

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

个人简介

- 教育经历
 - 华东师范大学，工学学士，专业：软件工程（2011-2015）
 - 华东师范大学，工学博士，专业：软件工程（方向：数据科学与工程）（2015-2020）
 - 获得国家奖学金三次，博士研究生国家奖学金三次、校长奖学金等
- 科研情况
 - 研究兴趣包括知识抽取、知识图谱、计算语言学、自然语言理解等
 - 共计发表学术论文30余篇，包括第一作者CCF-A类论文6篇、第一作者CCF-B类论文7篇等

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）（中国互联网络发展状况报告）
 - 互联网总数据量：33ZB（2018年）→175ZB（2025年）（国际数据公司）
- 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

This screenshot shows a Google search results page for the query "grizzly". The top result is a link to the Grizzly.com website, which sells tools and machinery. Below the link are several snippets of information about grizzly bears, including their scientific name (*Ursus arctos horribilis*), habitat (North America), and some hunting statistics. There are also links to a parts store and a free catalog.

This screenshot shows a Google search results page for the query "how much does a bear weigh". The top result is a snippet from Wikipedia comparing the weights of various bear species. It includes images of giant pandas, polar bears, brown bears, and American black bears, along with their respective weight ranges.

This screenshot shows a Google search results page for the query "salmon". The top result is a snippet from Wikipedia about salmon species. It includes images of different types of salmon like wild Alaskan salmon and copper river salmon. Below the snippet is a shopping section showing various salmon products for sale, such as "Wild Alaskan Salmon Fillets" and "Smoked Salmon".

This screenshot shows a Google search results page for the query "tourist spot in vancouver". The top result is a snippet from Wikipedia about tourist spots in Vancouver. It includes images of Grouse Mountain, the Museum of Anthropology, Stanley Park, and Lynn Canyon Park. Each spot has a brief description and a rating.



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.
- Summary card: "2001年 · 奇幻电影/虚构作品 · 2 小时 39 分钟" (2001 Fantasy movie/Fictional work · 2 hours 39 minutes).
- Video thumbnail: "播放YouTube上的预告片" (Play trailer on YouTube).
- Rating: "95% 的用户顶了这部电影" (95% of users liked this movie).
- Description: "《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。"
- Image: Canadian and American flags.
- Image: Logo for "CITY OF VANCOUVER".
- Image: Portraits of Albert Einstein and Vincent van Gogh.

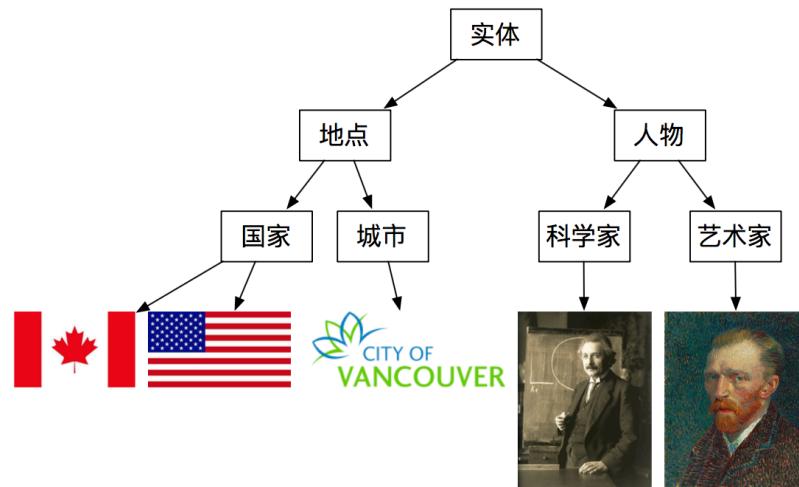
问题推荐

A screenshot of entity linking results for the query "watch harry potter". It shows a list of entities and their types:

- "哈利·波特与魔法石" (Movie)
- "国家" (Country)
- "城市" (City)
- "科学家" (Scientist)
- "艺术家" (Artist)

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）

ID	Pattern
1	$NP \text{ such as } \{NP,\}^* \{(or and)\} NP$
2	$\text{such } NP \text{ as } \{NP,\}^* \{(or and)\} NP$
3	$NP\{,\} \text{ including } \{NP,\}^* \{(or and)\} NP$
4	$NP\{,NP\}^* \{,\} \text{ and other } NP$
5	$NP\{,NP\}^* \{,\} \text{ or other } NP$
6	$NP\{,\} \text{ especially } \{NP,\}^* \{(or and)\} NP$

Probbase系统中使用的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

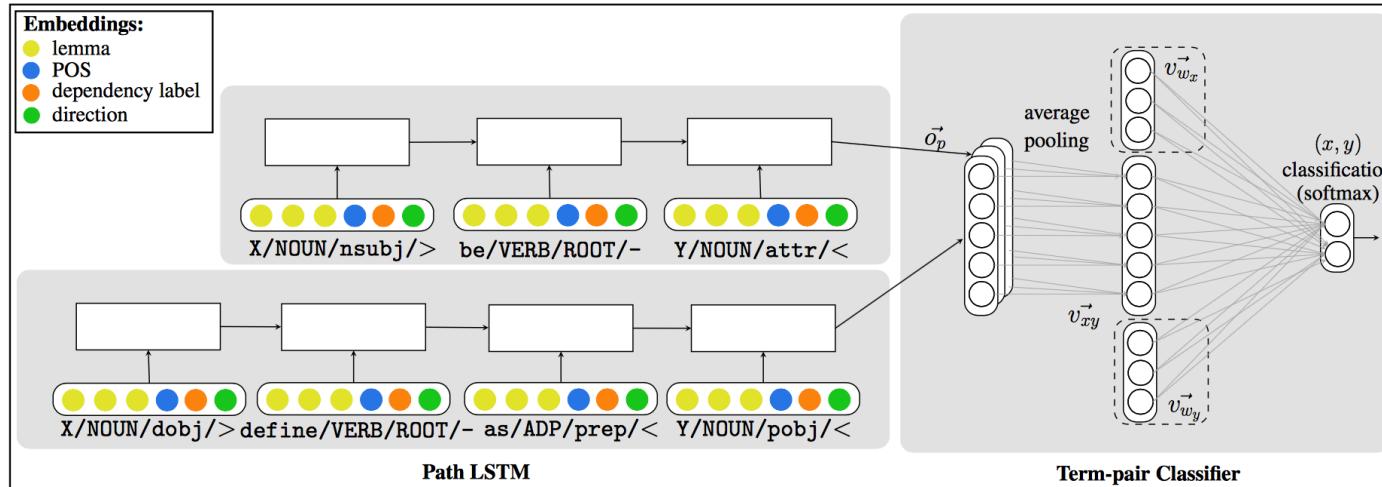
重要参考文献

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

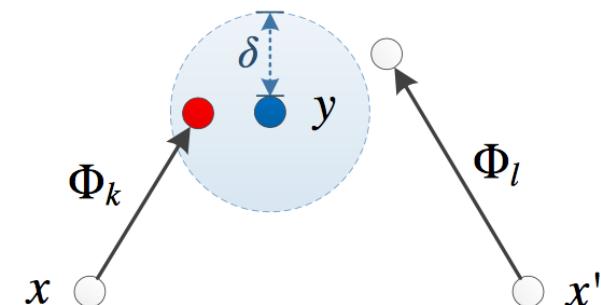
Fu et al. Learning Semantic Hierarchies via Word Embeddings. ACL 2014

分类体系与上下位关系抽取（3）

- 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，利用词嵌入丰富语义的同时，减轻“词汇记忆”问题
 - 充分结合中文语言模式和语言学规则



