

# **Algorithmic Studies on Relation Extraction from Chinese Short Texts**

## **面向中文短文本的关系抽取算法设计**

**答辩人：汪诚愚**

**导师：何晓丰**

**华东师范大学**

**East China Normal University**

# 知识图谱

- 海量碎片化知识造成“信息过载”问题
  - 中国网站数量：191万个（2010年）→518万个（2019年）
  - 互联网总数据量：33ZB（2018年）→175ZB（2025年）
- 知识图谱将海量互联网数据结构化，对知识推荐、查询理解、个性化搜索等有重要作用
- 构建知识图谱的核心任务：知识抽取

Google search results for "how much does a bear weigh". The snippet from "Bear > Mass" states: "Giant panda: 150 – 220 lbs".

Google search results for "grizzly". The snippet from Grizzly.com's website provides information about shop tools and machinery.

Freebase™



# 分类体系与上下位关系抽取（1）

- 分类体系（Taxonomy）：知识图谱概念的层次化表示
  - 上下位关系（Hypernymy）：“is-a”，分类体系的基础组成部分（猫-哺乳动物、桌子-家具）
  - 应用：自然语言推理、用户查询理解、语义搜索

查询：watch harry potter 人物？书籍？电影？

A screenshot of a Google search results page for the query "watch harry potter". The results include:

- Top result: "哈利·波特与魔法石" (Harry Potter and the Philosopher's Stone) with a thumbnail image.
- Information about the movie: "2001年 · 奇幻电影/虚构作品 · 2 小时 39 分钟".
- Call-to-action: "播放YouTube上的预告片" (Play trailer on YouTube).
- User feedback: "95% 的用户顶了这部电影" (95% of users liked this movie).
- Summary: "《哈利·波特与魔法石》是一部于2001年上映，英美合拍的奇幻电影，剧情改编自畅销作家J·K·罗琳所著，同名奇幻小说《哈利·波特》第一册《哈利·波特与魔法石》，且为“哈利·波特系列电影”中的第一部作品。维基百科"

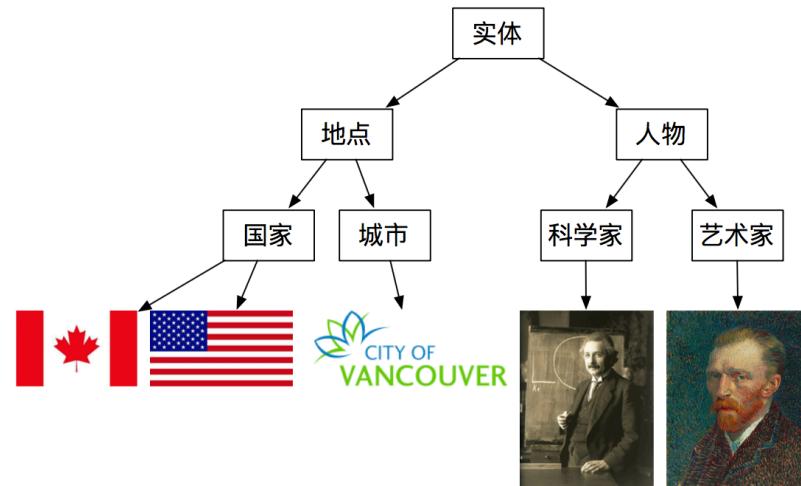
问题推荐

A screenshot of entity linking results for the query "watch harry potter". The results include:

- Result: "哈利·波特与魔法石" (Harry Potter and the Philosopher's Stone) with a thumbnail image.
- Information: "2001年 · 奇幻电影/虚构作品 · 2 小时 39 分钟".
- Call-to-action: "播放YouTube上的预告片" (Play trailer on YouTube).
- User feedback: "95% 的用户顶了这部电影" (95% of users liked this movie).
- Summary: "《哈利·波特与魔法石》是一部于2001年上映，英美合拍的奇幻电影，剧情改编自畅销作家J·K·罗琳所著，同名奇幻小说《哈利·波特》第一册《哈利·波特与魔法石》，且为“哈利·波特系列电影”中的第一部作品。维基百科"

知识推荐

分类体系示例



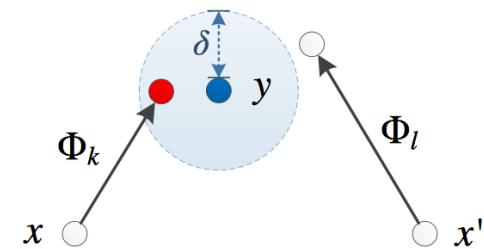
# 分类体系与上下位关系抽取（2）

- 模式匹配法
  - 人工订制模式：精度较高，但是覆盖率比较低（例如Hearst模式）
  - 自动生成模式：精度略微降低，提高了覆盖率
  - 中文上下位关系模式：覆盖率很低，无法得到广泛应用

Pattern	Translation
w 是[一个 一种] h	w is a [a kind of] h
w [、] 等 h	w[,] and other h
h [, ] 叫[做] w	h[,] called w
h [, ] [像]如 w	h[,] such as w
h [, ] 特别是 w	h[,] especially w

	P(%)	R(%)	F(%)
$M_{Wiki+CilinE}$	92.41	60.61	73.20
$M_{Pattern}$	97.47	21.41	35.11
$M_{Snow}$	60.88	25.67	36.11
$M_{balApinc}$	54.96	53.38	54.16
$M_{invCL}$	49.63	62.84	55.46
$M_{Fu}$	87.40	48.19	62.13

- 结合分布式表示的模型
  - 结合模式匹配法与分布式表示作为特征
- 本研究的出发点
  - 学习下位词到上位词的映射
  - 充分结合中文语言模式和语言学规则



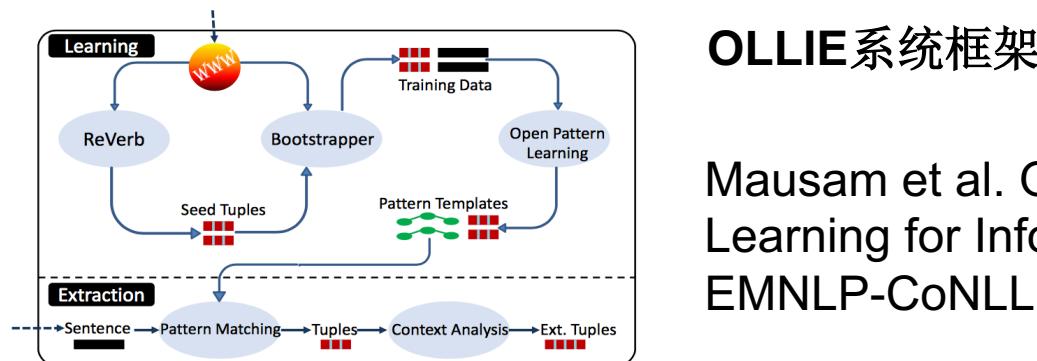
重要参考文献

Wu et al. Probase: A Probabilistic Taxonomy for Text Understanding. SIGMOD 2012

Fu et al. Learning Semantic Hierarchies via Word Embeddings. ACL 2014

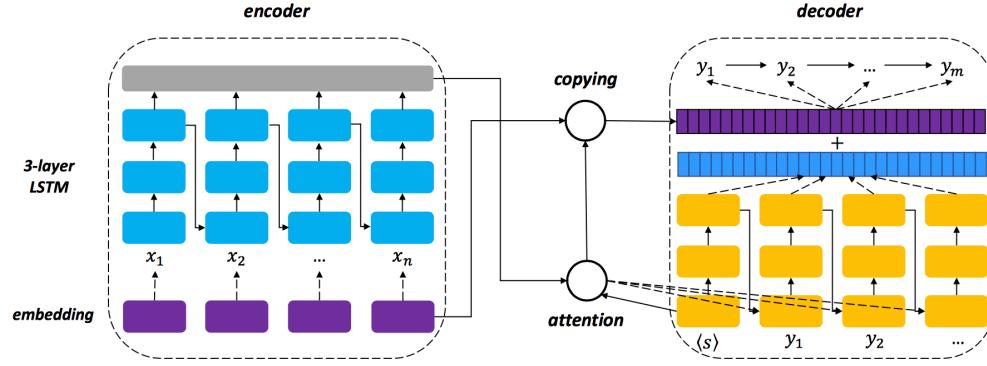
# 通用语义关系抽取（1）

- 关系分类：采用深度神经网络编码关系上下文
  - 缺点：只能对固定若干种关系进行预测，需要大量人工标注数据
- 开放关系抽取（Open Relation Extraction）
  - 传统方法：从未标注的文本中抽取“主语-谓语-宾语”结构，作为候选关系元组，无需定义待抽取的关系类别
  - 经典系统：ReVerb、WOE、OLLIE等



Mausam et al. Open Language Learning for Information Extraction.  
EMNLP-CoNLL 2012

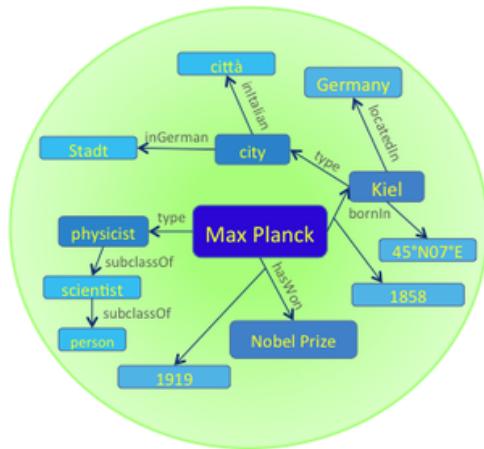
- 基于神经网络的方法：例如Encoder-Decoder架构等



Cui et al. Neural Open Information Extraction.  
ACL 2018

# 通用语义关系抽取 (2)

- 基于短文本的关系抽取
  - 基于维基百科类别系统的关系抽取：基于语言规则



典型系统



Suchanek et al. Yago - A Core of Semantic Knowledge.  
WWW 2007

American singers of German origin (*Pre-modifier + Head + Post-modifier*)

- 基于名词短语的开放关系抽取：处理场景比较单一，难以扩展至中文环境

Phrase	RELNOUN 1.1	RELNOUN 2.2
“United States President Obama”		(Obama, [is] President [of], United States)
“Seattle historian Feliks”	(Feliks, [is] historian [of], Seattle)	(Feliks, [is] historian [from], Seattle)
“Japanese foreign minister Kishida”		(Kishida, [is] foreign minister [of], Japan)
“GM Deputy Chairman Lutz”		(Lutz, [is] Deputy Chairman [of], GM)

Yahya et al. ReNoun: Fact Extraction for Nominal Attributes. EMNLP 2014

Pal and Mausam. Demonyms and Compound Relational Nouns in Nominal Open IE.

