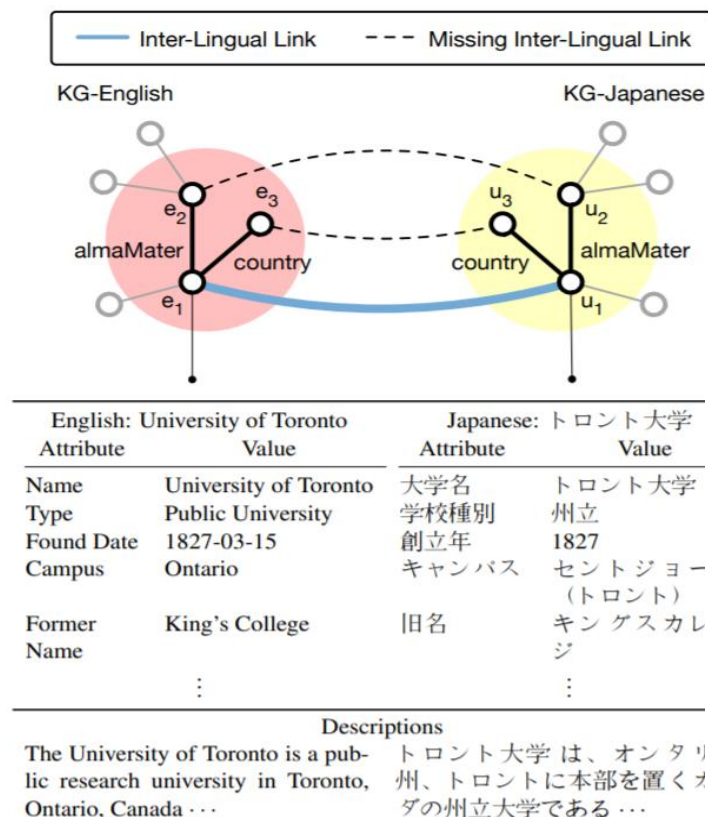
1	Michael_Jordan được lựa_chọn bởi Chicago_Bulls trong vòng thứ ba của NBA Draft năm 1984
2	Anh đã chơi cho Chicago_Bulls và Washington_Wizards trong sự_nghịệp của mình
3	Anh được chọn là Tân_bình NBA của năm trong mùa giải tân_bình của mình .





基于多视角的多语言实体对齐

- 任务**: 将源语言实体对齐到目标语言中
- 问题**: 实体存在歧义性
- 思路**: 将实体映射到高维语义空间, 将实体对齐任务转换为**相似度比较任务**。
利用实体的**结构信息**、**关系信息**以及**属性信息**实现多视角的特征嵌入。

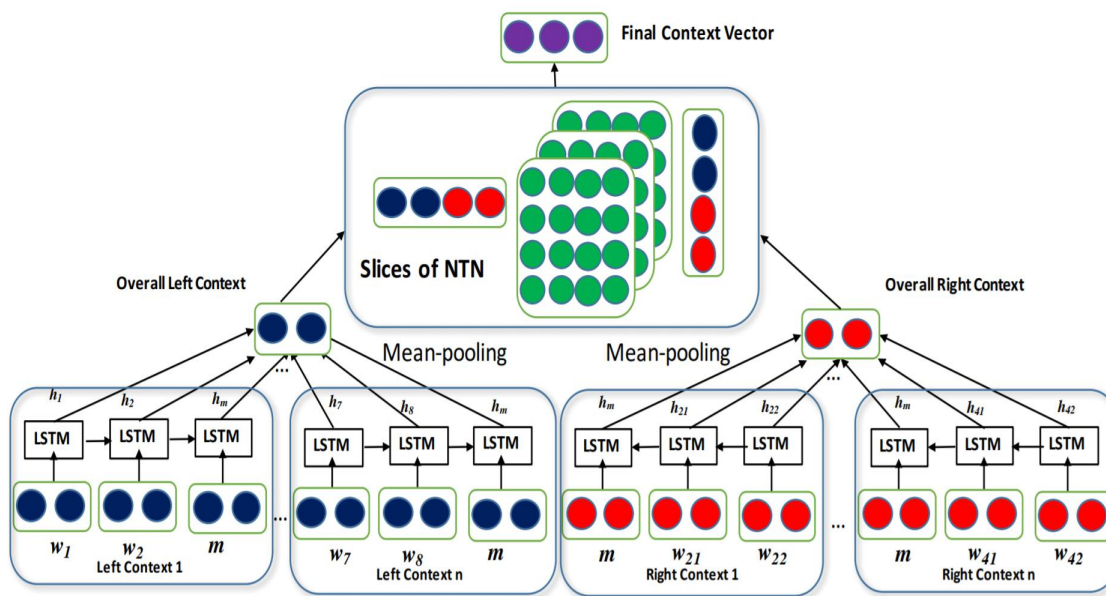
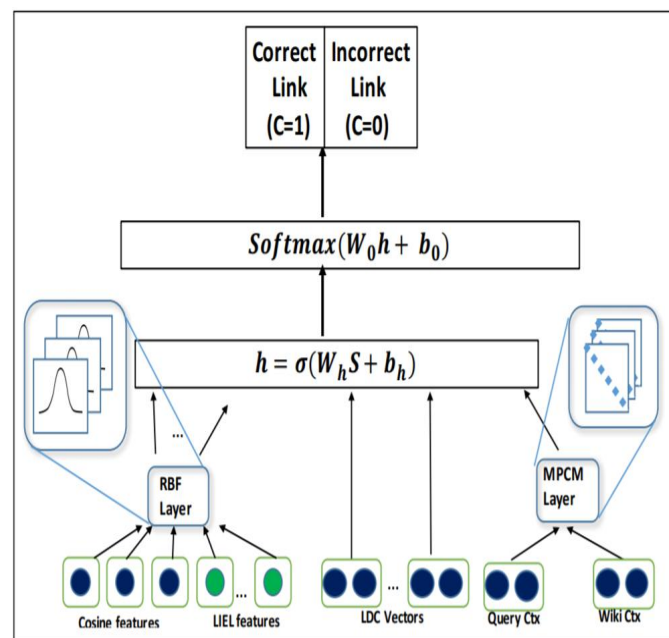


HMAN	56.2	85.1	93.4	53.7	83.4	92.5	56.7	86.9	94.5
HMAN w/o TE	3.2	16.7	38.3	3.5	17.2	38.5	5.4	22.3	45.5
HMAN w/o RE	50.2	78.4	86.5	49.3	78.6	87.0	52.6	81.6	89.1
HMAN w/o AE	49.2	81.0	89.8	48.8	80.9	90.0	52.2	83.3	91.6
HMAN w/o HW	46.8	76.1	84.1	46.0	76.2	84.6	50.5	79.5	87.5



基于多特征融合的跨语言实体链接

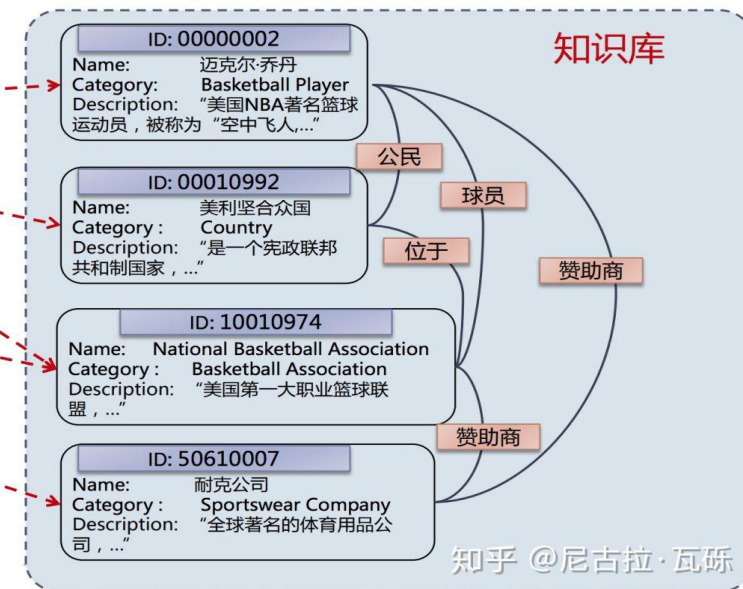
- **任务**: 将源语言抽取的实体, 链接到目标语言知识库
- **问题**: 实体与知识库中多个候选实体对应
- **思路**: 将跨语言实体链接任务转换为**实体和候选实体相似度比较**的任务。基于双语词嵌入, 利用实体的上下文语义进行实体消歧。



文本

乔丹
...
美国
NBA
著名
篮球
运动员, 他为
联盟
带来至少100
亿的收入, 也
把
耐克公司
从一家小公司
...

