- 词语相似度计算
  - 基于维基百科
  - 基于语义关系图
  - 基于相似度矩阵
- 层次关系和同义关系抽取
  - 基于词形规则模板
  - 基于词向量和规则
  - 基于词向量和相对余弦相似度
  - 基于词向量和谱聚类
  - 基于去除反义词
- 关键词抽取
  - 基于TF-IDF
  - 基于PC-Value
- 别名抽取
  - 基于BERT
  - 基于条件随机场
  - 基于Bi-LSTM+CRF
  - 基于翻译

目录:

- 中文维基百科的结构化信息抽取及词语相关度计算方法
- 基于词形规则模板的术语层次关系抽取方法
- 基于词语分布信息的TFIDF关键词抽取方法研究
- 一种基于语义关系图的词语语义相关度计算模型
- 专利技术术语的抽取方法
- 使用Bert搭建关系抽取模型进行别名抽取:
- 基于条件随机场的方志古籍别名自动抽取模型构建
- 旅游场景下的实体别名抽取联合模型
- Word Embedding Approach for Synonym Extraction of Multi-Word Term
- A Measure of Similarity between Graph Vertices: Applications to Synonym Extraction and Web Searching
- A Minimally Supervised Approach for Synonym Extraction with Word Embeddings
- Automatic Synonym Extraction Using Word2Vec and Spectral Clustering
- Automatic Relation Extraction - Can Synonym Extraction Benefit from Antonym Knowledge?
- Automatic translation of scholarly terms into patent terms using synonym extraction techniques

中文维基百科的结构化信息抽取及词语相关度计算方法

中文维基百科的结构化信息抽取及词语相关度计算方法

涂新辉[1,2]，张红春[1,2]，周琨峰[1,2]，何婷婷[1,2]

（1. 华中师范大学 计算机科学系，湖北 武汉 430079；
2. 国家语言资源监测与研究中心 网络媒体语言分中心，湖北 武汉 430079）

2012 中文信息学报

如何从海量的语言材料中自动地获取语义知识并利用这些语义知识来提高计算机的自然语言理解水平，已成为一个重要研究课题。维基百科官方只提供了一些基本数据文件，很多有用的结构化信息不能直接获取和利用。

文中计算语义相关度的方法分为两步：

1. 分别把词语w_A和w_B映射到维基百科主题页面P_A和P_B；
2. 计算维基百科主题P_A和P_B的语义相关度$SIM(P_A, P_B)$，将这个值作为w_A和w_B的语义相关度。

在英文维基百科中，可以利用这种思想来计算语义相关度：如果存在很多主题页面都指向那两个主题页面，那么这两个主题页面的相关度会较高。**然而中文维基百科存在一定的稀疏性，使用这种方法效果较差。**

文中提出一种利用主题页面间链接所对应的主题页面的类别特征来计算主题相关度的方法。两个维基百科主题页面P_a和P_b，构建向量时，有

$$w(P_m \to P_a) = |P_m \to P_a| \times \log\left\{\sum_{x=1}^{t} \frac{t}{|P_x \to P_a|}\right\} \quad (1)$$

这里的 $w(P_m \to P_a)$ 表示主题页面 P_m 到 P_a 的链接 $m \to a$ 的权重，$|P_m \to P_a|$ 表示主题页面 P_m 到 P_a 的链接的数量，$\log\left\{\sum_{x=1}^{t} \frac{t}{|P_x \to P_a|}\right\}$ 表示主题页面 P_a 的反链接概率，和逆文档频率 IDF 的作用是类似的，主要的作用是消除常用的维基百科主题存在的大量的入链接对相关度计算的影响。

$$V_{a\_in} = (w(P_1 \to P_a), w(P_2 \to P_a), \cdots, w(P_n \to P_a))$$
$$V_{b\_in} = (w(P_1 \to P_b), w(P_2 \to P_b), \cdots, w(P_n \to P_b))$$

对于维基百科主题页面 P_a 和 P_b，它们的所有入连接所对应的主题页面集合为 $\{P_i | i=1\cdots n\}$，这些链接所从属的类别集合为 $\{C_j | j=1\cdots m\}$，则对于主题页面 P_a 和类别 C_j，权重为：

$$w_C(P_a, C_j) = \sum_{i=1}^{n} w(P_i \to P_a) \times b(P_i, C_j) \quad (4)$$

$$V_{a\_in\_cat} = (w_C(P_a, C_1), w_C(P_a, C_2), \cdots, w_C(P_a, C_m))$$
$$V_{b\_in\_cat} = (w_C(P_b, C_1), w_C(P_b, C_2), \cdots, w_C(P_b, C_m))$$