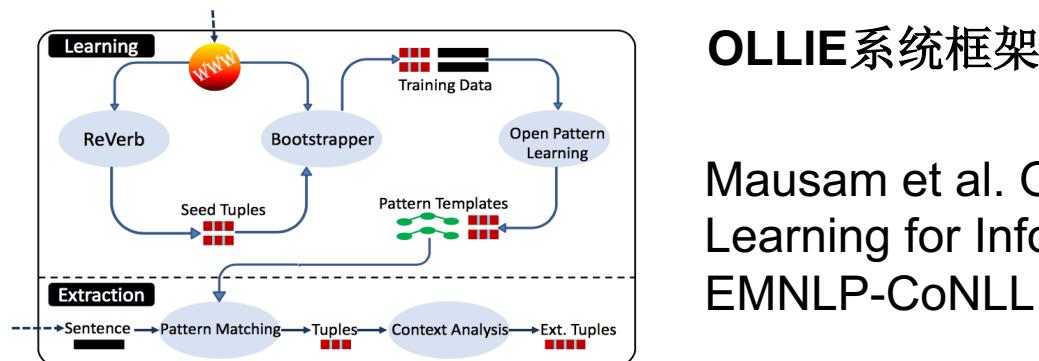
重要参考文献

Shwartz et al. Improving Hyponymy Detection with an Integrated Path-based and Distributional Method. ACL 2016

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

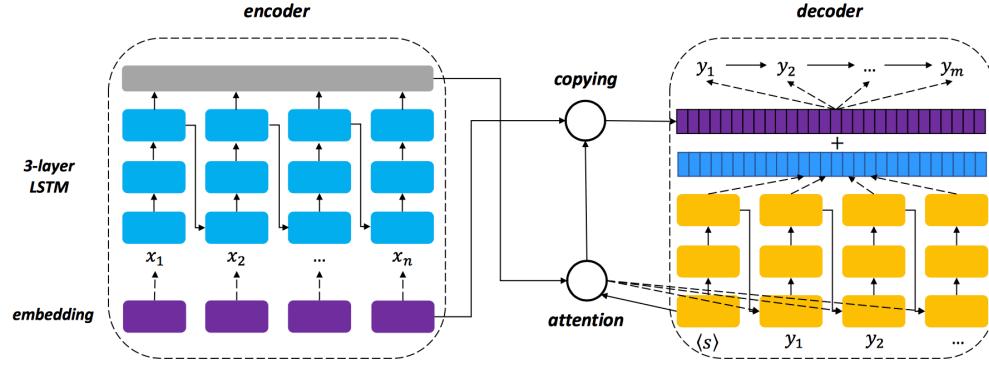
通用语义关系抽取（1）

- **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

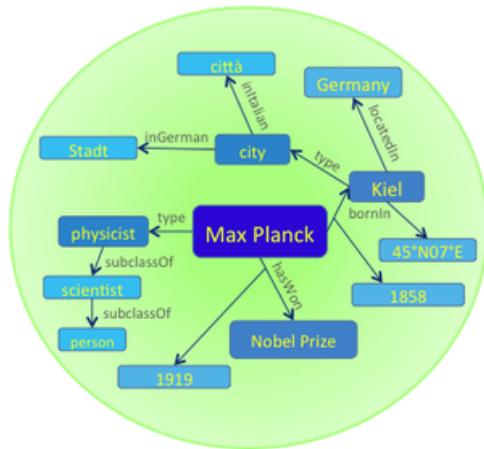
- 基于神经网络的方法：例如**Encoder-Decoder**架构等



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

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”，藏语译音为“耐力咯苏”，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 [\[隐藏\]](#)

- 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。马来熊毛色黑色（雄性比雌性大10–45%），前胸通常有一块明显的“U”型斑纹，斑纹呈浅棕黄或白色。马来熊头部比较宽，口鼻不突出，裸露无毛，呈浅棕或灰色，耳朵圆而小，位置较低。马来熊舌头很长，便于吞食**白蚁**或其他**昆虫**。脚掌爪钩呈**镰刀**型，脚掌内翻。尾巴是30–70毫米（1.2–2.8英寸）长。

 摄于上海动物园

习性 [\[编辑\]](#)

马来熊通常在清晨和傍晚活动，白天则在树洞或洞穴中休息。它们主要以植物为食，如浆果、嫩芽、根茎等，也会捕食小型哺乳动物、鸟类、鱼类等。它们游泳能力很强，经常在水中嬉戏。马来熊是独居动物，只有在繁殖季节才会短暂地聚集。

保护状况 [\[编辑\]](#)

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 | 中国哺乳动物 | 云南动物 | 老挝动物 | 孟加拉动物 | 印尼动物 | 柬埔寨动物 | 越南动物 | 泰国动物 | 马来西亚哺乳动物 | 印度哺乳动物 | 缅甸动物

(b) 用户标签 (User Tags)

帕丁顿熊2 Paddington 2 (2017)



导演: 保罗·金
编剧: 保罗·金 / 迈克尔·邦德 / 西蒙·法纳比
主演: 本·卫肖 / 休·格兰特 / 休·博纳维尔 / 莎莉·霍金斯 / 萨穆尔·乔斯林 / 更多...
类型: 喜剧 / 动画 / 家庭
官方网站: paddington.com/global/home
制片国家/地区: 英国 / 法国 / 美国
语言: 英语
上映日期: 2017-12-08(中国大陆) / 2017-11-10(英国)
片长: 103分钟
又名: 柏灵顿2(港) / 柏灵顿熊熊出任务(台)
IMDb链接: [tt4468740](https://www.imdb.com/title/tt4468740)



豆瓣成员常用的标签 · · · · ·

英国 喜剧 温情 温馨 动画 家庭
伦敦 2017

来源: <https://movie.douban.com>

(c) 关键词 (Keywords)

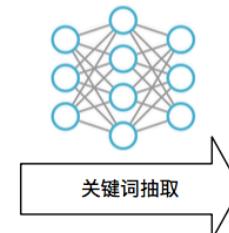
北极熊 [\[编辑\]](#)