AKBC@NAACL-HLT 2016

# 中文短文本数据源

- 特点：蕴含大量实体相关的知识，难以被传统关系抽取算法抽取

(a) 维基 (Wikis)

来源: <https://zh.wikipedia.org>

马来熊 [\[编辑\]](#)

维基百科，自由的百科全书

马来熊（学名：*Helarctos malayanus*），英文名为“Sun Bear”，[维语](#)译音为“耐力喀苏”，是[熊科](#)马来熊属（*Helarctos*）的唯一一种生物，生活在[东南亚](#)的热带雨林中。

目录 [\[展开\]](#)  
1 特征  
2 习性  
3 分布  
4 参考文献  
5 外部链接

**特征** [\[编辑\]](#)

 马来熊属[熊科](#)[熊亚科](#)马来熊属，是熊科动物中体型最小的成员。成年体高约120~150厘米，体重27~65公斤。马来熊全身黑色（雄性比雌性大10~45%），前胸通常有一块明显的“U”型斑纹，斑纹呈浅棕或浅黄白色。马来熊头部比较宽，口鼻不突出，裸露无毛，是浅棕或灰褐色。耳朵圆而小，位置较低。马来熊舌头长，便于吞食[白蚁](#)或其他**昆虫**。指甲所钩呈镰刀型，脚掌内缩，尾巴是30~70毫米（1.2~2.6英寸）长。

图片来自上海动物园

**习性** [\[编辑\]](#)

马来熊属[熊科](#)[熊亚科](#)马来熊属，是熊科动物中体型最小的成员。成年体高约120~150厘米，体重27~65公斤。马来熊全身黑色（雄性比雌性大10~45%），前胸通常有一块明显的“U”型斑纹，斑纹呈浅棕或浅黄白色。马来熊头部比较宽，口鼻不突出，裸露无毛，是浅棕或灰褐色。耳朵圆而小，位置较低。马来熊舌头长，便于吞食[白蚁](#)或其他**昆虫**。指甲所钩呈镰刀型，脚掌内缩，尾巴是30~70毫米（1.2~2.6英寸）长。

分类: IUCN易危物种 | 熊科 | 中国国家一级保护动物 | 华盛顿公约附录一之动物 | 中国城乳动物 | 云南动物 | 老挝动物 | 孟加拉动物 | 印尼动物 | 黑猩猩动物 | 越南动物 | 泰国动物 | 马来西亚哺乳动物 | 印度哺乳动物 | 棕榈动物

(b) 用户标签 (User Tags)

帕丁顿熊2 Paddington 2 (2017)



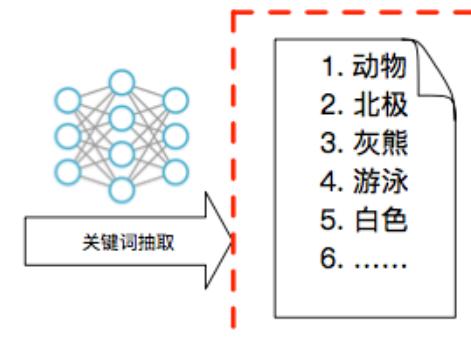
好于 88% 喜剧片  
好于 72% 动画片

来源: <https://movie.douban.com>

(c) 关键词 (Keywords)



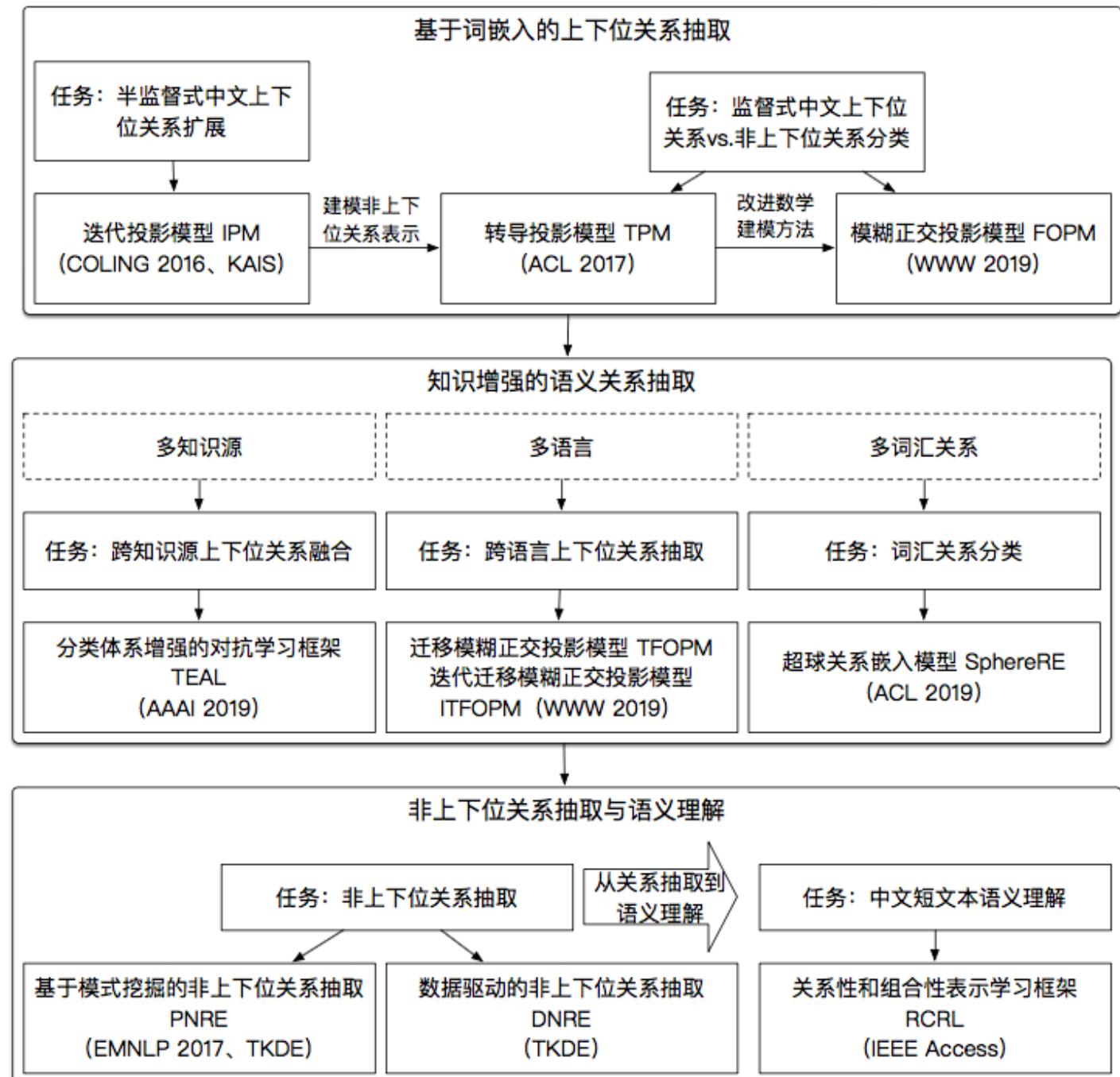
北极熊，又称为白熊或冰熊，是熊属的一个种，是北极地区的典型动物，可能是六十多万年前由灰熊演化出来。在所生存的空间里，北极熊位于食物链的最顶层。健康的北极熊会拥有极厚的脂肪及毛发，以在北极这种极端严寒的气候中生存。其中白色的外表在雪白的雪地上是良好的保护色。北极熊是游泳健将，主要在海冰上捕捉海豹为食。北极熊是一种能在恶劣酷寒的环境下生存的动物，其活动范围主要在北冰洋、即北极圈附近，而最南则可以在有浮冰出没的地方找到它们。



