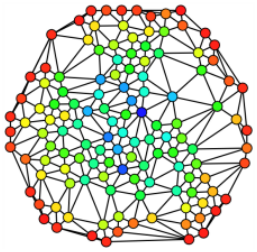
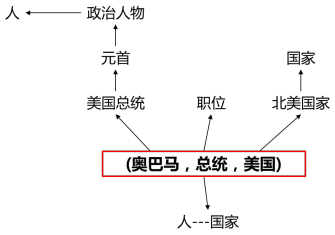
目录

- 知识建模
  - 知识体系概述
  - 典型知识体系
  - 知识体系手工建模方法
  - 知识体系自动建模方法
- 知识融合
  - 知识融合概述
  - 知识体系融合方法
  - 知识实例融合方法
- 大模型中的知识融合
  - 大模型对齐技术概述
  - 大模型对齐方法



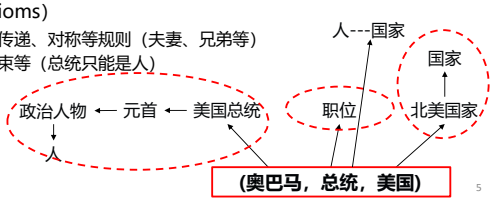
知识体系

- 知识体系（Knowledge Schema）：对于知识数据的描述和定义，是描述知识数据的“元数据”（metadata）
- 知识图谱：三元组为基本单元，以有向标签图为数据结构，从知识本体和知识实例两个层次，对世界万物进行体系化、规范化描述，并支持高效知识推理和语义计算的大规模知识系统。

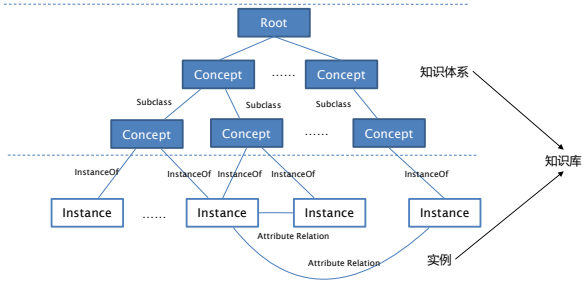


知识体系

- 知识体系主要包含三方面内容
  - 词汇、类别/概念的定义和描述
    - 词汇/术语（Term）、概念（Concept）
  - 概念之间的相互关系（Relation）
    - 分类关系（Taxonomic Relation）
      - Subclass: Is_A、Part_of
    - 非分类关系（Non-Taxonomic Relation）
      - Property/Attribute
  - 公理（Axioms）
    - 反向、传递、对称等规则（夫妻、兄弟等）
    - 值的约束等（总统只能是人）

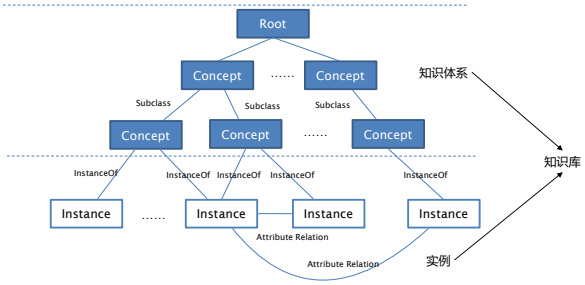


知识体系 vs. 知识库



知识体系的重要性

- 概念是人类能将万物准确归类的前提
  - 将事物准确归类（又称为范畴化）是人类认知世界的前提，没有概念（类别、范畴），人类就无法归类事物。

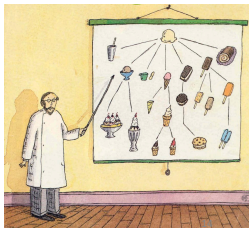


常用的知识组织形式

- Ontology（本体）
- Taxonomy（分类体系）
- Folksonomy/Metadata （开放标签）

Ontology

- Ontology is the study of the nature of being
- Ontologies 是（特定领域）信息组织的一种形式，是领域知识规范的抽象和描述，是表达、共享、重用知识的一种方法
- Ontology通过对于概念(Concept)、术语(Terminology)及其相互关系(Relation, Property)的规范化(Conceptualization)描述，勾画出某一领域的基本知识体系和描述语言
  - 真实世界的一个描述模型
    - 引入领域相关的术语集合
      - 概念/类别、属性、关系
  - 使用合适的逻辑来形式化



Heart is a muscular organ that is part of the circulatory system

Heart ⊑ MuscularOrgan ⊓ ∃isPartof.CirculatorySystem

Ontology

Definition 2.2 (Ontology). *An ontology is a tuple* $o = \langle C, I, R, T, V, \leq, \perp, \in, = \rangle$, *such that:*

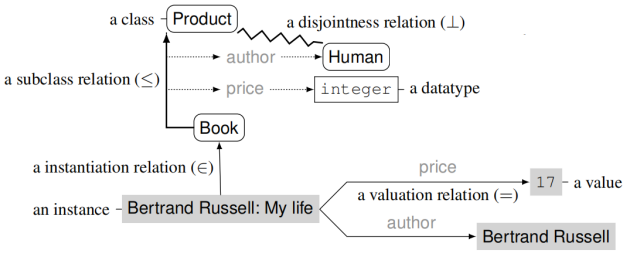
C *is the set of classes;*
I *is the set of individuals;*
R *is the set of relations;*
T *is the set of datatypes;*
V *is the set of values (C, I, R, T, V being pairwise disjoint);*
$\leq$ *is a relation on* $(C \times C) \cup (R \times R) \cup (T \times T)$ *called specialisation;*
$\perp$ *is a relation on* $(C \times C) \cup (R \times R) \cup (T \times T)$ *called exclusion;*
$\in$ *is a relation over* $(I \times C) \cup (V \times T)$ *called instantiation;*
$=$ *is a relation over* $I \times R \times (I \cup V)$ *called assignment.*

Euzenat et al. Ontology Matching

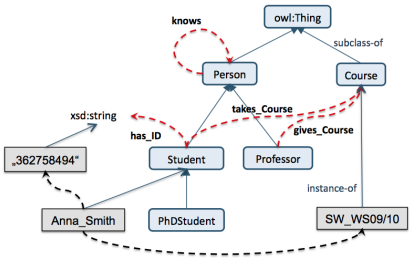
Ontology的特点

- 概念化（Conceptualization）
  - Abstract model (concepts)
  - 指某一概念系统所蕴涵的语义结构，它是对某一事实结构的一组非正式的约束规则。它可以理解和/或表达为一组概念（如实体、属性、过程）及其定义和相互关系
- 显式化（Explicit）
  - The concepts are explicitly defined
- 规范化（Formal）
  - Machine readable
- 公理化（Shared）
  - Accepted by a group and not private to some individual

Ontology 例子

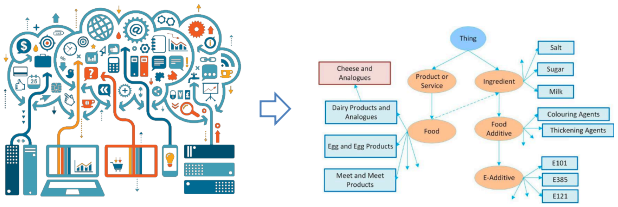


Ontology 例子



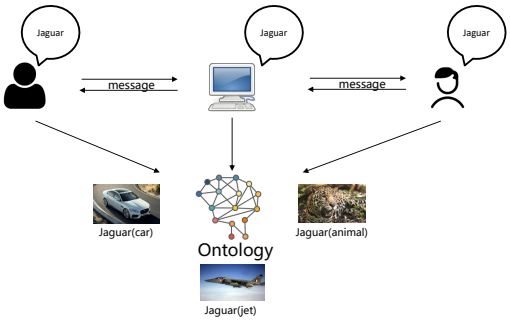
Ontology 应用

- 管理知识（定义、存储、分类）

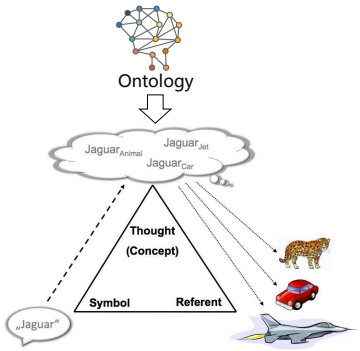


Ontology 应用

- 减少歧义

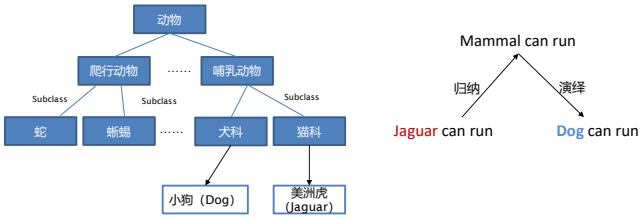


Ontology 应用



Ontology应用

- 推理



Ontology应用

- 基于实例化的应用
  - 任务：实体搜索
    - 定义：给定一个概念作为查询，检索该概念的典型实体
    - 例子：搜索“北京知名高校”，返回“清华大学”、“北京大学”、“中国科学院大学”等
  - 任务：样本增强
    - 定义：利用概念下的实例，增强样本
    - 例子：对样本“清华大学在哪个城市”替换实体，生成“国科大在哪个城市”作为新样本

Ontology应用

- 基于概念化的应用
  - 任务：文本分类
    - 定义：根据文本中实体的概念，将文本分为不同类别
    - 例子：包含“足球”“篮球”的文本大概率与包含“宝马”“奔驰”的文本属于不同类别
  - 任务：主题分析
    - 定义：给定文本，分析文本属于什么主题
    - 例子：包含“足球”“篮球”的文本大概率属于体育类主题
  - 任务：用户画像
    - 定义：给定用户信息，为用户生成显式的概念
    - 例子：根据用户描述“精通Java、Android开发”，可以为其打上“面向对象编程”“手机App开发”的标签