保护状况

绝灭 EX
易危 EW
受威胁 CR
濒危 EN
VU
无危 NT
LC

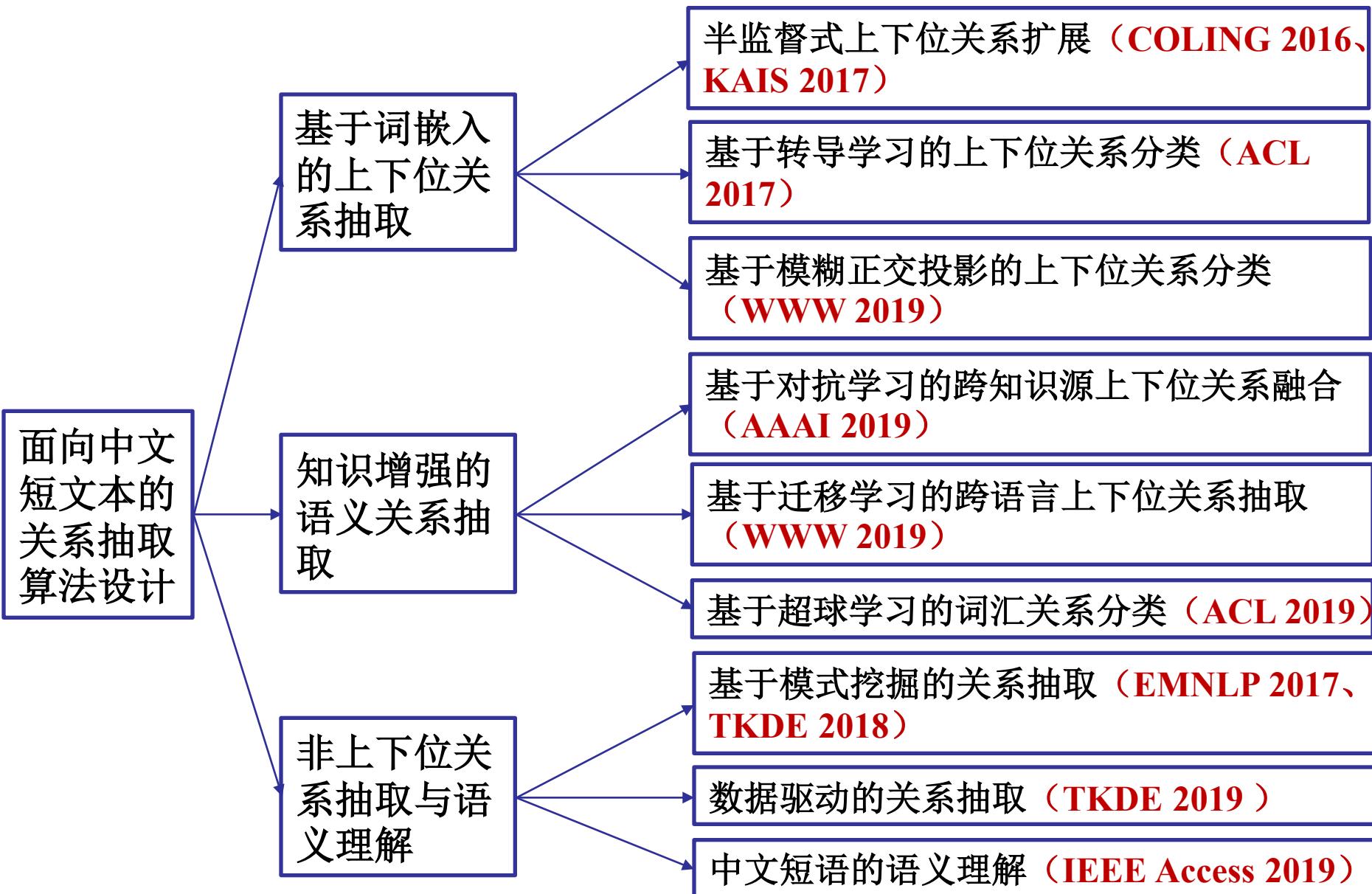
易危 (IUCN 3.1) [1]

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。



1. 动物
2. 北极
3. 灰熊
4. 游泳
5. 白色
6.

算法研究的整体框架



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)

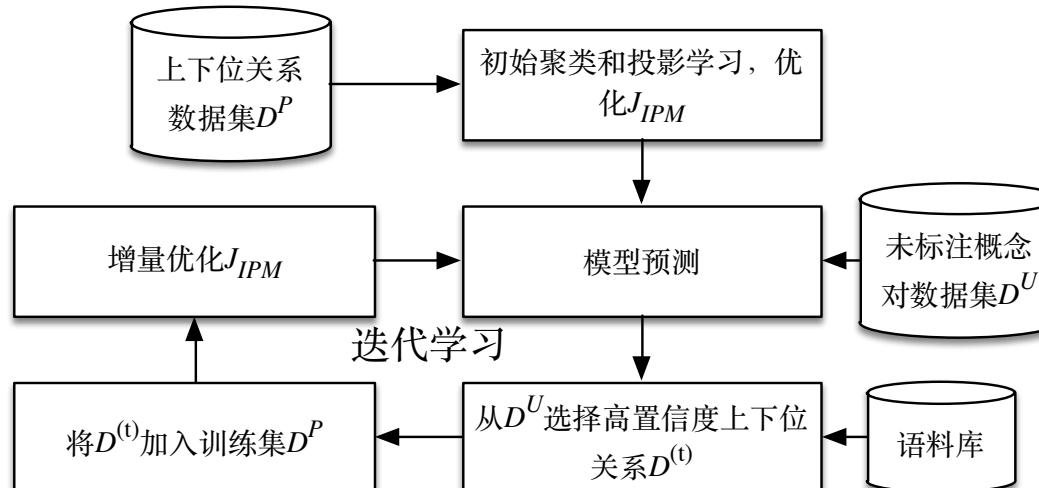
输入数据集： D^P, D^N, D^T

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元组作为“种子”，自动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



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）

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

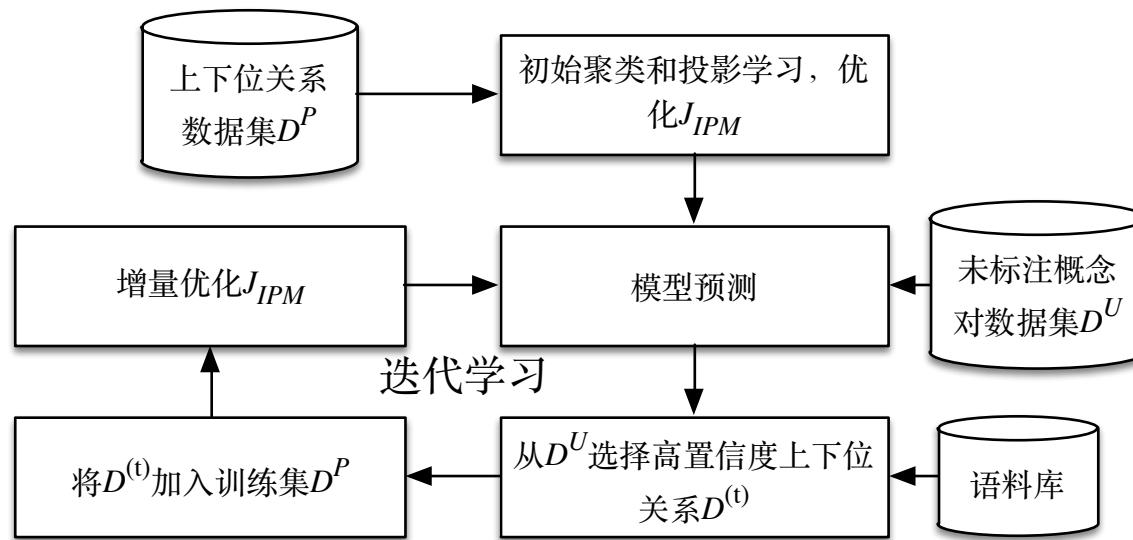
↑ ↑ ↑

将上下位关系 投影矩阵 偏置向量
聚成多个簇

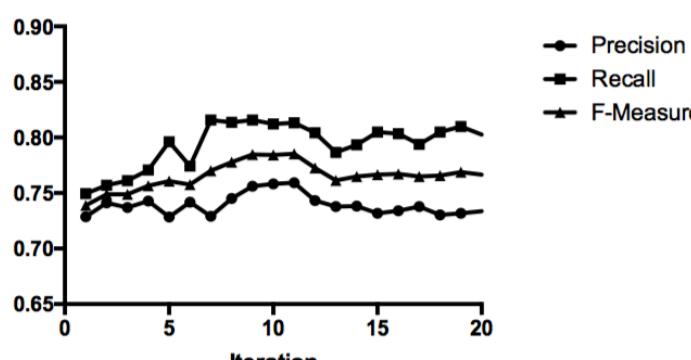
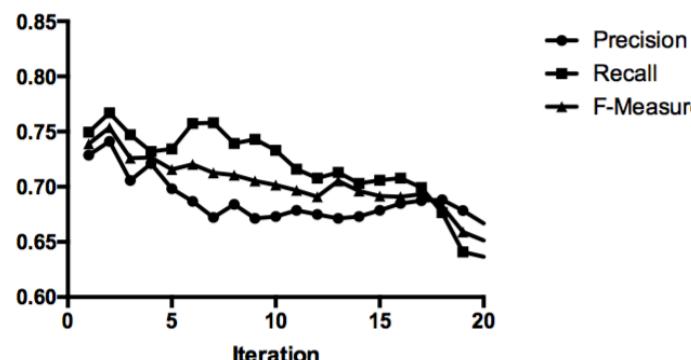
半监督式上下位关系扩展 (3)