# 中文短文本关系抽取的困难性

- 中文基础 NLP 分析的低准确度
- 常识性关系的上下文稀疏性
- 短文本的语法结构和语义不完整性
- 标注数据集的缺乏

# 算法研究整体框架



# 第一部分：基于词嵌入的上下位关系抽取

## 先前工作：投影模型

输入数据集： $D^P$

学习范式：监督（归纳）学习

学习线性投影矩阵，将下位词的词向量投影至上位词

方法：从未标注数据和语料库学习上下位关系的语义

目的：学习中文不同领域的上下位关系的复杂语义表达

## 迭代投影模型 (IPM)

输入数据集： $D^P, D^U$

学习范式：半监督学习

方法：建模非上下位关系表示，加入语言规则

目的：更加地学习上下位vs. 非上下位学习的分类决策边界

## 模糊正交投影模型 (FOPM)

输入数据集： $D^P, D^N$

学习范式：监督（归纳）学习

方法：改进投影学习的数学建模方法

目的：提升模型在不同语言上预测上下位关系的能力，提高模型精度

## 转导投影模型 (TPM)

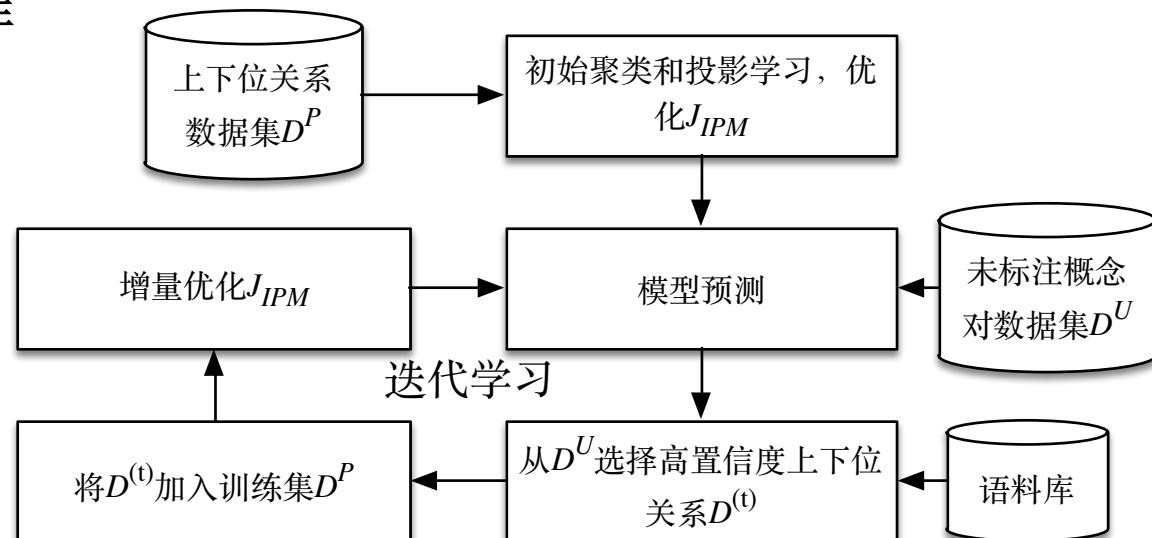
输入数据集： $D^P, D^N, D^T$

学习范式：监督（转导）学习

# 半监督式上下位关系扩展（1）

- 半监督式迭代投影模型

- 任务目标：给定少量中文上下位关系作为“种子”，进行上下位关系扩展
- 技术难点
  - 中文上下位关系相关语言模式的低覆盖率→利用词嵌入模型
  - 中文训练集大小有限→半监督迭代学习
- 方案：迭代学习下位词到上位词在词嵌入空间的投影，使用中文语言模式监督迭代学习过程



Chengyu Wang, Xiaofeng He. Chinese Hypernym-Hyponym Extraction from User Generated Categories. **COLING 2016 (CCF-B)**

Chengyu Wang, Yan Fan, Xiaofeng He, Aoying Zhou. Predicting Hypernym-Hyponym Relations for Chinese Taxonomy Learning. **KAIS 58(3): 585–610 (2019) (CCF-B)**

# 半监督式上下位关系扩展 (2)

- 初始模型训练

- 观察：上下位关系有复杂的语义，**不同粒度、不同领域**的上下位关系在词嵌入空间有不同的关系表示

类别	示例	$\ \vec{x}_i - \vec{y}_i\ _2$
真正例	$\vec{v}(\text{日本}) - \vec{v}(\text{国家}) \approx \vec{v}(\text{澳大利亚}) - \vec{v}(\text{国家})$	$1.03 \approx 0.99$
现象 1	$\vec{v}(\text{日本}) - \vec{v}(\text{国家}) \not\approx \vec{v}(\text{日本}) - \vec{v}(\text{亚洲国家})$	$1.03 \not\approx 0.71$
现象 2	$\vec{v}(\text{日本}) - \vec{v}(\text{国家}) \not\approx \vec{v}(\text{主权国}) - \vec{v}(\text{国家})$	$1.03 \not\approx 1.32$
现象 3	$\vec{v}(\text{日本}) - \vec{v}(\text{国家}) \not\approx \vec{v}(\text{西瓜}) - \vec{v}(\text{水果})$	$1.03 \not\approx 0.39$

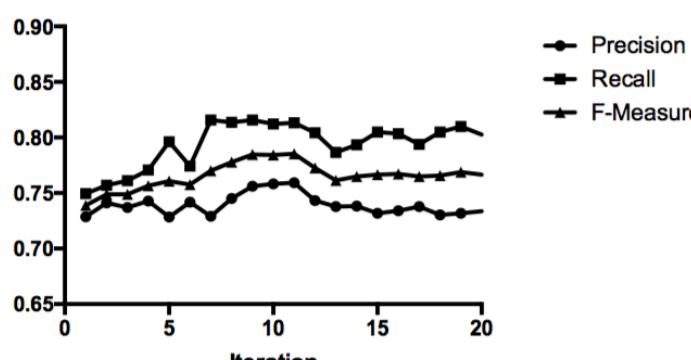
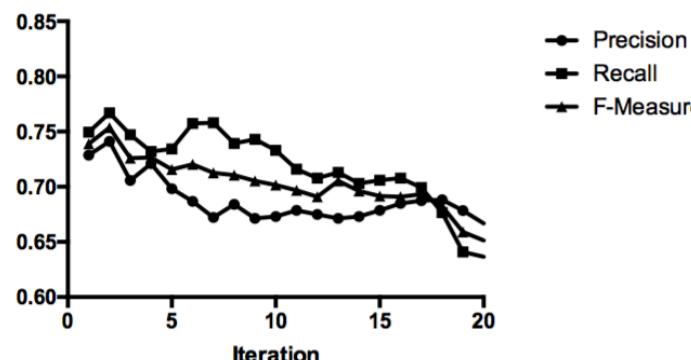
- 模型设计：**分段线性投影模型**（学习多个投影矩阵，将下位词的词向量投影到上位词）

$$J_{IPM} = \frac{1}{K} \sum_{k=1}^K \sum_{(x_i, y_i) \in C_k} \|\mathbf{M}_k \vec{x}_i + \vec{b}_k - \vec{y}_i\|^2$$

- 迭代模型训练（迭代次数  $t = 1, \dots, T$ ）

1. 采样：从  $D^U$  **采样** 部分未标注数据  $U^{(t)}$
2. 预测：使用第  $t$  个迭代**投影模型**筛选出正例  $U_+^{(t)}$
3. 选择：使用基于**中文语言模式**的选择算法，从  $U_+^{(t)}$  选择**高置信度正例**  $U_*^{(t)}$
4. 更新：**增量更新**训练集，得到第  $t + 1$  个迭代**投影模型**

# 半监督式上下位关系扩展（3）

- 模型迭代分析
  - 模型在每个迭代的实验结果（先提升，后稳定）
  - 去除关系选择的结果
- 整体实验结果
  - 采用词嵌入建模下位词到上位词的投影，取得较好结果
  - 中文语言模式精度有限，但可有效“监督”半监督学习过程

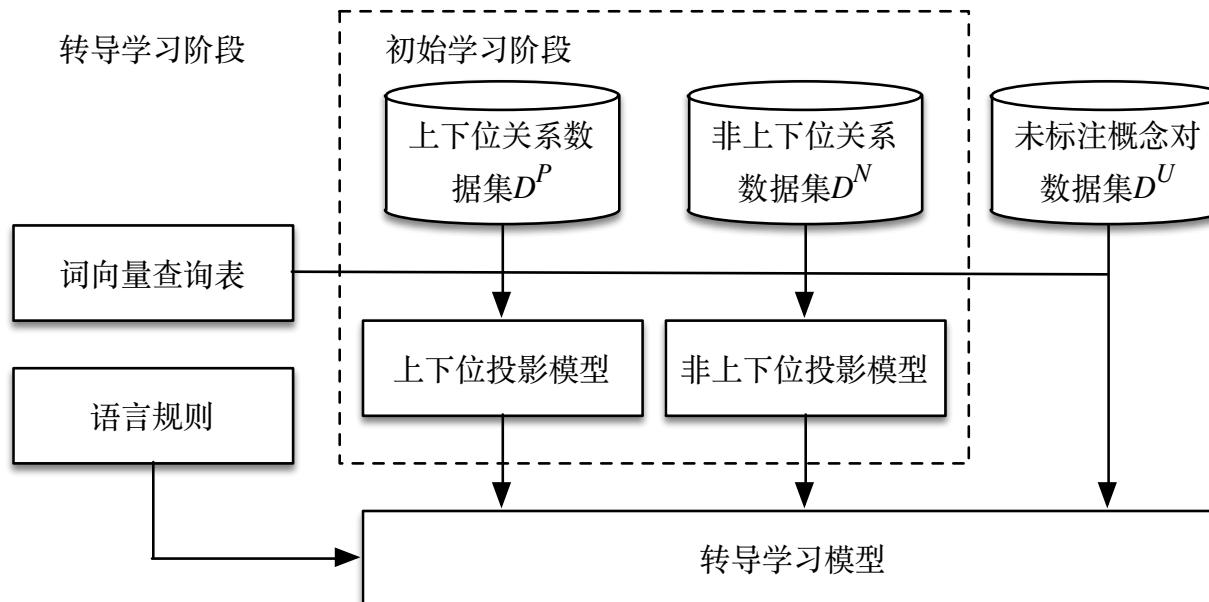
方法	精准度	召回率	F 值
基线方法			
Hearst [38]	0.962	0.198	0.328
Snow 等人 [53]	0.673	0.281	0.396
CN-WikiTaxonomy [21]	<b>0.985</b>	0.254	0.404
invCL [64]	0.485	0.581	0.529
Fu 等人 [25]	0.717	0.749	0.733
IPM 及其变体			
IPM-Initial	0.741	0.767	0.753
IPM-Random	0.690	0.757	0.722
IPM-Positive	0.754	0.801	0.776
IPM	0.758	0.814	0.786
IPM&CN-WikiTaxonomy	0.788	<b>0.847</b>	<b>0.816</b>

