哈爾濱工業大學
社会计算与信息检索研究中心

基于表示学习的开放域中文知识推理

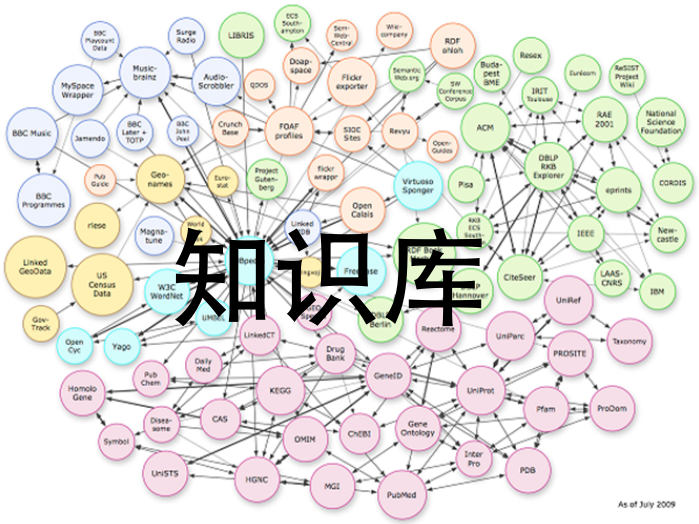
CCKS2016-全国知识图谱与语义计算大会

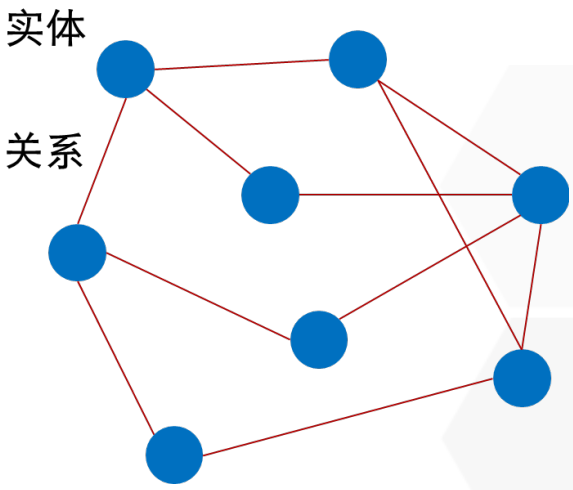
姜天文，秦兵，刘挺

1 研究背景与目的

1.研究背景与目的

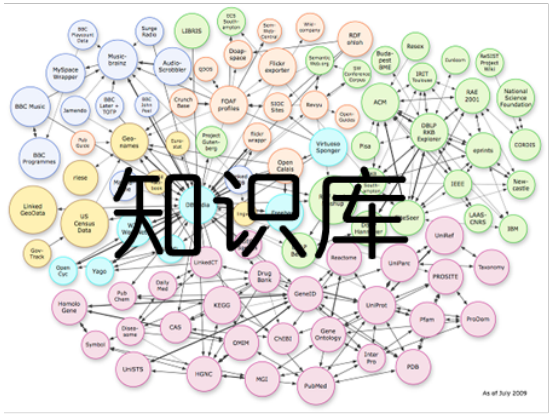
网状的表示形式：计算效率较低、无法很好应对数据稀疏问题。以符号为基础的网状形式无法应对连续空间里的数值计算。