- 迭代模型训练 (迭代次数 $t = 1, \dots, T$)

- 采样: 从 D^U 采样部分未标注数据 $U^{(t)}$
- 预测: 使用第 t 个迭代投影模型筛选出正例 $U_+^{(t)}$
- 选择: 使用基于中文语言模式的关系选择算法, 从 $U_+^{(t)}$ 选择高置信度正例 $U_*^{(t)}$
- 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（4）

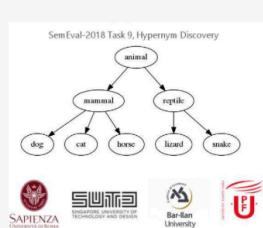
- 模型迭代分析
 - 模型在每个迭代的实验结果（先提升，后稳定）
 - 去除关系选择的结果
- 整体实验结果
 - 采用词嵌入建模下位词到上位词的投影，取得较好结果
 - 中文语言模式精度有限，但可有效“监督”半监督学习过程

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 [52]	0.673	0.281	0.396
CN-WikiTaxonomy [21]	0.985	0.254	0.404
invCL [63]	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	0.847	0.816

半监督式上下位关系扩展（5）

• 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



SemEval-2018 Task 9: Hypernym Discovery

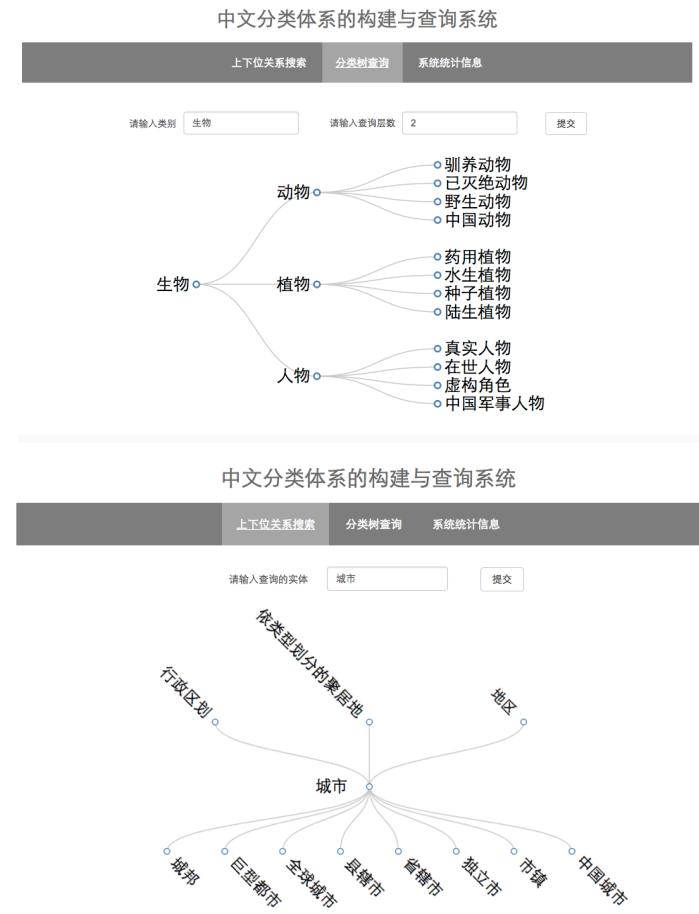
Organized by CamachoCollados - Current server time: Nov. 22, 2019, 3:03 a.m. UTC

Previous ▶ Current Post-Evaluation

Evaluation Jan. 12, 2018, midnight UTC Jan. 30, 2018, midnight UTC

• 系统研究

- 已研发中文分类体系的构建与查询系统



Chengyu Wang, Xiaofeng He, Aoying Zhou. A Short Survey on Taxonomy Learning from Text Corpora: Issues, Resources and Recent Advances. **EMNLP 2017 (CCF-B)**

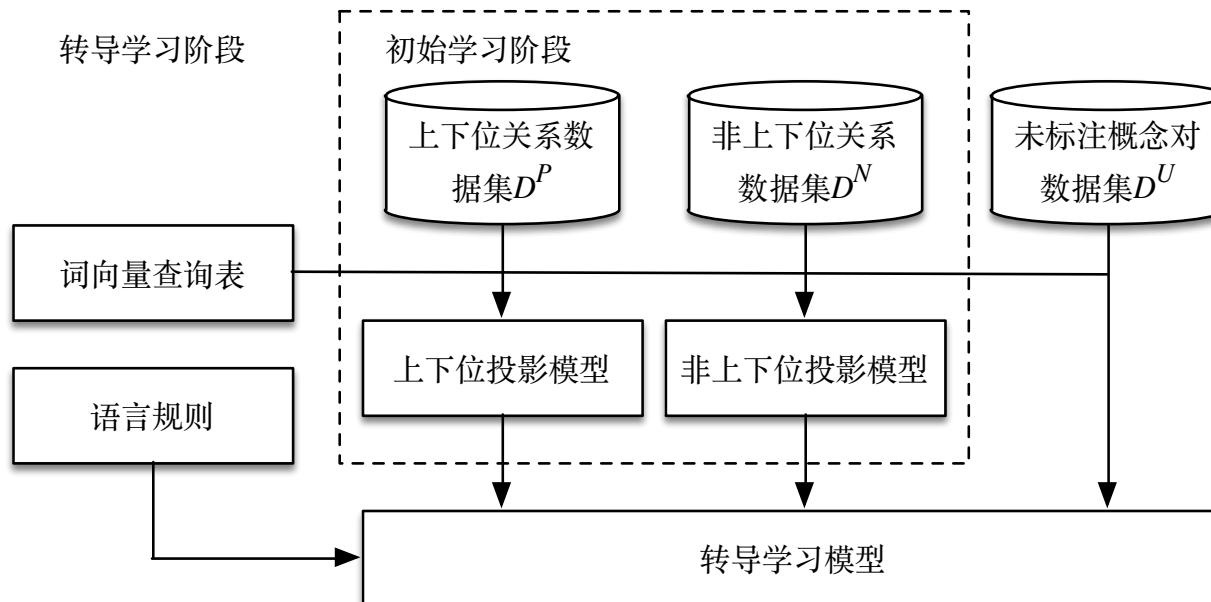
Chengyu Wang, Yan Fan, Xiaofeng He, Aoying Zhou. Predicting Hypernym-Hyponym Relations for Chinese Taxonomy Learning. **KAIS 58(3): 585–610 (2019) (CCF-B)**

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

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

↑ ←
最终预测分数 语言规则预测分数
(仅涉及部分数据)

规则 描述

P1 若 x_i 的核心词与候选上位词 y_i 相匹配， y_i 很可能是 x_i 的上位词。例如，“动物”是“哺乳动物”的上位词。

N1 若 x_i 的核心词与候选上位词 y_i 的非核心词相匹配， y_i 很可能不是 x_i 的上位词。例如，“动物学”不是“哺乳动物”的上位词。

N2 若候选上位词 y_i 的核心词与扩展自 [21] 的中文主题词词典相匹配， y_i 很可能不是 x_i 的上位词。该词典包括了 184 个非概念主题词，例如政治、军事等。