# 基于转导学习的上下位关系分类（1）

- 监督式转导投影模型

- 模型改进点：

- 同时利用**正例**（上下位关系）和**非正例**（非上下位关系）的训练数据，进行关系分类
    - 建模上下位关系投影的**非线性投影分量**
    - 支持**语言规则**的模型注入



# 基于转导学习的上下位关系分类 (2)

## • 转导学习

- 初始预测损失: 转导学习的最终结果与初始模型预测结果接近

$$\mathcal{O}_s = \|\mathbf{W}(\mathbf{F} - \mathbf{S})\|_2^2$$

↑  
初始预测置信度（作为转导学习阶段权重）      最终预测分数      初始预测分数

- 违反语言规则损失: 转导学习的最终结果应该尽可能不违反语言规则

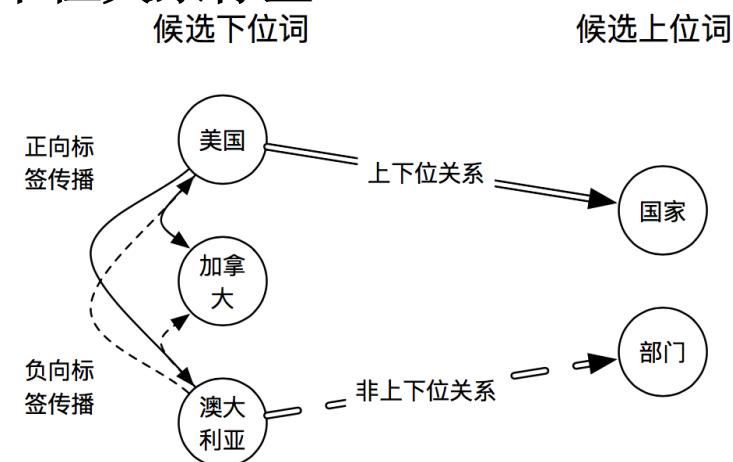
$$\mathcal{O}_r = \|\mathbf{F} - \mathbf{R}\|_2^2$$

- 非线性映射损失: 如果两个实体 $x_i$ 和 $x_j$ 的语义相似, 则对于某概念 $y$ ,  $(x_i, y)$ 和 $(x_j, y)$ 有相似的上下位或非上下位关系标签

$$sim(p_i, p_j) = \begin{cases} \cos(\vec{x}_i, \vec{x}_j) & y_i = y_j \\ 0 & \text{其他} \end{cases}$$

$$\mathcal{O}_n = \mathbf{F}^T \mathbf{P}^{-1} \mathbf{F}$$

协方差相似度矩阵



# 基于转导学习的上下位关系分类 (3)

- 中文实验结果
  - 整体实验结果

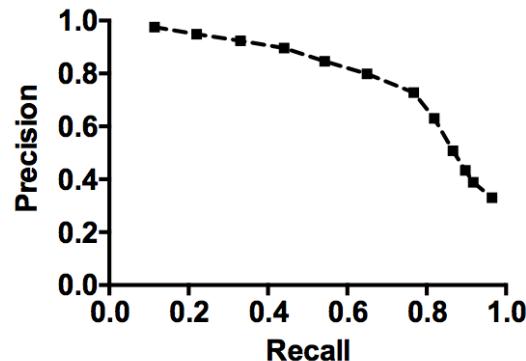
数据集	FD			BK		
	方法	精准度	召回率	F 值	精准度	召回率
Fu-S [25]	0.641	0.560	0.598	0.714	0.648	0.679
Fu-P [25]	0.664	0.593	0.626	0.727	0.675	0.700
$\vec{x}_i \oplus \vec{y}_i$	0.677	<b>0.752</b>	0.697	0.803	0.759	0.780
$\vec{x}_i + \vec{y}_i$	0.653	0.607	0.629	0.727	0.656	0.689
$\vec{x}_i - \vec{y}_i$	0.719	0.606	0.657	0.784	0.607	0.684
IPM	0.693	0.645	0.669	0.739	0.698	0.718
TPM-Initial	0.707	0.692	0.699	0.817	0.785	0.800
TPM	<b>0.728</b>	0.705	<b>0.716</b>	<b>0.836</b>	<b>0.806</b>	<b>0.821</b>

- 算法细节分析

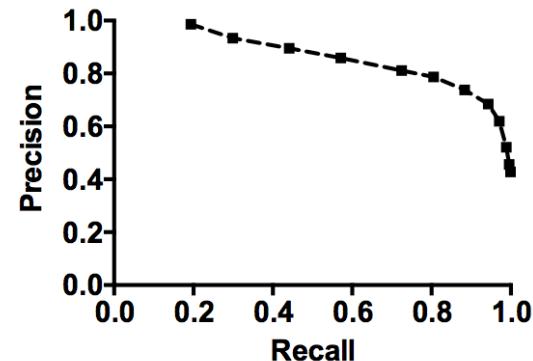
真正率/真负率	P1	N1	N2
数据集 FD	0.986	0.923	0.941
数据集 BK	0.976	0.968	0.973

语言规则的有效性

参数 $\theta$ 分析



(a) 数据集 : FD



(b) 数据集 : BK

# 基于模糊正交投影的上下位关系分类 (1)

## • 模糊正交投影模型

- 模型改进点：同时建模上下位与非上下位关系的**复杂投影**关系，提升模型语言独立性
  - 上下位关系：**不同粒度、不同领域**的上下位关系
  - 非上下位关系：**近义词关系、反义词关系、整体部分关系等**
- 上下位关系投影建模方法
  - 正交投影

$$\min \sum_{(x_i, y_i) \in D^P} \|\mathbf{M}^P \vec{x}_i - \vec{y}_i\|^2 \text{ s. t. } (\mathbf{M}^P)^T \cdot \mathbf{M}^P = \mathbf{I}$$

↑  
归一化词向量

加入**正交性约束**，使得投影后的词向量也是归一化的

- 模糊正交投影

$$\tilde{J}(\mathcal{M}^P) = \frac{1}{2} \sum_{k=1}^K \sum_{(x_i, y_i) \in D^P} a_{i,k}^P \|\mathbf{M}_k^P \vec{x}_i - \vec{y}_i\|^2$$

$a_{i,k}^P$ : 训练数据 $(x_i, y_i)$ 对第 $k$ 个上下位关系投影分量的权重  
(启发式方法学习)

$$\text{s. t. } (\mathbf{M}_k^P)^T \cdot \mathbf{M}_k^P = \mathbf{I}, \quad \sum_{(x_i, y_i) \in D^P} a_{i,k}^P = 1, k = 1, \dots, K$$

# 基于模糊正交投影的上下位关系分类 (2)

- 上下位关系vs.非上下位关系分类

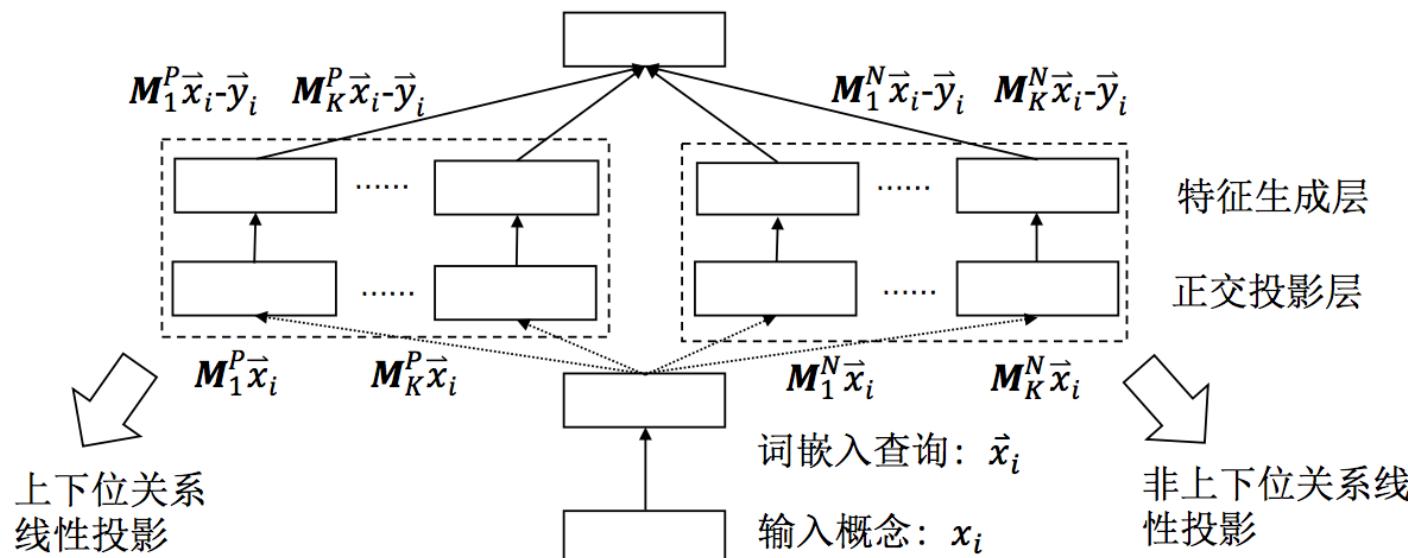
- 特征：分析计算两组映射（基于 $2K$ 个投影矩阵）的残差

$$\mathcal{F}^P(\vec{x}_i, \vec{y}_i) = (\mathbf{M}_1^P \vec{x}_i - \vec{y}_i) \oplus (\mathbf{M}_2^P \vec{x}_i - \vec{y}_i) \oplus \cdots \oplus (\mathbf{M}_K^P \vec{x}_i - \vec{y}_i) \quad \text{正例较小, 反例较大}$$

$$\mathcal{F}^N(\vec{x}_i, \vec{y}_i) = (\mathbf{M}_1^N \vec{x}_i - \vec{y}_i) \oplus (\mathbf{M}_2^N \vec{x}_i - \vec{y}_i) \oplus \cdots \oplus (\mathbf{M}_K^N \vec{x}_i - \vec{y}_i) \quad \text{正例较大, 反例较小}$$