知识库



知乎 @尼古拉·瓦砾

Systems	In-KB acc. %
TAC Rank 1	79.2
TAC Rank 2	71.6
Sil & Florian (2016)	78.6
He et al. (2013)	81.0
Chisholm & Hachey (2015)	80.7
Sun et al. (2015)	83.9
Yamada et al. (2016)	85.2
Globerson et al. (2016)	87.2
This work	85.0
This work+CtxLSTMs	86.3
This work+CtxLSTMs+LDC	86.9
This work+CtxLSTMs+LDC+MPCM	87.4



基于枢轴语言的跨语言实体链接

- 任务：**极低资源下的跨语言实体链接
- 问题：**缺乏双语词嵌入或者双语词典等资源
- 思路：**利用高资源语言作为枢轴语言，借助枢轴语言和低资源语言的相似性。构建高资源和低资源相似语言的字符级相似度模型，基于此训练实体链接模型。

Model	bn	jv	lo	mr	pa	te	ti	uk	ug	Avg.
EXACT	.00	.63	.02	.00	.00	.00	.02	.02	.03	.08
TRANS	.00	.63	.02	.17	.00	.00	.46	.02	.03	.15
ENCODE										
MANUAL	.36 (hi)	.70 (id)	.07 (th)	.46 (hi)	.31 (hi)	.20 (ta)	.44 (am)	.25 (ru)	.16 (tr)	.33
BEST-53	.38 (ms)	.70 (id)	.07 (th)	.46 (hi)	.36 (te)	.36 (pa)	.44 (am)	.41 (kk)	.16 (tr)	.37
PBEL										
MANUAL	.48 (hi)	.86 (id)	.28 (th)	.62 (hi)	.49 (hi)	.33 (ta)	.69 (am)	.50 (ru)	.32 (tr)	.51
BEST-53	.48 (hi)	.86 (id)	.28 (th)	.62 (hi)	.49 (hi)	.47 (hi)	.69 (am)	.54 (kk)	.32 (tr)	.53
MULTI	.53	.87	.28	.62	.48	.46	.69	.56	.40	.54

Marathi

[पोलंड] हा मध्य युरोपातील एक देश आहे

Gloss: [Poland] is a country in Central Europe.

Cross-lingual Entity Linking

पोलंड
Marathi

Poland

Grapheme Pivoting

पोलंड
Marathi

पोलैंड
Hindi

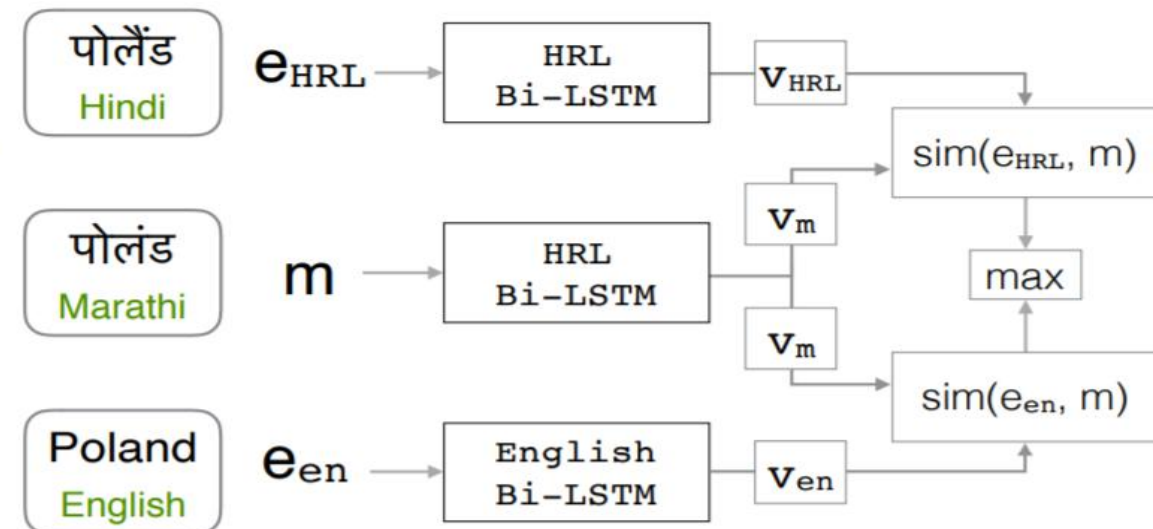
Poland

Phoneme Pivoting

poləndə
Marathi IPA

polæ:ndə
Hindi IPA

powlənd
English IPA





多语言知识图谱



NELL



- 多语言知识图谱对于多语言理解、多语言事件分析等任务有着重要作用。

- 目前代表性多语言知识图谱有：

DBpedia: 127种语言、2800万实体、95亿三元组

Yago: 45种语言、459万实体、2400万三元组

Freebase: 6800万实体、24亿三元组

XLORE: 清华大学开发，大规模中英知识图谱

XLORE Core数据统计

		百度百科	中文维基	英文维基	总数
整体	概念	32,009	150,241	326,518	508,768
	实例	1,629,591	640,622	1,235,178	3,505,391
	属性	157,370	45,190	26,723	229,283
		百度百科	中文维基	英文维基	总数
上下位关系	InstanceOf	7,584,931	1,449,925	3,032,515	12,067,371
	SubClassOf	2,784	191,577	555,538	749,899
		百度百科	中文维基	英文维基	
跨语言链接	百度百科	-	10,216/336,890	4,846/303,108	
	中文维基	10,216/336,890	-	28,921/454,579	
	英文维基	4,846/303,108	28,921/454,579	-	

