



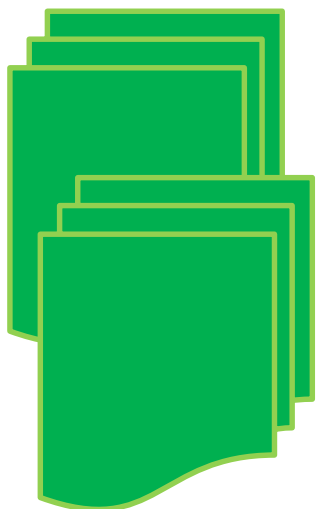
哈尔滨工业大学
社会计算与信息检索研究中心

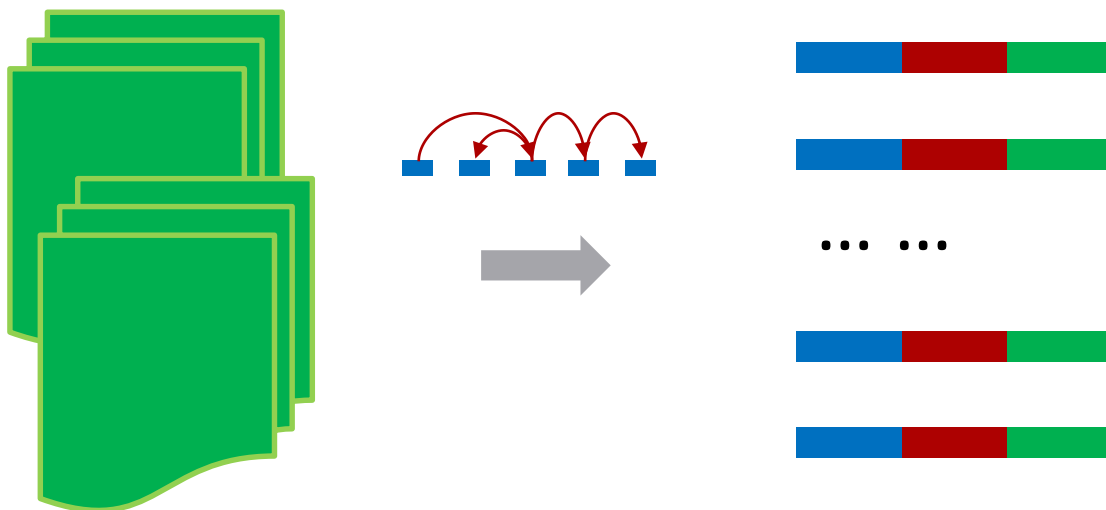


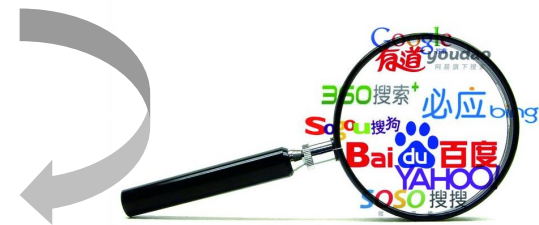
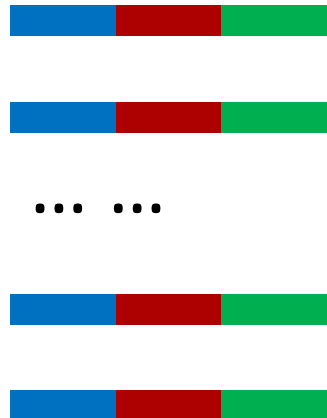
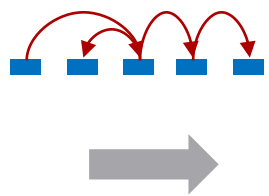
融入搜索引擎的启发式开放域三元组抽取

CCKS2017-全国知识图谱与语义计算大会

刘勇杰，姜天文，秦兵，刘铭，刘挺









哈尔滨工业大学
社会计算与信息检索研究中心

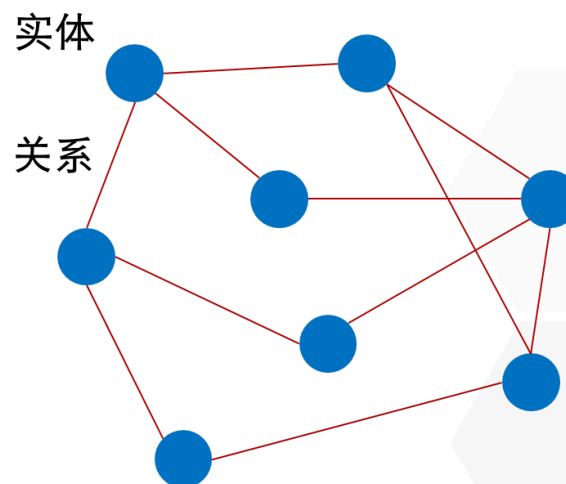
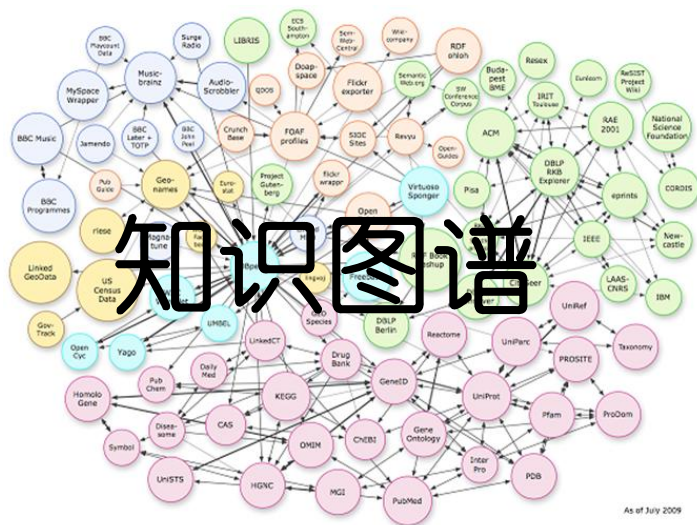


1

研究背景与目的

1.研究背景与目的

作为知识图谱构建与补全的关键步骤，实体关系的抽取成为知识图谱相关研究工作的重点。





1.研究背景与目的

早期的实体关系抽取任务（from ACE），研究工作可以大致分为四类：

基于模式匹配、基于特征、基于核函数，以及基于神经网络的方法。

1.研究背景与目的

早期的实体关系抽取任务（from ACE），研究工作可以大致分为四类：

基于模式匹配、基于特征、基于核函数，以及基于神经网络的方法。

有极高的准确率，但这种方法需要领域专家设计专门的模板，领域性较强，不能很好的移植到其他领域，局限性较为明显。另外，由于难以建立完整而准确的模式集合，基于模式识别的方法很难取得理想的召回率。

1.研究背景与目的

早期的实体关系抽取任务（from ACE），研究工作可以大致分为四类：

基于模式匹配、基于特征、基于核函数，以及基于神经网络的方法。

基于特征的方法是将关系实例通过一定粒度的词法分析和句法分析转换为平面特征向量，然后采用机器学习模型比较特征向量之间的相似性并分类。相对于建立模式集合，特征的提取简单、有效，不需要具有专业知识的专家进行大量人工操作。

1.研究背景与目的

早期的实体关系抽取任务（from ACE），研究工作可以大致分为四类：

基于模式匹配、基于特征、基于核函数，以及基于神经网络的方法。

基于核函数的方法是通过构造核函数,隐式地计算特征向量内积,从而得到关系实例之间的相似性。

1.研究背景与目的

早期的实体关系抽取任务（from ACE），研究工作可以大致分为四类：

基于模式匹配、基于特征、基于核函数，以及基于神经网络的方法。

对词语或句子进行嵌入式表示学习，利用神经网络自动获取其不同深度的特征向量，进而进行关系分类的学习。

1.研究背景与目的

伴随着互联网技术的发展，其数据类型更加多样化，关系抽取受到语义单元类型的限定以及关系类型的限制，难以与当下网络数据快速，多样化的增长趋势相适应。

1.研究背景与目的

伴随着互联网技术的发展，其数据类型更加多样化，关系抽取受到语义单元类型的限定以及关系类型的限制，难以与当下网络数据快速，多样化的增长趋势相适应。



传统关系抽取

1.研究背景与目的

伴随着互联网技术的发展，其数据类型更加多样化，关系抽取受到语义单元类型的限定以及关系类型的限制，难以与当下网络数据快速，多样化的增长趋势相适应。

— ? —

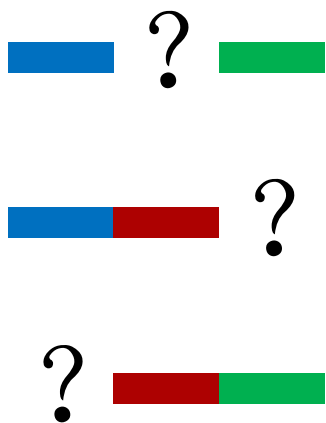
— ?