一词多义的情况可以通过以下两个步骤消歧：

1. 找到词语对应的维基百科主题；
2. 从候选主题集合中找到可能性最大的主题。

数据集：中文维基百科；结果：

1. 基于入链接或出链接的方法在中文维基百科中效果较差，说明稀疏性问题对相关性计算影响较大；
2. 利用链接主题所属类别的方法得到的相关度更接近人工的相关度值。

优点：一定程度上缓解了数据稀疏性问题；

未来工作：结合维基百科中的文本内容提供更好的语义相关度计算。

基于词形规则模板的术语层次关系抽取方法

基于词形规则模板的术语层次关系抽取方法[1)]

韩红旗[1] 徐 硕[1] 桂 婕[1] 乔晓东[1] 朱礼军[1] 安小米[2]

（1. 中国科学技术信息研究所，北京 100038；
2. 数据工程与知识工程教育部重点实验室（中国人民大学），中国人民大学信息资源管理学院，北京 100872）

2013 情报学报

在张巍[1]的研究中，定义了一组通用的模板，如下表所示。

表 1 术语层次关系模板

模板编号	模板	模板类型	模板实例
T1	（C，A + C）	IS-A	（情报领域，图书情报领域）
T2	（A + C，B + C）	IS-A	（信息技术，数据挖掘技术） （人力资源开发，数字信息资源开发）
T3	（C + B，C）	PART-OF	（多媒体信息资源管理，多媒体信息资源）
T4	（C + A，C + B）	PART-OF	（信息资源管理，信息可视化） （信息资源管理教育，信息资源管理开发）
T5	（A + B + C，A + C）	IS-A	（信息资源管理技术，信息管理技术） （现代信息资源管理，现代企业信息资源管理）