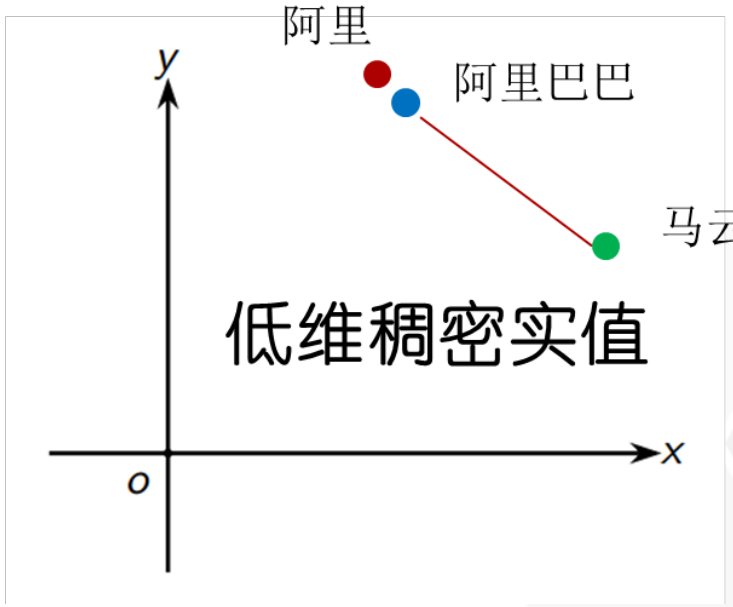




1.研究背景与目的

表示学习旨在将网状的语义信息表示为稠密低维的实值向量。



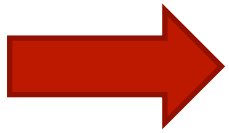


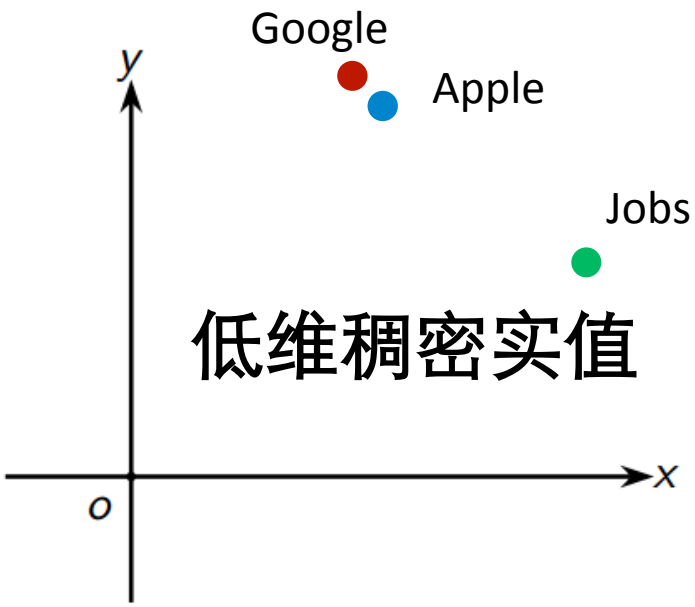
1.研究背景与目的

距离模型、能量模型、张量模型、翻译模型……

WordNet
A lexical database for English

Freebase™





1.研究背景与目的

传统类型的知识库 → 开放域知识库

（Google，_雇佣关系_，拉里·佩奇）

↓

（Google，创始人，拉里·佩奇）

（百度公司，_雇佣关系_，李彦宏）

↓

（百度公司，CEO，李彦宏）

1.研究背景与目的

本文的主要研究内容

本课题的主要研究：对开放域中文知识库进行低维实值向量表示以提高知识库应用时的计算效率、使语义数值化；