Ontology应用

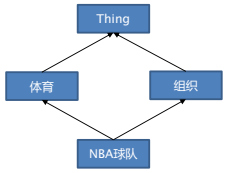
- 基于概念化的应用
  - 任务：基于概念的解释
    - 定义：根据概念信息，为事件提供解释
    - 例子：特斯拉Model S的加速性能很好，因为它是“电动汽车”，电动汽车通常具有较好的加速性能
  - 任务：概念归纳
    - 定义：从实体集合或者词袋归纳概念标签
    - 例子：给定“清华大学”、“北京大学”、“中国科学院大学”，可以归纳出“中国高校”、“知名大学”等概念
  - 任务：语义表示
    - 定义：利用概念集合表达实体、词汇的语义
    - 例子：iphone 13的语义可以表达为其概念集合{“全面屏手机”、“智能手机”、“苹果手机”}

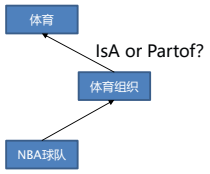
Ontology应用

- 综合使用实例化和概念化的应用
  - 任务：实体推荐
    - 定义：根据一些实体的概念信息为其推荐其他实体
    - 例子：当用户搜索“iPhone15”时，为其推荐“华为 P70”， “小米14”等手机
  - 任务：规则挖掘
    - 定义：利用概念从大量数据中挖掘出具有一定泛化能力的规则
    - 例子：健身的人吃蔬菜、吃燕麦等食物，可以挖掘出健身的人主要吃“高纤维食物”

Ontology 问题

- 在Ontology中对于层级体系（Taxonomic Relation）定义严格
  - IsA 或者 Part of关系
  - 概念的二义性问题



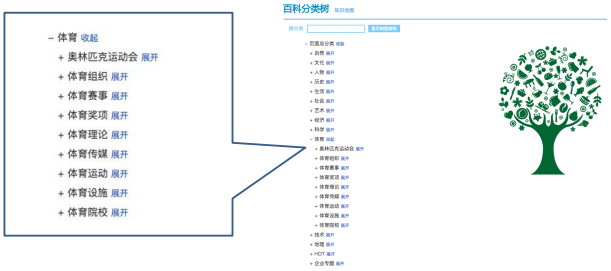


常用的知识组织形式

- Ontology（本体）
- Taxonomy（分类体系）
- Folksonomy/Metadata （开放标签）

Taxonomy

- Taxonomic Relation： 领域相关



Taxonomy

Folksonomy/Metadata

- 取消Taxonomic Relation
- 概念类别冗余
  - 可以存在多个表征同一概念的类别语义标签
  - 类别标签由用户提供

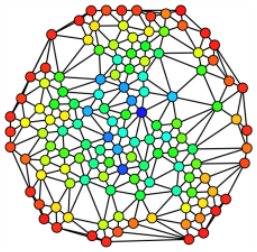
Folksonomy(Metadata)

不同知识组织形式的比较

- Ontology
  - 树状结构，上下位节点之间是严格的IsA关系
  - 优点：可以适用于知识的推理
  - 缺点：无法表示概念的二义性（运动员：体育？人物？）
- Taxonomy
  - 树状结构，上下位节点之间非严格的IsA关系
  - 优点：可以表示概念的二义性（体育—>运动员）
  - 缺点：不适用于推理，无法避免概念冗余（餐厅：美食？机构？地点？）
- Folksonomy
  - 类别标签，更加开放
  - 优点：能够涵盖更多的概念
  - 缺点：如何进行标签管理？

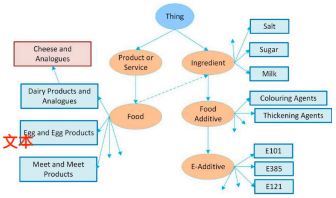
目录

- 知识建模
  - 知识体系概述
  - 典型知识体系
  - 知识体系手工建模方法
  - 知识体系自动建模方法
- 知识融合
  - 知识融合概述
  - 知识体系融合方法
  - 知识实例融合方法
- 大模型中的知识融合
  - 大模型对齐技术概述
  - 大模型对齐方法



知识体系构建的目标

- 术语、概念/类别
- 概念、类别层级体系
- 属性关系
- 约束
  - 属性的定义域（Domain）
  - 属性的值域（Range）



文本

公理

$\forall x, y, z\ (father(x, y) \wedge father(y, z) \rightarrow grandpa(x, z))$ Axioms (A^O)

售价（Domain:汽车型号，Range: 价格） Relationships (R)

IsA（吉利，汽车品牌） Concept hierarchies (H^C)

汽车品牌[国家、创办时间、...] Concepts (C)

{Benz、奔驰、宾士、大奔} Synonyms 同义词

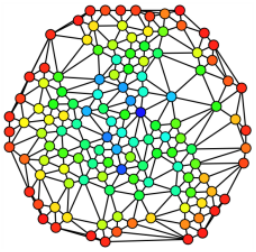
汽车、丰田、宝马、汽车、汽车品牌 Terms

人工构建方法

- 知识体系构建过程
  - 确定领域及任务 cyc:被称为是"人工智能历史上最有争议的项目"之一
    - 吸取cyc的教训，通用知识库难以实用
    - 限定领域内知识可以穷举
  - 体系复用
    - 站在巨人的肩膀上
  - 定义术语、概念/类别（Term、Concept）
  - 确定分类体系（ Relation）
  - 确定关系、属性（Non-Taxonomy Relation）
  - 定义约束
    - 总统（Entity1： 人物，Entity2：国家）

目录

- 知识建模
  - 知识体系概述
  - 典型知识体系
  - 知识体系手工建模方法
  - 知识体系自动建模方法
- 知识融合
  - 知识融合概述
  - 知识体系融合方法
  - 知识实例融合方法
- 大模型中的知识融合
  - 大模型对齐技术概述
  - 大模型对齐方法

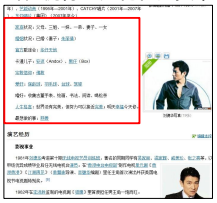


自动构建方法

- 任务
  - 挖掘术语、概念/类别
  - 构建概念、类别层级体系
  - 挖掘属性关系
- 方法 通过网络挖掘获取概念、关系
  - 基于结构化、半结构化数据的知识体系构建
  - 基于非结构化数据（纯文本）的知识体系构建
- Note:
  - 目前还不能直接使用，但是可以节省人力
  - 不同于实例知识抽取，知识体系构建通常只需要构建一次

基于半结构化数据的知识体系挖掘

- 基本假设：同一网站中的页面具有相似性，所设定模板具有复用性
- 关键核心：模板学习与挖掘
  - 模板学习：如何通过自动学习，挖掘模板（Pattern）和构建抽取器（Wrapper）
  - 噪声滤除：如何在网页中识别/分割出所需的信息块



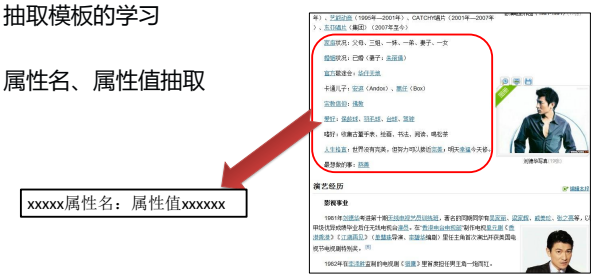
xxxxx属性名：属性值xxxxxx

半结构文本中的属性名、属性值抽取

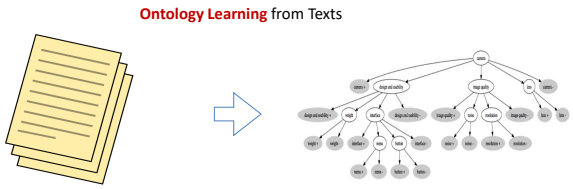
- 目标：从百科普通条目半结构化网页中自动学习模板，抽取实体属性及相关的属性值

基本步骤

- 半结构化信息块的识别/定位
- 抽取模板的学习
- 属性名、属性值抽取



基于非结构化数据（纯文本）的知识体系构建



- 术语、概念抽取
- 同义词挖掘
- 关系挖掘
  - 上下位关系
  - 属性

概念抽取：术语（Terminology）抽取

- 概念/术语质量评估
  - 频率
    - 一般来说，一个N-Gram在给定的文档集合中要出现的足够频繁才能被视作高质量短语。 N-Gram本身就是单词的序列组合，因此一个单词序列被使用得越多，就越可能是一个高质量短语。如果一个N-Gram出现的次数过少，那么它极有可能只是一个拼写错误。
  - 一致性
    - 一致性是指N-Gram的搭配频率明显高于其各部分偶然组合在一起的可能性，即反映的是N-Gram中不同单词的搭配是否合理或者是否常见。搭配越常见，相应的N-Gram越有可能是一个高质量短语

概念抽取：术语（Terminology）抽取

- 概念/术语质量评估
  - 信息量
    - 一般来说，一个高质量短语应该传达一定的信息，即表达一定的主题或者概念。比如， ”机器学习”与“这篇论文”在机器学习论文语料中出现的频率都很高，单词之间的搭配也都很合理，但是显然后者没有太多信息量
  - 完整性
    - 一个高质量短语必须在特定的上下文中是一个完整的语义单元，比如“vector machine”在机器学习论文语料中很少单独出现，更多的是以“support vector machine”的完整形式出现，那么它自身就不能算是一个高质量短语

概念抽取：术语（Terminology）抽取

- Step1: 生成术语候选（Terms Extraction）
  - N-Grams
  - 基于模板进行抽取
    - POS based Patterns: N-N, Adj-N
- Step2：候选排序，过滤噪声（Ranking）
  - 基于频率统计的方法
  - 基于主题模型的方法
  - 基于图排序的方法

频率统计
- **TF-IDF**
- **C-value/NC-value**
- **PMI (Pointwise Mutual Information，点互信息)**
- **Search Engine**
- **Domain Relevance (抽选领域相关的候选)**
- **Domain Consensus(Information Entropy)**