目 录

1

研究背景

2

多语言知识图谱构建

3

多语言文本语义表征

4

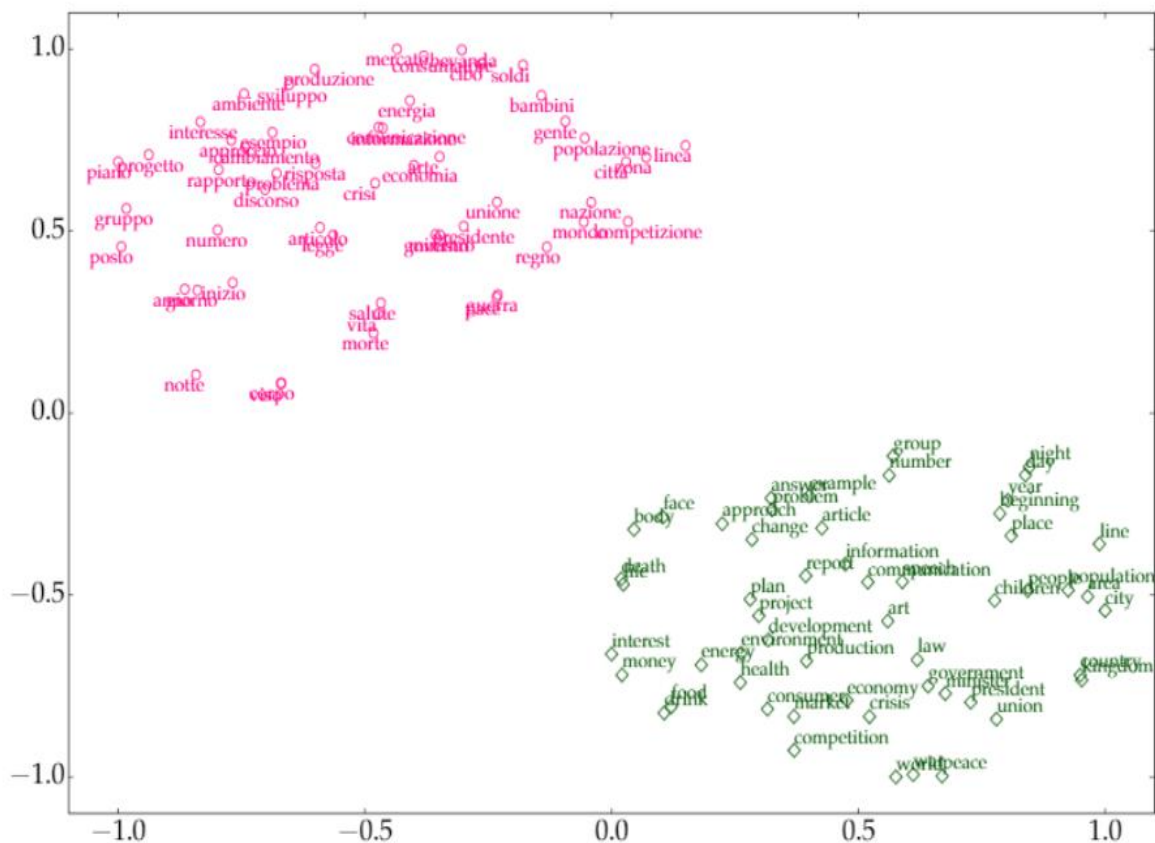
多语言事件分析及观点挖掘

5

研究趋势

多语言文本语义表征

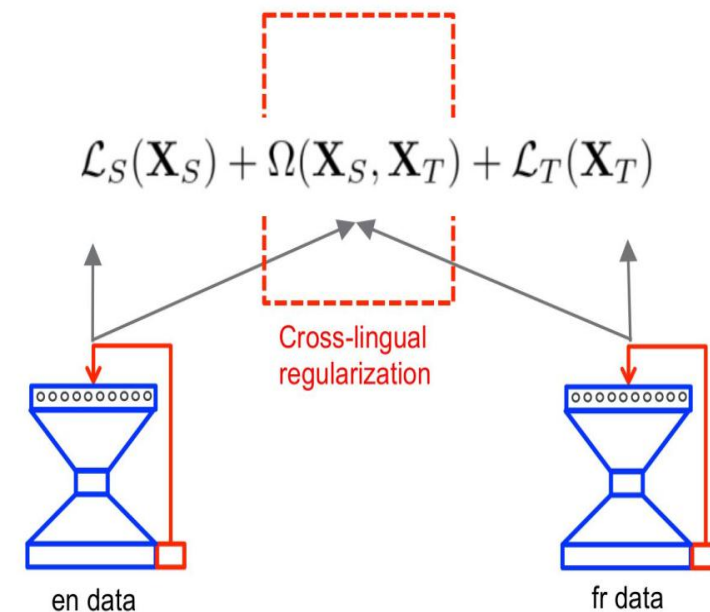
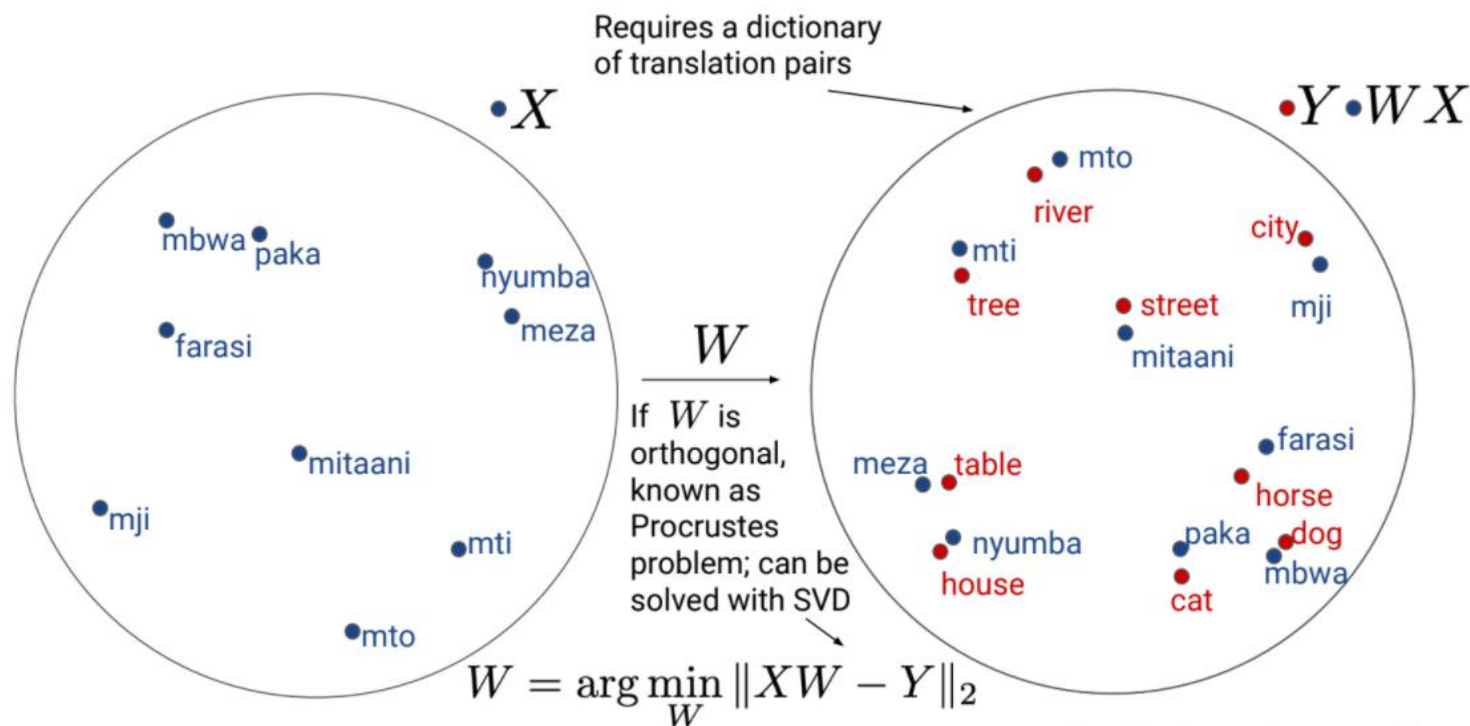
- 多语言文本语义表征的核心任务就是在高维的空间中实现不同语言的对齐。
- 目前主要包括多语言词嵌入和多语言深度语义表征任务。





基于映射的跨语言词嵌入方法

- 任务**: 利用双语词典实现多语言词嵌入
- 思路**: 将跨语言词嵌入看做一个高维空间对齐的任务。
首先在单语上分别训练单语词嵌入，然后基于**双语词典**，学习一个**映射函数**，实现源语言和目标语言词向量在高维语义空间上的对齐。



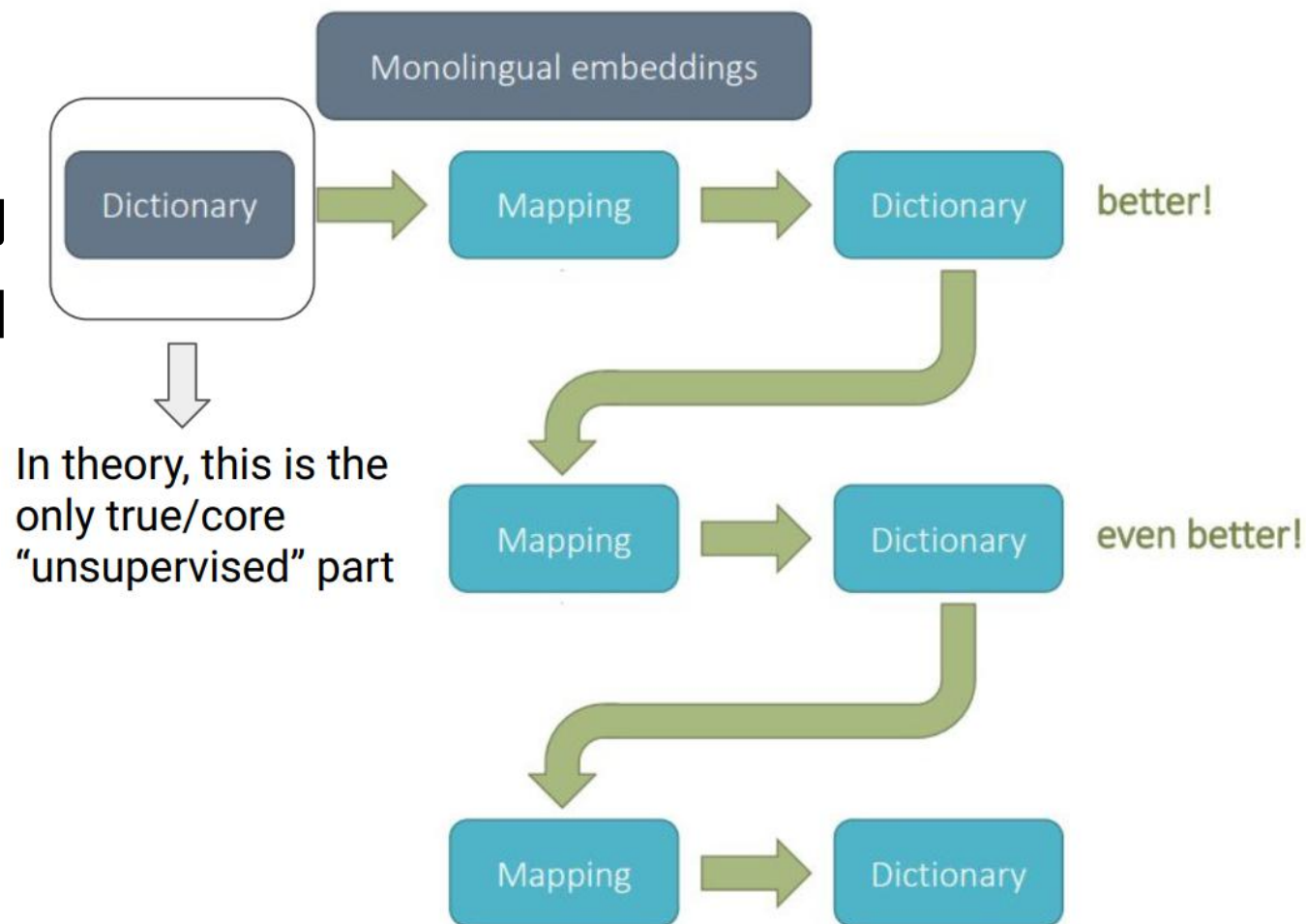
Model	Dict	EN-DE	EN-FR	EN-TR	EN-RU	Avg
<i>Supervised</i>						
PROC	1K	0.561	0.504	0.534	0.544	0.536
PROC	5K	0.607	0.534	0.568	0.585	0.574
PROC-B	1K	0.613	0.543	0.568	0.593	0.579
PROC-B	3K	0.615	0.532	0.573	0.599	0.580
DLV	5K	0.614	0.556	0.536	0.579	0.571
RCSLS	1K	0.376	0.357	0.387	0.378	0.374
RCSLS	5K	0.390	0.363	0.387	0.399	0.385



基于自学习的多语言词嵌入方法

- 任务**: 实现低资源场景下的词对齐
- 问题**: 双语词典规模较小
- 思想**: 首先利用已有的种子词典实现语义空间映射, 然后检索语义相近词扩充词典。迭代训练这个过程, 得到更好的映射矩阵和词典。