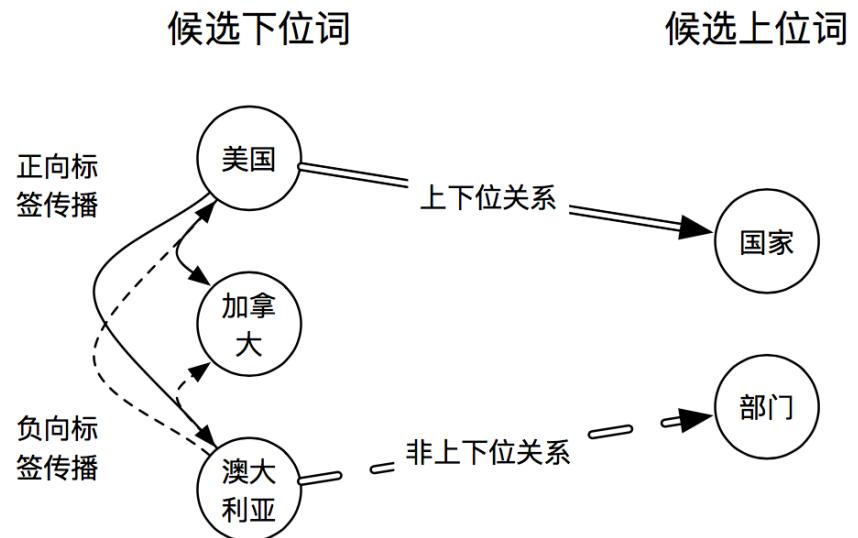
基于转导学习的上下位关系分类 (3)

- 转导学习
 - 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



- 整体优化目标

$$J(\mathbf{F}) = \mathcal{O}_s + \mathcal{O}_r + \frac{\mu_1}{2} \mathcal{O}_n + \frac{\mu_2}{2} \|\mathbf{F}\|_2^2$$

- 预测 (x_i, y_i) 有上下位关系指标： $F_i > \theta$

基于转导学习的上下位关系分类 (4)

- 中文实验结果
 - 整体实验结果

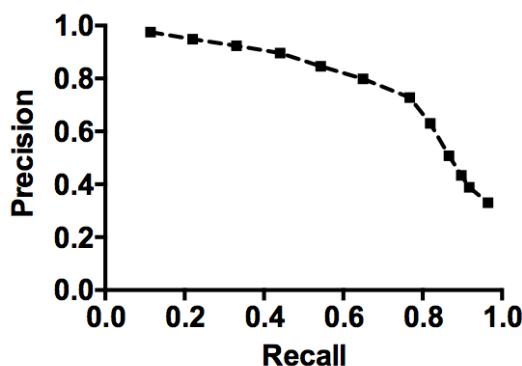
数据集	FD			BK		
方法	精准度	召回率	F 值	精准度	召回率	F 值
Fu-S [25]	0.641	0.560	0.598	0.714	0.648	0.679
Fu-P [25]	0.664	0.593	0.626	0.727	0.675	0.700
$\vec{x}_i \oplus \vec{y}_i$	0.677	0.752	0.697	0.803	0.759	0.780
$\vec{x}_i + \vec{y}_i$	0.653	0.607	0.629	0.727	0.656	0.689
$\vec{x}_i - \vec{y}_i$	0.719	0.606	0.657	0.784	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	0.728	0.705	0.716	0.836	0.806	0.821

- 算法细节分析

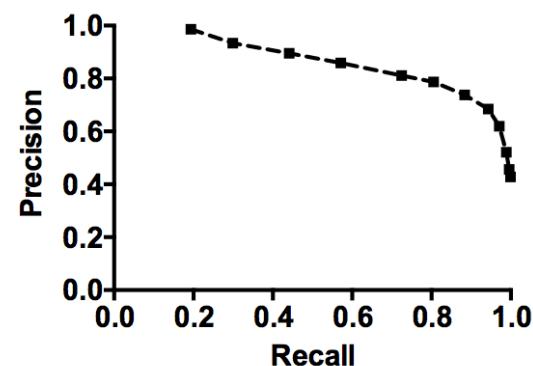
真正率/真负率	P1	N1	N2
数据集 FD	0.986	0.923	0.941
数据集 BK	0.976	0.968	0.973

语言规则的有效性

参数 θ 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

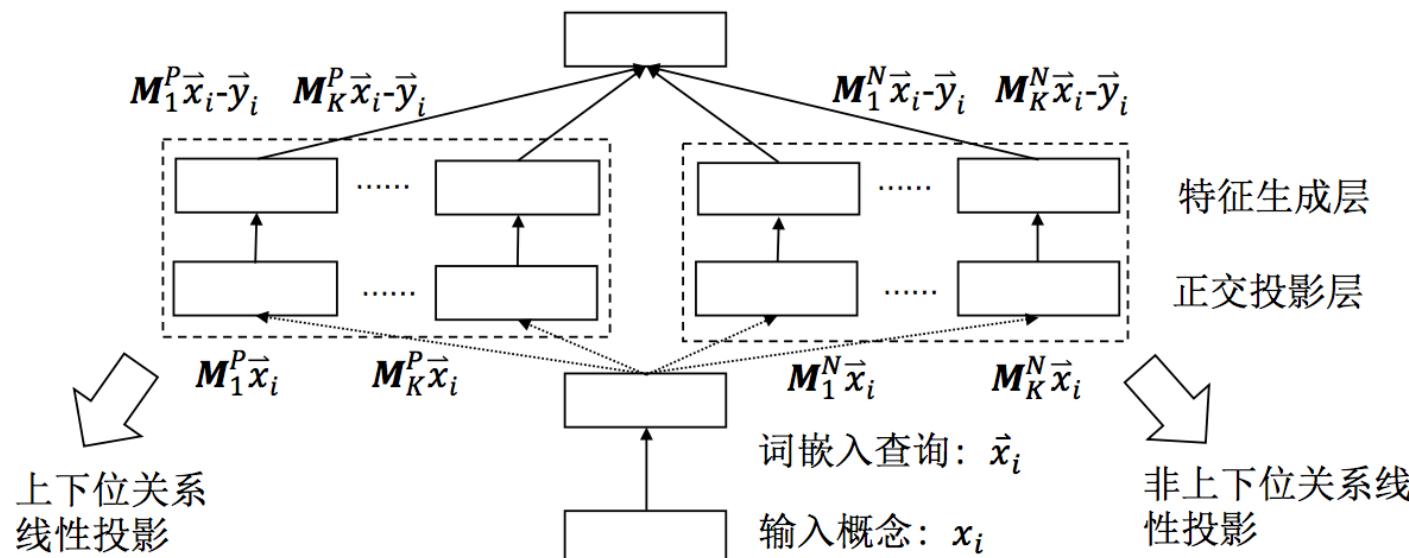
- 特征：分析计算两组映射（基于 $2K$ 个投影矩阵）的残差

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

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

- 整体网络架构

上下位关系/非上下位关系分类器： f

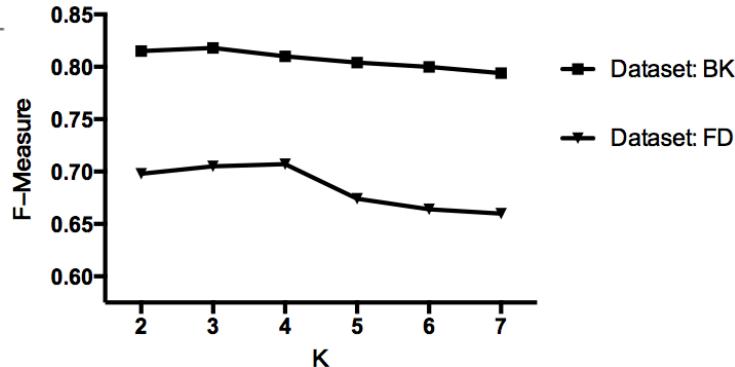


基于模糊正交投影的上下位关系分类（3）

- 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	0.752	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	0.728	0.705	0.716	0.836	0.806	0.821
FOPM	0.713	0.698	0.705	0.825	0.812	0.818