- 整体网络架构

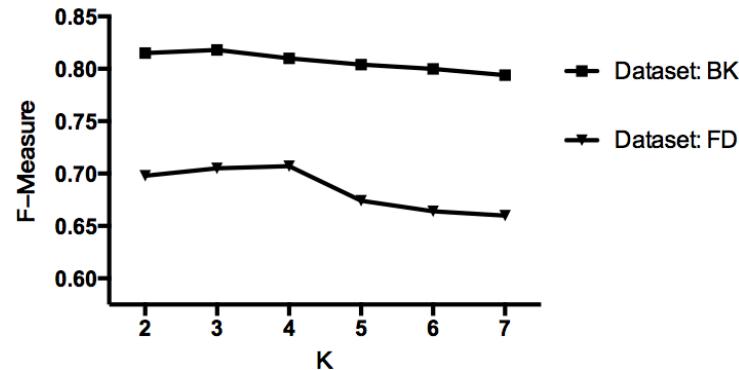
上下位关系/非上下位关系分类器： $f$



# 基于模糊正交投影的上下位关系分类（3）

- 中文数据集的实验结果

数据集	FD			BK		
	方法	精准度	召回率	F 值	精准度	召回率
Fu-S [25]	0.641	0.560	0.598	0.714	0.648	0.679
Fu-P [25]	0.664	0.593	0.626	0.727	0.675	0.700
$\vec{x}_i \oplus \vec{y}_i$	0.677	<b>0.752</b>	0.697	0.803	0.759	0.780
$\vec{x}_i + \vec{y}_i$	0.653	0.607	0.629	0.727	0.656	0.689
$\vec{x}_i - \vec{y}_i$	0.719	0.606	0.657	0.784	0.607	0.684
IPM	0.693	0.645	0.669	0.739	0.698	0.718
TPM	<b>0.728</b>	0.705	<b>0.716</b>	<b>0.836</b>	<b>0.806</b>	<b>0.821</b>
FOPM	0.713	0.698	<b>0.705</b>	<b>0.825</b>	<b>0.812</b>	<b>0.818</b>



参数K对实验结果的影响  
(需要调整的唯一参数)

- 综述研究

- 针对上下位关系预测和分类体系构建，综述100+篇近期论文
- 成为SemEval-2018 Task 9: Hypernym Discovery官方参考文献

# 第二部分：知识增强的语义关系抽取

前序研究：基于词嵌入投影模型的上下位关系预测

方法：利用对抗学习，使利用原始训练集学习的投影神经网络学习到分类体系中的知识

目的：融合分类体系和训练集中的上下位关系知识，提升投影模型的效果

多知识源

分类体系增强的对抗学习框架（TEAL）

输入数据集： $D^P$ 、 $D^U$ 、 $T^P$ 、 $T^U$   
学习范式：监督（归纳）学习+对抗学习

方法：利用深度迁移学习和双语词典生成技术，将前述FOPM模型扩展到跨语言投影学习的情况

目的：实现跨语言上下位关系预测，在小样本学习情景下，提高小语种上下位关系预测精度

多语言

迁移模糊正交投影模型（TFOPM）、  
迭代迁移模糊正交投影模型（ITFOPM）

输入数据集： $D_S^P$ 、 $D_S^N$ 、 $D_T^P$ 、 $D_T^N$   
学习范式：监督（归纳）学习+迁移学习

方法：提出超球学习方法，使得具有不同词汇关系的关系元组更容易区分

目的：建模的不同词汇关系的表示，实现多种词汇关系（包括上下位关系）的分类

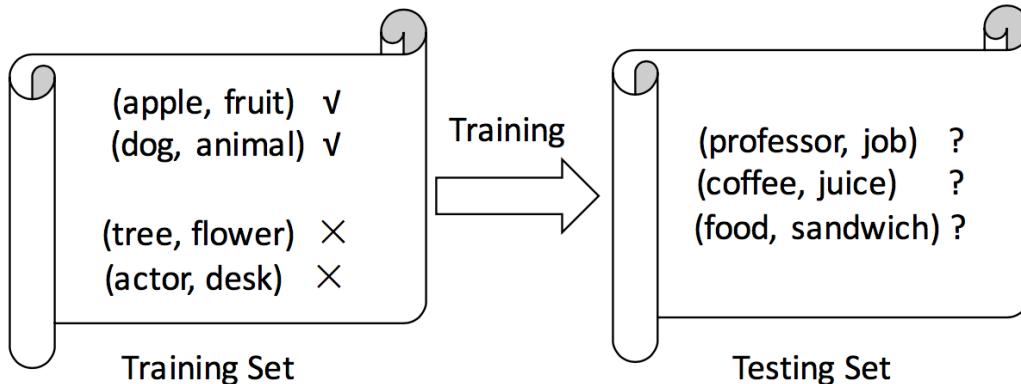
多词汇关系

超球关系嵌入学习（SphereRE）

输入数据集： $D$ 、 $U$   
学习范式：监督（转导）学习

# 分类体系增强的对抗学习框架（1）

- 基本思路



分类体系	上下位关系数量
WikiTaxonomy ( 英语 ) [19]	105418
YAGO ( 英语 ) [8]	8277227
WiBi ( 英语 ) [96]	2736022
Probbase ( 英语 ) [10]	16285393
Probbase+ ( 英语 ) [97]	21332357
CN-WikiTaxonomy ( 中文 ) [21]	1317956
CN-Probbase ( 中文 ) [20]	32925306

采用**目标训练数据**训练

- 数据量小
- 数据质量高
- 数据领域窄

采用**分类体系采样**数据训练

- 数据量大
- 数据有噪音
- 数据领域广

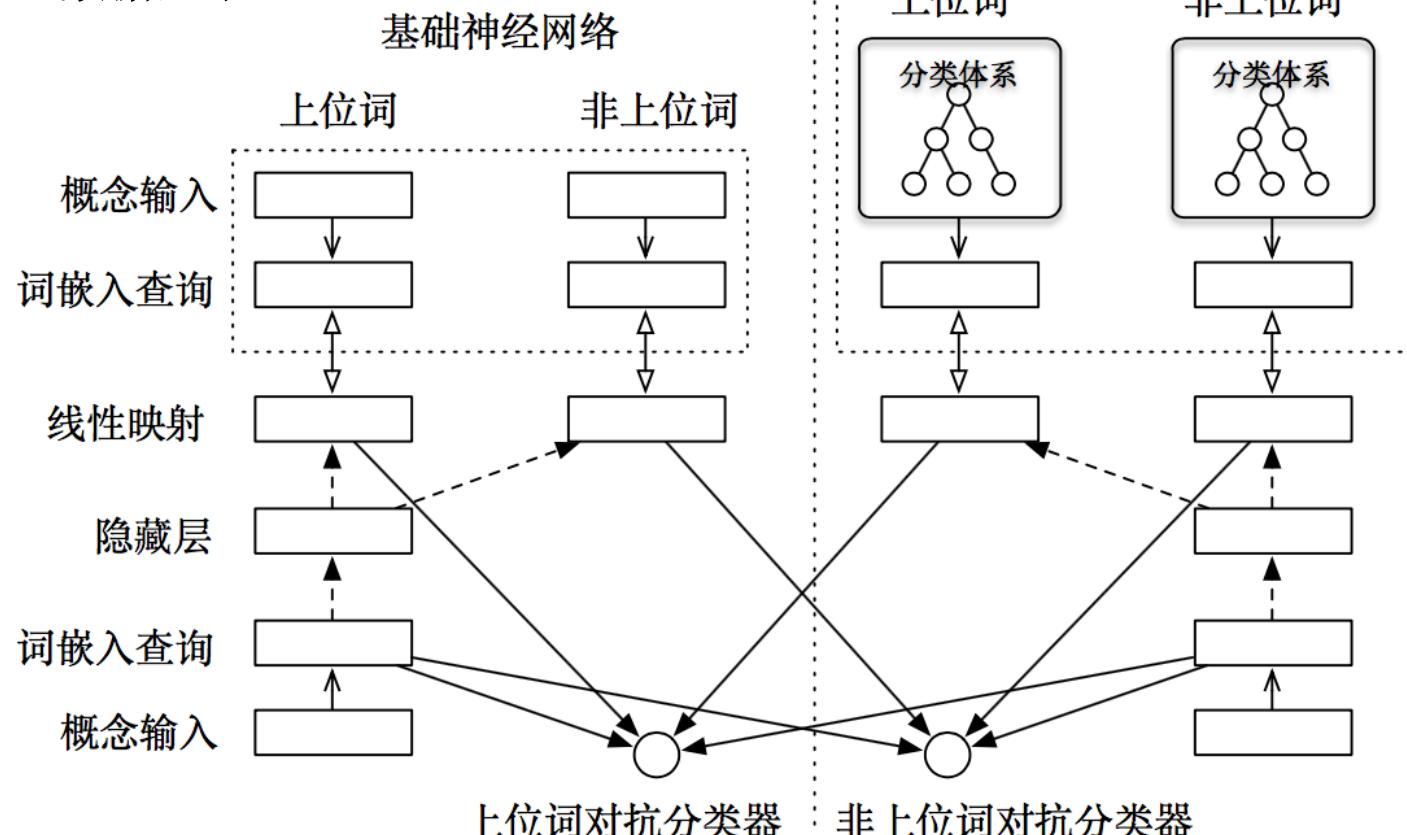
- 采用基于**深度神经网络**的对抗学习技术，将**分类体系**中的海量上-下位关系知识融入到基础神经网络中，提高基础神经网络的学习能力