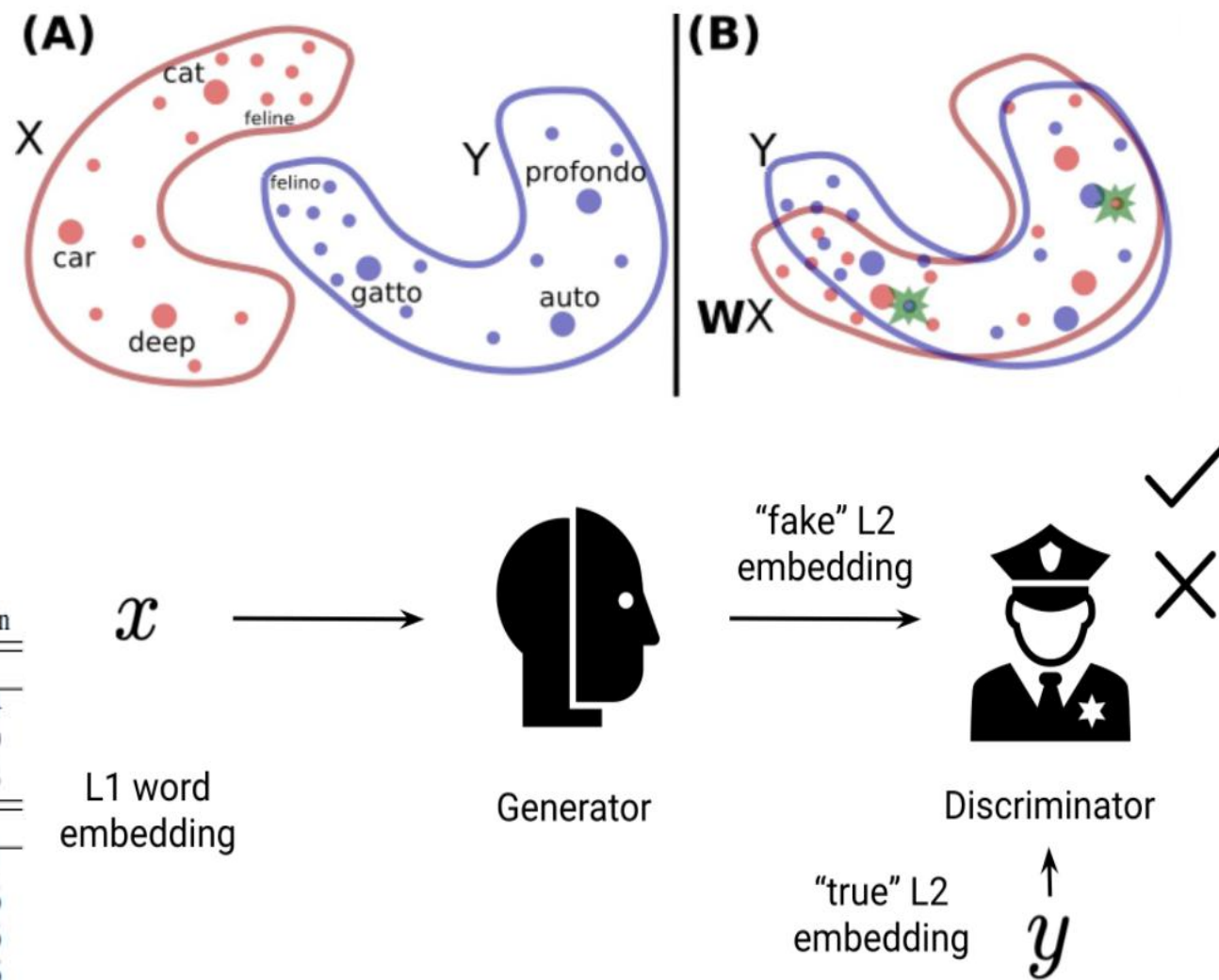
Model	ES-EN	NL-EN	IT-EN
BICVM (TYPE 1)	0.532	0.583	0.569
BILBOWA (TYPE 2)	0.632	0.636	0.647
BWESG (TYPE 3)	0.676	0.626	0.643
BNC+GT (Type 4)	0.677	0.641	0.646
ORTHO	0.233	0.506	0.224
BNC+HYB+ASYM	0.673	0.626	0.644
BNC+HYB+SYM (3388; 2738; 3145)	0.681	0.658*	0.663*
HFQ+HYB+ASYM	0.673	0.596	0.635
HFQ+HYB+SYM	0.695*	0.657*	0.667*





基于对抗网络的无监督多语言词嵌入方法

- 任务**: 无监督多语言词对齐方法
- 问题**: 低资源语言中双语词典资源缺乏。
- 思想**: 基于**对抗学习方法**, 通过无监督的方式实现对齐。利用生成器将源语言映射到目标语言, 利用鉴别器判断其是来自于源语言的映射还是目标语言, 实现多语言双语对齐。



	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en	en-eo	eo-en
<i>Methods with cross-lingual supervision and fastText embeddings</i>												
Procrustes - NN	77.4	77.3	74.9	76.1	68.4	67.7	47.0	58.2	40.6	30.2	22.1	20.4
Procrustes - ISF	81.1	82.6	81.1	81.3	71.1	71.5	49.5	63.8	35.7	37.5	29.0	27.9
Procrustes - CSLS	81.4	82.9	81.1	82.4	73.5	72.4	51.7	63.7	42.7	36.7	29.3	25.3
<i>Methods without cross-lingual supervision and fastText embeddings</i>												
Adv - NN	69.8	71.3	70.4	61.9	63.1	59.6	29.1	41.5	18.5	22.3	13.5	12.1
Adv - CSLS	75.7	79.7	77.8	71.2	70.1	66.4	37.2	48.1	23.4	28.3	18.6	16.6
Adv - Refine - NN	79.1	78.1	78.1	78.2	71.3	69.6	37.3	54.3	30.9	21.9	20.7	20.6
Adv - Refine - CSLS	81.7	83.3	82.3	82.1	74.0	72.2	44.0	59.1	32.5	31.4	28.2	25.6

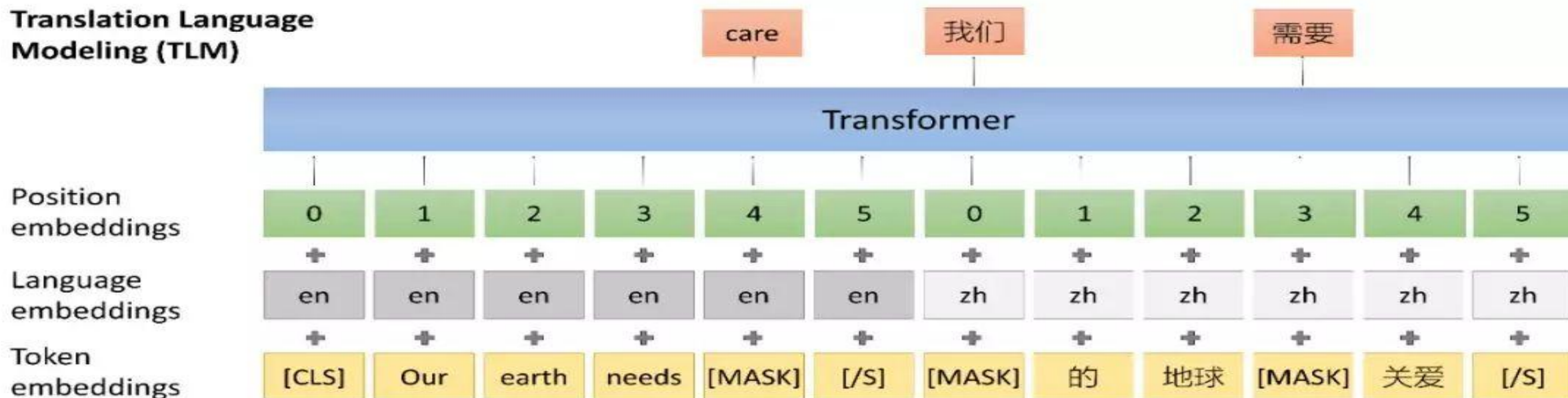


基于预训练模型的多语言深度语义表征

- 任务**：基于深度学习框架，实现多语言句子及文档的统一表征
- 思想**：基于预训练语言模型，利用大规模单语语料，基于Transformer架构来处理不同的语言，学习到两个语言之间的对应关系，使得模型具有跨语言迁移能力。

Model	train	#M	en	nl	es	de	Avg
Lample et al. (2016)	each	N	90.74	81.74	85.75	78.76	84.25
Akbik et al. (2018)	each	N	93.18	90.44	-	88.27	-
mBERT [†]	each	N	91.97	90.94	87.38	82.82	88.28
	en	1	91.97	77.57	74.96	69.56	78.52
XLM-R _{Base}	each	N	92.25	90.39	87.99	84.60	88.81
	en	1	92.25	78.08	76.53	69.60	79.11
	all	1	91.08	89.09	87.28	83.17	87.66
XLM-R	each	N	92.92	92.53	89.72	85.81	90.24
	en	1	92.92	80.80	78.64	71.40	80.94
	all	1	92.00	91.60	89.52	84.60	89.43

Translation Language Modeling (TLM)