Ranking：TF，TF-IDF

- 词频 (term frequency, TF)：给定的词语在篇章中出现的次数。这个数字通常会被归一化(一般是词频除以篇章/句子总词数), 以防止它偏向长的篇章和句子

$$TF_w = \frac{\text{词}w\text{出现的次数}}{\text{该篇章中所有词条数据}}$$

- 逆向文件频率 (inverse document frequency, IDF)：如果包含词语t的文档越少，即IDF越大，则说明词条具有很好的区分能力。

$$IDF_w = \log(\frac{\text{语料中的总篇章数}}{\text{包含了词}w\text{的篇章数}+1})$$

$$TF-IDF_w = TF_w * IDF_w$$

- 以汽车领域为例：雷克萨斯 vs. 教师
- TF，TF-ID衡量了目标词（短语）在语料中的重要性

Ranking：C-value/NC-value

- 对于短语（multiword expression），利用C/NC值估计构成该短语的置信度
  - C-value：衡量语料中出现的高频的最长短语，考虑了词汇长度和父子短语的相互影响

候选词长度　候选在语料中出现的频率　预设候选最长长度，即没有父短语

$$Cvalue(a) = \begin{cases} log_2|a| f(a) & if\ |a| = g \\ log_2|a| \left(f(a) - \frac{1}{|C(a)|} \sum_{k=1}^{C(a)} f(k) \right) & otherwise \end{cases}$$

所有包含a的候选集合，所有的父短语

Ranking：C-value/NC-value

- NC-value：相对于C-value，除了考虑候选出现的频率之外还考虑上下文信息

t在语料中出现的频率　语料中所有词个数

$$NCvalue(a) = \alpha Cvalue(a) + (1-\alpha)(\sum_{t \in C_a} f_a(t) \frac{f(t)}{n})$$

候选a上下文中词集合　t在候选a上下文出现的频率

- 上下文在候选排序过程中有重要作用，例如：我正在学习 知识图谱课程，学习后面，课程前面一般是一个学科、技术的名称

PMI

- PMI（Pointwise Mutual Information，点互信息），也是在抽取领域短语时常用的指标，刻画了短语组成部分之间的一致性程度

$$PMI(x;y) = \log \frac{p(x,y)}{p(x)p(y)}$$

Y单独出现的概率

候选X和Y联合出现的概率

X单独出现的概率

- “的电影”与“电影院”都有较高的频次，但是通过PMI可以识别出“电影院”相对于“的电影”是质量更高的短语

Ranking：Search Engine

- 利用搜索引擎验证当前候选词
- 按照短语进行搜索：加双引号

Domain Relevance

- Domain Relevance (DR): 抽取领域相关的候选

$$DR_{(t,k)} = \frac{p(t|D_k)}{\sum_{i=1}^{m} p(t|D_i)}$$

$p(t|D_k)$：候选t在领域D_k出现的概率，m表示领域个数

- 基于DR术语选择：
  - 非术语（通用词）：不同领域中的分布类似
  - 术语（专业词）：在目标领域内的分布具有显著性

Domain Consensus (Information Entropy)

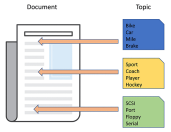
- Entropy：描述事物无序性的一个重要的参数，熵越大则无序性越强，同时，熵表示一个随机变量的不确定性。
- 术语选择：
  - 非术语（通用词）：不同领域中出现概率类似，没什么区分性
  - 术语（专业词）：不同领域中词语概率不一样，具有明显的语义特性

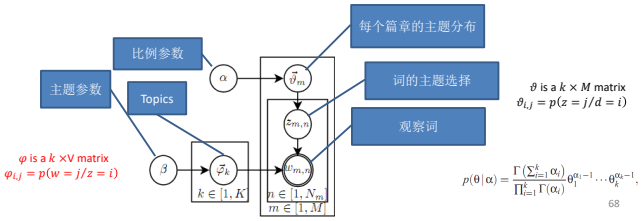
$$DR_{(t,k)} = \sum_{d \in D_k} p(t|d) log \frac{1}{p(t|d)}$$

$p(t|d)$：候选t在文档d出现的概率

Ranking：Topic Model

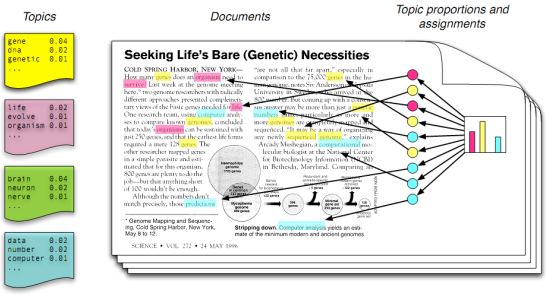
- 主题模型（Topic-model）
  - 篇章是由主题组成；
  - 篇章中的词，是以一定概率从主题中选取生成的；
  - 不同的主题，词语出现的概率分布是不同；
- 术语发现：
  - 提取不同主题中出现概率较大的词语





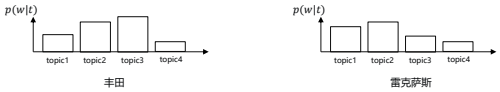
$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1},$$

Topic Model (LDA)



利用主题分布计算词之间的相似度

- 通过主题模型，我们可以获取每个词对于主题上的分布，在此基础上计算候选与种子词的语义相似度



$$score_{\cos}(word_1, word_2) = \cos ine(V_{word_1}, V_{word_2})$$

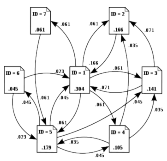
$$score_{euc}(word_1, word_2) = \sqrt{\sum_{k=1}^{N} (V_{word_1,k} - V_{word_2,k})^2}$$

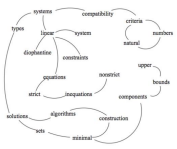
Ranking：TextRank

- PageRank：如果一个网页被很多其他网页链接到，说明这个网页比较重要；如果一个网页被一个权值很高的网页链接到，则其重要性也会相应增加
- TextRank：判断两个词间是否存在相关关系，则根据词语的共现关系。实际处理时，取一定长度的窗，在窗内的共现关系则视为有效。
- 术语选择：随机游走计算之后得到每个词语的重要度，按照重要度选择术语

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

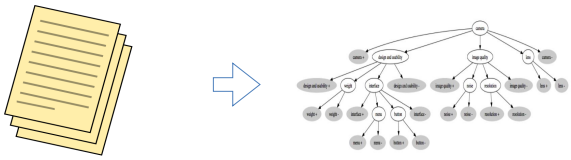
$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$





基于非结构化数据（纯文本）的知识体系构建

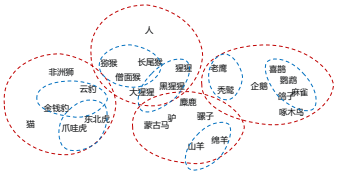
Ontology Learning from Texts



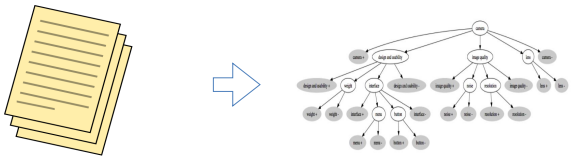
- 术语、概念抽取
- 同义词挖掘
- 关系挖掘
  - 上下位关系
  - 属性

同义词挖掘

- 目标：具有相同/相近语义的术语聚合在一起，形成概念
- 核心：语义相似度计算
  - 基于网络挖掘的同义词挖掘方法
    - 在网页中通过网络结构特性挖掘可能的同义词
  - 基于词典挖掘的同义词挖掘方法
    - 利用已有的词典（WordNet，Hownet，大词林，FrameNet等）获取同义词
  - 基于聚类的同义词挖掘
    - 利用聚类算法将同义词聚成簇，同时根据类别大小构建上下位关系
    - 核心：词的语义表示



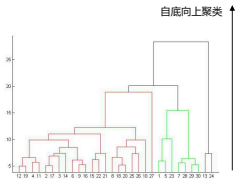
基于非结构化数据（纯文本）的知识体系构建



- 术语、概念抽取
- 同义词挖掘
- 关系挖掘
  - 上下位关系
  - 属性

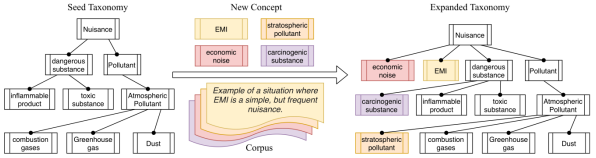
层次聚类（Hierarchical clustering）

- 算法(自底向上)基本步骤
  - 1. 将每个待聚类样本看作一类，计算两两样本之间的最小距离；
  - 2. 将距离最小（或者距离小于某一阈值）的两个类合并成一个新类；
  - 3. 更新类间距离（对于新类，重新计算类别中心）；
  - 4. 重复2、3，直到所有类最后合并成一类。
- 核心：如何计算类间距离
  - 欧氏距离
  - 样本语义表示（空间投射）
    - 基于词典
    - 基于上下文
    - 基于词向量
    - ……



介绍

- **分类体系扩展 Taxonomy Expansion（上下位关系）**
  - 给定一个现有分类体系 和 一些新的待插入的节点，要得到新的更完整的分类体系。



数据集

- Semeval-2016
- 环境、食品和科学

Language	Domain	Source	|V|	|E|	#i.i.	#c.c.	Cycles
English	Environment	Eurovoc	261	261	60	1	no
	Food	Combined	1556	1587	70	1	no
		WordNet	1486	1576	302	1	no
	Science	Combined	453	465	54	1	no
		Eurovoc	125	124	31	1	no
		WordNet	429	452	117	1	no

1	croquette	dish
2	cake mix	ready mix
3	grapefruit peel candied citrus peel	
4	tomato paste	ingredient
5	lime juice	juice
6	soda	beer
7	compote	dessert
8	puff paste	pastry
9	high tea	dinner
10	brunswick stew	stew
11	pudding	dish
12	decaffeinated coffee	coffee

基于模板学习的上下位关系抽取

- 上下位关系抽取

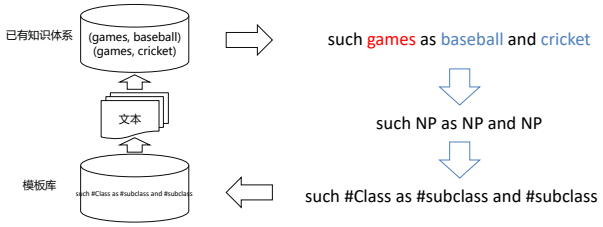
It is viewed as an advanced battery technology matching the images of many **European luxury vehicle brands** such as **Mercedes**, **BMW**, and **Audi**.