并对这种分布式知识表示在实际应用价值的探索。

（Google，_雇佣关系_，拉里·佩奇）

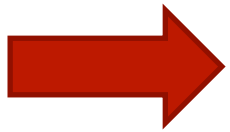
↓

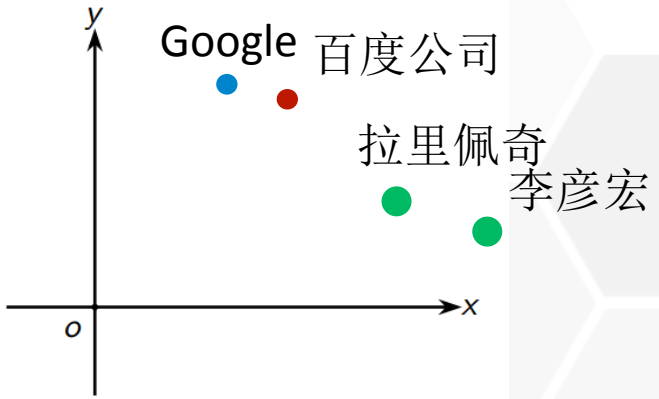
（Google，创始人，拉里·佩奇）

（百度公司，_雇佣关系_，李彦宏）

↓

（百度公司，CEO，李彦宏）





哈爾濱工業大學
社会计算与信息检索研究中心

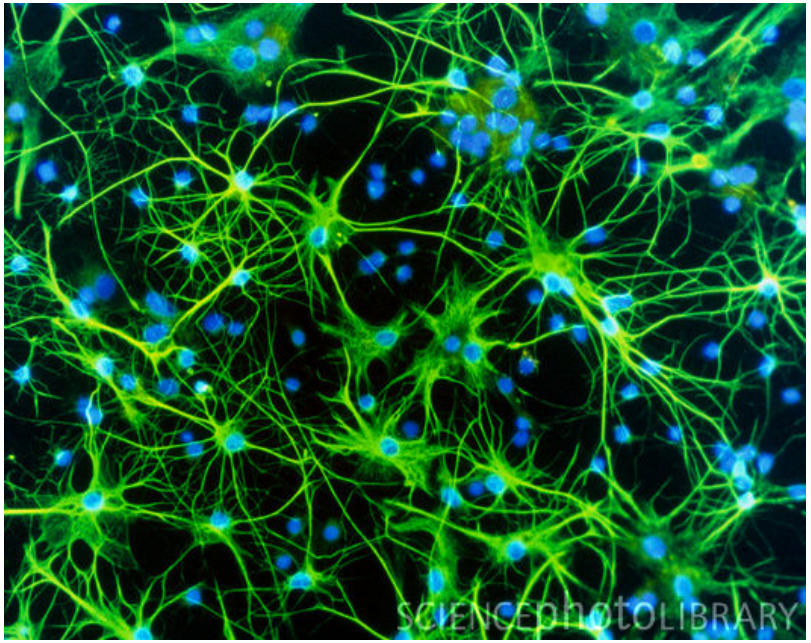
2

基于翻译模型的知识库表示学习方法

2.基于翻译模型的知识库表示学习方法

表示学习的概念及其理论基础

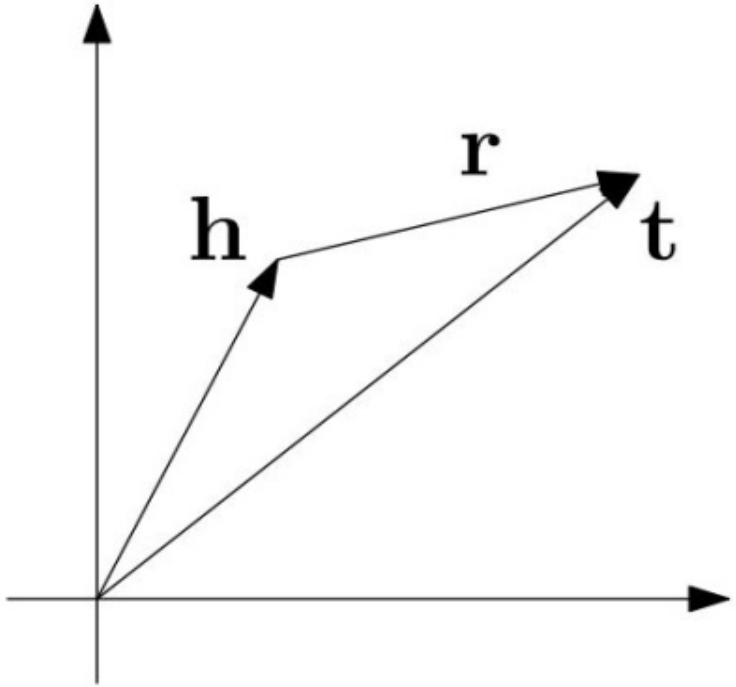
通过使用机器学习的方法将研究对象的语义信息表示为低维稠密的实值向量（分布式表示）