整体实验结果



参数K对实验结果的影响
(需要调整的唯一参数)

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

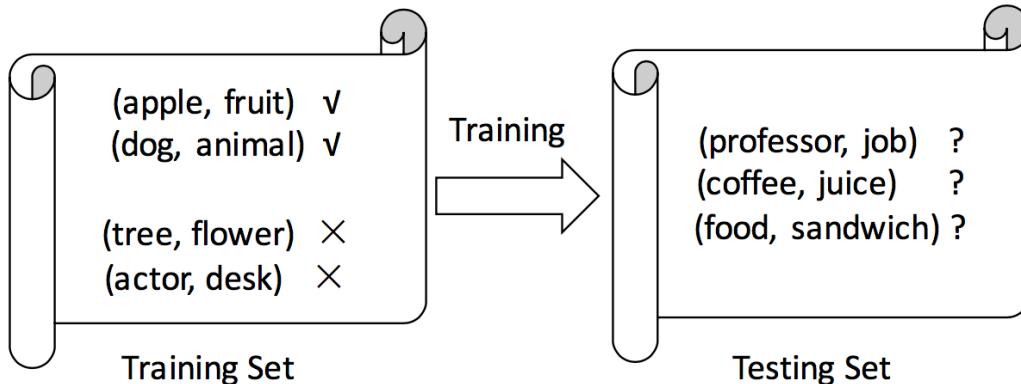
多词汇关系

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) [95]	2736022
Probbase (英语) [10]	16285393
Probbase+ (英语) [96]	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

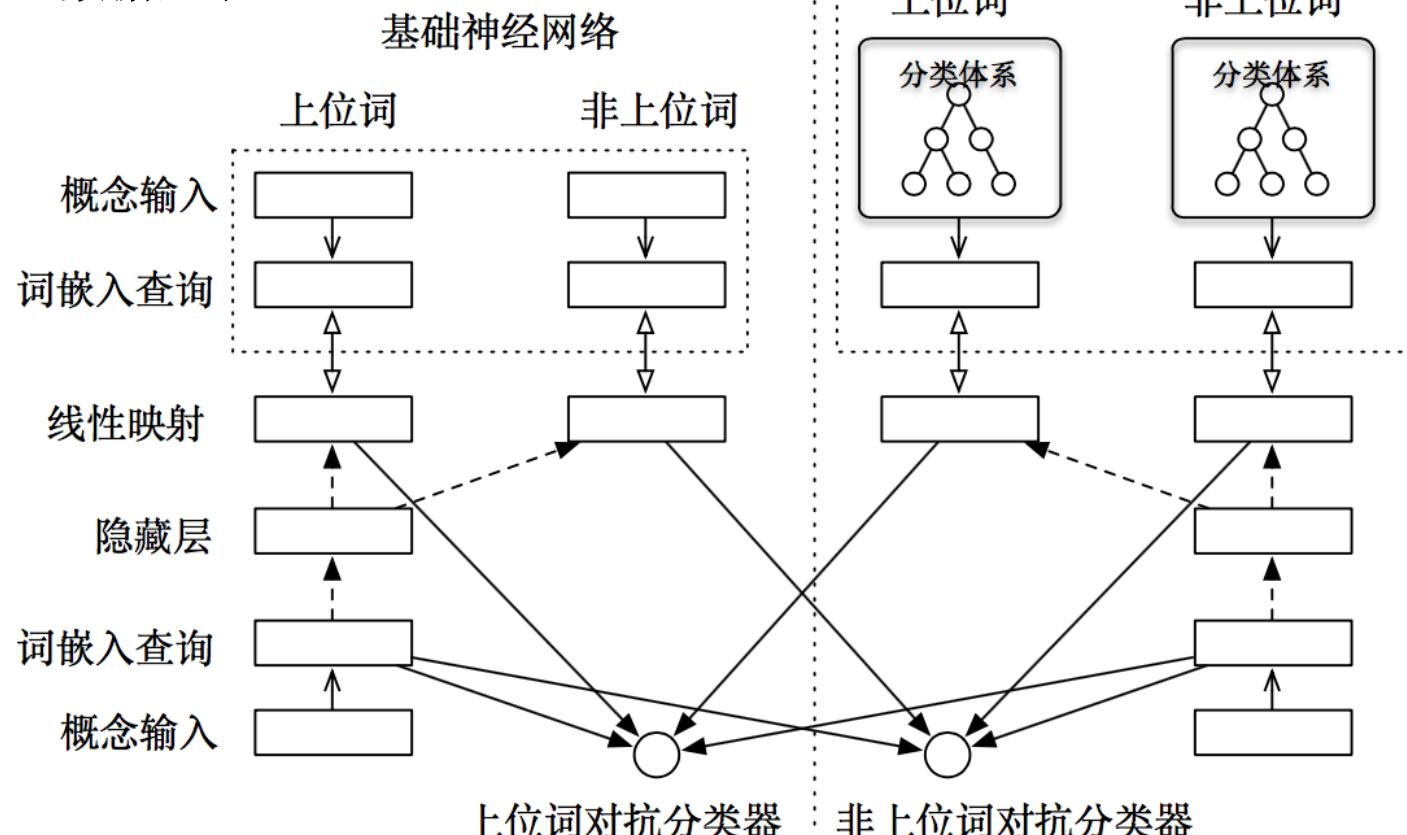
- 采用基于深度神经网络的对抗学习技术，将分类体系中的海量上下位关系知识融入到基础神经网络中，提高基础神经网络的学习能力

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	0.752	0.697	0.803	0.759	0.780
本文先前工作的最佳结果						
TPM	0.728	0.705	0.716	0.836	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	0.728	0.791	0.870	0.829