- 启发式规则：Heuristic Patterns

- such NP as {NP,}*{or|and} NP
  - such games as baseball and football
- NP {,NP}* {,} {and|or} other NP
  - rabbits and other animals (rabbits and other pets)
- NP {,} including {NP,}* {or|and} NP
  - fruits including apples and pears
- NP {,} especially {NP,}* {or|and} NP
  - Europeans, especially Italians (US presidents, especially democrats)
- NP is a NP
  - A kangaroo is an animal living in Australia
- a NP {named|called} NP
  - Japanese people like to play a game called Go
- NP, The NP
  - John loves his Audi. The car ...
- NP and other NP
  - universities and other institutions
- NP such as NP
  - sports such as basketball
- Among NP, NP
  - Among all musical instruments, piano are ...
- NP {except for|other than} NP
  - Employees except for managers suffer from ...

基于模板学习的上下位关系抽取

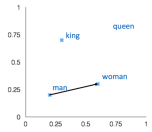
- 基于bootstrapping的模板学习
  - Step1：模板学习
  - Step2：模板过滤
  - Step3：基于习得模板的概念上下位关系抽取



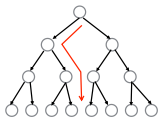
基于词表示学习的上下位关系抽取

- 基于词的向量表示，计算词之间的上下位关系

$$V_{man} - V_{woman} \approx V_{king} - V_{queen}$$



$$V_{虾} - V_{对虾} \approx V_{猫科} - V_{老虎}$$



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

属性抽取（No-taxonomic Relation)

- 属性抽取：针对某一概念、实体，抽取其属性关系，即No-taxonomic关系。通常情况下，属性抽取常面向限定领域或者限定类别
- 基于词性、句法的模板抽取方法

NP + Prep + CP:

battery of the **camera**

CP + with + NP

mattress with a *cover*

CP Verb(have, include, contain, consist, comprise..) NP

the **car** has a powerful *engine*

属性抽取（No-taxonomic Relation)

- 面对特定领域内文本的属性名抽取，可以利用额外信息。
- 例如：从商品评论中抽取商品属性，可以利用用户的评价词

"I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, and wanted me to return it to the shop. …"

Feature 1: **Touch screen**

Positive: 212

- *The touch screen was really **cool**.*
- *The touch screen was so **easy** to use and can do amazing things.*

…

Negative: 6

- The screen is easily **scratched**.
- I have a lot of **difficulty** in removing finger marks from the touch screen.

…

Feature 2: **battery life**

…

属性抽取（No-taxonomic Relation)

- 抽取策略：用户在商品评论的评价对象往往是某一商品或其属性词，评价词与商品属性词往往在同一句话中共现，利用评价词指示属性词

核心问题：获取评价关系

- 基于词共现的商品属性词抽取方法
- 基于句法模板的商品属性词抽取
- ……

屏幕的确*不错* ⇨ **这款手机**的信号*不错*

⇩

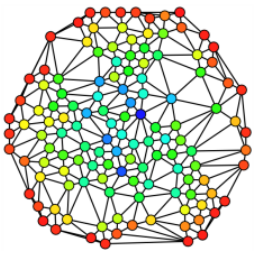
Iphone**外观***没的说* ⇦ **这款手机**信号*没的说*

Pattern#1: <OC> $\xrightarrow{mod}$ <TC> Example: This phone has an **amazing design**
Pattern#2: <OC> $\xrightarrow{pmod}$ <TC> Example: the **buttons easier** to use
Pattern#3: <OC> $\xrightarrow{remod}$ <TC> Example: 漂亮 的 外观 (**beautiful design**).
Pattern#4: <OC> $\xrightarrow{nsubj}$ <TC> Example: 这款 手机 不错 (This **phone** is **good**)
Pattern#5: <OC> $\xrightarrow{mod}$ (NN) $\xleftarrow{subj}$ <TC> Example: **IPhone** is a **revolutionary** smart phone
Pattern#6: <OC> $\xrightarrow{pmod}$ (NN) $\xleftarrow{subj}$ <TC> Example: **S3** is the phone **cheaper** to obtain.

句法模板

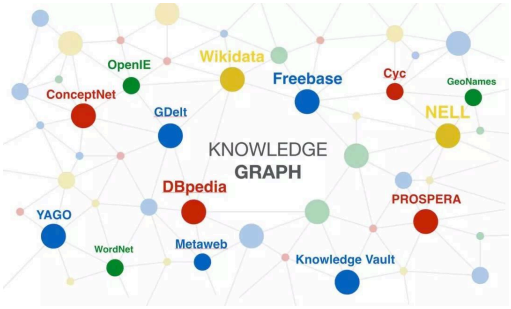
目录

- 知识建模
  - 知识体系概述
  - 典型知识体系
  - 知识体系手工建模方法
  - 知识体系自动建模方法
- 知识融合
  - 知识融合概述
  - 知识体系融合方法
  - 知识实例融合方法
- 大模型中的知识融合
  - 大模型对齐技术概述
  - 大模型对齐方法



知识融合

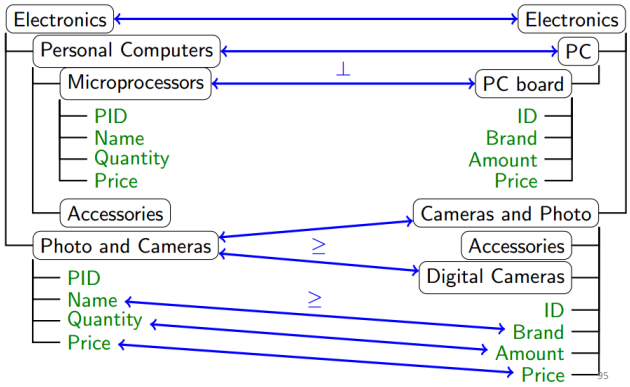
- 单一知识图谱难以覆盖各个领域，需要整合不同领域、不同语言、不同结构、不同模态的知识资源



知识图谱融合

- 任务定义
  - 给定两个知识图谱O和O'，I和I'分别是两个图谱中元素（概念、类别、实体，关系等），知识图谱融合的目标是建立三元组<I, I', r>，r为两个图谱中元素之间的关系，例如：$\leq \geq = \perp$
- 目标：将不同知识图谱融合为一个统一、一致、简洁的形式，为使用不同知识图谱的应用程序之间的交互建立互操作性

例子



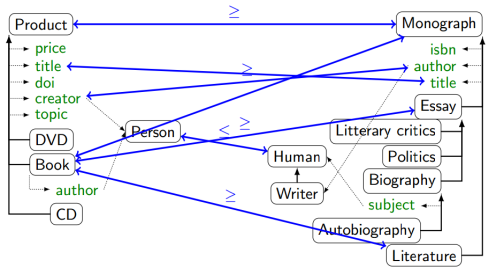
评测：OAEI (Ontology Alignment Evaluation Initiative)

- Ontology Alignment Evaluation Initiative
  - 本体对齐竞赛，目的是评估、比较、交流及促进本体对齐工作
  - 每年举办一次，结果公布在官网上
  - http://oaei.ontologymatching.org/

2021 OAEI-2021 at the ISWC ontology matching workshop, Online
2020 OAEI-2020 at the ISWC ontology matching workshop, Online
2019 OAEI-2019 at the ISWC ontology matching workshop, Auckland
2018 OAEI-2018 at the ISWC ontology matching workshop, Monterey
2018 OAEI-2017.5, somewhere in cyberspace
2017 OAEI-2017 at the ISWC ontology matching workshop, Wien
2016 OAEI-2016 at the ISWC ontology matching workshop, Kobe
2015 OAEI-2015 at the ISWC ontology matching workshop, Bethlehem
2014 OAEI-2014 at the ISWC ontology matching workshop, Riva del Garda
2013 OAEI-2013 at the ISWC ontology matching workshop, Sydney
2012 OAEI-2012 at the ISWC ontology matching workshop, Boston
2012 OAEI-2011.5 at ESWC 2012 iWEST workshop, Hersounissos
2011 OAEI-2011 at the ISWC ontology matching workshop, Bonn
2010 OAEI-2010 at the ISWC ontology matching workshop and IWEST workshop, Shanghai
2009 OAEI-2009 at the ISWC ontology matching workshop, Fairfax
2008 OAEI-2008 at the ISWC ontology matching workshop, Karlsruhe
2007 OAEI-2007 at the ISWC ontology matching workshop, Busan

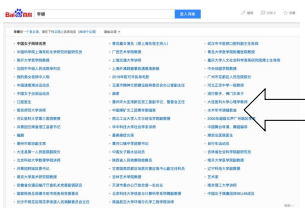
任务分解：本体匹配

- 本体匹配 (ontology matching, OM)
  - 侧重发现（模式层）等价或相似的类、属性或关系
  - 本体映射、本体对齐 ……



任务分解：实例对齐

- 实体对齐 (entity alignment, EA)
  - 侧重发现指称真实世界相同对象的不同实例
  - 实体消解、实例匹配 ……

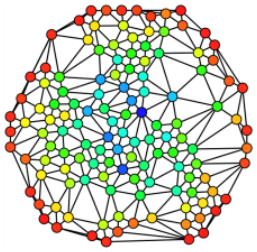


李娜（百度百科）

李娜（互动百科）

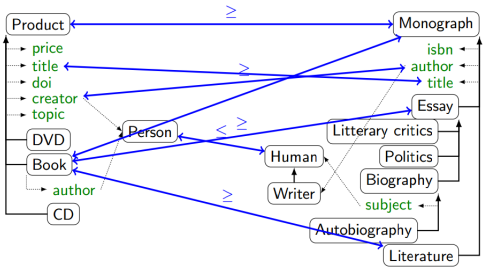
目录

- 知识建模
  - 知识体系概述
  - 典型知识体系
  - 知识体系手工建模方法
  - 知识体系自动建模方法
- 知识融合
  - 知识融合概述
  - 知识体系融合方法
  - 知识实例融合方法
- 大模型中的知识融合
  - 大模型对齐技术概述
  - 大模型对齐方法



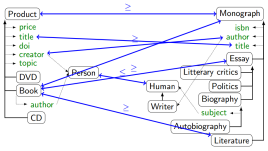
知识体系融合：本体匹配

- 本体匹配 (ontology matching, OM)
  - 侧重发现（模式层）等价或相似的类、属性或关系
  - 本体映射、本体对齐 .....