模板中的T2规则不能确保一定满足IS-A关系，T4规则不能确保一定满足PART-OF关系。而且这些模板不能确定哪个术语是上位概念，哪个术语是下位概念。

文中提出泛化度指标和相关度指标来解决上述问题。泛化度指标用来测量两个术语在概念层次树上的相对位置，相关度指标用于测量两个术语在语义上的相关性。

文中把一个术语在一个领域中的专用性称为术语的泛化度，借鉴信息论中的信息量定义公式得到。采用点互信息来衡量两个概念间的关联性。

抽取流程如下图：

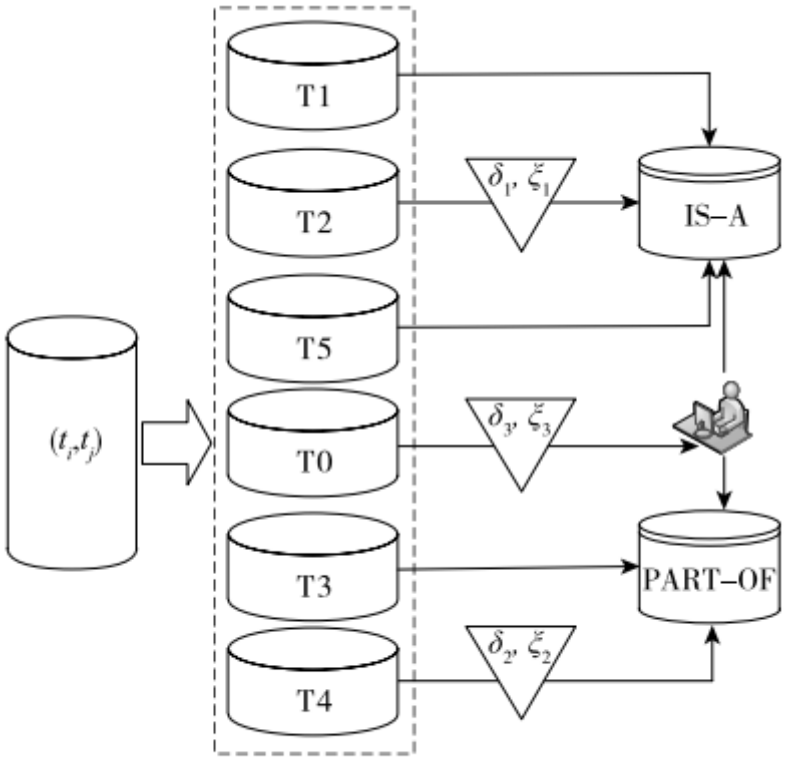

图 2　术语层次关系抽取原理示意图

其中，T0表示未匹配到模板的术语对，仅通过相关度和泛化度进行判断。T2和T4需要经过相关度和泛化度进一步判断。

数据集：万方数据中获取到的信息资源管理领域的科技论文数据。实验效果如下：

表 5　各模板下最终抽取的层次关系数量分布

模板	抽取数	IS-A	PART-OF	正确率(%)
T1	471	471	0	100.00
T2	48	38	0	79.17
T3	515	0	515	100.00
T4	145	0	127	87.59
T5	5	5	0	100.00
T0	122	3	49	42.98
总计	1306	517	691	92.50

优点：提高了T2和T4模板抽取的正确率。

缺点：在不符合模板匹配的术语对层次关系抽取上效果仍然较差。采用互信息的测量相关度可能会带来一些错误。

> 互信息在其他条件相等的情况下，由低频词组成的二元组的互信息要大于高频词组成的二元组。

[1]张巍，于洋，游宏梁．面向词汇知识库自动构建的概 念术语关系识别[J]．现代图书情报技术，2009(11):10-16. [2]张勇. 中文术语自动抽取相关方法研究[D]. 华中师范大学,2006

基于词语分布信息的TFIDF关键词抽取方法研究

结合词语分布信息的 TFIDF 关键词抽取方法研究

徐振强[1,2]，李保利[1,2]

(1.河南工业大学 信息科学与工程学院，郑州 450001;2.数字出版技术国家重点实验室，北京 100871)

2014 中原工学院学报

基于TD-IDF的抽取算法，只考虑了词语以及词语的常用程度等信息，却忽略了词语在文本中的分布信息，如词语的分布规律、词语出现的位置等。

文中结合词语的分布信息，进一步提高TF-IDF的性能。改进的TF-IDF算法将$RFPos$以及STD_{dist}引入TF-IDF值的计算中。结合词语分布信息的算法如下：

$$(TF*(1-STD_{\text{dist}})+RFPos)*IDF$$

$$RFPos=1-\frac{\text{在本文中首次出现位置}}{\text{文本的长度}}$$

$$STD_{\text{dist}}=\begin{cases}\alpha & (TF=1)\\ \sqrt{\dfrac{TF\times\sum\limits_{i=1}^{TF}(RD(i,i+1))^2-(\sum\limits_{i=1}^{TF}RD(i,i+1))^2}{TF^2}} & (TF\neq 1)\end{cases}$$