目 录

1

研究背景

2

多语言知识图谱构建

3

多语言文本语义表征

4

多语言事件分析及观点挖掘

5

研究趋势



多语言事件关联分析及观点挖掘任务

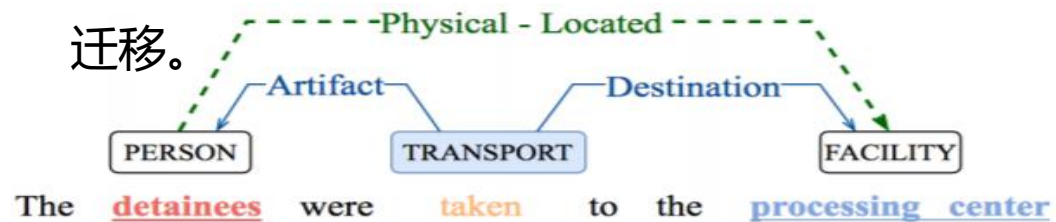
多语言事件关联分析及观点挖掘

- 多语言事件抽取
- 多语言事件关联关系
- 多语言观点句识别
- 多语言情感倾向性分析
- 多语言摘要

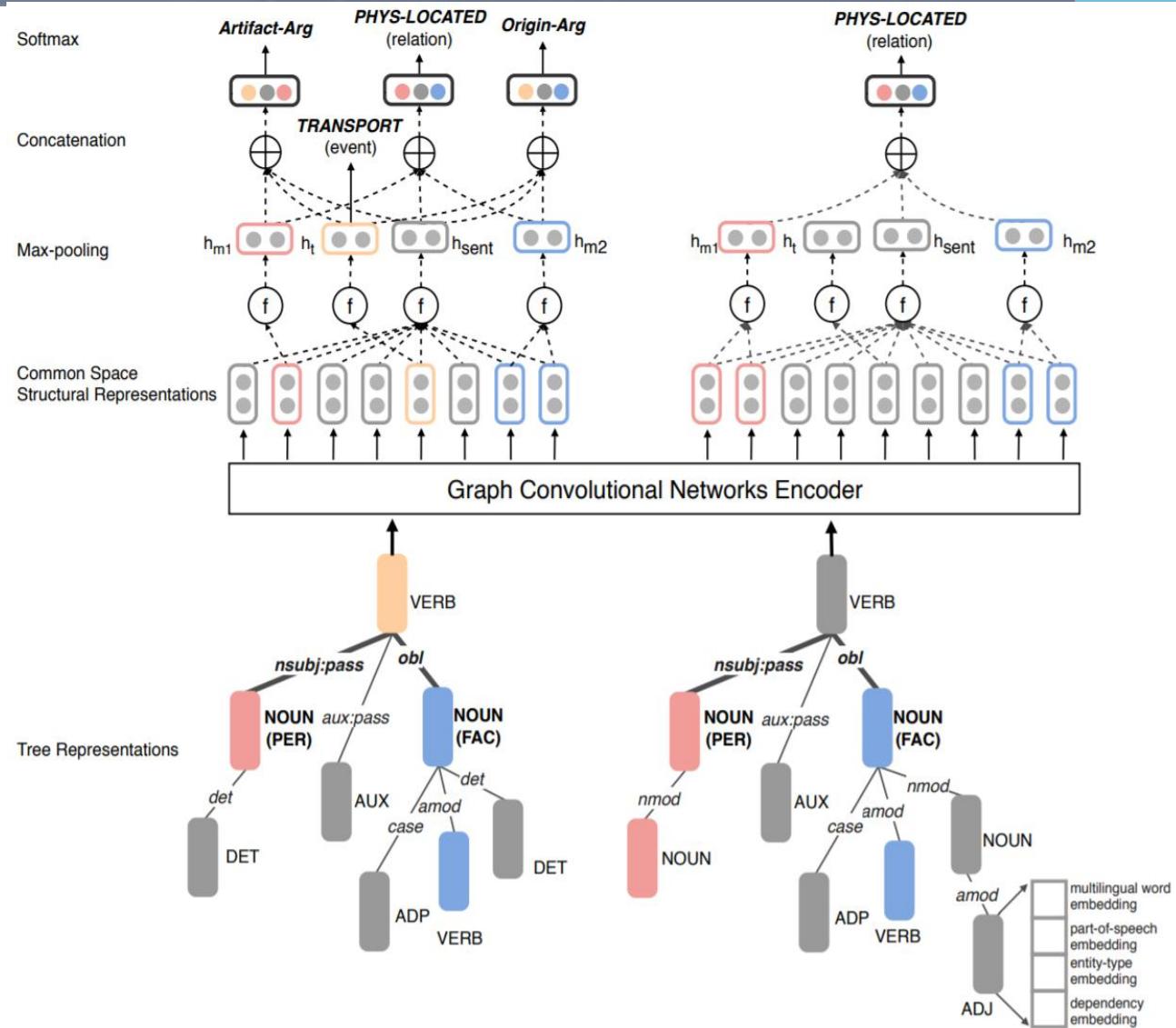


基于结构迁移的多语言事件抽取

- 任务**: 从多语言文本中抽取事件要素及关系
- 问题**: 低资源语言缺乏标注语料
- 思路**: 利用**多语言词向量**实现不同语言的统一表征, 同时借助**图卷积网络**, 编码不同语言的依存句法结构获得不同语言的结构化表征, 实现**特征和结构**信息的迁移。



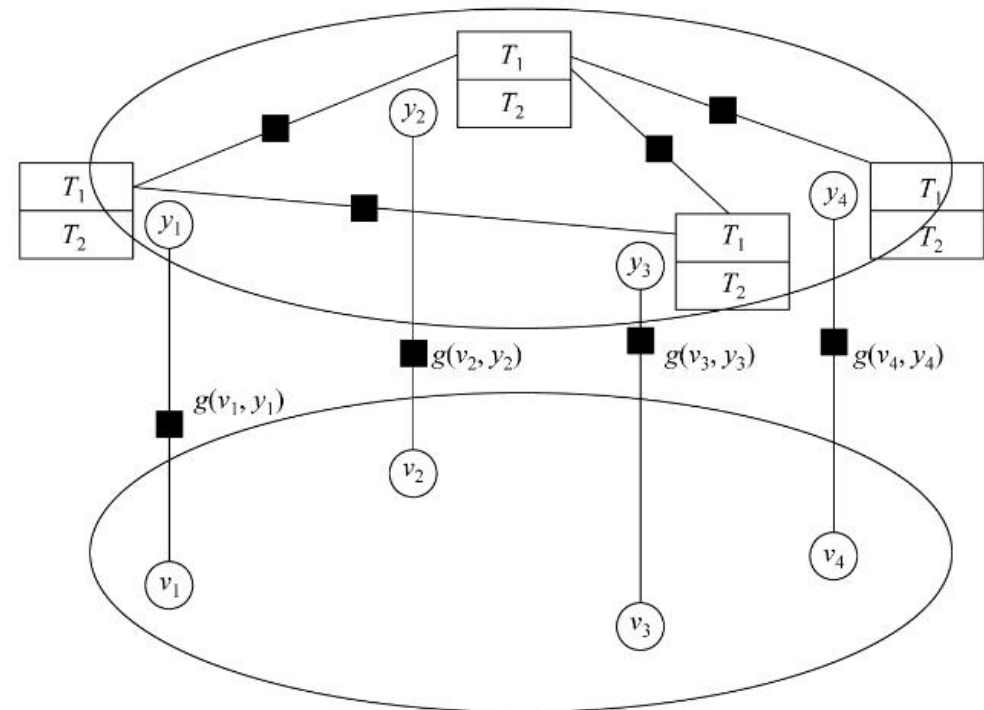
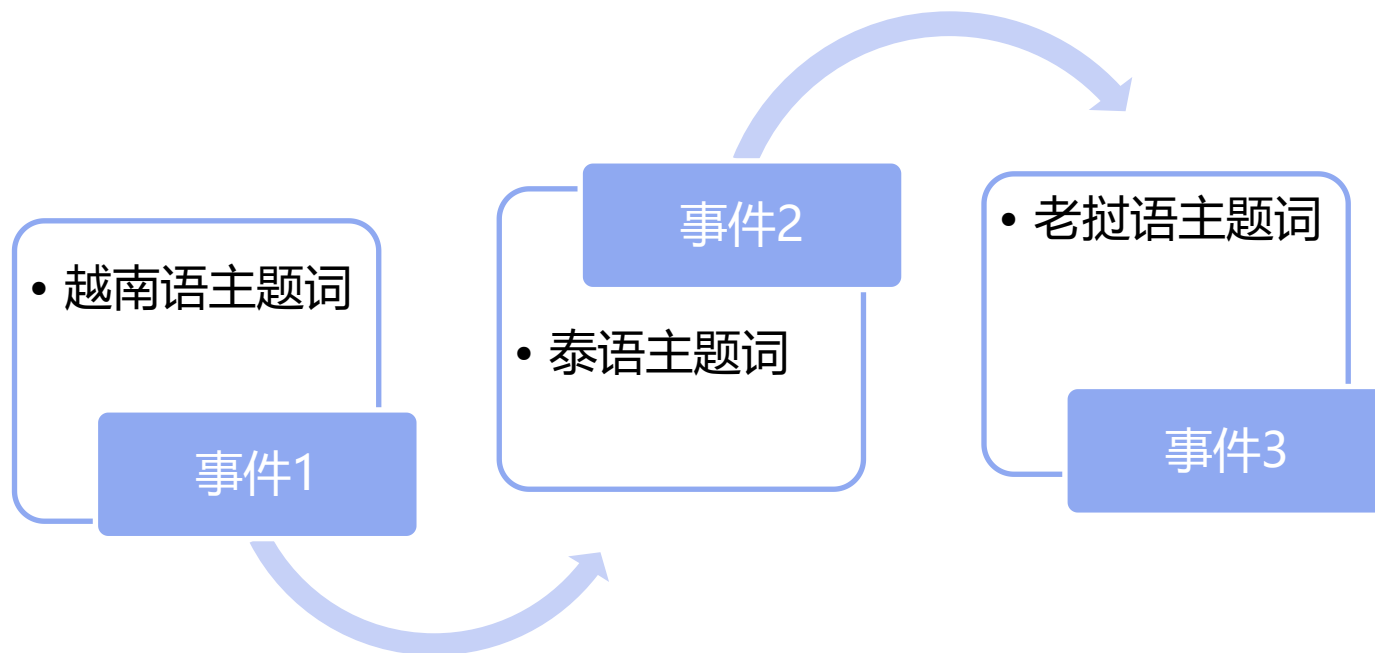
Train \ Test			
	English	Chinese	Arabic
English	63.9	59.0	61.8
Chinese	51.6	59.3	60.6
Arabic	43.1	50.1	64.0
English + Chinese	—	—	63.1
English + Arabic	—	60.1	—
Chinese + Arabic	51.9	—	—