⟺ [0.23,0.43,…,0.12]

2.基于翻译模型的知识库表示学习方法

TransE模型简介



2.基于翻译模型的知识库表示学习方法

TransE模型简介

TransE 优化的目标是对于满足关系的(h,r,t)，有:

$$h+r\approx t$$

模型的代价函数为:

$$\mathcal{L}=\sum_{(h,r,t)\in S}\sum_{(h',r,t')\in S_{(h,r,t)}'}[\gamma+d(h+r,t)-d(h'+r,t')]_{+}$$

其中$[x]_{+}$ 代表x的正数部分，$\gamma>0$是一个边界值，另外，

$$S_{(h,r,t)}'=\{(h',r,t)|h'\in E\}\cup(h,r,t')t'\in E$$

其中E代表实体集合。

2.基于翻译模型的知识库表示学习方法

TransE模型的改进

改进后的TransE模型命名为TransE_ipv，TransE_ipv的训练过程中的代价函数为：

$$\mathcal{L}=\sum_{(h,r,t)\in S}\sum_{(h',r',t')\in S'_{(h,r,t)}}[\gamma+d(h+r,t)-d(h'+r',t')]_+$$

其中$[x]_+$代表x的正数部分，$\gamma>0$是一个边界值，另外，

$$S'_{(h,r,t)}=\{(h',r,t)|h'\in E\}\cup\{(h,r,t')|t'\in E\}\cup\{(h,r',t)|r'\in R\}$$

其中E代表实体的集合，R代表关系指示词的集合。

哈爾濱工業大學
社会计算与信息检索研究中心

3

实验结果与分析

3. 实验结果与分析

实验数据的获取

从互联网中抽取开放域实体关系三元组作为实验数据

“infobox”：包含属性-值对结构化文档

3. 实验结果与分析

实验数据的获取

中文名	哈尔滨工业大学计算机科学与技术学院	现任校长	周玉院士
英文名	哈工大计算机学院	知名校友	陈光熙 王天然 怀进鹏
创办时间	2000年	校　训	规格严格，功夫到家
类　别	国家示范性软件学院	专职院士	方滨兴 人
学校类型	工科	主要院系	计算机科学与技术系 信息安全 生物信息
属　性	211工程 985工程 C9	主要奖项	国家科技进步一等奖（2002年）

3. 实验结果与分析

实验数据的获取

锚文本 实体词

哈尔滨工业大学（*Harbin Institute of Technology*），简称"哈工大（HIT）"，坐落于中国北方冰城哈尔滨市，中华人民共和国工业和信息化部直属重点大学，首批"211工程"、"985工程"重点建设院校，"九校联盟(C9)"、"中俄工科大学联盟"、"中国-西班牙大学联盟"主要成员，国家首批"111计划"、"2011计划"、"千人计划"、"卓越计划"入选高校，中管副部级建制，由工业和信息化部、教育部、黑龙江省人民政府三方重点共建。

哈尔滨工业大学溯源于1920年创办的哈尔滨中俄工业学校，建校初衷为培养铁路工程技术人才；而后历经"中俄工业大学校"、"哈尔滨工业大学校"、"哈尔滨高等工业学校"等多个阶段，学校在1938年1月正式定名为哈尔滨工业大学，沿用至今。[1]

截止2015年7月，哈工大已有材料科学、工程学、物理学、化学、计算机科学、环境与生态学、数学、生物学与生物化学等8个学科进入ESI全球前1%的研究机构行列，其中材料科学、工程学已进入全球前1‰的研究机构行列。该校拥有哈尔滨本部及哈尔滨工业大学（威海）、哈尔滨工业大学深圳研究生院三个校区，共有全日制学生31903人，其中本科生16718人、研究生13263人（含硕士生7585人、博士生5678人），留学生1922人。