? —

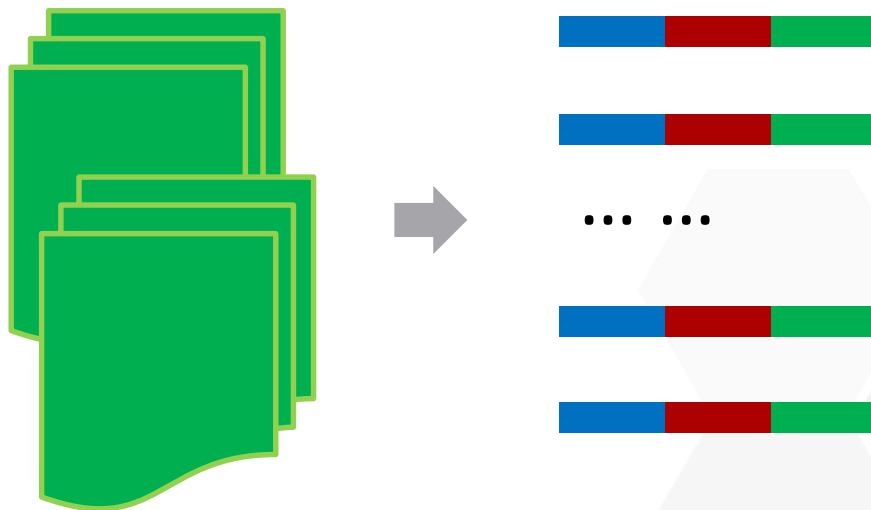
传统关系抽取

1.研究背景与目的

伴随着互联网技术的发展，其数据类型更加多样化，关系抽取受到语义单元类型的限定以及关系类型的限制，难以与当下网络数据快速，多样化的增长趋势相适应。



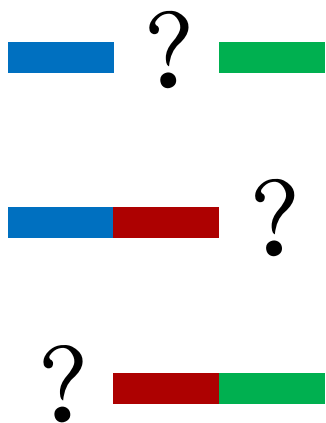
传统关系抽取



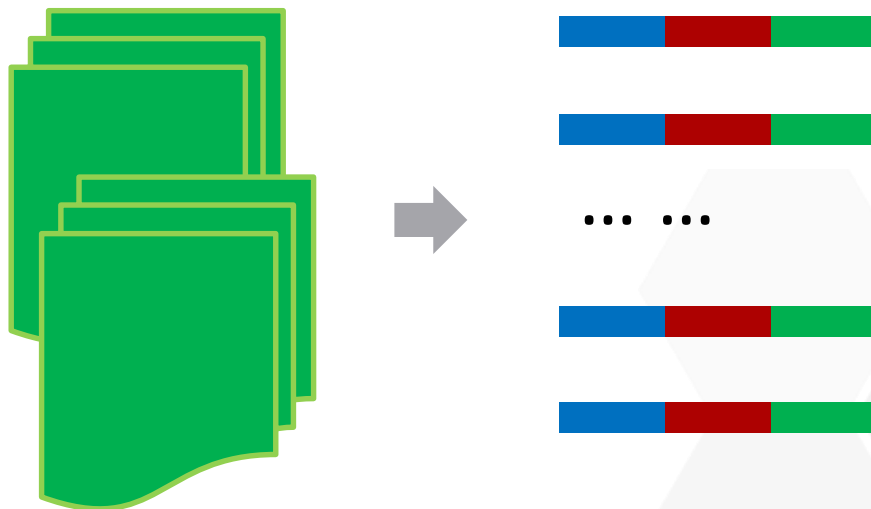
开放域三元组抽取

1.研究背景与目的

开放域三元组抽取系统更加注重文本实体对之间的语义表达,而不再强调类别关系。



传统关系抽取



开放域三元组抽取



1.研究背景与目的



1.研究背景与目的

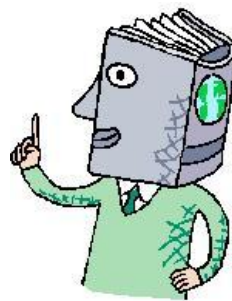


TextRunner

1.研究背景与目的



TextRunner



ReVerb

Ollie

1.研究背景与目的



TextRunner

PATTY Relation Mining

MPI-INF|Databases

Thesaurus Relations Taxonomy

▼ DBpedia Relations

- academicAdvisor
- affiliation
- album
- almaMater
- anthem
- appointer
- architect
- artist
- assembly
- associate
- associatedBand
- associatedMusicalArtist
- author

Relation: dbpedia:author

1-26 of 26

Pattern

- [[adj]] book by;
- [[det]] novel by;
- was written to;
- [[num]] book written by;
- [[det]] collection of [[num]] my
- [[adj]] lyrics by;