基于双语主题和因子图的多语言事件关联分析

- **任务**: 判断不同语言下的事件**是否关联**
- **思路**: 将多语言事件关联分析转换为**多语言文本相似度任务**, 首先基于双语主题模型表征双语文本, 在此基础上, 依据多语言新闻事件的文本相似度**构建事件因子图**, 实现事件关联强度分析。



ϵ	召回率 R	准确率 P	F 值
0.2	0.895 8	0.498 8	0.640 7
0.3	0.845 2	0.648 6	0.434 1
0.4	0.802 3	0.778 7	0.492 5
0.5	0.775 4	0.868 7	0.819 4
0.6	0.725 6	0.852 6	0.784 0



基于双向交叉注意力的多语言事件关联分析

- 任务**：判断不同语言下的事件内在逻辑关系类型
- 思路**：基于深度学习框架，将时序关系识别转化为**基于上下文语义的分类问题**。首先利用双语词向量实现不同语言映射到同一语义空间，然后基于**双向交叉注意力**，综合考虑不同语言之间以及同一语言句子内部的上下文信息，最后实现时序关系分类。

模型	P (%)	R (%)	F
注意力机制	85.12	82.43	83.26
交叉注意力机制	86.04	83.51	84.19
本文模型	87.32	84.07	85.98

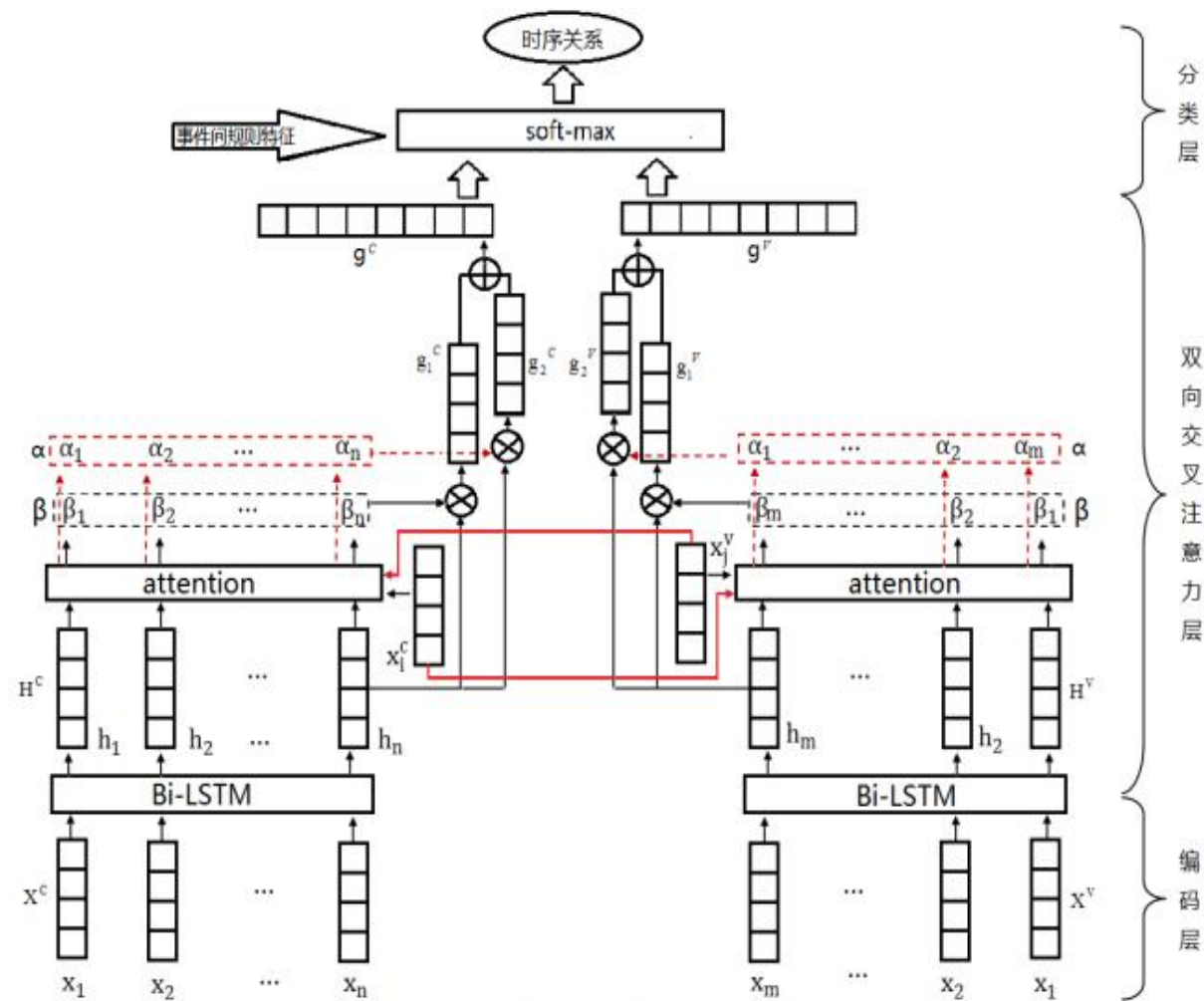
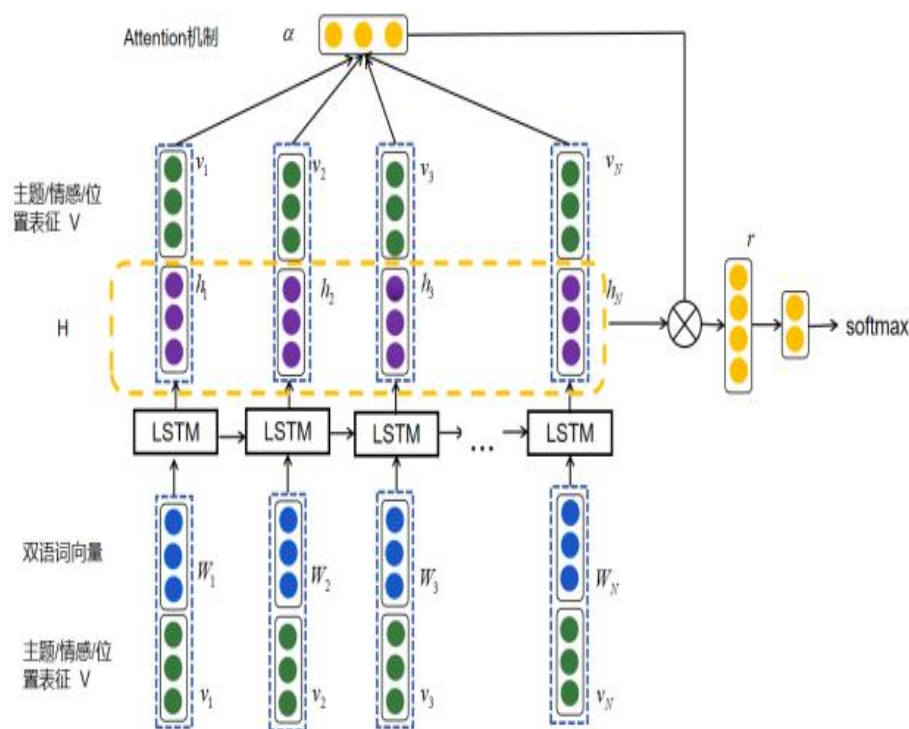