核心问题

- 核心问题：语义匹配
  - 不同字符：Michael Jordan vs. M.J.
  - 不同语言：中国 vs. China
  - 不同结构：$\text{hasSon}(x,y)$ vs. $\text{hasChild}(x,y) \sqcap \text{Male}(y)$
  - 不同表达：生日 vs. 出生年月
- 方法
  - 文本相似性
    - 字形 、语言 、语义
    - 文本本身、借助外部资源
  - 结构相似性
    - 上下位层级关系、Domain、Range等
    - 单对匹配 vs 集体 (collective) 匹配



方法分类

- 基于文本相似度的映射方法
  - 基于字符串匹配的映射方法
  - 基于语言处理的映射方法
  - 基于语义匹配的映射方法
- 基于结构相似度的映射方法
  - 基于内部结构的映射方法
  - 基于外部结构的映射方法
  - 基于网络表示学习的映射方法

基于字符串匹配的映射方法

- 汉明距离（Hamming Distance）：存在字符串x,t, 则它们之间的距离$\delta(x,t)$ 定义为：

$$\delta(s,t)=\frac{\left(\sum_{i=1}^{\min(|s|,|t|)} s[i]\neq t[i]\right)+||s|-|t||}{\max(|s|,|t|)}$$

people
↕
person

$$\delta(people,person)=\frac{4}{6}$$

基于字符串匹配的映射方法

- Substring Similarity：存在字符串x,y, t是x和y的最长公共子串，则它们之间的相似度$\sigma(x,t)$ 定义为

$$\sigma(x,y)=\frac{2|t|}{|x|+|y|}$$

$$\sigma(article,aricle)=\frac{2*4}{13}=8/13 \qquad \sigma(net,network)=\frac{2*3}{10}=6/10$$

- N-gram Similarity： $ngram(x,n)$为字符串x中长度为n的子串集合，则对于字符串x,t， 它们之间的相似度$\bar{\sigma}(x,t)$ 定义为

$$\bar{\sigma}(s,t)=\frac{|ngram(s,n)\cap ngram(t,n)|}{\min(|s|,|t|)-n+1}$$

$ngram(article,2)=\{ar,rt,ti,ic,cl,le\}$
$ngram(aricle,2)=\{ar,ri,ic,cl,le\}$

$$\bar{\sigma}(article,aricle)=\frac{4}{6-2+1}=4/5$$

基于字符串匹配的映射方法

- 编辑距离（Levenshtein Distance）：将一个字符串转成另一个字符串的最少编辑操作数（插入、删除、替换）

$$'Lvensshtain' \xrightarrow{\text{插入 'e'}} 'Levensshtain'$$
$$'Levensshtain' \xrightarrow{\text{删除 's'}} 'Levenshtain'$$
$$'Levenshtain' \xrightarrow{\text{替换 'a'→'e'}} 'Levenshtein'$$

- 将Lvensshtain 转换成Levenshtein，总共操作 3 次，编辑距离是3。这是典型的动态规划问题，可通过动态规划算法计算。给定两个字符串A, B，i, j分别为字符串A, B的下标，则它们之间的编辑距离为$\delta_{A,B}(|A|, |B|)$，不失一般性，$\delta_{A,B}(i, j)$可以计算为：

$$\delta_{A,B}(i,j) = min\begin{cases} \delta_{A,B}(i-1,j)+1 & \\ \delta_{A,B}(i,j-1)+1 & if\ A[i]=B[j],\ flag=0 \\ \delta_{A,B}(i-1,j-1)+flag & if\ A[i]\neq B[j], flag=1 \end{cases}$$

基于语言处理的映射方法

- 语言规范化（Normalization）
  - 词切分（Tokenization）
    - Hands-Free_Kits → <hands, free, kits>
  - 词形还原（Lemmatization）
    - Kits → Kit
  - 短语中的停用词去除（Stop words elimination）
    - Stop words list: a, the, by, type of, their, from

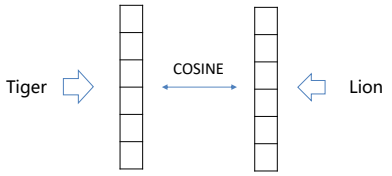
	Papers-on-the-theory ⟷	Theoretical paper
Tokenization:	papers on the theory	theoretical paper
Lemmatization:	paper on the theory	theory paper
Stop words elimination:	paper theory	theory paper

基于语义匹配的映射方法

- 获取/学习两个图谱中元素（概念、关系、实体等）的语义向量表示，然后利用距离度量函数（例如：欧式距离）计算他们之间的相似度

$$\sigma_V(s,t) = \frac{\sum_{i\in|V|} \overrightarrow{s}_i \times \overrightarrow{t}_i}{\sqrt{\sum_{i\in|V|} \overrightarrow{s}_i^2 \times \sum_{i\in|V|} \overrightarrow{t}_i^2}}$$

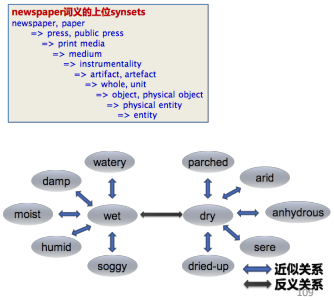
- 核心问题：如何得到语义向量？



利用外部资源：WordNet，Hownet

- WordNet中将英语的名词、动词、形容词、和副词组织为Synsets，每一个Synset表示一个基本的词汇概念

- 概念关系
- 同义关系
- 反义关系
- 上位关系
- 下位关系
- 整体关系（名词）
- 部分关系（名词）
- 蕴含关系（动词）
- 因果关系（动词）
- 近似关系（形容词）



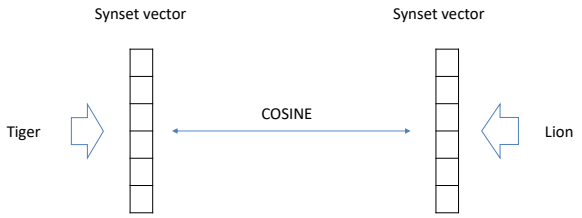
利用WordNet中的直接标注信息

- 直接利用WordNet中的上下位、同义等标注信息
  - 上下位关系
    - A ⊑ B if A is a hyponym or meronym of B
      - Brand ⊑ Name
    - A ⊒ B if A is a hypernym or holonym of B
      - Europe ⊒ Greece
  - 同义关系
    - A = B if they are synonyms
      - Quantity = Amount
  - 反义关系
    - A ⊥ B if they are antonyms or the siblings in the part of hierarchy
      - Microprocessors ⊥ PC Board

基于WordNet的相似度计算

- 直接利用WordNet中对于词的Synset标注，构建其语义向量。对于词（概念）

$$\sigma(s,t) = \frac{|\Sigma(s) \cap \Sigma(t)|}{|\Sigma(s) \cup \Sigma(t)|}$$



基于搜索引擎的语义相似度计算

- Google distance
  - 利用搜索引擎的返回文档数，估计两个概念的相关度/相似度
  - 给定两个概念x, y，Normalized Google-based Distance:

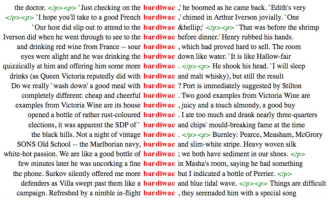
$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

  - $f(x)$：搜索引擎对于查询x返回的文档数
  - $f(y)$：搜索引擎对于查询y返回的文档数
  - $f(x,y)$：搜索引擎对于查询(x,y)返回的文档数
  - M：搜索引擎中所有的文档数($M \approx 10^{10}$)

NGD(中国, 中华人民共和国)=0.799

基于词向量表示的语义匹配

- 利用上下文信息学习词的词表示向量
  - 基本原则：具有相同(类似)上下文信息的词应该具有相同(类似)的词表示

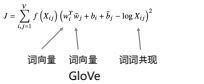


	glass	drink	grape	red	meal
bardiwac	10	22	43	16	29
car	5	0	0	10	0

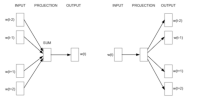
$$\vec{v} = (c_1, c_2, ..., c_n)$$

词向量表示学习方法

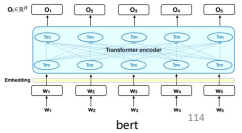
- 传统词向量表示学习：基于共现信息进行统计
  - 主题模型
    - Latent Semantic Analysis (LSA)
    - Probabilistic Semantic Analysis (PLSA)
    - LDA
  - 聚类模型
    - Brown Clustering
    - Hyperspace Analogue to Language
    - GloVe
- 神经语言模型：利用上下文预测目标词
  - NNLM
  - COBW
  - Word2Vec
  - BERT
  - GPT-3
  - XLNet



GloVe

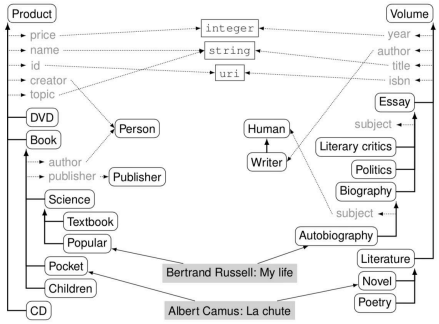


word2vec



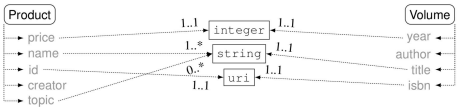
bert

基于结构相似度的映射方法



基于内部结构的映射方法

- 面对实体内部结构信息
  - 属性的domain和range
  - 常用于对齐前的预处理，去掉明显不能对齐的实体
  - ...