ReVerb

Ollie

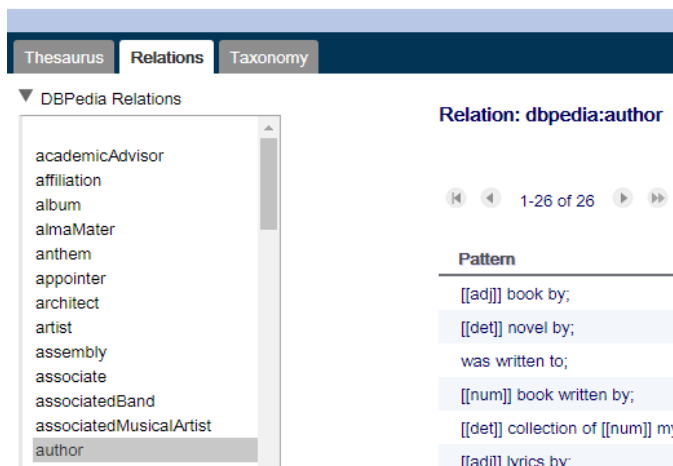
1.研究背景与目的



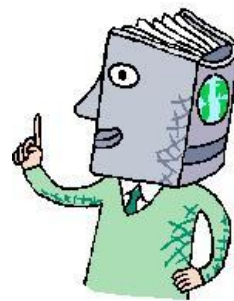
TextRunner

PATTY Relation Mining

MPI-INF|Databases

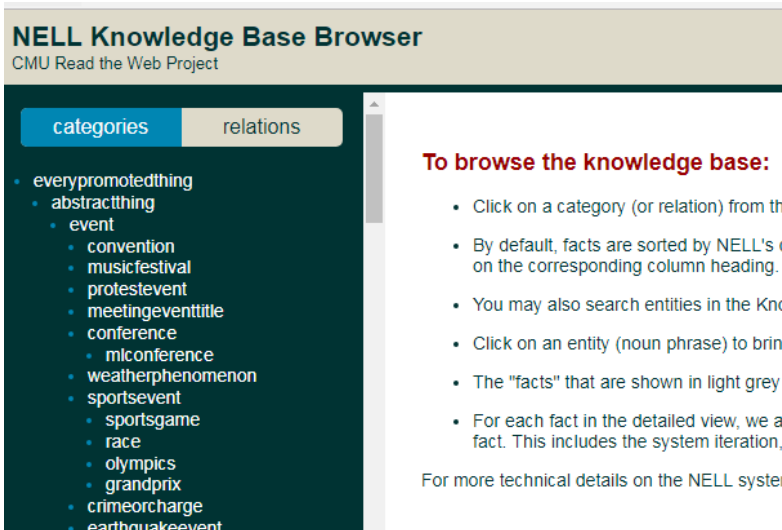


The interface shows a tabbed view with 'Relations' selected. A list of DBpedia Relations is displayed, including 'academicAdvisor', 'affiliation', 'album', 'almaMater', 'anthem', 'appointer', 'architect', 'artist', 'assembly', 'associate', 'associatedBand', 'associatedMusicalArtist', and 'author'. The 'author' relation is selected, showing a list of patterns such as '[[ad]] book by;', '[[det]] novel by;', 'was written to;', '[[num]] book written by;', '[[det]] collection of [[num]] my', and '[[ad]] lyrics by;'.



ReVerb

Ollie



The interface shows a sidebar with 'categories' and 'relations' tabs. The 'categories' tab is active, displaying a list of categories including 'everypromotedthing', 'abstractthing', 'event', 'convention', 'musicfestival', 'protestevent', 'meetingeventtitle', 'conference', 'mlconference', 'weatherphenomenon', 'sportsevent', 'sportsgame', 'race', 'olympics', 'grandprix', 'crimeorcharge', and 'earthquakeevent'.

To browse the knowledge base:

- Click on a category (or relation) from the sidebar.
- By default, facts are sorted by NELL's confidence on the corresponding column heading.
- You may also search entities in the Knowledge Base.
- Click on an entity (noun phrase) to bring up a detailed view.
- The "facts" that are shown in light grey.
- For each fact in the detailed view, we also show the system iteration, the confidence, and the source.

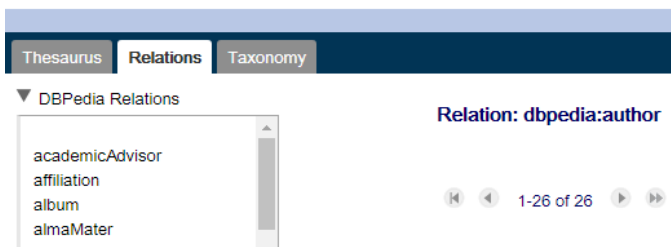
For more technical details on the NELL system, see the NELL project page.

1.研究背景与目的



PATTY Relation Mining

MPI-INF|Databases



Te

这些系统中有的需要自动构造训练语料，从标注语料中提取关系模版或训练分类器，之后再抽取三元组，有的则根据语法特征直接从句法分析结果中抽取三元组。

NELL Knowledge Base

CMU Read the Web Project



- Click on a category (or relation) from the list.
- By default, facts are sorted by NELL's confidence on the corresponding column heading.
- You may also search entities in the Knowledge Base.
- Click on an entity (noun phrase) to bring up a detailed view.
- The "facts" that are shown in light grey.
- For each fact in the detailed view, we allow you to click on the system iteration, to see the original source.

For more technical details on the NELL system



ReVerb

Ollie

1.研究背景与目的

本文从中文句子中词语之间的依存特点入手，提出了一种融入搜索引擎的启发式开放域三元组抽取方法。

中文句法依存特点

搜索引擎

启发式规则

无监督



哈尔滨工业大学
社会计算与信息检索研究中心



2

方法



2.方法





2.方法

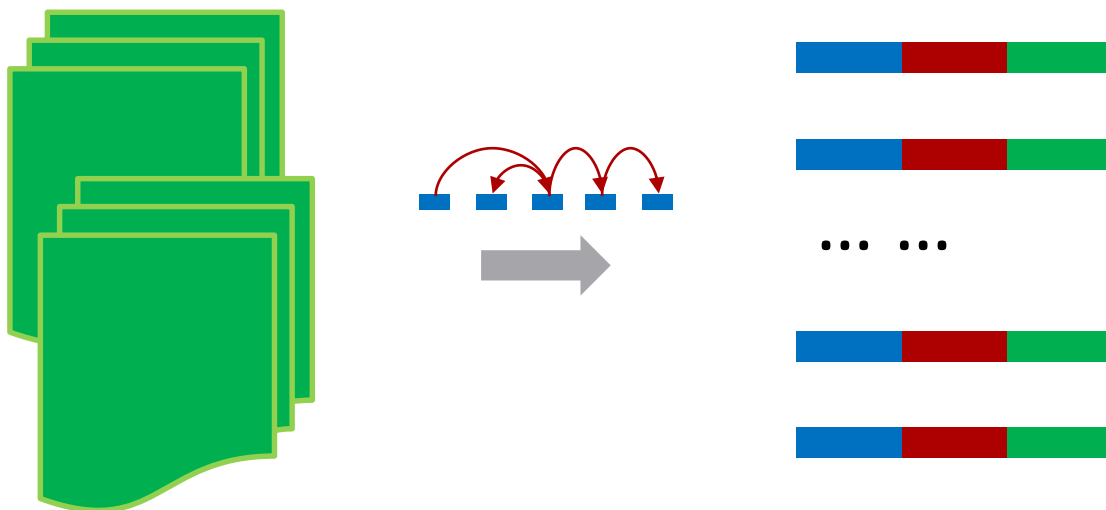
基于启发式规则模版的开放域候选三元组抽取

2.方法

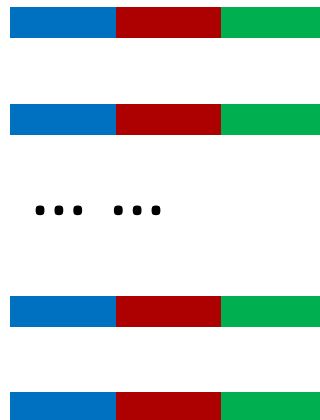
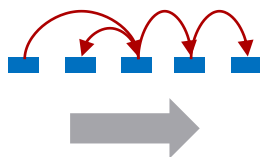
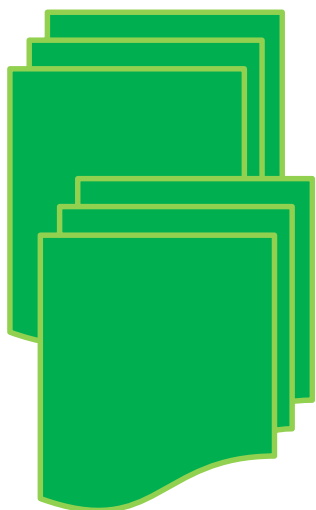