3. 实验结果与分析

实验数据的获取

据此方法，共从百度百科的“infobox”中共获取2,438,145条开放域实体关系三元组

-	Small数据集	All数据集
实体数量	333,007	1,551,231
关系指示词数量	21,649	57,235
关系三元组数量	524,676	2,438,145
训练数据（目标知识库）	519,676	2,428,145
测试数据	5,000	10,000

3. 实验结果与分析

基于知识库表示学习的关系指示词推理

如知识库中有以下两个实体关系三元组：

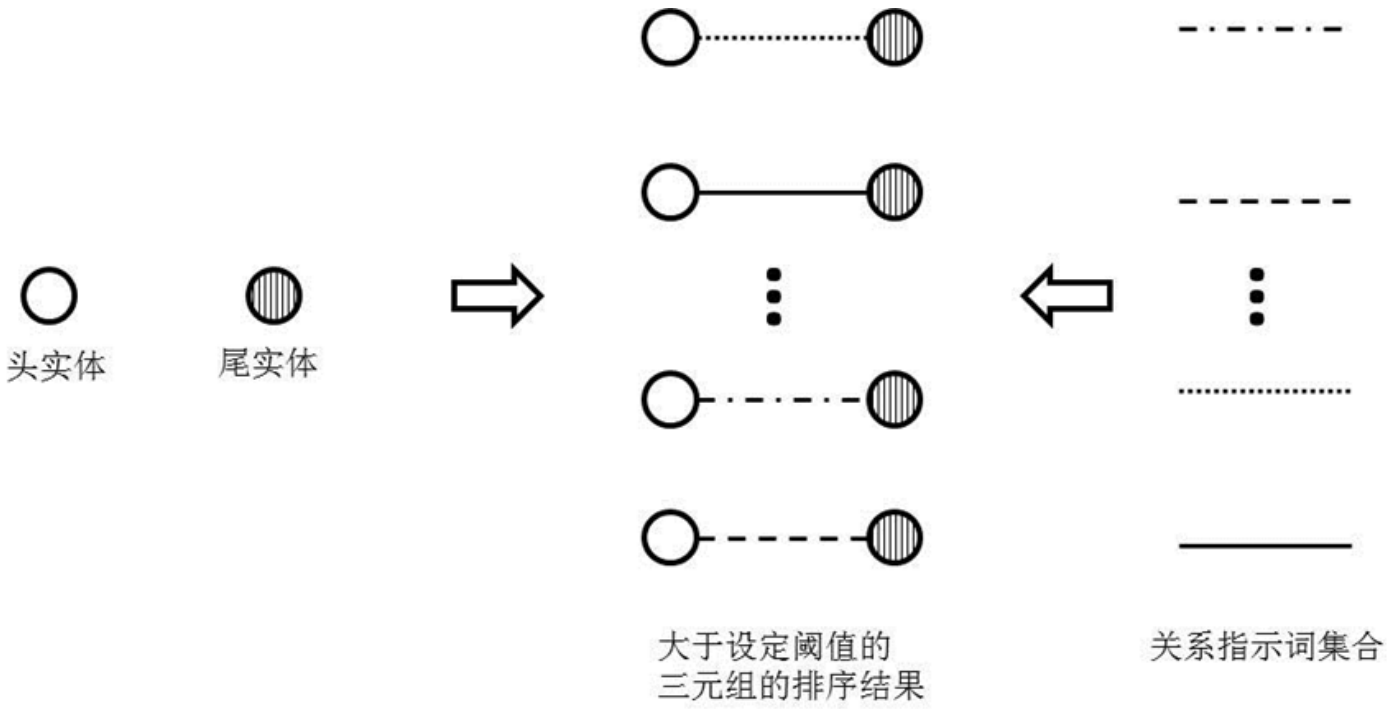
（泰坦尼克号，主要角色，杰克），（莱昂纳多，饰演，杰克）

那么，我们希望推理出如下关系以补全到现有知识图谱中：

（泰坦尼克号，主演，莱昂纳多）

3. 实验结果与分析

基于知识库表示学习的关系指示词推理



3. 实验结果与分析

基于知识库表示学习的关系指示词推理

关系指示词推理测试的实验结果

data set	@hit_10	recall_hit_10	F1_hit_10	@hit_1	recall_hit_1	F1_hit_1
small(TransE)	48.05%	9.34%	15.64%	39.81%	7.74%	12.96%
samll(TransE_ipv)	92.41%	36.06%	51.88%	83.03%	32.40%	46.61%
all(TransE_ipv)	90.65%	48.27%	63.00%	72.29%	41.16%	52.45%