# 分类体系增强的对抗学习框架 (2)

- 分类体系增强的神经网络的设计：训练**Student Network**同时，模仿**Teacher Network**的行为

## Student Network:

采用训练集训练  
(数据量小)



## Teacher Network:

采用分类体系采样  
数据集训练 (数据量大)

**上位词对抗分类器：**  
给定下位词，区分生成的上位词词向量从哪个神经网络学习到

**非上位词对抗分类器：**给定下位词，区分生成的非上位词词向量从哪个神经网络学习到

# 分类体系增强的对抗学习框架（3）

- 中文实验结果

数据集	FD			BK		
方法	精准度	召回率	F 值	精准度	召回率	F 值
现有工作中的强基线算法						
$\vec{x}_i \oplus \vec{y}_i$	0.677	<b>0.752</b>	0.697	0.803	0.759	0.780
本文先前工作的最佳结果						
TPM	<b>0.728</b>	0.705	0.716	<b>0.836</b>	0.806	0.821
TEAL 框架的实验结果						
TEAL-S	0.695	0.684	0.689	0.788	0.869	0.827
TEAL-AS	0.721	0.736	<b>0.728</b>	0.791	0.870	<b>0.829</b>

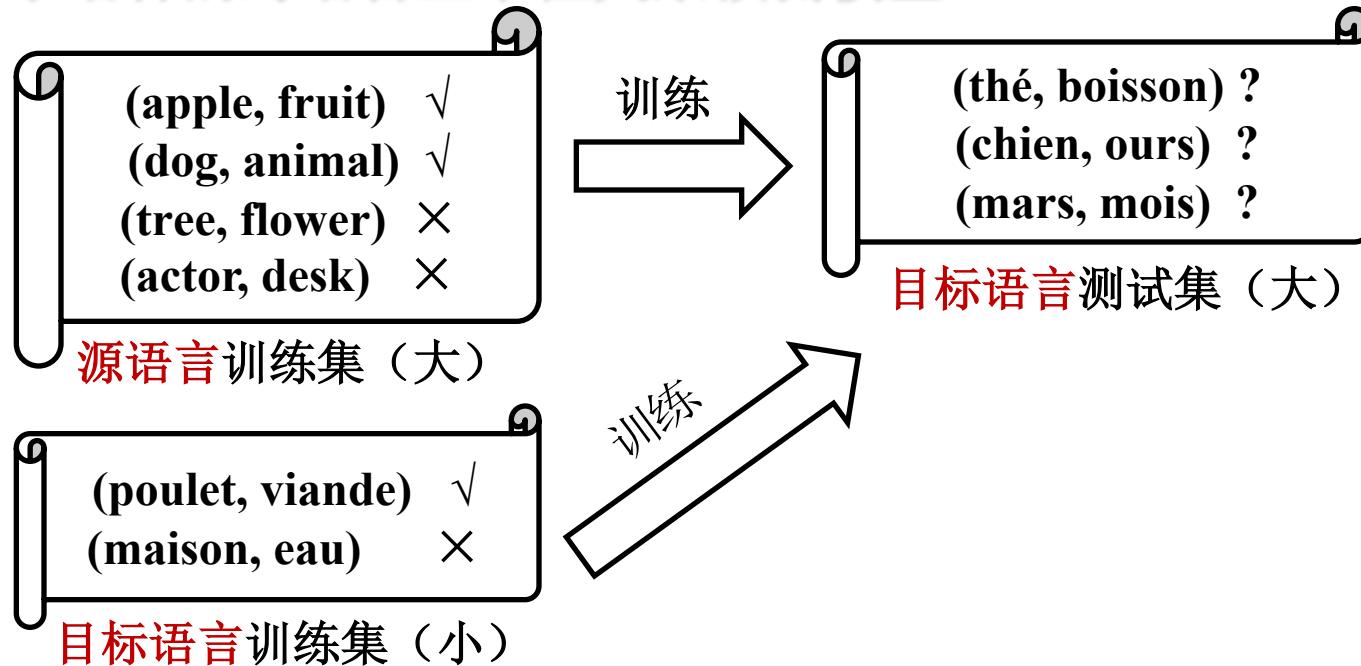
- 应用：对Microsoft Concept Graph的扩展

- 给定某上位词，利用本模型和基于Word2Vec的KNN搜索，获得新的下位词

上位词	# 正确/# 总数	准确率	上位词	# 正确/# 总数	准确率
material	78/102	0.76	goods	20/20	1.00
person	17/19	0.89	sector	18/20	0.90
group	37/43	0.86	component	76/80	0.95
technology	12/14	0.86	individual	24/24	1.00
provision	15/15	1.00	location	8/9	0.89
合计	302/346	0.87			

# 基于迁移学习的跨语言上下位关系抽取（1）

- 面向小语种的跨语言上下位关系预测模型



- 小语种（非英语）上下位关系抽取的训练集比较小，难以直接获得或人工标注
- 扩展**模糊正交投影模型**，结合**双语字典生成**（**Bilingual lexicon induction**）和**深度迁移学习**，实现跨语言的上下位关系预测

# 基于迁移学习的跨语言上下位关系抽取（2）

- **迁移模糊正交投影模型（TFOPM）的基本任务**
  - 输入：源语言上下位/非上下位关系数据集 $D_S^P$ 和 $D_S^N$ 、目标语言上下位/非上下位关系数据集 $D_T^P$ 和 $D_T^N$
  - 预测目标：目标语言术语对数据集 $U_T$
- 跨语言上下位关系映射学习
  - $S$ : 将源语言词向量映射到目标语言（双语字典生成）
  - $\gamma$ : 控制源语言单个训练数据对目标语言重要性（启发式规则）
  - $\beta$ : 控制源语言整体训练损失（可调）

$$\tilde{J}(\mathcal{M}^P) = \frac{\beta}{2} \sum_{(x_i, y_i) \in D_S^P} \sum_{k=1}^K a_{i,k}^P \gamma_{i,k}^P \|\mathbf{M}_k^P \cdot \mathbf{S}\vec{x}_i - \mathbf{S}\vec{y}_i\|^2$$

$$+ \frac{1-\beta}{2} \sum_{(x_i, y_i) \in D_T^P} \sum_{k=1}^K a_{i,k}^P \|\mathbf{M}_k^P \vec{x}_i - \vec{y}_i\|^2$$

$$\text{s. t. } (\mathbf{M}_k^{P^T}) \cdot \mathbf{M}_k^P = \mathbf{I}, \quad \sum_{(x_i, y_i) \in D_S^P} a_{i,k}^P \gamma_{i,k}^P = 1, \quad \sum_{(x_i, y_i) \in D_T^P} a_{i,k}^P = 1$$

$$k = 1, \dots, K$$

- **迭代迁移模糊正交投影模型（ITFOPM）**
  - 采用半监督迭代学习，扩展目标语言的训练集

# 基于迁移学习的跨语言上下位关系抽取（3）

## • 整体实验分析

- 数据集：从**Open Multilingual Wordnet**计划生成训练集和测试集
- 任务一：跨语言上下位关系方向分类
  - 上下位关系vs.反上下位关系
- 任务二：跨语言上下位关系检测
  - 上下位关系vs.非上下位关系

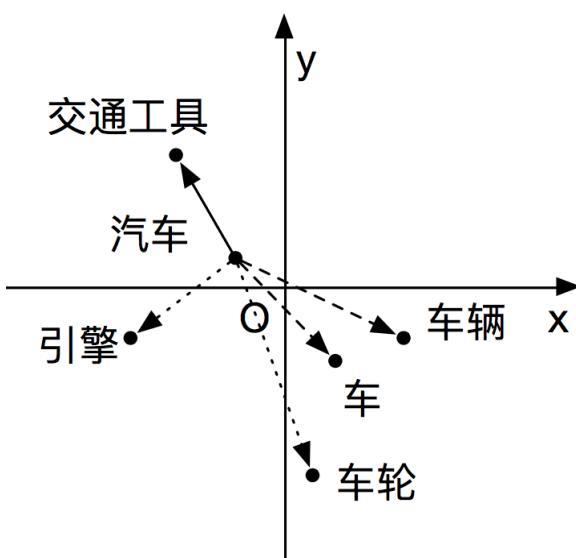
关系 ↓ 语言 →	fr	zh	ja	it	th	fi	el
# 上下位关系	4,035	2,962	1,448	3,034	1,156	7,157	2,612
# 非上下位关系	8,947	6,382	3,203	6,081	1,977	9,433	1,454

方法	fr	zh	ja	it	th	fi	el
任务：跨语言上下位关系方向分类							
Santus 等人 [66]	0.65	0.65	0.68	0.61	0.63	0.70	0.62
Weeds 等人 [74]	0.76	0.71	0.77	0.76	0.72	0.77	0.70
Kiela 等人 [160]	0.67	0.65	0.71	0.68	0.65	0.70	0.62
Shwartz 等人 [50]	0.79	0.67	0.71	0.72	0.66	0.75	0.66
<b>TFOPM-N</b>	0.78	0.71	0.75	0.76	0.73	0.76	0.71
<b>TFOPM</b>	0.80	0.72	0.76	0.78	0.75	0.78	0.73
<b>ITFOPM-N</b>	<b>0.82</b>	0.72	0.76	0.78	0.75	<b>0.81</b>	0.72
<b>ITFOPM</b>	0.81	<b>0.74</b>	<b>0.78</b>	<b>0.81</b>	<b>0.78</b>	<b>0.81</b>	<b>0.75</b>
任务：跨语言上下位关系检测							
Santus 等人 [66]	0.67	0.63	0.67	0.62	0.64	0.62	0.64
Weeds 等人 [74]	0.74	0.66	0.68	0.71	0.62	0.68	0.69
Kiela 等人 [160]	0.70	0.61	0.65	0.68	0.57	0.61	0.67
Shwartz 等人 [50]	0.72	0.66	0.69	0.64	0.66	0.69	0.70
<b>TFOPM-N</b>	0.72	0.67	0.70	0.70	0.68	0.71	0.70
<b>TFOPM</b>	0.75	0.71	0.76	0.72	0.69	0.72	0.71
<b>ITFOPM-N</b>	0.72	<b>0.74</b>	0.77	<b>0.74</b>	0.67	0.71	0.72
<b>ITFOPM</b>	<b>0.76</b>	0.73	<b>0.78</b>	<b>0.74</b>	<b>0.72</b>	<b>0.73</b>	<b>0.73</b>