针对中文句子中词语之间的依存关系特点，对语句中实体关系表述的方式进行分析，获取了四类具有一定泛化能力的启发式规则。

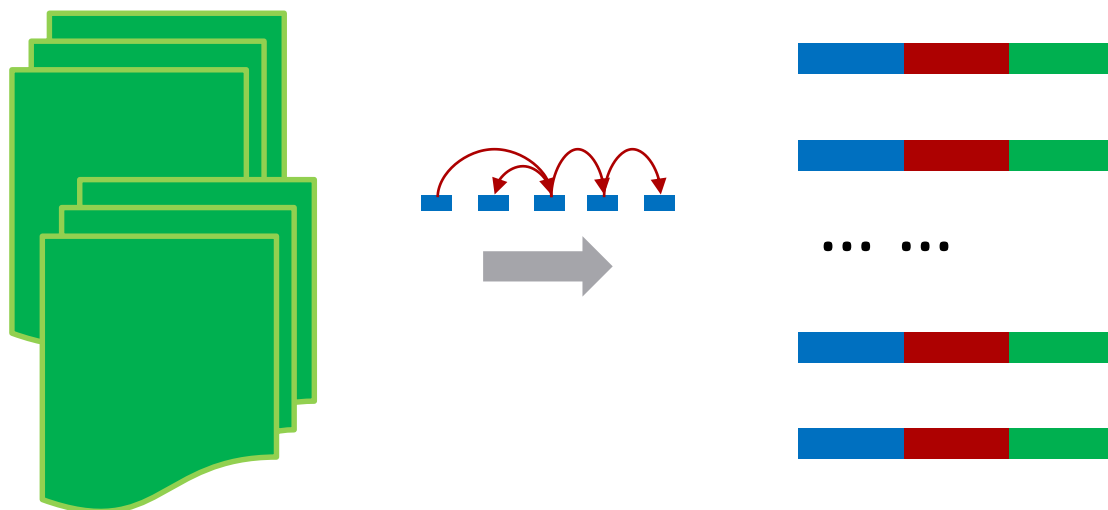
2.方法



2.方法

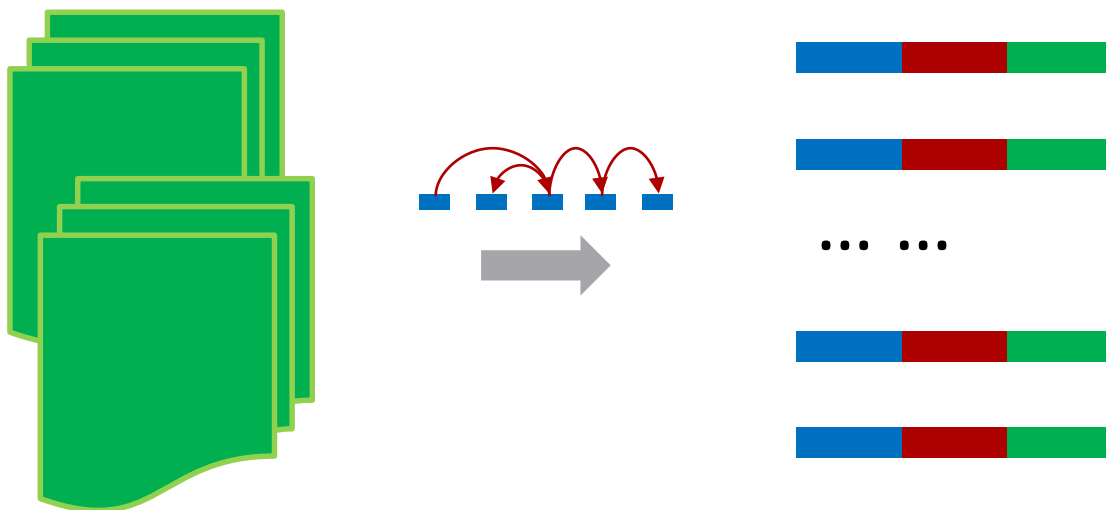


2.方法



基于搜索引擎的开放域三元组置信度计算

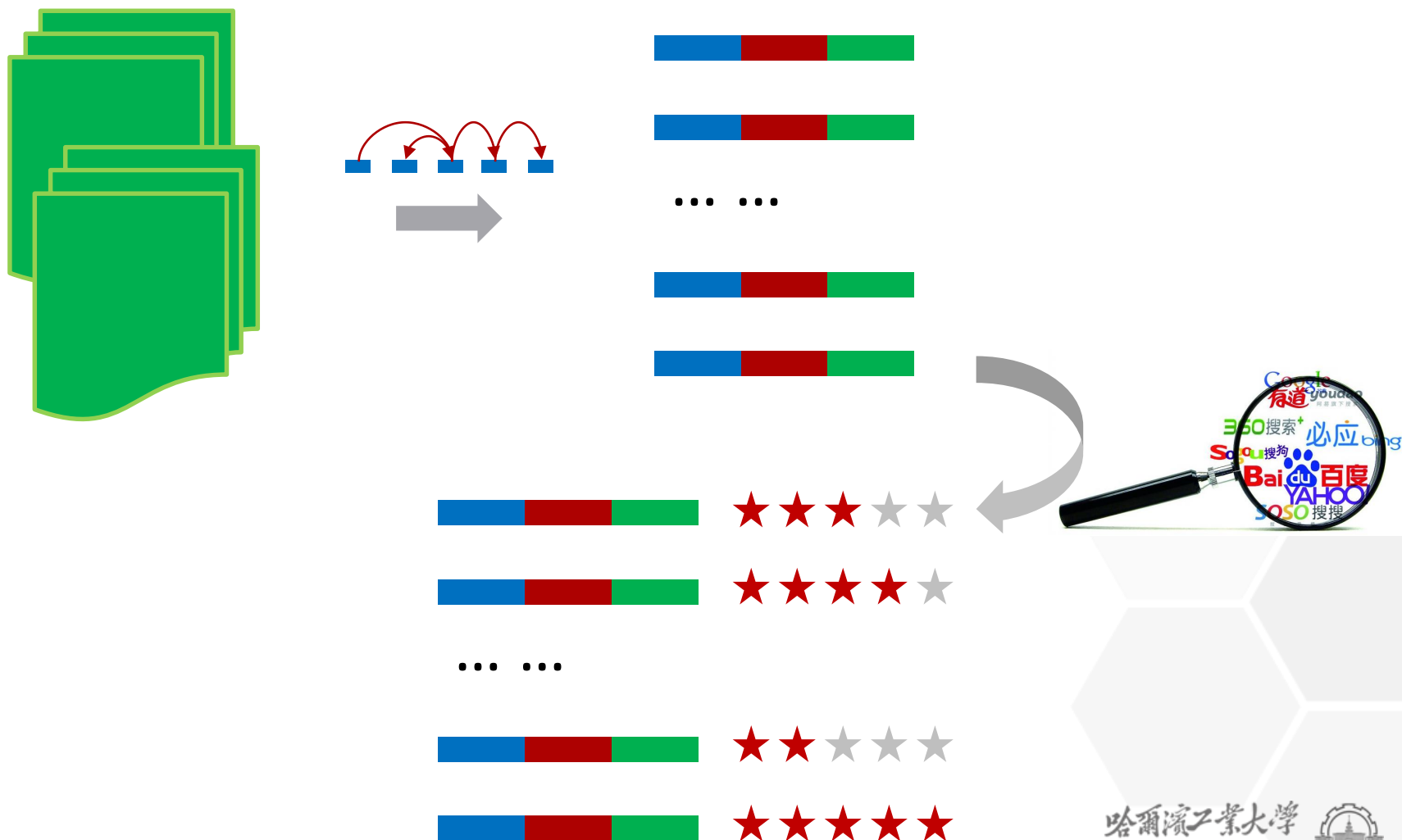
2.方法



将抽取到的候选关系三元组放入搜索引擎中，为每一个候选三元组获取置信度，基于置信度筛选掉错误的和不常见的关系三元组。

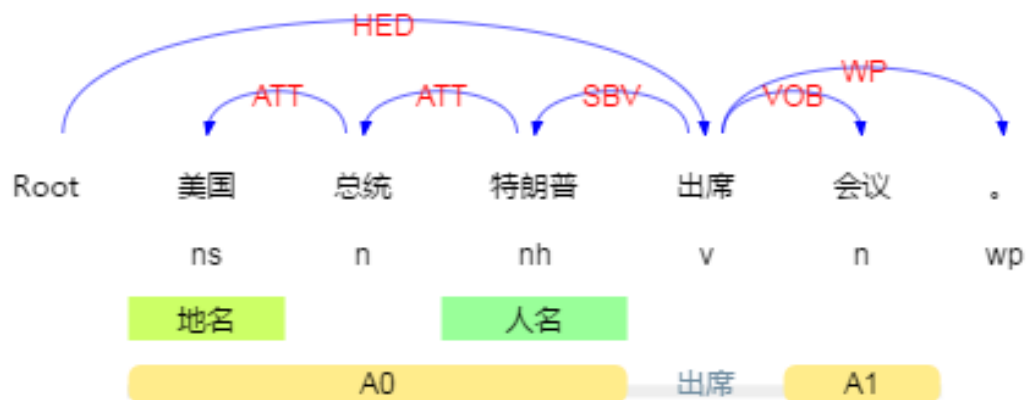


2.方法



2.1 基于启发式规则模版的开放域候选三元组抽取

启发式规则模板的获取 定中关系三元组



- 头实体与尾实体皆为命名实体；
- 中间指示词必须为名词；
- 命名实体与中间指示词均为ATT依存关系，且依次指向前者。

美国 总统 特朗普

2.1 基于启发式规则模版的开放域候选三元组抽取

启发式规则模板的获取 定语后置关系三元组



- 指示词为动词；