图 4.5 BLCATT 模型

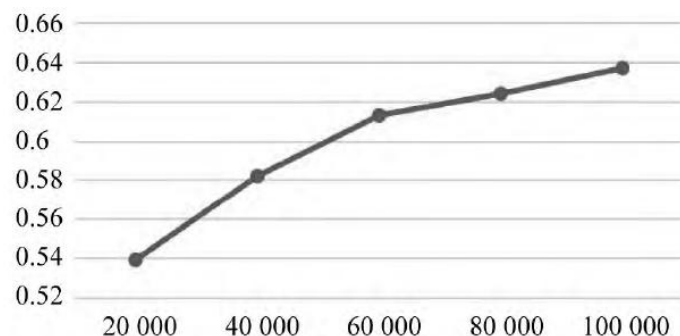


基于多特征融合的多语言事件观点句识别



任务：输入不同语言的新闻事件文档，从中识别出事件观点句。

问题：新闻事件中观点句的情感表达不明显；不同语言差异较大，句子难以对齐；低资源语言的标记语料的缺失，导致新闻观点句识别效果较差。

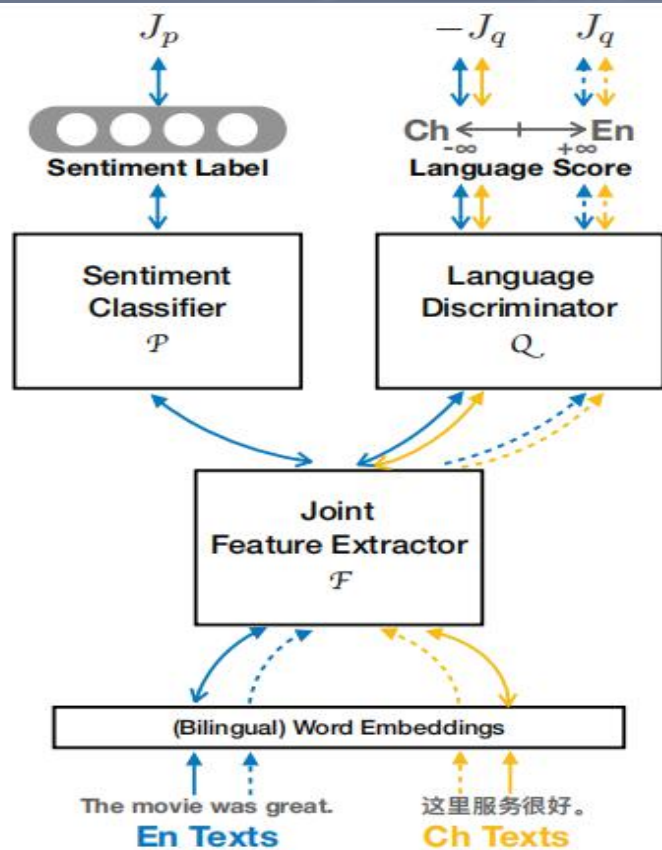


融入位置	观点句特征	P	R	F ₁
无	无	0.550	0.622	0.584
词向量	主题	0.560	0.645	0.600
	位置	0.565	0.637	0.599
	情感	0.576	0.640	0.606
	主题+位置+情感	0.581	0.655	0.616
词向量+注意力机制	主题	0.578	0.647	0.611
	位置	0.582	0.638	0.609
	情感	0.615	0.636	0.625
	主题+位置+情感	0.637	0.654	0.645

思路：构建双语词嵌入模型，将句子的主题、情感和位置等特征融入编码层和注意力机制，得到句子在不同语言文档中的综合表征，并结合句子的语义特征进行观点句识别。



基于对抗学习的多语言情感倾向性分析



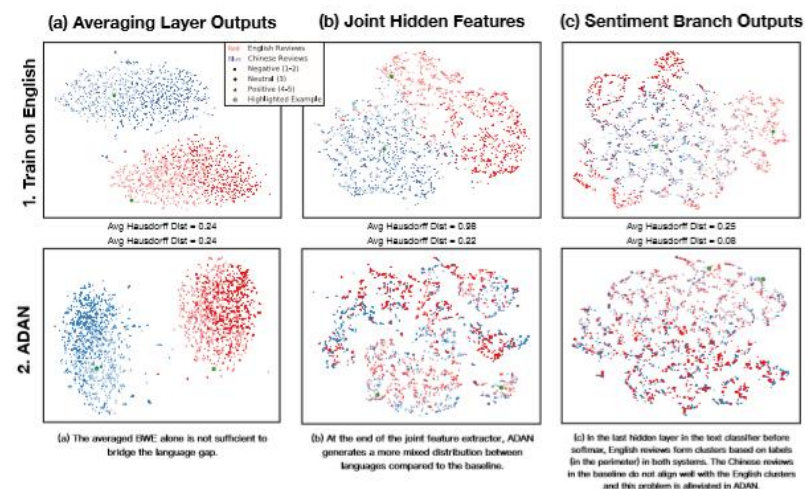
任务：输入不同语言的主观句或观点句，判断其表达的**情感倾向性**。

问题：如何构建双语的**情感表达映射或关联**，以及如何获得情感知识。

Algorithm 1 ADAN Training

Require: labeled SOURCE corpus $\mathbb{X}_{src}$; unlabeled TARGET corpus $\mathbb{X}_{tgt}$; Hyperparameter $\lambda > 0$, $k \in \mathbb{N}$, $c > 0$.

```
1: repeat
2:    $\triangleright Q$  iterations
3:   for  $qiter = 1$  to  $k$  do
4:     Sample unlabeled batch  $x_{src} \sim \mathbb{X}_{src}$ 
5:     Sample unlabeled batch  $x_{tgt} \sim \mathbb{X}_{tgt}$ 
6:      $f_{src} = F(x_{src})$ 
7:      $f_{tgt} = F(x_{tgt})$   $\triangleright$  feature vectors
8:      $loss_q = -Q(f_{src}) + Q(f_{tgt})$   $\triangleright$  Eqn (2)
9:     Update  $Q$  parameters to minimize  $loss_q$ 
10:     $ClipWeights(Q, -c, c)$ 
11:   $\triangleright$  Main iteration
12:  Sample labeled batch  $(x_{src}, y_{src}) \sim \mathbb{X}_{src}$ 
13:  Sample unlabeled batch  $x_{tgt} \sim \mathbb{X}_{tgt}$ 
14:   $f_{src} = F(x_{src})$ 
15:   $f_{tgt} = F(x_{tgt})$ 
16:   $loss = L_p(P(f_{src}); y_{src}) + \lambda(Q(f_{src}) - Q(f_{tgt}))$   $\triangleright$  Eqn (4)
17:  Update  $F, P$  parameters to minimize  $loss$ 
18: until convergence
```



Methodology	Approach	Accuracy	
		Chinese	Arabic
Train-on-SOURCE-only	Logistic Regression	30.58%	45.83%
	DAN	29.11%	48.00%
Domain Adaptation	mSDA (Chen et al., 2012)	31.44%	48.33%
Machine Translation	Logistic Regression + MT	34.01%	51.67%
	DAN + MT	39.66%	52.50%
CLD-based CLTC	CLD-KCNN (Xu and Yang, 2017)	40.96%	52.67% [†]
	CLDFA-KCNN (Xu and Yang, 2017)	41.82%	53.83% [†]
Ours	ADAN	42.49% $\pm$ 0.19%	54.54% $\pm$ 0.34%

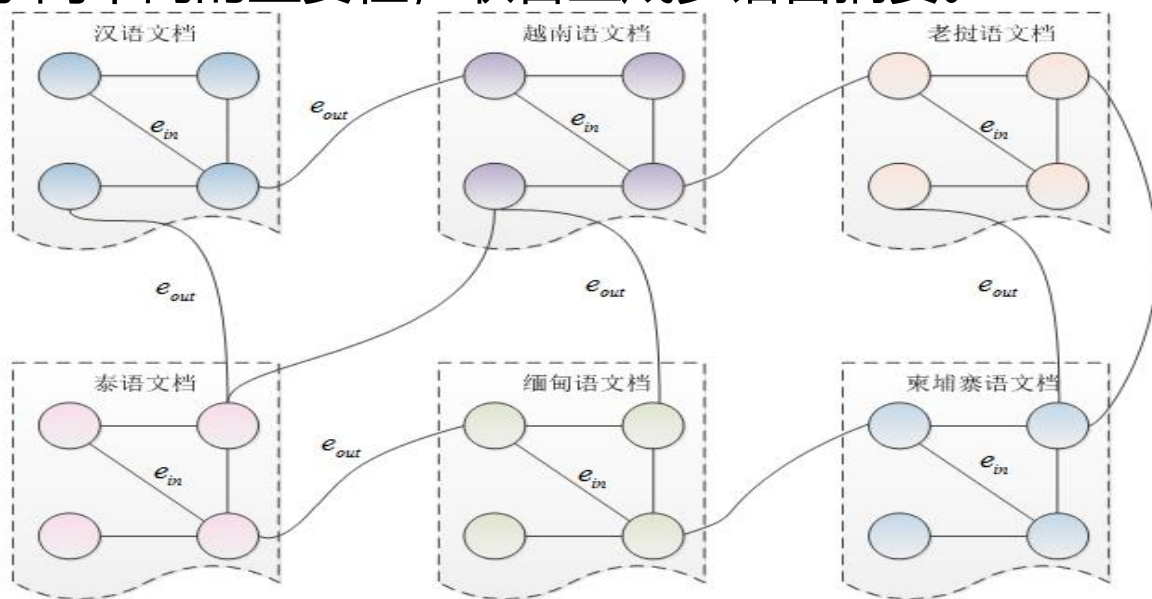
思路：基于对抗学习，将从资源丰富语言上学习到的知识迁移到未标记数据的低资源语言上。模型输入带标注的源语言句子和未标注的目标语言句子，利用共同特征提取器提取两个句子特征，进行情感倾向性分析。