> TF-IDF的主要思想：如果某个特征项在文档中出现的频率高且不常用，则具有很好的文档代表能力。

数据集：Inspec、DUC2001、NUS。结果：

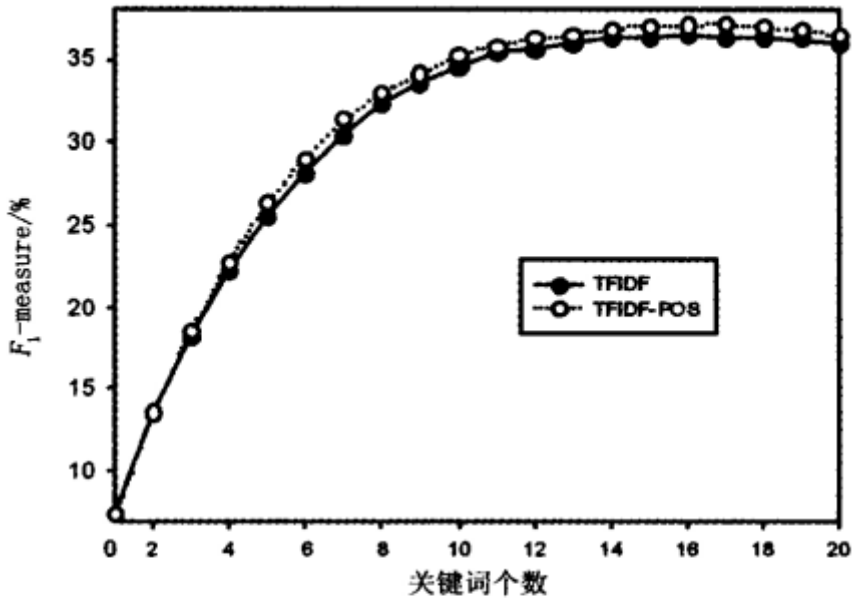

图 1　在 Inspec 语料上 F_1 测度值的变化

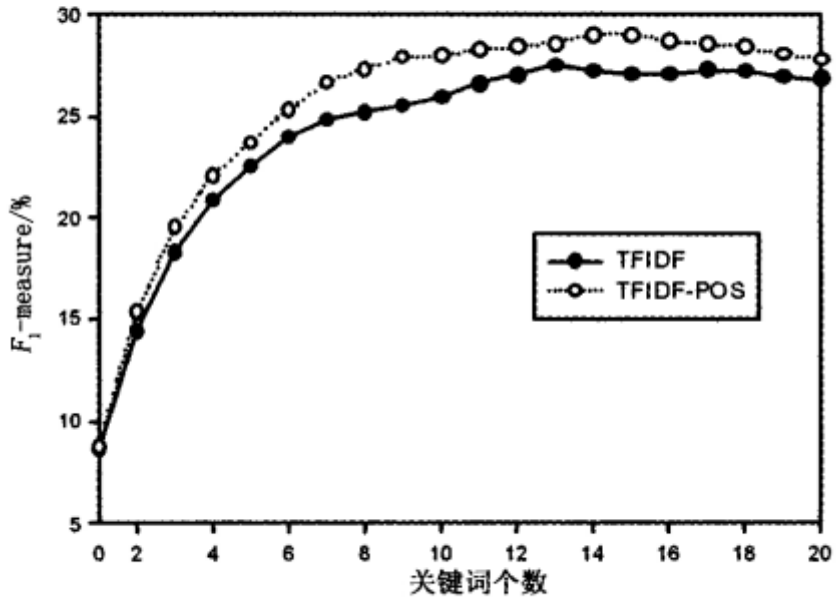

图 2　在 DUC2001 语料上 F_1 测度值的变化

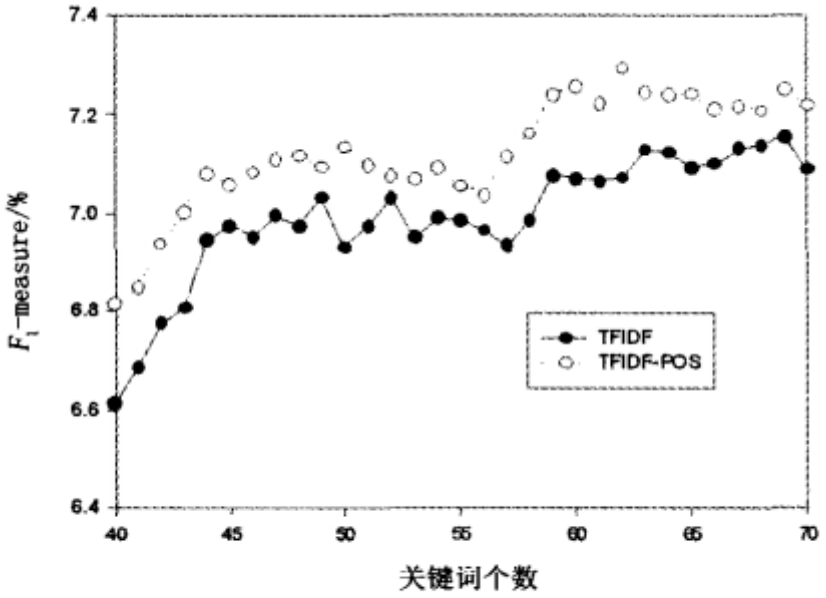

图 3　在 NUS 语料上 F_1 测度值的变化

优点：关键词抽取的准确率有所提高；

未来工作：进一步考虑结合位置及顺序等信息；充分利用多种信息特征来提高抽取准确率。

一种基于语义关系图的词语语义相关度计算模型

一种基于语义关系图的词语语义相关度计算模型

张仰森[1]　郑 佳[1]　李佳媛[1]