- 指示词同时拥有ATT关系（定中关系）、VOB关系（动宾关系）、RAD关系（右附加关系）；
- 指示词的ATT关系父节点和VOB关系子节点为命名实体。

北海公园 位于 北京

2.1 基于启发式规则模版的开放域候选三元组抽取

启发式规则模板的获取

主谓宾关系三元组



- 指示词为动词；
- 指示词的SBV和VOB依存关系指向两个命名实体。

青海

地处

青藏高原

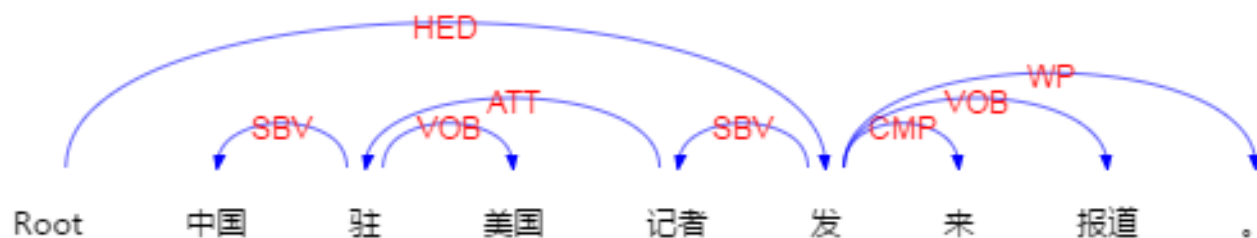
2.1 基于启发式规则模版的开放域候选三元组抽取

启发式规则模板的获取 主谓动补关系三元组

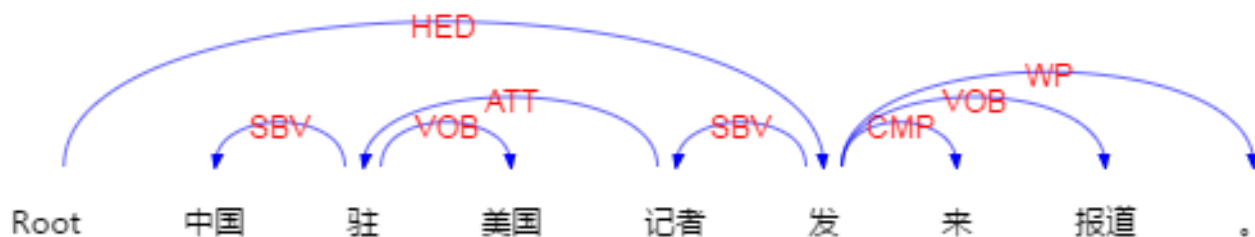


- 指示词为动词；
- 指示词拥有CMP（动补关系），并且补语还拥有POB（介宾关系）；
- 指示词的SBV关系指向命名实体，补语的POB关系子节点为命名实体。

2.2 基于搜索引擎的开放域三元组置信度计算

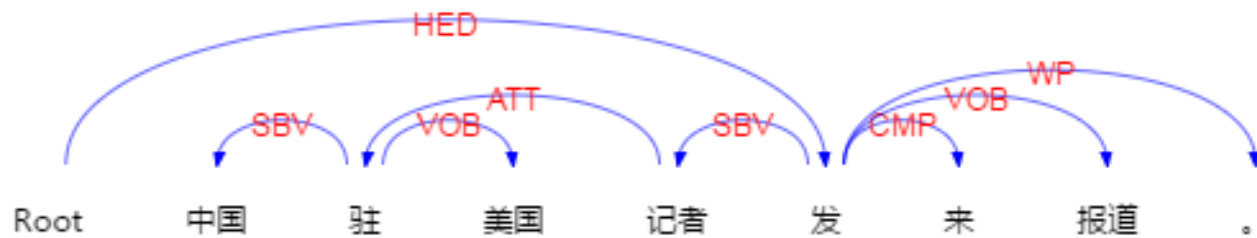


2.2 基于搜索引擎的开放域三元组置信度计算



中国 驻 美国

2.2 基于搜索引擎的开放域三元组置信度计算



中国 驻 美国

?