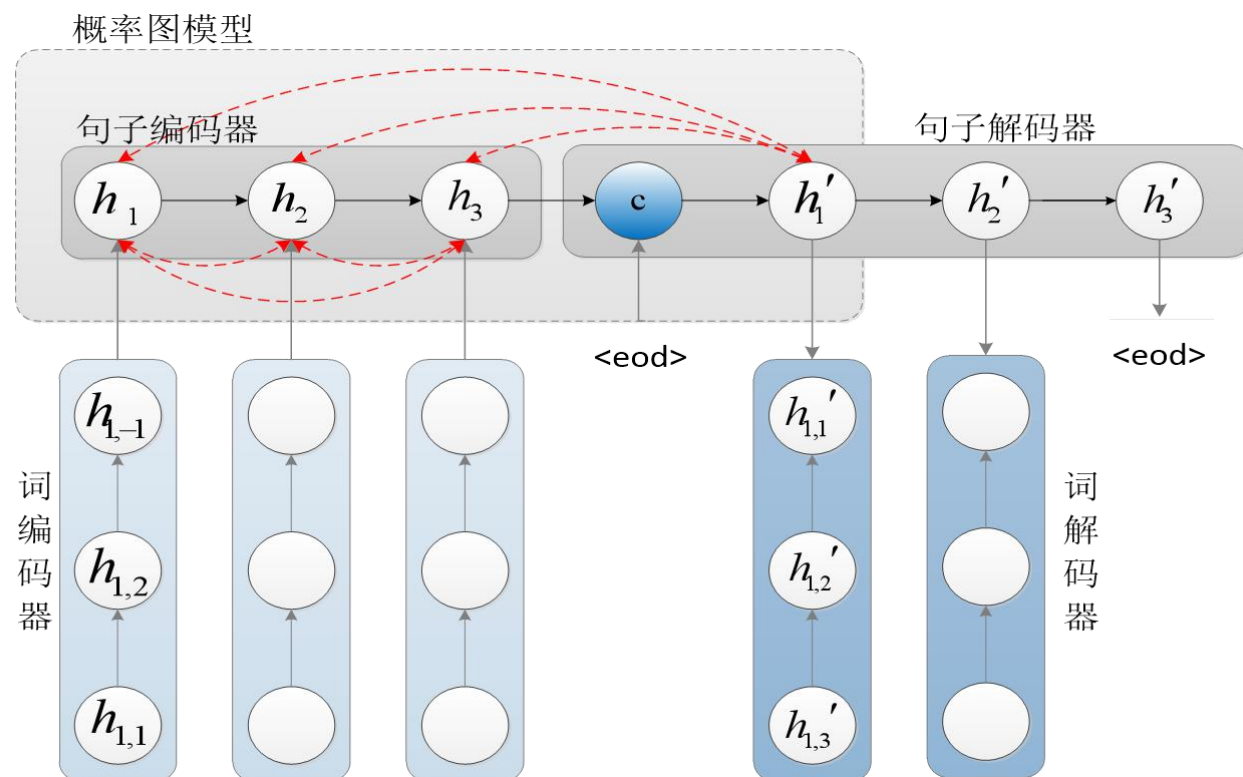
基于句子关联图的跨语言文本摘要

任务：实现描述同一事件的多种不同语言新闻的联合摘要

思路：构造多语言不同文档的**句子关联图**，并将句子关联图融入到序列到序列模型的注意力机制中，利用句子间不同的重要性，联合生成多语言摘要。



	ROUGE-1	ROUGE-2	ROUGE-3
TextRank	0.3166	0.1352	0.2519
LReg	0.3227	0.1470	0.2650
LSTM-Feature	0.3824	0.2270	0.3169





基于序列到序列模型的跨语言文本摘要

任务：输入源语言得到目标语言摘要

思路：利用Transformer模型强大的表征能力，直接将文本翻译的能力引入到跨语言摘要的模型中，同时实现摘要和翻译。

佛山一对小夫妻帮农民工从网上订火车票，每张收取10元手续费，日前小夫妻被刑拘。农民工叫冤称，“他们也没漫天要价”，比黄牛好多了。律师称按照法律，变相加价倒卖等行为构成倒卖车票情节严重，将被刑事处罚。

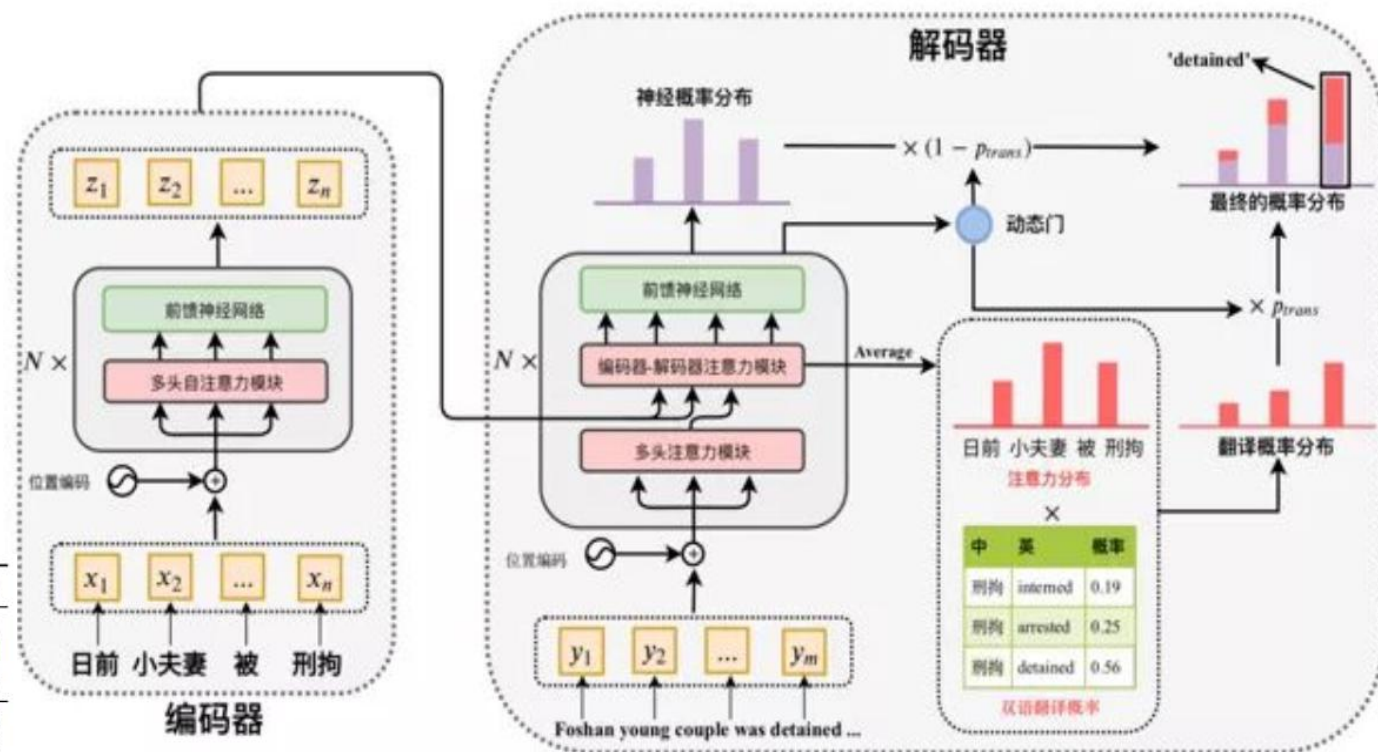
原文 (中文)

佛山—Foshan
小夫妻—young couple
火车票—train tickets
10元—10 yuan
刑拘—detained
农民工—migrant workers

翻译对照表

摘要 (英文)

Foshan young couple was detained for charging 10 yuan for buying train tickets online for migrant workers



	Model	RG-1	RG-2	RG-L		Model	RG-1	RG-2	RG-L	MVS
Baseline	GETran	28.19	11.40	25.77	Baseline	GETran	24.34	9.14	20.13	0.64
	GLTran	32.17	13.85	29.43		GLTran	35.45	16.86	31.28	16.90
	TNCLS	36.82	18.72	33.20		TNCLS	38.85	21.93	35.05	19.43
Baseline +Extra Data	CLSMS	38.25	20.20	34.76	Baseline +Extra Data	CLSMS	40.34	22.65	36.39	21.09
	CLSMT	40.23	22.32	36.59		CLSMT	40.25	22.58	36.21	21.06
ATS	Naive	40.19	21.84	36.46	ATS	Naive	40.40	23.82 [†]	36.63	21.86*
	Equal	39.98	21.63	36.29		Equal	40.10	23.36*	36.22	21.41
	Adapt	40.47	22.21	36.89		Adapt	40.68	24.12[†]	36.97	22.15

目 录

1

研究背景

2

多语言知识图谱构建

3

多语言文本语义表征

4

多语言事件分析及观点挖掘

5

研究趋势

研究趋势

- 面向低资源的多语言知识图谱构建
- 面向低资源及差异化语言的语义空间对齐方法
- 面向知识驱动的多语言事件分析方法
- 面向多语言的话题关联、演化以及态势分析方法
- 面向多语言的细粒度观点挖掘方法
- 面向多语言的对比摘要方法



昆明理工大学
KUNMING UNIVERSITY OF SCIENCE AND TECHNOLOGY



谢谢！

昆明理工大学信息工程与自动化学院 余正涛
ztyu@hotmail.com