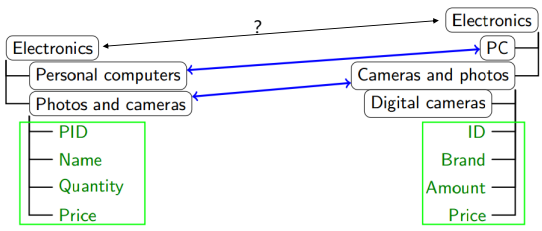


	char	fixed	enumeration	int	number	string
string	0.7	0.4	0.7	0.4	0.5	1.0
number	0.6	0.9	0.0	0.9	1.0	0.5

examples in data type similarity table

基于外部结构的映射方法

- 基于图的拓扑结构计算两个节点的相似度
- 基本假设
  - 如果两个不同节点的邻居节点（父类、子类、属性等）是类似的，则这两个节点也是类似的

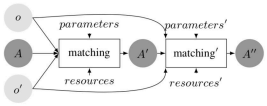


方法融合

- 单一方法并不能取得最优效果，需要将不同方法结果进行融合
- 线性融合

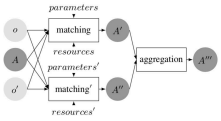
$$Similarity(x,y) = w_1 \times Similarity_{string} + w_2 \times Similarity_{semantic} + w_3 \times Similarity_{structure} + \cdots$$

- 迭代式融合



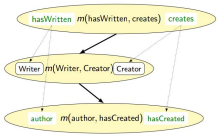
冲突解决

- 投票策略



- 全局寻优、联合推断
  - 贝叶斯网

$$P(X_1,\ldots,X_n) = \prod_i P(X_i|parents(X_i)), i = 1,\ldots,n$$



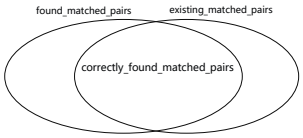
评价

- Precision, Recall, F1

$$\text{Precision} = \frac{\#\ correctly\_found\_matched\_pairs}{\#\ found\_matched\_pairs}$$

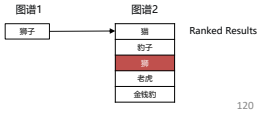
$$\text{Recall} = \frac{\#\ correctly\_found\_matched\_pairs}{\#\ existing\_matched\_pairs}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



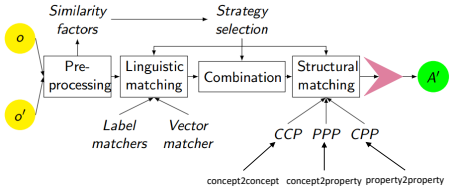
- 基于检索的评价指标
  - 给定一个concept/entity，根据打分获取其TopN可能对齐的结果
  - P@N, MAP (平均准确率)，MRR (平均倒数排名)...

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}$$



已有的知识融合系统： RiMOM

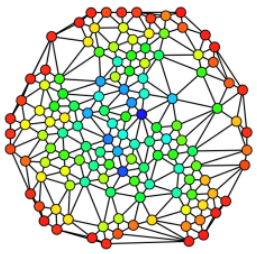
- 动态多策略对齐框架
  - 基于语言学的相似度
  - 基于结构的相似度



Euzenat et al. Ontology Matching

目录

- 知识建模
  - 知识体系概述
  - 典型知识体系
  - 知识体系手工建模方法
  - 知识体系自动建模方法
- 知识融合
  - 知识融合概述
  - 知识体系融合方法
  - 知识实例融合方法
- 大模型中的知识融合
  - 大模型对齐技术概述
  - 大模型对齐方法



实体消歧定义

- 命名实体的歧义指的是一个实体指称项可对应到多个真实世界实体，例如，给定如下的四个实体指称项 "*Michael Jordan*"

MJ1: Michael Jordan is a researcher in machine learning.

MJ2: Learning in Graphical Models: Michael Jordan

MJ3: M. Jordan wins NBA MVP.

MJ4 : Michael Jordan plays basketball in Chicago Bulls.

- 确定一个实体指称项所指向的真实世界实体，这就是命名实体消歧

意义：知识图谱

- 对于一段自然语言文本，“迈克尔·乔丹教授昨天访问了CMU”。要从自然语言文本中抽取信息构成知识图谱，处理流程如下：
  - 命名实体识别：
    - “[迈克尔·乔丹]/PER教授昨天访问了[CMU]/ORG”。
  - 关系抽取：
    - (迈克尔·乔丹，visit, CMU)
- 抽取出三元组并不能直接构造知识图谱，因为我们不知道迈克尔·乔丹到底是哪个迈克尔·乔丹，CMU到底指的是哪个机构。因此，需要利用实体链接技术。

迈克尔·乔丹 → 

CMU → 

实体歧义的来源

- 自然语言任务的多样性：同一意义不同表达
- Variability of Natural Language
- Name variations

科比
黑曼巴
小飞侠
蜗壳
81先生
洛城铁匠

实体消歧的来源

- 自然语言任务的歧义性：同一表达不同意义
- Ambiguity of Natural Language
- Name ambiguity

王强

维基百科，自由的百科全书

王强是非常常见的中文人名，较出名的有：

- 王文 (明朝)，明朝大臣，初名**王强**
- 王强 (解放军)：曾任中国人民解放军济南军区空军副参谋长等职，2014年7月晋升空军上将[1]。现任西部战区副司令员。
- 王强 (书法家)：中央财经大学文化与传媒学院教授、院长
- 王强 (足球运动员)
- 王强 (罪犯)
- 王强 (台湾歌手)：本名王百照，1970年代台湾男性美声团体“原野三重唱”的主唱（19
- 王强 (中国大陆歌手)：中国大陆男歌手，代表作品《秋天不回来》。
- 王强 (摔跤运动员)
- 王强 (演员)
- 王强 (企业家)：真格基金联合创始人，新东方教育科技集团联合创始人

普通词的歧义

- 一词多义
  - 打水，打电话，打毛衣，打哈欠，打拍子
  - 拍子坏了，打拍子
  - 看电影，看病
  - 炒菜，炒外汇，炒鱿鱼
  - Bank：银行，河岸
  - Plant：植物，工厂
- 一义多词
  - 计算机，电脑
  - WordNet
  - 同义词词林

词义排歧 vs 实体消歧

- 相同点：
  - 实体消歧和词义排歧都在解决语言中词汇歧义的问题。
- 不同点：
  - 普通词及其义项通常是比较固定的，可以由词典列举；实体词及其义项无法列举。
  - 实体词的义项数目要比普通词多很多。
  - 实体词消歧的场景比普通词消歧要丰富。
  - 实体词消歧可利用的特征比普通词要丰富。

Culture and the arts [edit]

Literature [edit]
- *Chicago* (magazine), a lifestyle periodical about the Illinois city
- *Chicago* (manga), by Yumi Tamura
- *Chicago* (novel), by Egyptian author Alaa-Al-Aswany
- "Chicago" (poem), by Carl Sandburg
- *The Chicago Manual of Style*, for American English

Music [edit]
- Chicago (band), a rock band founded in 1967
  - *Chicago* (album), 1970
- "Chicago" (Graham Nash song), 1970
- "Chicago" (Michael Jackson song)
- "Chicago song", a composition by Marcus Miller on David Sanborn's *A Change of Heart*, 1987
- "Chicago" (Sufjan Stevens song), 2005
- "Chicago (That Toddlin' Town)", a 1922 song by Fred Fisher
- Chicago house, a genre of electronic dance music
- "Chicago", a 1977 single by Kiki Dee
- "Chicago", a song by Groove Armada from the 1999 album *Vertigo*
- "Chicago", a song by Kate Voegele from the 2007 album *Don't Look Away*
- "Chicago", a song by The Devil Wears Prada from the 2011 album *Dead Throne*
- "Chicago", a song by Tom Waits from the 2011 album *Bad as Me*
- Chicago (trance artist), stage name for David Christman

Theatre, film and television [edit]
- *Chicago* (play), 1926, written by Maurine Dallas Watkins
  - *Chicago* (1927 film), based on the 1926 play
  - *Chicago* (musical), first performed in 1975, based on the 1926 play
    - *Chicago* (2002 film), based on the 1975 musical
- "Chicago" (*Prison Break*), a 2007 TV episode

实体消歧分类

- 基于无监督聚类的实体消歧
  - 把所有实体指称项按其指向的目标实体进行聚类
  - 每一个实体指称项对应到一个单独的类别

MJ1: Michael Jordan is a researcher in machine learning.

MJ2: Research in Graphical Models: Michael Jordan

MJ3: M. Jordan wins NBA MVP.

MJ4 : Michael Jordan plays basketball in Chicago Bulls

MJ1 MJ2

MJ3 MJ4

- 基于实体链接的实体消歧
  - 将实体指称项与目标实体列表中的对应实体进行链接实现消歧

MJ4 : Michael Jordan plays basketball in Chicago Bulls

基于聚类的实体消歧

- 基本思路
  - 指向相同实体的实体指称项有相似的上下文
  - 利用聚类算法进行消歧
  - 核心问题：选取何种特征对指称项进行表示
    - 词袋模型(Bagga et al., COLING, 1998)
    - 语义特征(Pederson et al., CLITP, 2005)
    - 社会化网络(Bekkerman et al., WWW, 2005)
    - 维基百科的知识(Han and Zhao, CIKM, 2009)
    - 多源异构语义知识融合(Han and Zhao, ACL, 2010)

基于聚类的实体消歧：评测

- WePS： Web People Search Evaluation
  - WePS1是SEMEVAL2007的子任务
  - WePS2是WWW的一个workshop
  - 任务：Web环境中的人名消歧，即给定一个包含某个歧义人名的网页集合，按照网页中人名指称项所指向的人物概念来对网页进行聚类，以及抽取一个网页中关于某个人的特定属性来辅助进行人名消歧。
  - 评测方法

$$\text{Purity} = \sum_i \frac{|C_i|}{n} \max \text{Precision}(C_i, L_j)$$

$$\text{Inverse Purity} = \sum_i \frac{|L_i|}{n} \max \text{Precision}(L_i, C_j)$$

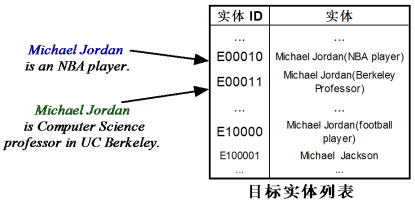
$$\text{F} = \frac{1}{\alpha \frac{1}{\text{Purity}} + (1-\alpha)\frac{1}{\text{Inverse Purity}}}$$