2.2基于搜索引擎的开放域三元组置信度计算

候选三元组正负例对比分析





2.2基于搜索引擎的开放域三元组置信度计算

候选三元组正负例对比分析

中国

首都

北京





2.2基于搜索引擎的开放域三元组置信度计算

候选三元组正负例对比分析

中国

首都

北京

城市

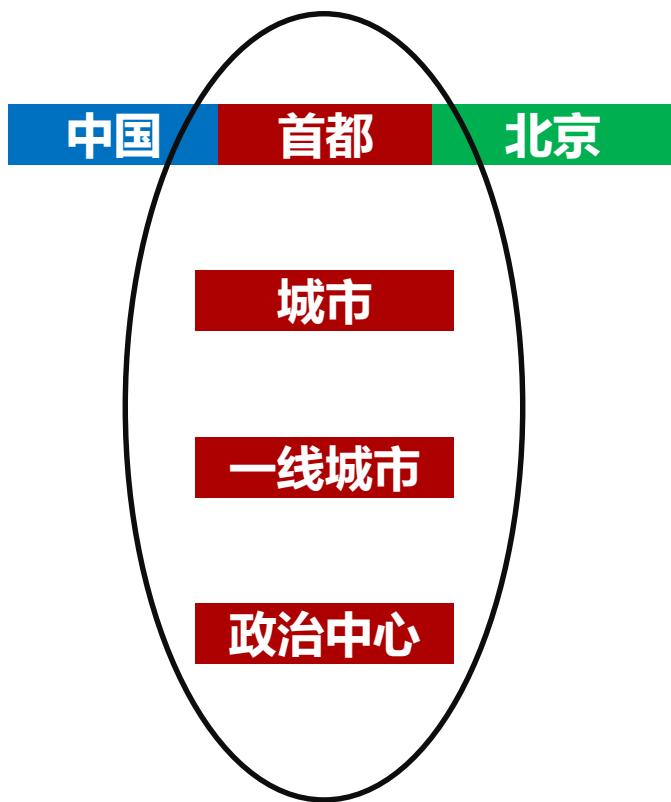
一线城市

政治中心



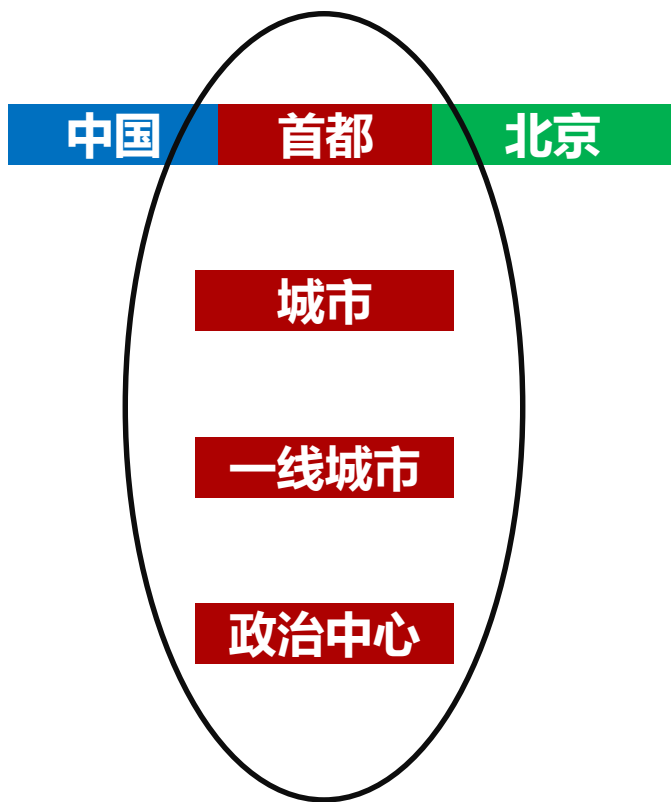
2.2 基于搜索引擎的开放域三元组置信度计算

候选三元组正负例对比分析



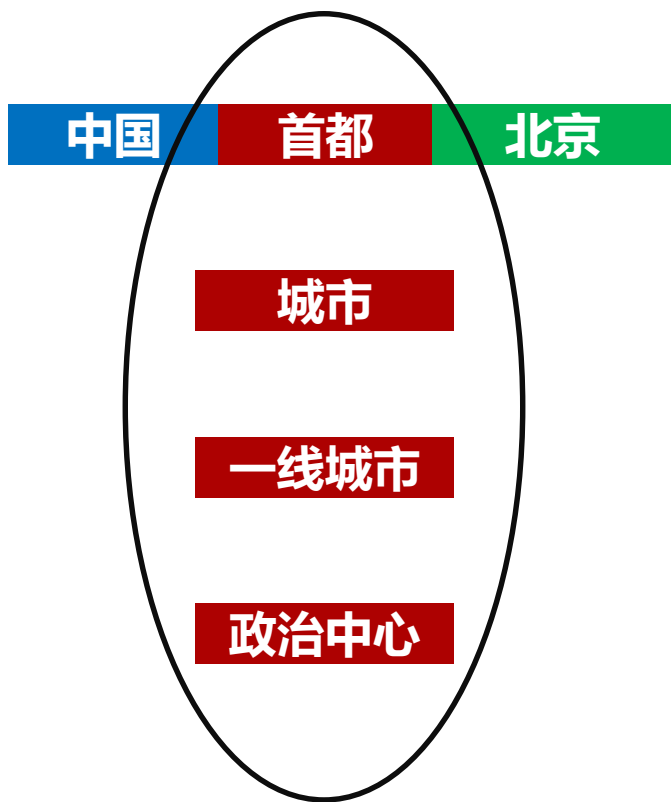
2.2 基于搜索引擎的开放域三元组置信度计算

候选三元组正负例对比分析



2.2基于搜索引擎的开放域三元组置信度计算

候选三元组正负例对比分析



中国 驻 美国

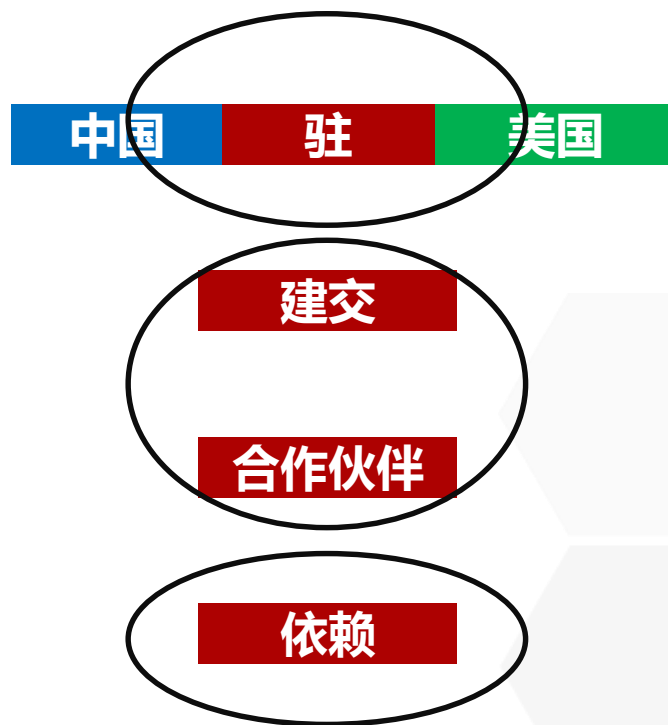
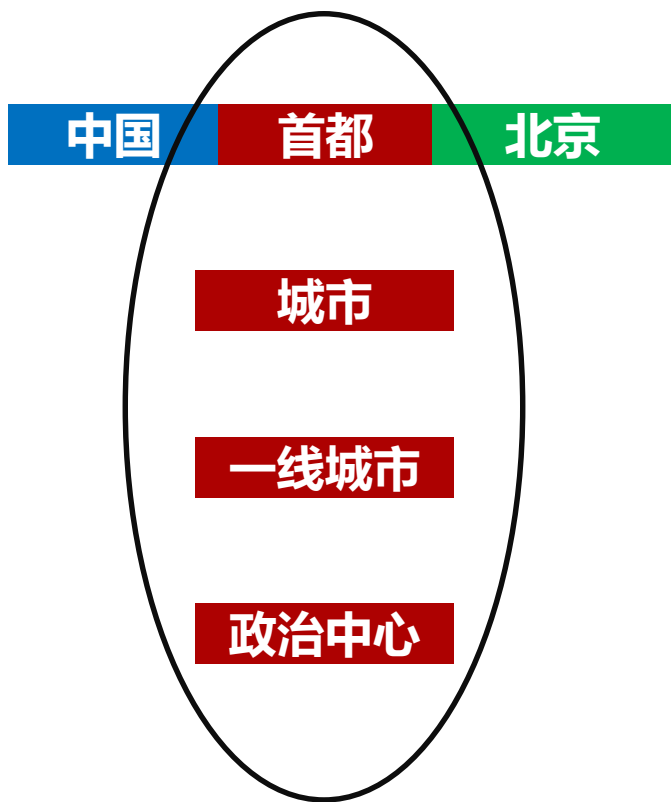
建交

合作伙伴

依赖

2.2 基于搜索引擎的开放域三元组置信度计算

候选三元组正负例对比分析





2.2 基于搜索引擎的开放域三元组置信度计算

候选三元组正负例对比分析

中国

首都

北京

中国

驻

美国

启发于此，可以把三元组关系指示词与其可能的关系指示词之间的语义相似度作为它的置信度，而一个三元组可能的关系指示词，我们可借助搜索引擎获取。





2.2基于搜索引擎的开放域三元组置信度计算

基于搜索引擎的候选关系指示词获取

2.2 基于搜索引擎的开放域三元组置信度计算

基于搜索引擎的候选关系指示词获取



2.2 基于搜索引擎的开放域三元组置信度计算

基于搜索引擎的候选关系指示词获取



中共中央

?

习近平

2.2基于搜索引擎的开放域三元组置信度计算

基于搜索引擎的候选关系指示词获取



中共中央

?

习近平



2.2基于搜索引擎的开放域三元组置信度计算

基于搜索引擎的候选关系指示词获取



总书记

同志

...

的

中共中央

?

习近平



2.2基于搜索引擎的开放域三元组置信度计算

基于搜索引擎的候选关系指示词获取



总书记

同志

...

的

中共中央

?