实体分布式表示可以通过计算高效地推理出实体对中潜在的关系

3. 实验结果与分析

基于知识库表示学习的尾实体推理

头实体
关系指示词

尾实体

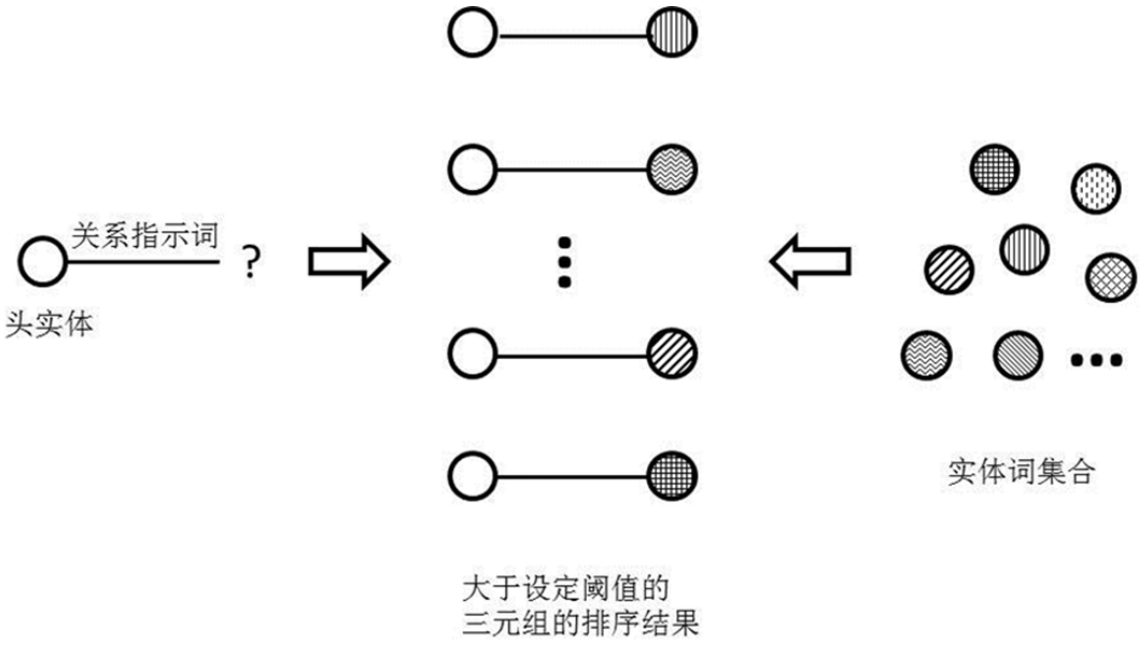
周杰伦
妻子

？

3. 实验结果与分析

基于知识库表示学习的尾实体推理

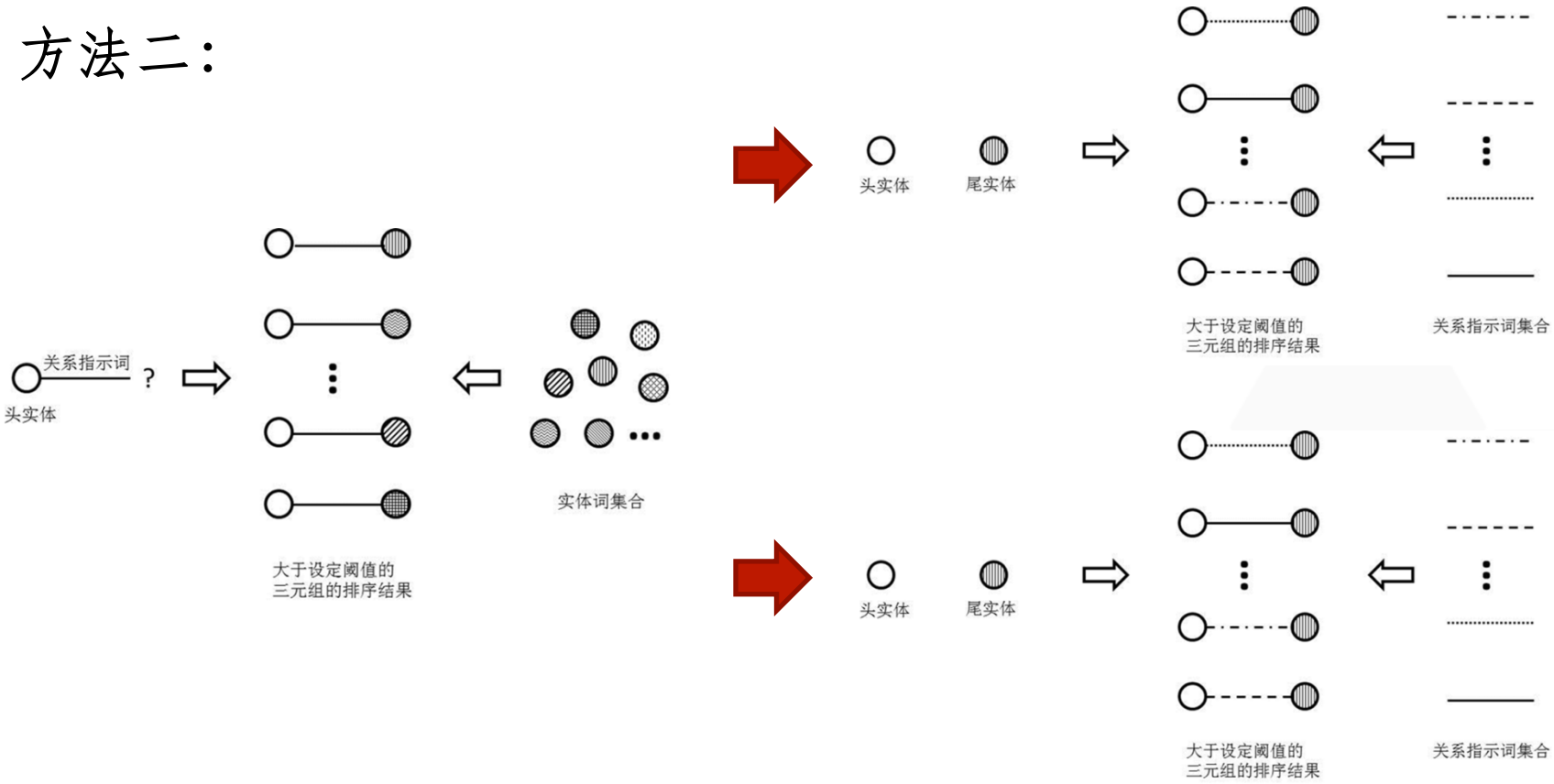
方法一：



3. 实验结果与分析

基于知识库表示学习的尾实体推理

方法二：



3. 实验结果与分析

基于知识库表示学习的尾实体推理

利用方法一、二做尾实体推理测试的实验结果（small数据集）

model	method	threshold	@hit_10	@hit_1	recall_hit_1
TransE_ipv	first	1.0	30.45%	15.46%	14.44%
TransE_ipv	second	1.0	38.49%	20.83%	15.22%