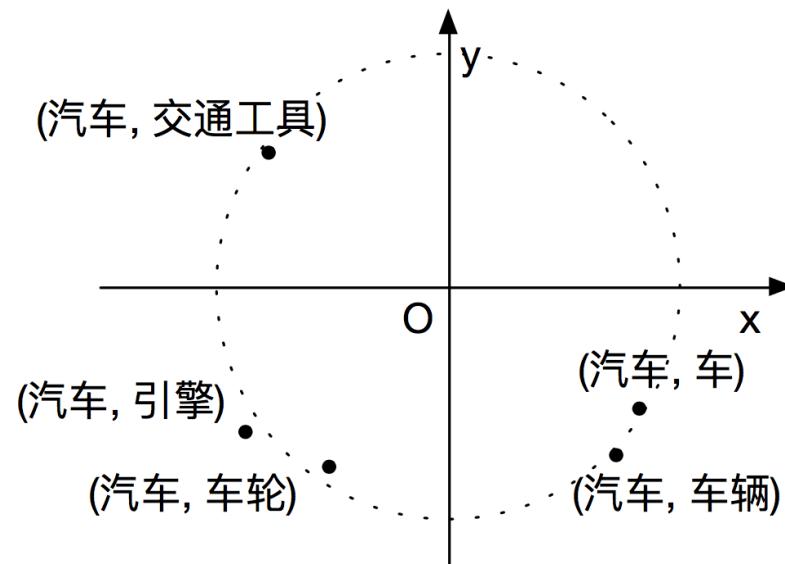
# 基于超球学习的词汇关系分类 (1)

- 从二元关系分类到多词汇关系分类

- 词汇关系：上下位关系、近义词关系、反义词关系、整体部分关系等
- 问题：词汇关系大多为常识性知识，语义模式覆盖率低，语言模式的表达比较模糊
- 方案：将具有相似词汇关系的术语对映射到超球空间下相近的位置



(a) 词向量空间

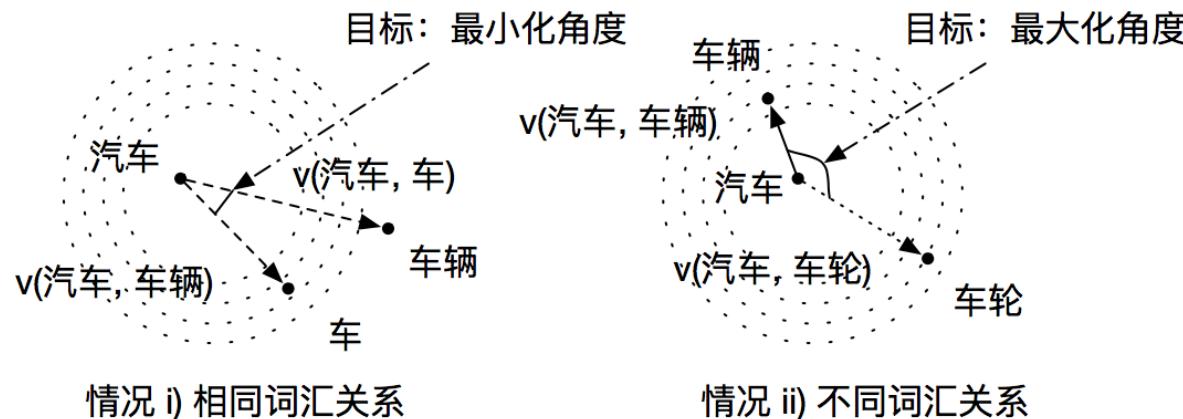


(b) 关系超球空间

# 基于超球学习的词汇关系分类 (2)

- **SphereRE: 超球关系嵌入学习**  $J(\Theta) = J_f + \lambda_1 J_g + \lambda_2 \|\Theta\|^2$ 
  - **投影学习**  $f_m(\vec{x}_i)$ : 将具有词汇关系  $r_m$  关系主语  $x_i$  在词嵌入空间投影至关系主语  $y_i$ 
$$J_f = \sum_{i=1}^{|D|} \sum_{r_m \in R} I(r_i = r_m) \|f_m(\vec{x}_i) - \vec{y}_i\|^2$$
  - **超球关系学习**: 有相同词汇关系的术语对在超球空间内距离最小化, 有不同词汇关系的术语对在超球空间内距离最大化

$$J_g = \sum_{i,j}^{D \cup U} \delta(r_i, r_j) g(f_i(\vec{x}_i) - \vec{x}_i, f_j(\vec{x}_j) - \vec{x}_j)$$



# 基于超球学习的词汇关系分类 (3)

- 整体实验结果（四个公开数据集）

方法	K&H+N			BLESS		
	准确度	召回率	F 值	准确度	召回率	F 值
$\vec{x}_i \oplus \vec{y}_i$	0.909	0.906	0.904	0.811	0.812	0.811
$(\vec{x}_i \oplus \vec{y}_i)_h$	0.983	0.984	0.983	0.891	0.889	0.889
$\vec{x}_i - \vec{y}_i$	0.888	0.886	0.885	0.801	0.803	0.802
$(\vec{x}_i - \vec{y}_i)_h$	0.941	0.942	0.941	0.861	0.859	0.860
NPB	0.713	0.604	0.55	0.759	0.756	0.755
LexNET	0.985	0.986	0.985	0.894	0.893	0.893
LexNET <sub>h</sub>	0.984	0.985	0.984	0.895	0.892	0.893
NPB+Aug	-	-	0.897	-	-	0.842
LexNET+Aug	-	-	0.970	-	-	0.927
<b>SphereRE</b>	<b>0.990</b>	<b>0.989</b>	<b>0.990</b>	<b>0.938</b>	<b>0.938</b>	<b>0.938</b>
方法	ROOT09			EVALution		
	准确度	召回率	F 值	准确度	召回率	F 值
$\vec{x}_i \oplus \vec{y}_i$	0.636	0.675	0.646	0.531	0.544	0.525
$(\vec{x}_i \oplus \vec{y}_i)_h$	0.712	0.721	0.716	0.57	0.573	0.571
$\vec{x}_i - \vec{y}_i$	0.627	0.655	0.638	0.521	0.531	0.528
$(\vec{x}_i - \vec{y}_i)_h$	0.683	0.692	0.686	0.536	0.54	0.539
NPB	0.788	0.789	0.788	0.53	0.537	0.503
LexNET	0.813	0.814	0.813	0.601	0.607	0.6
LexNET <sub>h</sub>	0.812	0.816	0.814	0.589	0.587	0.583
NPB+Aug	-	-	0.778	-	-	0.489
LexNET+Aug	-	-	0.806	-	-	0.545
<b>SphereRE</b>	<b>0.860</b>	<b>0.862</b>	<b>0.861</b>	<b>0.62</b>	<b>0.621</b>	<b>0.62</b>

# 中文短文本关系挖掘与语义理解

- 概念短文本：上下位关系
  - 中国企业家、杭州师范大学教师、未来全球领袖、阿里巴巴人物
  - 应用上下位关系预测算法
- 关系短文本：非上下位关系
  - 杭州师范大学教师：“任职”
  - 1964年出生：“出生”
  - 长江商学院校友：“毕业”
- 中文短文本处理的困难性
  - 语义上下文稀疏、表达灵活、缺乏大量训练数据