习近平





2.2基于搜索引擎的开放域三元组置信度计算

关系指示词与候选关系指示词的相似度计算





2.2 基于搜索引擎的开放域三元组置信度计算

关系指示词与候选关系指示词的相似度计算

总书记

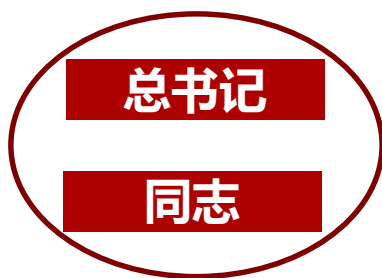
同志





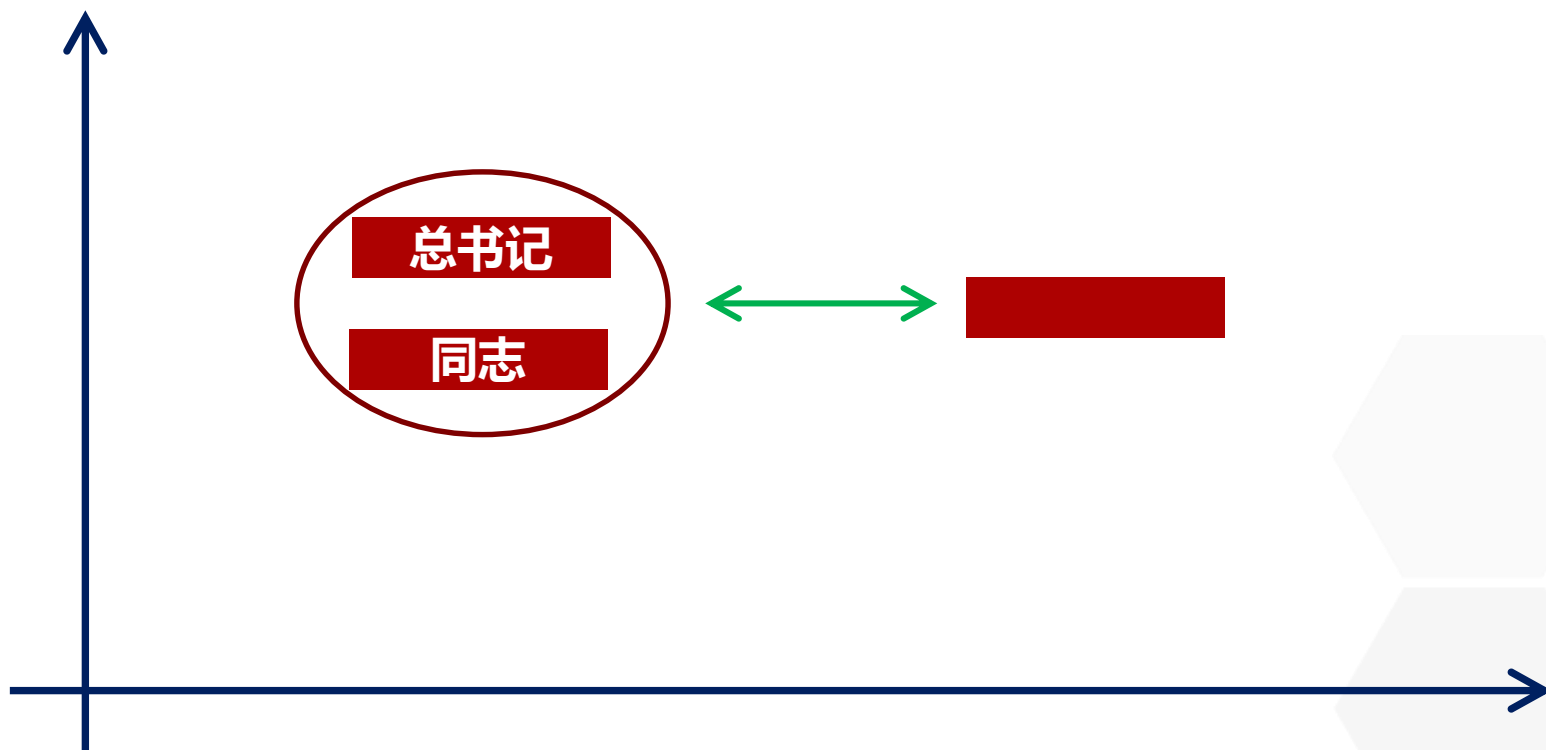
2.2 基于搜索引擎的开放域三元组置信度计算

关系指示词与候选关系指示词的相似度计算



2.2基于搜索引擎的开放域三元组置信度计算

关系指示词与候选关系指示词的相似度计算





哈尔滨工业大学
社会计算与信息检索研究中心



3

实验与结果分析



3. 实验与结果分析

数据集与评估方法

从各大门户网站共爬取18000条新闻报道，涉及军事、娱乐、体育等九个不同领域。



3. 实验与结果分析

数据集与评估方法

从各大门户网站共爬取18000条新闻报道，涉及军事、娱乐、体育等九个不同领域。

新闻领域	文章数量	句子数量	新闻领域	文章数量	句子数量
财经	2,000	202,611	教育	2,000	143,505
军事	2,000	144,505	科技	2,000	139,186
汽车	2,000	119,084	时尚	2,000	58,278
体育	2,000	123,489	娱乐	2,000	78,255
政治	2,000	105,350	——	——	——

3. 实验与结果分析

数据集与评估方法

从各大门户网站共爬取18000条新闻报道，涉及军事、娱乐、体育等九个不同领域。

新闻领域	文章数量	句子数量	新闻领域	文章数量	句子数量
财经	2,000	202,611	教育	2,000	143,505
军事	2,000	144,505	科技	2,000	139,186
汽车	2,000	119,084	时尚	2,000	58,278
体育	2,000	123,489	娱乐	2,000	78,255
政治	2,000	105,350	——	——	——

在评估阶段，本文使用准确率、召回率、以及F1值作为评估标准，以评估方法的性能。



3. 实验与结果分析

基于启发式规则模板抽取候选关系三元组



3. 实验与结果分析

基于启发式规则模板抽取候选关系三元组

新闻领域	候选三元组	定中关系	主谓宾关系	定语后置关系	主谓动补关系
财经	1,745	202,611	270	34	32
教育	888	144,505	90	19	12
军事	2,223	1,373	741	40	69
科技	918	779	113	13	13
汽车	532	468	45	12	7
时尚	376	264	94	11	7
体育	2,676	1,524	1,027	62	63
娱乐	748	519	201	13	15
政治	2,537	2,197	281	24	35