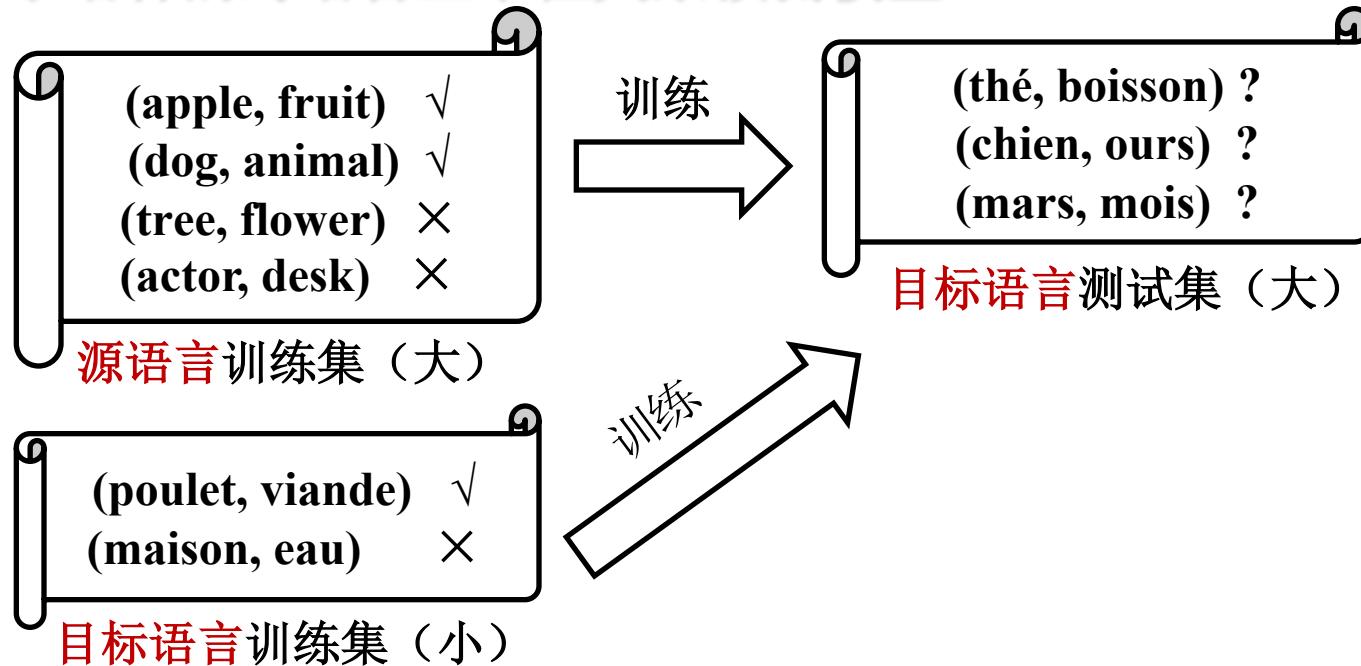
- 应用：对Microsoft Concept Graph的扩展

- 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

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

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

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

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

- S : 将源语言词向量映射到目标语言（双语字典生成）
- γ : 控制源语言单个训练数据对目标语言重要性（启发式规则）
- β : 控制源语言整体训练损失（可调）

基于迁移学习的跨语言上下位关系抽取（3）

- 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

Algorithm 8 ITFOPM 训练算法

```
1: 利用算法 7, 在数据集  $D_S^P$ 、 $D_S^N$ 、 $D_T^P$  和  $D_T^N$  上训练 TFOPM
2: while 算法不收敛 do
3:   for 每一个目标语言术语对  $(x_i, y_i) \in U_T$  do
4:     if  $conf(x_i, y_i) > \tau$  then
5:       if 分类器  $f$  预测  $(x_i, y_i)$  为上下位关系 then
6:         更新  $D_T^P = D_T^P \cup \{(x_i, y_i)\}$ 
7:       else
8:         更新  $D_T^N = D_T^N \cup \{(x_i, y_i)\}$ 
9:       end if
10:      更新  $U_T = U_T \setminus \{(x_i, y_i)\}$ 
11:    end if
12:  end for
13:  利用算法 7, 在数据集  $D_S^P$ 、 $D_S^N$ 、 $D_T^P$  和  $D_T^N$  上更新 TFOPM
14: end while
```

基于迁移学习的跨语言上下位关系抽取（4）

• 整体实验分析

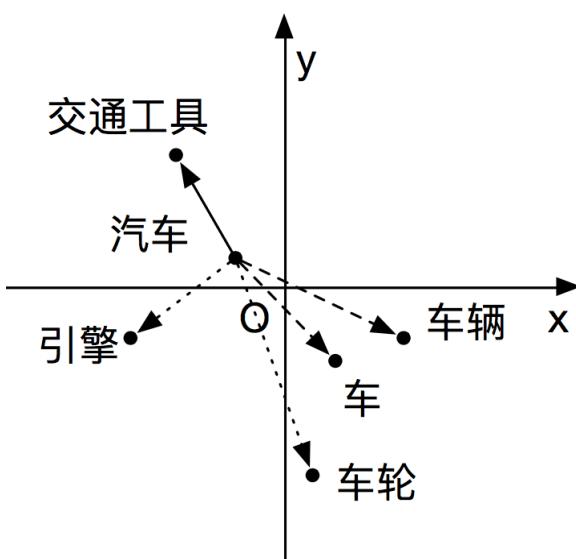
- 数据集：从 **Open Multilingual Wordnet** 计划生成训练集和测试集
- 任务一：跨语言上下位关系方向分类
 - 上下位关系vs.反上下位关系
- 任务二：跨语言上下位关系检测
 - 上下位关系vs.非上下位关系

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							
任务：跨语言上下位关系方向分类	fr	zh	ja	it	th	fi	el
Santus 等人 [65]	0.65	0.65	0.68	0.61	0.63	0.70	0.62
Weeds 等人 [73]	0.76	0.71	0.77	0.76	0.72	0.77	0.70
Kiela 等人 [159]	0.67	0.65	0.71	0.68	0.65	0.70	0.62
Shwartz 等人 [49]	0.79	0.67	0.71	0.72	0.66	0.75	0.66
TFOPM-N	0.78	0.71	0.75	0.76	0.73	0.76	0.71
TFOPM	0.80	0.72	0.76	0.78	0.75	0.78	0.73
ITFOPM-N	0.82	0.72	0.76	0.78	0.75	0.81	0.72
ITFOPM	0.81	0.74	0.78	0.81	0.78	0.81	0.75
任务：跨语言上下位关系检测							
Santus 等人 [65]	0.67	0.63	0.67	0.62	0.64	0.62	0.64
Weeds 等人 [73]	0.74	0.66	0.68	0.71	0.62	0.68	0.69
Kiela 等人 [159]	0.70	0.61	0.65	0.68	0.57	0.61	0.67
Shwartz 等人 [49]	0.72	0.66	0.69	0.64	0.66	0.69	0.70
TFOPM-N	0.72	0.67	0.70	0.70	0.68	0.71	0.70
TFOPM	0.75	0.71	0.76	0.72	0.69	0.72	0.71
ITFOPM-N	0.72	0.74	0.77	0.74	0.67	0.71	0.72
ITFOPM	0.76	0.73	0.78	0.74	0.72	0.73	0.73

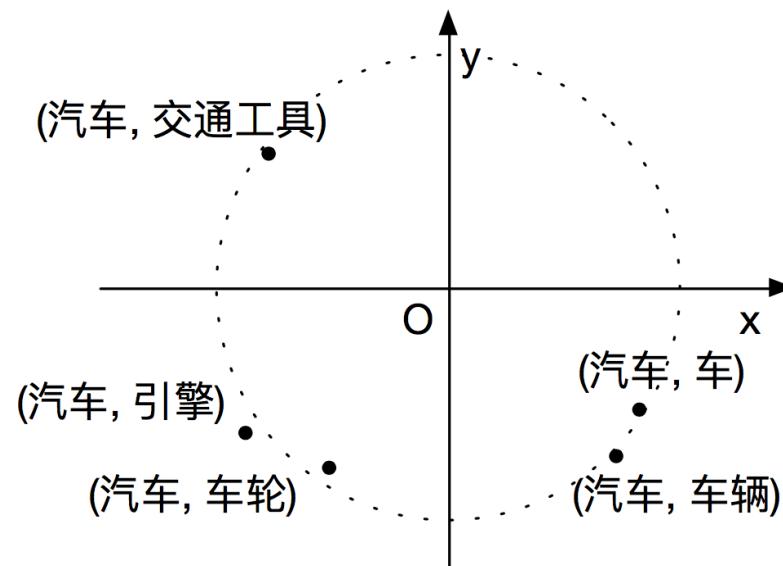
基于超球学习的词汇关系分类 (1)