3. 实验结果与分析

基于知识库表示学习的尾实体推理

尾实体推理测试的实验结果（TransE_ipv）

data set	@hit_10	recall_hit_10	F1_hit_10	@hit_1	recall_hit_1	F1_hit_1
samll	38.49%	28.12%	32.50%	20.83%	15.22%	17.59%
all	26.69%	21.40%	23.75%	11.15%	8.94%	9.92%

通过观察发现尾实体推理的准确率远不如关系推理

3. 实验结果与分析

基于知识库表示学习的尾实体推理

尾实体推理测试的实验结果（TransE_ipv）

data set	@hit_10	recall_hit_10	F1_hit_10	@hit_1	recall_hit_1	F1_hit_1
samll	38.49%	28.12%	32.50%	20.83%	15.22%	17.59%
all	26.69%	21.40%	23.75%	11.15%	8.94%	9.92%

实体具有长尾分布的特点，这些长尾部分的实体和其他实体有极少的关系联系在一起。

3. 实验结果与分析

基于知识库表示学习的尾实体推理

在每次迭代中对每个训练三元组构造50个腐败三元组进行训练

model	@hit_10	recall_hit_10	F1_hit_10	@hit_1	recall_hit_1	F1_hit_1
TransE_ipv	38.49%	28.12%	32.50%	20.83%	15.22%	17.59%
TransE_ipvn	41.47%	27.34%	32.95%	27.76	18.30%	22.06%

尝试增加腐败三元组的数量对尾实体推理有较好的影响。

4 结论

4. 结论

- 当知识库的知识规模不断扩大，基于网状结构知识库的推理很难较好地满足实时计算的需求。
- 基于知识库表示学习的关系指示词推理准确率可以达到80%以上，且无需设计复杂的算法。
- 知识分布表示的尾实体推理测试中，使用增加训练过程中三元组负例的方法可以将准确率提升7个百分点，无需设计复杂算法即可实现对尾实体的推理。

参考文献

[1]. Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.

[2].Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph da-tabase for structuring human knowledge[C]

[3].Miller E. An introduction to the resource description framework[J]. Bulletin of the American Society for Information Science and Technology, 1998, 25(1): 15-19.

[4].刘知远, 孙茂松, 林衍凯, 等. 知识表示学习研究进展[J]. 计算机研究与发展, 53(2): 247-261.

参考文献

[5].Bengio Y, Courville A, Vincent P. Representation learning: A review and new per-spectives[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2013, 35(8): 1798-1828.

[6].Bordes A, Weston J, Collobert R, et al. Learning structured embeddings of knowledge bases[C]

[7].Bordes A, Glorot X, Weston J, et al. A semantic matching energy function for learning with multi-relational data[J]. Machine Learning, 2014, 94(2): 233-259.

[8].Bordes A, Glorot X, Weston J, et al. Joint learning of words and meaning representa-tions for open-text semantic parsing[C]

[9].Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion[C]

参考文献

[10].Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C]

[11].Wang Z, Zhang J, Feng J, et al. Knowledge Graph Embedding by Translating on Hy-perplanes[C]//AAAI. 2014: 1112-1119.

[12].Lin Y, Liu Z, Sun M, et al. Learning Entity and Relation Embeddings for Knowledge Graph Completion[C]//AAAI. 2015: 2181-2187.

[13].Ji G, He S, Xu L, et al. Knowledge Graph Embedding via Dynamic Mapping Ma-trix[C]//Proceedings of ACL. 2015: 687-696.

[14].Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning[C]

谢谢各位聆听!