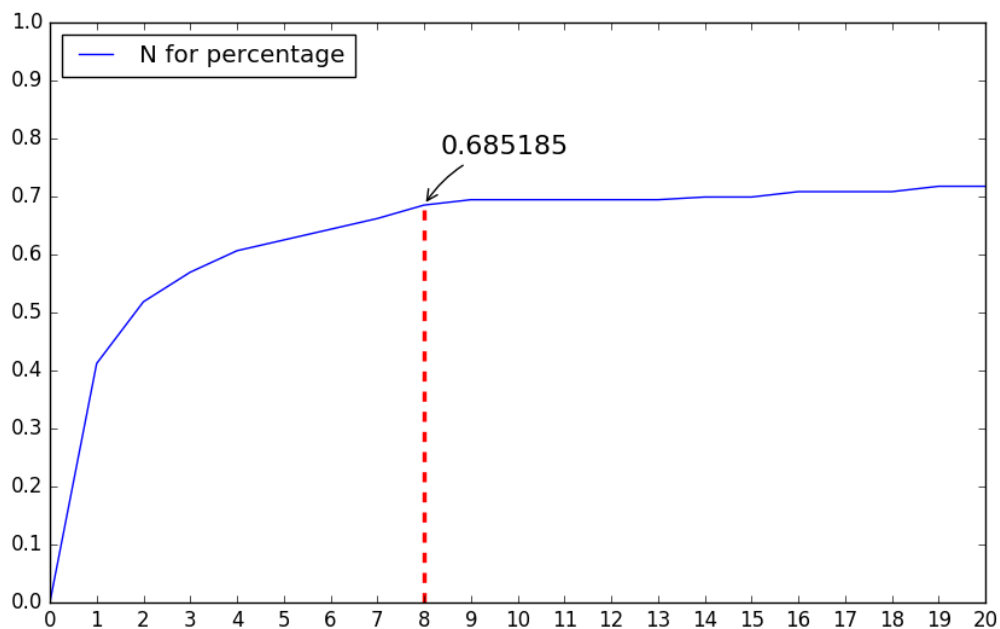
3. 实验与结果分析

基于启发式规则模板抽取候选关系三元组



3. 实验与结果分析

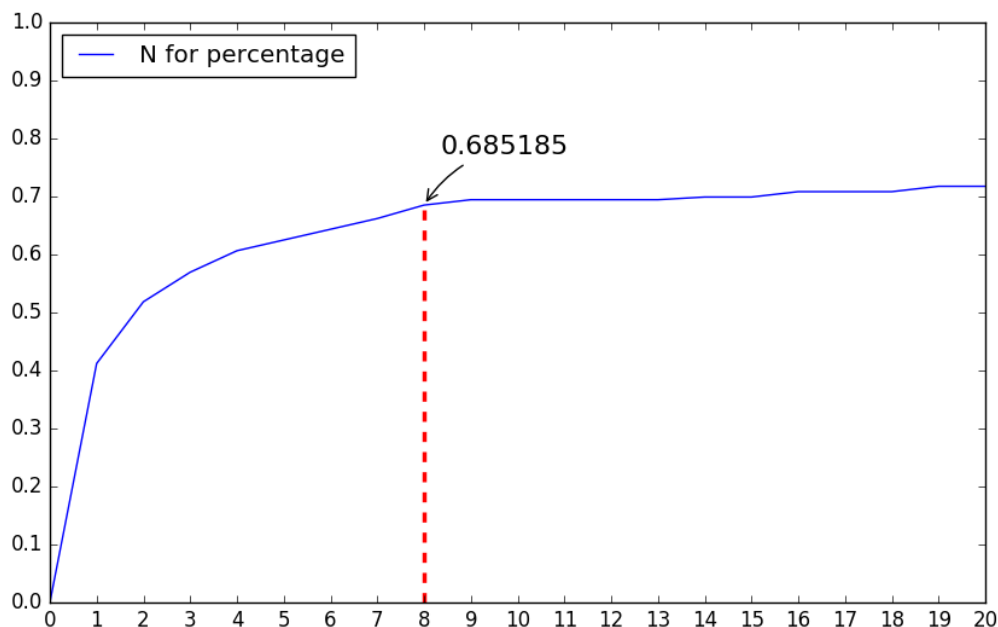
基于启发式规则模板抽取候选关系三元组



N与满足条件三元组所占比例关系图

3. 实验与结果分析

基于启发式规则模板抽取候选关系三元组



N与满足条件三元组所占比例关系图

在N取值小于8时，曲线上升非常迅速，而当N到达8之后，曲线渐渐趋于平缓。有大约70%的三元组在前8个相关词中能找到一个与关系指示词非常相近的词。

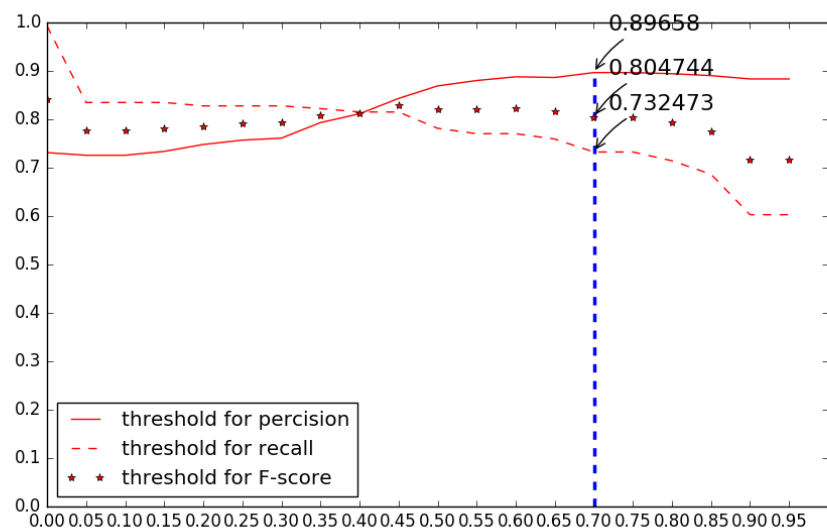


3. 实验与结果分析

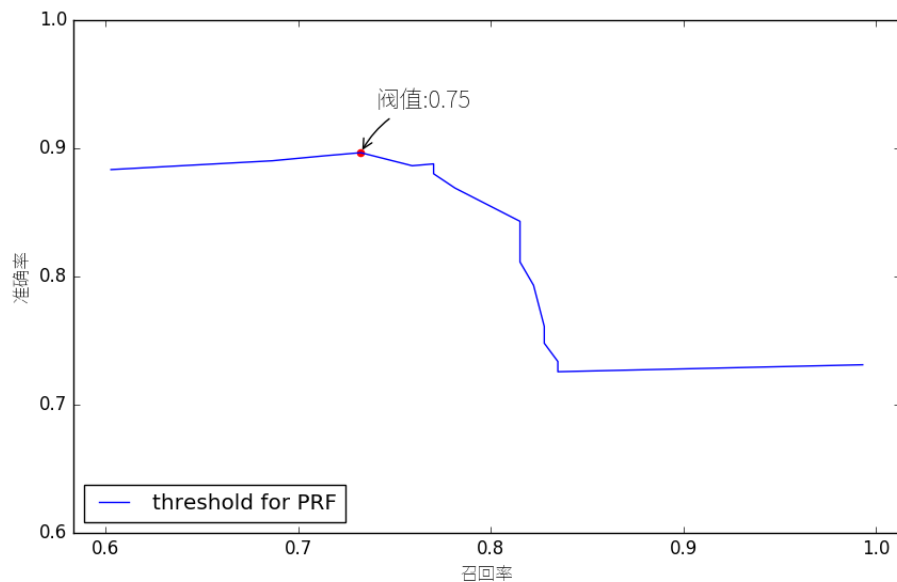
置信度阈值的确定与三元组准确率

3. 实验与结果分析

置信度阈值的确定与三元组准确率



阈值与PRF关系图



PR图

3. 实验与结果分析

方法	准确率	召回率	F1值	阈值
启发式规则	73.11	-	84.17	0
启发式规则+搜索引擎	89.65	73.24	80.47	0.7

3. 实验与结果分析

方法	准确率	召回率	F1值	阈值
启发式规则	73.11	-	84.17	0
启发式规则+搜索引擎	89.65	73.24	80.47	0.7

候选三元组 (基于启发式规则抽取)	候选关系指示词 (基于搜索引擎抽取)	置信度 (基于搜索引擎计算)	系统判断	人工标注
(伊拉克, 首都, 巴格达)	首都	100	True	True
(美国, 前财政部部长, 萨默德)	部长	99.99	True	True
(青海, 地处, 青藏高原)	位于	78.01	True	True
(蔡演威, 如愿到, 济南)	特批	29.92	False	False
(美国, 依赖, 俄国)	痛恨	36.44	False	False
(江苏, 油田, 高邮)	农产品	30.30	False	False



哈尔滨工业大学
社会计算与信息检索研究中心



4

结论

4. 结论

- 本文从中文句子中词语之间的依存特点入手，通过分析实体关系在语句中的表述方式，提出了四种启发式规则模板以辅助句法依存分析工具。
- 从新闻语料中抽取候选三元组，再利用搜索引擎计算三元组的置信度，进一步过滤三元组，获得最终的实体关系三元组。
- 实验结果显示，我们的方法简单而有效，获得了89%的准确率及高达到80%以上的F1值，满足实际要求，可以应用于知识图谱构建以及知识补全的任务上。



谢谢各位聆听！