- 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

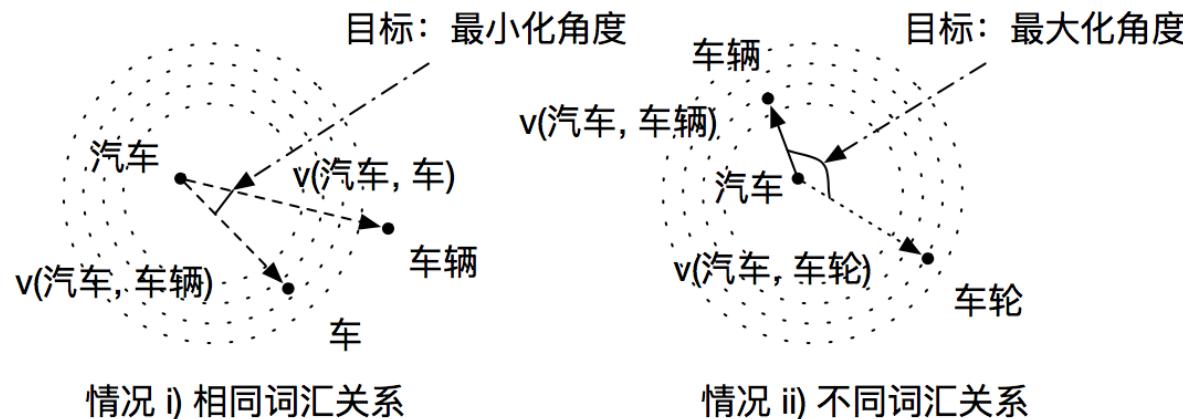


(b) 关系超球空间

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）

Method↓ Dataset→	K&H+N			BLESS			ROOT09			EVALution		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Concat	0.909	0.906	0.904	0.811	0.812	0.811	0.636	0.675	0.646	0.531	0.544	0.525
Concat _h	0.983	0.984	0.983	0.891	0.889	0.889	0.712	0.721	0.716	0.57	0.573	0.571
Diff	0.888	0.886	0.885	0.801	0.803	0.802	0.627	0.655	0.638	0.521	0.531	0.528
Diff _h	0.941	0.942	0.941	0.861	0.859	0.860	0.683	0.692	0.686	0.536	0.54	0.539
NPB	0.713	0.604	0.55	0.759	0.756	0.755	0.788	0.789	0.788	0.53	0.537	0.503
LexNET	0.985	0.986	0.985	0.894	0.893	0.893	0.813	0.814	0.813	0.601	0.607	0.6
LexNET _h	0.984	0.985	0.984	0.895	0.892	0.893	0.812	0.816	0.814	0.589	0.587	0.583
NPB+Aug	-	-	0.897	-	-	0.842	-	-	0.778	-	-	0.489
LexNET+Aug	-	-	0.970	-	-	0.927	-	-	0.806	-	-	0.545
SphereRE	0.990	0.989	0.990	0.938	0.938	0.938	0.860	0.862	0.861	0.62	0.621	0.62
Improvement	-	-	0.5%↑	-	-	1.1%↑	-	-	4.7%↑	-	-	2.0%↑

- CogALex-V Shared Task的实验结果

方法	同义词关系	反义词关系	上下位关系	部分关系	总计
GHHH [142]	0.204	0.448	0.491	0.497	0.423
LexNET [99]	0.297	0.425	0.526	0.493	0.445
STM [101]	0.221	0.504	0.498	0.504	0.453
SphereRE	0.286	0.479	0.538	0.539	0.471

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

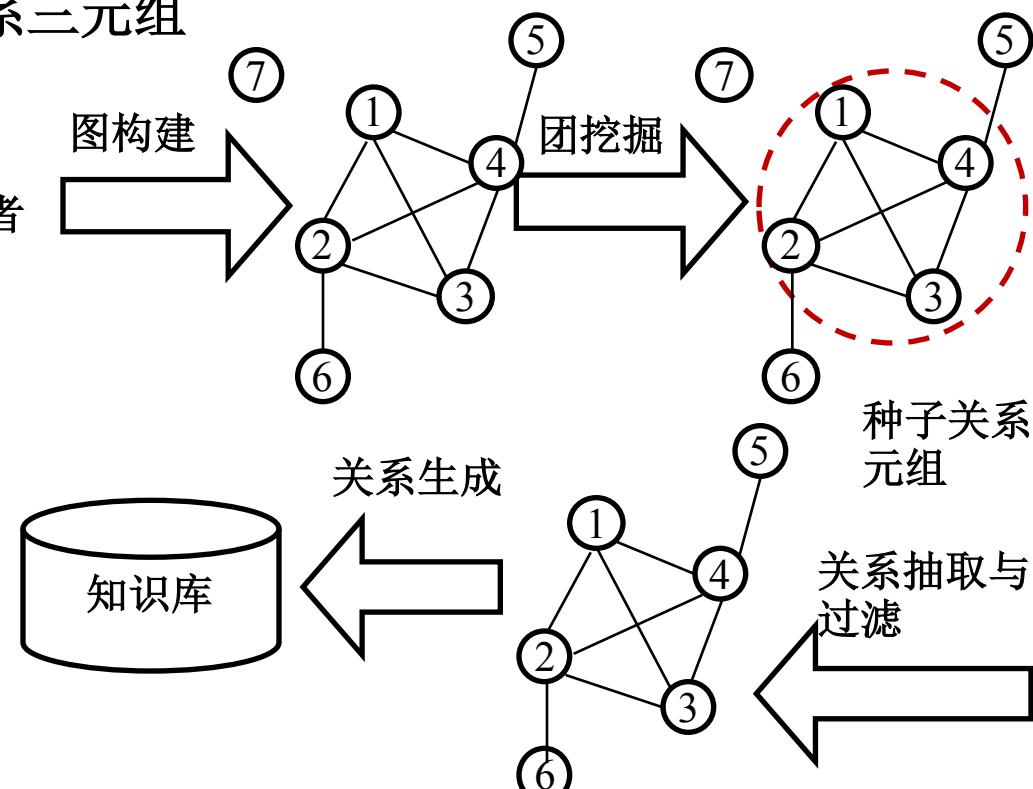
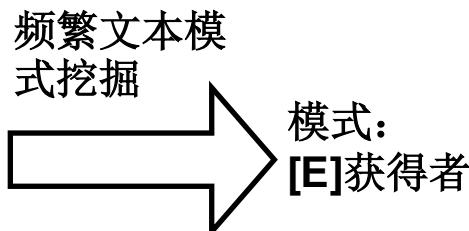
分类：法国荣誉军团骑士勋章持有人 | 1964年出生 | 在世人物 | 中国企业家 | 中华人民共和国亿万富豪 | 未来全球领袖 | 杭州师范大学校友
| 杭州电子工业学院教师 | 杭州师范大学教师 | 动物福利相关人士 | 中华人民共和国企业家 | 香港科技大学荣誉博士 | 时代百大人物 | 阿里巴巴人物
| 华谊兄弟人物 | 浙江企业家 | 中国首席执行官 | 中国共产党党员 | 长江商学院校友 | 杭州人 | 绍兴人 | 嵊州人 | 马姓
| 改革开放40年百名杰出民营企业家 | 改革先锋称号获得者

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%
PNRE	97.4%

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. TKDE³⁸

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	94.3%	182	193	95.9%	185	
DNRE	289	92.7%	268	314	93.9%	295	
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	95.1%	194	188	96.3%	181	
DNRE	324	92.3%	299	274	94.2%	258	
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 $N_1 N_2$ 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
RCRL	0.801	0.783	0.792	0.784	0.733	0.758

具体实验过程和**Baseline**详见论文

结论

- 算法：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
 - 不适合采用基于上下位模式的深度学习算法
 - 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
- 数据集及实验：仍然缺乏高质量的标注数据集、评测标准等

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

THANK YOU!

Questions & Answers?