纯净度Purity：评价的聚类结果中每个类别中指称项的平均准确率。

倒纯净度Inverse Purity：评价的聚类结果中每个类别中指称项的平均召回率。

实体链接的任务

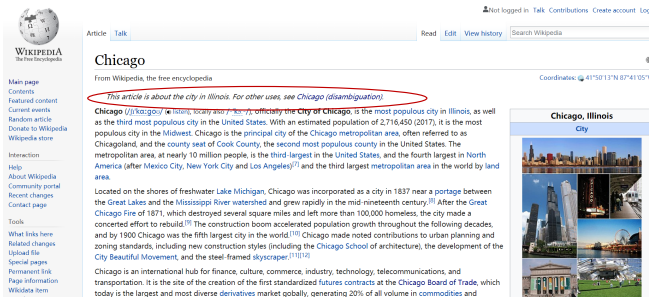
- 任务：给定实体指称项和它所在的文本，将其链接到给定知识库中的相应实体上。



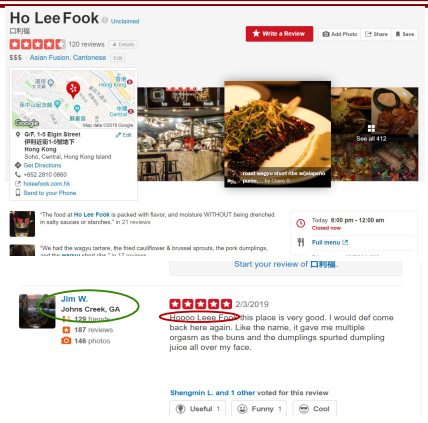
实体链接的输入输出

- 输入
  - 目标实体知识库：目前最常用的是Wikipedia，在其他一些任务中可能是特定领域的知识库，比如说社交媒体中的Yelp，电影领域的IMDB等。
  - 待消歧实体指称项及其上下文信息。
- 输出
  - 文本中实体指称项映射到的知识库中的实体。

实体链接的输入：链接到Wikipedia

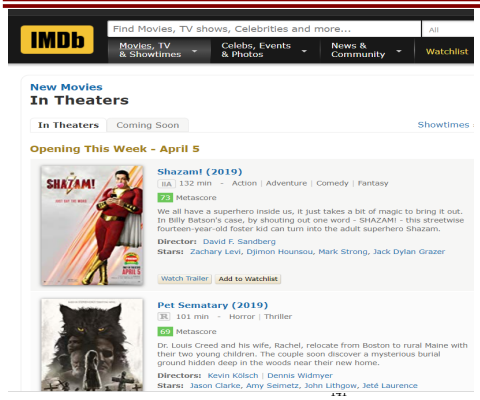


实体链接的输入：链接到Yelp



- Yelp是美国最大点评网站，用户可以在Yelp网站中给商户打分，提交评论，交流购物体验等。
- Yelp中的很多实体并不出现在Wikipedia中，比如说“口利福 Ho Lee Fook”。
- Yelp中的很多普通用户也不会出现在Wikipedia中，但是它们都在Yelp这类平台上有账号，也是一个实体。
- Yelp中的评论中也存在实体歧义，需要链接到Yelp本身来消歧。

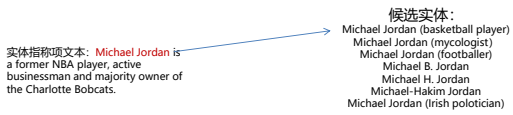
实体链接的输入：链接到IMDb



- Internet Movie Database
- 互联网电影资料库
- 是一个关于电影演员、电影、电视节目、电视明星和电影制作的在线数据库。
- 包括了影片的众多信息、演员、片长、内容介绍、分级、评论等。
- 对于电影的评分目前使用最多的就是IMDb评分。

实体链接主要步骤

- 主要步骤
  - 候选实体的发现
    - 给定实体指称项，链接系统根据知识、规则等信息找到实体指称项的候选实体
  - 候选实体的链接
    - 实体链接的核心
    - 系统根据指称项和候选实体之间的相似度等特征，选择实体指称项的目标实体
  - 无链接实体（NIL）的聚类



候选实体发现

- 如何根据实体指称项找出候选实体
  - 利用Wikipedia的信息
  - 利用上下文信息

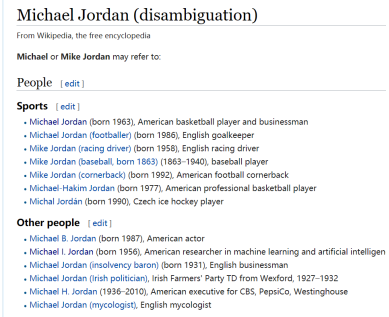
实体指称项	候选实体
Michael Jordan	Michael Jordan (basketball) Michael Jordan (mycologist) Michael Jordan (football) Michael B. Jordan (American actor) …
AI	Artificial intelligence Ai (singer) …
…	…

候选实体发现：Wikipedia锚文本

- 超链接是指文本内由一文件连接至另一文件的链接
- 蓝色的字表示网页中的超链接

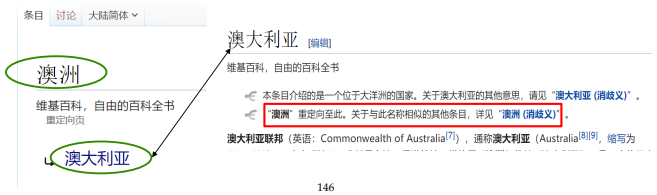
候选实体发现：Wikipedia消歧页面

- 维基百科中用于消除“一词多义”所引起的歧义的页面。



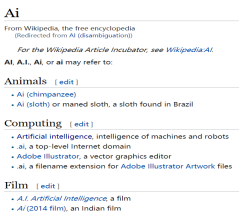
候选实体发现：Wikipedia重定向页面

- 重定向是一种特殊的页面，它提供一种运作机制，使得人们在输入该名称进入条目时，系统能够自动导航到重定向页面内部指定的另一相关页面中，从而实现相关页面可以以多个名称进行访问。
- 例如：如果设定了“名称为‘澳洲’”，而内容指向“澳大利亚”的重定向页之后，任何人都可以用“澳洲”这一名称进入到澳大利亚条目中。



缩略语候选实体发现：利用上下文

- 问题
  - 缩略语在实体指称项中十分常见，据统计，在KBP2009的测试数据，在3904个实体指称项中有827个为缩略语，缩略语指称项具有很强的歧义性。


Ai

From Wikipedia, the free encyclopedia
(Redirected from AI (disambiguation))

For the Wikipedia Article Incubator, see Wikipedia:AI.

AI, A.I., Ai, or ai may refer to:

Animals [edit]
- Ai (chimpanzee)
- Ai (sloth) or maned sloth, a sloth found in Brazil

Computing [edit]
- Artificial intelligence, intelligence of machines and robots
- .ai, a top-level internet domain
- Adobe Illustrator, a vector graphics editor
- .ai, a filename extension for Adobe Illustrator Artwork files

Film [edit]
- *A.I. Artificial Intelligence*, a film
- *Ai* (2014 film), an Indian film

张涛，刘康，赵军. 基于缩略语扩展与网络数据挖掘的新型实体链接系统. 第七届全国信息检索学术会议. 济南，2011年.

缩略语候选实体发现：利用上下文

- 动机
  - 缩略语指称项具有很强的歧义性，但它的全称往往是没有歧义的
    - ABC和American Broadcasting Company
    - AI和Artificial Intelligence
  - 在实体指称项文本中，缩略语的全称出现过。
- 解决方法
  - 利用人工规则抽取实体候选。

张涛，刘康，赵军. 基于缩略语扩展与网络数据挖掘的新型实体链接系统. 第七届全国信息检索学术会议. 济南，2011年.

候选实体链接

- 基本方法：计算实体指称项和候选实体的相似度，选择相似度最大的候选实体。
  - 利用先验知识做初始排序
  - 局部实体链接
  - 协同实体链接
- 三个步骤都可以单独进行候选实体链接。但是大多数模型是将三者组合使用。

候选实体链接

- 基本方法：计算实体指称项和候选实体的相似度，选择相似度最大的候选实体
  - 利用先验知识做初始排序
  - 局部实体链接
  - 协同实体链接
- 三个步骤都可以单独进行候选实体链接。但是大多数模型是将三者组合使用。

利用指称项指向实体的概率作为先验知识做初始排序

- 初始排序的目的
  - 实体指称项的候选实体可能非常多，比如“芝加哥”的候选可能有5000多个候选实体。为了平衡计算精度和计算时间，利用指称项指向实体的概率作为先验知识对候选实体进行粗筛选是非常必要的。
- 指称项指向实体的概率

$$P(entity|mention) = \frac{count(mention \rightarrow entity)}{\sum_{entity' \in W} count(mention \rightarrow entity')}$$

利用指称项指向实体的概率作为先验知识做初始排序

- 利用Wikipedia中的超链接信息计算指称项指向实体的概率，对指称项Chicago候选进行初始排序：

Rank	Entity	P(Entity|“Chicago”)
1	Chicago(City)	0.76
2	Chicago(band)	0.41
3	Chicago(2002_film)	0.022
20	Chicago Maroons Football	0.00186
100	1985 Chicago Whitesox Season	0.000023448
5005	Chicago Cougars	0.00000528

  - 只保留P（Entity|“Chicago”）最大的K个实体作为指称项的候选实体，K是模型的超参数。
  - 这种利用指称项指向实体的概率思想最早是由Medelyan et al 在2008年提出的。

利用指称项指向实体的概率作为先验知识做初始排序

Generated Candidates K	Data sets			
	ACE	MSNBC	AQUAINT	Wiki
1	81.69	72.26	91.01	84.79
3	85.44	86.22	96.83	94.73
5	86.38	87.35	97.17	96.37
20	86.85	88.67	97.83	98.59

Ratinov et al. (2011)

- K=1，表示只保留概率最高的那个候选实体，那个候选实体是正确答案的比例除了MSNBC，其他三个数据集都大于80%。
- K=5，保留概率最大的5个候选实体，这5个候选实体中包含正确答案的比例在四个数据集中均大于86%。