马云

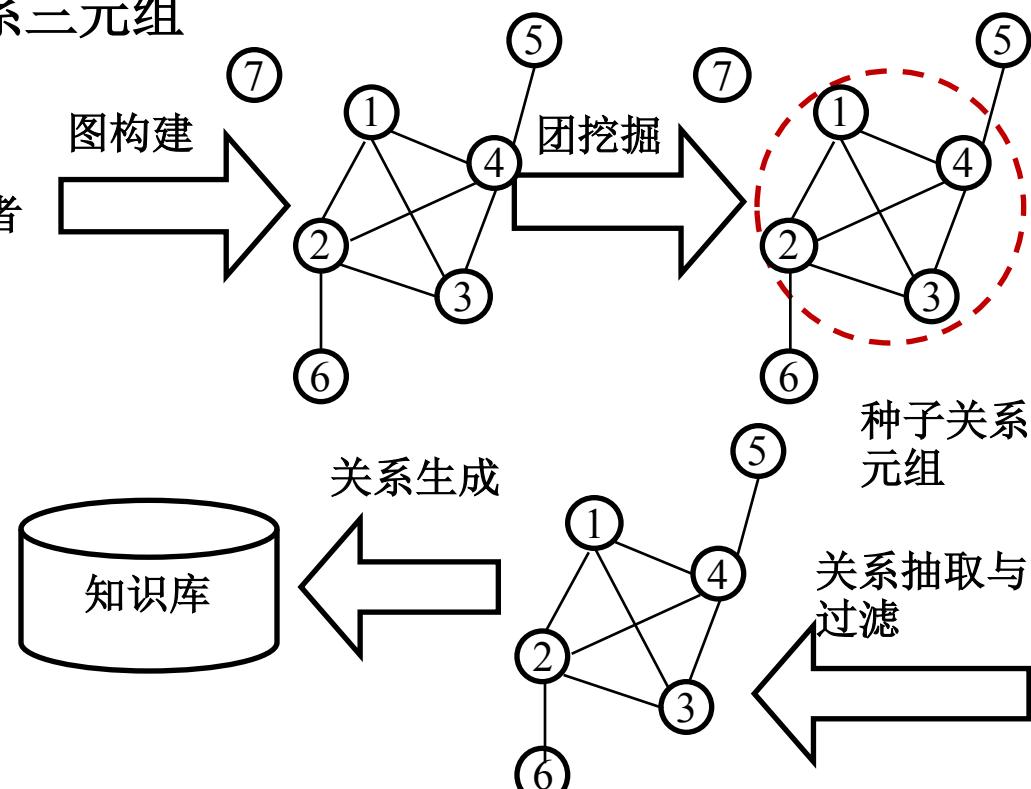
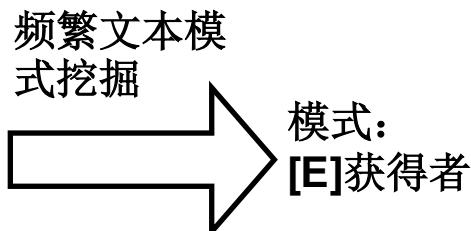
分类：法国荣誉军团骑士勋章持有人 | 1964年出生 | 在世人物 | 中国企业家 | 中华人民共和国亿万富豪 | 未来全球领袖 | 杭州师范大学校友  
| 杭州电子工业学院教师 | 杭州师范大学教师 | 动物福利相关人士 | 中华人民共和国企业家 | 香港科技大学荣誉博士 | 时代百大人物 | 阿里巴巴人物  
| 华谊兄弟人物 | 浙江企业家 | 中国首席执行官 | 中国共产党党员 | 长江商学院校友 | 杭州人 | 绍兴人 | 嵊州人 | 马姓  
| 改革开放40年百名杰出民营企业家 | 改革先锋称号获得者

# 基于模式挖掘的关系抽取（1）

## • 中文短文本的关系短文本挖掘

- 采用**频繁文本模式挖掘**获得关系描述模式
- 利用图的**团挖掘算法**得到关系**种子元组**
- 利用**语义相似度**抽取完整关系三元组

图灵奖获得者  
霍维茨奖获得者  
葛莱美奖获得者  
总统自由勋章获得者  
.....



Chengyu Wang, Yan Fan, Xiaofeng He, Aoying Zhou. Learning Fine-grained Relations from Chinese User Generated Categories. **EMNLP 2017 (CCF-B)**

Chengyu Wang, Yan Fan, Xiaofeng He, Aoying Zhou. Decoding Chinese User Generated Categories for Fine-grained Knowledge Harvesting. **TKDE 31(8): 1491–1505(2019) (CCF-A)**

# 基于模式挖掘的关系抽取（2）

## • 实验分析

- 实验数据：中文维基百科，约60万个中文实体和240万个**中文实体-关系类别对**，无人工标注
- 整体实验结果

关系类别	关系元组数量	准确度	覆盖度
毕业	44,118	98.0%	22.9%
位于	29,460	97.2%	8.5%
建立	20,154	95.0%	31.5%
出生	11,671	98.3%	41.4%
成员	8,445	96.0%	4.2%
启用	8,956	98.2%	21.6%
逝世	5,597	100.0%	18.4%
得奖	3,262	90.0%	27.3%

### 整体准确度评估

方法	预估准确度
PNRE-Conf	74.4%
PNRE-Filter	94.2%
<b>PNRE</b>	<b>97.4%</b>

人工抽样统计

抽取出的关系在**CN-Dbpedia V2.0**的覆盖率，  
**低覆盖率**表示抽取出的关系具有**高Novelty**，  
利于**知识图谱补全**

- 其他实验细节参见论文Wang et al. Decoding Chinese User Generated Categories for Fine-grained Knowledge Harvesting. TKDE<sup>32</sup>

# 数据驱动的关系挖掘（1）

- 数据驱动的中文短文本关系挖掘框架

- 解决问题：关系的表达有“长尾分布”，“长尾”的语义关系无法通过频繁模式挖掘抽取
- 三个模块：修饰词敏感的短语分割（MPS）、候选关系元组生成（CRG）、缺失关系谓词检测（MRPD）

输入

	概念	短文本（经过中文分词）
1	布鲁塞尔	欧盟 委员会 总部 城市
2	布鲁塞尔	10世纪 建立 的 西欧 城市



# 数据驱动的关系挖掘（2）

- 四个领域的实验效果

- 领域数据：某领域的中文维基实体，及相应关系类别短语

方法	# 关系	准确度	Yield	# 关系	准确度	Yield	
领域	通用			政治			
CN-WikiRe [33]	87	41.7%	41	84	57.1%	48	
CN-RELNOUN [174]	31	93.5%	29	35	88.6%	31	
ZORE [34]	28	75.0%	21	34	76.4%	26	
Cui 等人 [172]	52	51.9%	27	51	43.1%	22	
PNRE	193	<b>94.3%</b>	182	193	<b>95.9%</b>	185	
<b>DNRE</b>	<b>289</b>	92.7%	<b>268</b>	<b>314</b>	93.9%	<b>295</b>	
提升	+49.7%		+47.3%	+62.7%		+59.5%	
领域	娱乐			军事			
CN-WikiRe [33]	102	39.2%	40	76	53.9%	41	
CN-RELNOUN [174]	42	88.1%	37	34	82.3%	28	
ZORE [34]	21	76.2%	16	32	81.2%	26	
Cui 等人 [172]	54	48.1%	26	44	56.8%	25	
PNRE	204	<b>95.1%</b>	194	188	<b>96.3%</b>	181	
<b>DNRE</b>	<b>324</b>	92.3%	<b>299</b>	<b>274</b>	94.2%	<b>258</b>	
提升	+58.8%		+54.1%	+45.7%		+42.5%	

# 中文短语语义理解（1）

- 中文习语性（Idiomaticity）预测与关系推理

- 中文复合名词 ( $N_1N_2$ ) 的习语性分类

- 透明（Transparent）： $N_1$ 修饰 $N_2$ ，表示 $N_2$ 的一种属性
      - 示例：固体燃料
      - 推导上下位关系：（固体燃料, is-a, 燃料）
      - 推导属性：（固体燃料, has-property, 固体）
    - 部分模糊（Partly Opaque）： $N_1$ 和 $N_2$ 之间有动词性关系
      - 示例：办公用品
      - 推导上下位关系：（办公用品, is-a, 用品）
      - 推导语义关系：（办公用品, used-for, 办公）
    - 部分习语性（Partly Idiomatic）： $N_1$ 有暗喻含义
      - 示例：皮包公司
      - 推导上下位关系：（皮包公司, is-a, 公司）
    - 完全习语性（Completely Idiomatic）： $N_1$ 和 $N_2$ 完全不可分，表达某种概念
      - 示例：夫妻肺片

# 中文短语语义理解（2）

## • 中文习语性（Idiomaticity）预测模型

- 现象 1: 部分  $N_1$  和  $N_2$  的语言模式与  $N_1N_2$  的习语性程度相关
  - “固体的燃料”：“固体燃料”是透明的
- 现象 2: 具有相似组合性（Compositionality）的中文复合名词有相似的习语性程度
  - “固体燃料”中“固体”和“燃料”语义组合性低：“固体燃料”具有较低习语性
- 目标函数

$$\mathcal{J} = \sum_{x_i \in L} sl(f_i, \tilde{f}_i) + \lambda \sum_{x_i, x_j \in L \cup U} \mu_{i,j} ul(\tilde{f}_i, \tilde{f}_j)$$

- **Supervised Loss:** 分类损失，采用基于语言模式的特征
- **Unsupervised Loss:** 图损失，使相似组合性的数据预测标签相似

# 中文短语语义理解（3）

- 整体实验结果
  - CNCBaike、CNCWeb数据集：从百度百科和网络语料库采集中文复合名词，进行人工标注
  - 实验结果

数据集	CNCBaike			CNCWeb		
方法	精准度	召回率	F 值	精准度	召回率	F 值
$\vec{N}_1 + \vec{N}_2$	0.622	0.631	0.626	0.512	0.508	0.510
$\vec{N}_1 \oplus \vec{N}_2$	0.663	0.657	0.660	0.508	0.472	0.489
$\vec{N}_1 - \vec{N}_2$	0.567	0.606	0.586	0.597	0.478	0.531
King 和 Cook [198]	0.664	0.691	0.682	0.563	0.582	0.572
Salehi 等人 [202]	0.675	0.663	0.669	0.705	0.648	0.675
Cordeiro 等人 [204]	0.704	0.693	0.698	0.723	0.652	0.686
Pattern	0.770	0.766	0.768	0.745	0.687	0.715
RRL	0.785	0.776	0.780	0.762	0.703	0.731
<b>RCRL</b>	<b>0.801</b>	<b>0.783</b>	<b>0.792</b>	<b>0.784</b>	<b>0.733</b>	<b>0.758</b>

具体实验过程和Baseline详见论文

# 结论

- 面向中文短文本的关系抽取算法与主流关系抽取算法有显著差别
  - 使用词嵌入模型建模中文短文本的语义
  - 需要结合语言规则、部分语言模式、常识性知识，以及文本挖掘算法进行抽取
  - 深度自然语言理解仍然比较困难
- 未来研究展望
  - 融合异构知识源的中文关系抽取
  - 基于神经网络的复杂语义关系自动推理
  - 常识性知识的表示学习与关系补全
  - 编码中文语言学知识的神经网络模型
  - 基于深度语言模型的关系抽取与理解

**THANK YOU!**

**Questions & Answers?**