利用指称项指向实体的概率作为先验知识做初始排序

- 但是指称项指向实体的概率并不鲁棒
  - 在Tweets数据上，即使保留前50个候选实体，性能也只有77%。

Generated Candidates K	Tweets
1	60.21
50	77.75

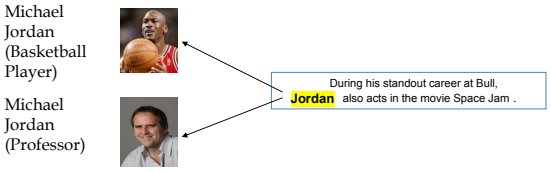
Meij et al. (2012)

- 所以需要在先验知识之上探索更加鲁棒的特征。

候选实体链接

- 基本方法：计算实体指称项和候选实体的相似度，选择相似度最大的候选实体
  - 利用先验知识做初始排序
  - ➡ 局部实体链接
  - 协同实体链接
- 三个步骤都可以单独进行候选实体链接。但是大多数模型是将三者组合使用。

候选实体链接：局部实体链接



对于文本中出现的Jordan这个实体指称项，通过计算指称项Jordan和MJ(篮球)和MJ(教授)之间的语义相似度，挑出相似度最大的实体作为实体指称项的链接实体。

候选实体链接：局部实体链接

- 传统特征的方法（2016-2017年之前）
  - 算法的核心是如何手工设计有效的特征
  - 实体的表示很简单，一般采用Wikipedia页面中的词条来表示实体
  - 论文：
    - BOW模型 (Honnibal TAC 2009, Bikel TAC 2009)
    - 加入实体流行度等特征（Han ACL 2011）
    - 加入候选实体的类别特征（Bunescu et al., EACL 2006）

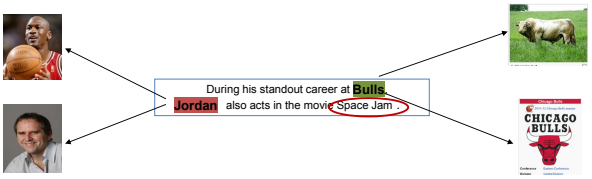
候选实体链接：局部实体链接

- 表示学习的方法（2017-至今）
  - 算法的核心是如何获得实体和实体指称项上下文的分布式表示
  - 实体表示比较复杂
    - 从不同粒度来表示实体
    - 可能会利用实体的类别（Entity Type）信息
    - 可能会利用Wikipedia中实体与实体共现关系
  - 论文：
    - 卷积神经网络模型（Francis-Landau et al., NAACL 2016）
    - 利用预训练实体向量表示实体（Ganea and Hofmann, EMNLP 2017）

候选实体链接

- 基本方法：计算实体指称项和候选实体的相似度，选择相似度最大的候选实体
  - 利用先验知识做初始排序
  - 局部实体链接
  - ➡ 协同实体链接
- 三个步骤都可以单独进行候选实体链接。但是大多数模型是将三者组合使用。

候选实体链接：协同实体链接



不仅要考虑实体指称项与目标实体的语义相似度，
还要考考虑目标实体之间的全局语义相似度

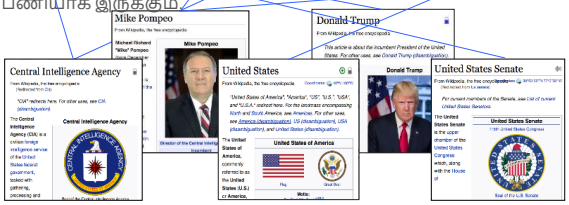
候选实体链接：协同实体链接

- 协同实体链接是在局部实体链接之上，增加了一个全局项（协同策略），来综合考虑目标实体之间一致性。
- 全局项计算方法：
  - 基于图的方法 (Han et al., SIGIR 2011)
  - 基于条件随机场的方法 (Ganea et al., EMNLP 2017)
  - 基于Pair-Linking的方法 (Phan et al., CIKIM 2017)

跨语言实体链接

泰米尔语：印度南部、斯里兰卡东北部

சிஜஏ இயக்குநர் மைக் பாம்பேயோ நியமனத்துக்கு அமெரிக்க செனட் சபை ஒப்புதல். ஆனால், சிஜஏ முகமை மற்றும் அமெரிக்க அதிபர் டிரம்ப் இடையே ஒரு பயனுள்ள அலுவல் ரீதியான உறவினை உருவாக்குவதே மைக் பாம்பேயோவின் உடனடி பணியாக இருக்கும்.

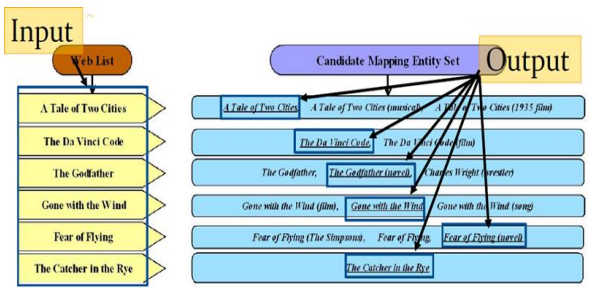


注：中央情报局局长蓬佩奥就任特朗普政府的国务卿，提名已获得参议院的通过。但蓬佩奥的当务之急是要在CIA和美国总统特朗普之间建立良好的关系。

跨语言实体链接

- 跨语言实体链接定义：
  - 将一种语言的表述的实体指称项链接到另一种语言的知识库中，例如实体指称项是泰米尔语描述的，链接到英文Wikipedia中。
  - 跨语言实体主要由 Text Analysis Conference（TAC）——Knowledge Base Population（KBP）——Entity Linking Tracks推动。
- 跨语言实体链接的难点：
  - 跨语言实体链接中要利用到不同语言知识库之间的对应关系，但很多语言Wikipedia并不完备，例如北梭托语只有4000个Wikipedia页面。
  - 带来两个问题：
    - 跨语言候选实体生成很难。
    - 跨语言文本相似度计算很难：神经网络跨语言实体链接需要解决实体指称项描述语言词向量和英文词向量的位于不同语义空间的问题。

实体列表中的实体链接



列表中的实体指称项应该拥有同一种类型

Shen W, Wang J, Luo P, et al. LIEGE: link entities in web lists with knowledge base. SIGKDD 2012

实体列表中的实体链接

- 基于”列表中的实体指称项应该拥有同一种类型”的假设，候选实体要满足：
  - 这个候选实体的先验概率较高
  - 这个候选实体的类型与同一个列表中其他列表项的对应实体的类型一致（语义相似）
- 建模语义相似的方法
  - 基于类型层次结构的相似性
  - 实体上下文分布相似性
  - 利用最大间隔方法自动学习特征的权值，为每个候选实体定义链接质量
  - 利用迭代替换算法对实体列表中所有相对应的实体进行联合优化

Shen W, Wang J, Luo P, et al. LIEGE: link entities in web lists with knowledge base. SIGKDD 2012

社交数据中的实体链接

- 社交媒体的特点（以Tweet为例）
  - 用户多，每个月的活跃用户超过3亿3千万人
  - 数目大，每天Tweet的数量超过5亿条, 主题从生活到突发的新闻
- Tweet文本的特点：
  - 字数限制，每条tweet不超过140个字，文本短
  - 噪音大，非正式的缩写，写作方式口语化，打字错误
  - 实时性强，tweet内容中包含了很多新发生的事件和新产生的实体

Shen W, Wang J, Luo P, et al. Linking named entities in tweets with knowledge base via user interest modeling. SIGKDD 2013

社交数据中的实体链接

- 社交数据中实体链接面临的挑战
  - 常用的实体链接方法中，实体指称项上下文和知识库中实体的描述之间的相似度是重要的特征。由于社交文本的特性，很难计算这一相似度。
  - 常用的实体链接方法中，协同链接是重要的部分，但在社交数据中协同链接可能没有用武之地。以Tweet为例，每条Tweet中平均只包含0.76个实体，实体个数少，不同实体之间的一致性这一重要特征很难利用。

Shen W, Wang J, Luo P, et al. Linking named entities in tweets with knowledge base via user interest modeling. SIGKDD 2013

实体链接：常用数据集

- AIDA（AIDA CoNLL-YAGO）：由马普研究所公开的数据集，是目前最大的手工标注实体链接数据集。它是基于CoNLL 2013 实体识别数据集上标注的，题材是路透社新闻。
- WNED：自动构建的数据，数据规模很大。WNED-CWEB是从ClueWeb中自动构建的，WNED-WIKI是从Wikipedia中自动构建的。由于是自动构建的数据，所以数据中噪音比较大，可信度较低。

数据集	实体指称项个数	文档的数目
AIDA-train	18448	946
AIDA-A(valid)	4791	216
AIDA-B(test)	4485	231
MSNBC	956	20
AQUAINT	727	50
ACE2004	257	36
WNED-CWEB	11154	320
WNED-WIKI	6821	320

实体链接：常用数据集

- TAC KBP数据集 2009-2018：TAC (Text Analysis Conference) KBP (Knowledge Base Population)是国际上知名的实体链接评测，由美国国防高级研究计划局（DARPA）资助。数据来源是新闻和论坛，是手工标注的数据集。以TAC KBP 2015年为例，介绍数据集的规模。

TAC KBP 2015标注的篇章个数

领域	中文	西班牙文	英文	总量
训练数据				
新闻	84	82	85	251
论坛	63	47	83	193
总量	147	129	168	444
测试数据				
新闻	84	82	85	251
论坛	63	47	83	193
总量	147	129	168	444

TAC KBP 2015标注的实体指称项个数

	训练	测试
总量	30838	32533
中文	13116	11066
西班牙文	4177	5822
英文	13545	15645

实体链接：评测

- TAC-KBP (2009-2013)：**Entity Linking**
  - 任务：将文本中的实体指称项链接到Wikipedia中的真实概念，达到消歧的目的。
  - 评测方法：

以指称项为单位计算的准确率

$$Accuracy_{micro} = \frac{NumCorrect}{NumQueries}$$

$$Accuracy_{macro} = \frac{\sum_{i}^{NumEntities} \frac{NumCorrect(E_i)}{NumQueries(E_i)}}{NumEntities}$$

以实体为单位计算的准确率