2018 自动化学报

> 语义相关度反映的是两个词语之间的相互关联程度，指词语之间的组合特点，即看到一个词会联想到另一个语义相关的词，可以用这两个词在统一语境中共现的可能性来衡量。
>
> 语义相似度只词语之间的相似程度，通常指语义本身具有某些相似的特性，反映的是词语之间的聚合特点，即一个词可以用另一个词替换。

基于语义词典的方法有如下问题：

1. 自然语言中的词语往往具有很强的模糊性；

2. 词语语义知识含量丰富，人工构建的语义词典很不完备；
3. 自然语言随着时间的变化存在语义漂移现象。

基于统计的方法有如下问题：对语料库依赖性大、计算量大、数据稀疏问题严重、数据噪声多、存储需求大等缺陷。

针对这些不足，文中提出一种基于语义词典和语料库相结合的词语语义相关度计算模型。首先，在分析HowNet语义表示的基础上，提取丰富的语义关系，以语义关系三元组为基础，构建语义关系图。然后对大规模语料进行依存句法分析，提取出依存句法关系，筛选后加入到语义关系树。最后，利用图论的相关理论对语义关系图中的语义关系信息进行处理，提出一种基于语义关系图的词语语义相关度计算模型。在给定两个词语后，采用图论的遍历算法遍历语义关系图，得到两个词语的语义联通路径数目n和每条路径的长度$L_i(1 \leq i \leq n)$，通过n和L_i来计算词语相关度。

Spearman系数可以估计两个变量之间的相关性，取值在$[0,1]$之间，值越大表示相关性越大。定义如下：

假设存在两个随机变量 $\boldsymbol{X}$、$\boldsymbol{Y}$, 它们的元素个数均为 n, 其中 X_i、Y_i 分别表示两个随机变量的第 i 个值 $(1 \leq i \leq n)$. 对 $\boldsymbol{X}$、$\boldsymbol{Y}$ 进行排序 (同时为升序或降序), 得到 $\boldsymbol{X}$、$\boldsymbol{Y}$ 的排序集合 $\boldsymbol{x}$、$\boldsymbol{y}$, 其中元素 x_i、y_i 分别为 X_i、Y_i 在 $\boldsymbol{x}$、$\boldsymbol{y}$ 中的排序序号, 令 $d_i = x_i - y_i$ $(1 \leq i \leq n)$. 则随机变量 $\boldsymbol{X}$、$\boldsymbol{Y}$ 之间的 Spearman 系数的计算如式 (13) 所示:

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(d_i^2)}{n(n^2-1)} \tag{13}$$

数据集采用[1]中构建的WordSimilarity-353(WS353)翻译成中文。结果如下：

表 2　不同方法的 Spearman 系数比较

Table 2　The comparison of Spearman in different methods

模型		Spearman 系数	模型		Spearman 系数
Knowledge-based	LIU[23]	0.4202	Knowledge-based	WUP[24]	0.3390
	WU[23]	0.3205		J&C[24]	0.3180
Corpus-based	TFIDF[17]	0.4030		Lin[24]	0.3480
	COMB[17]	**0.5150**		Resnik[24]	0.3530
	ICLinkBased[23]	0.2786	Corpus-based	LSA[24]	**0.5810**
	ICSubCategoryNodes[23]	0.2803		ESA[24]	**0.6290**
	WLM[23]	0.4984		SSA[24]	**0.5370**
	WLT[23]	**0.5126**	Knowledge + Corpus-based	WTMGW[24]	**0.7500**
Our methods	HN	0.4389			
	DSR	**0.5012**			
	HN+DSR	**0.5358**			

优点：模型性能在中文语料上为最优的。

缺点：有些相关度不太准确，原因有：

1. HowNet中有些义原的描述不合理导致语义关系产生偏差；
2. 依存分析器可能会分析出一些不合理的语义依存搭配关系；
3. 存在数据资源有限，数据稀疏，语义漂移等问题。

未来工作：增大语料库的数量；探索更为直接的语义三元组获取方法，避免语义词典和语义依存分析的错误传递导致词语语义相关度计算的偏差。

[1]Finkelstein L, Gabrilovich E, Matias Y, et al. Placing search in context: The concept revisited[C]//Proceedings of the 10th international conference on World Wide Web. 2001: 406-414.

专利技术术语的抽取方法

专利技术术语的抽取方法[1)]

韩红旗[1,2]　朱东华[3]　汪雪锋[3]

（1. 数据工程与知识工程教育部重点实验室（中国人民大学），北京　100872；
2. 中国人民大学信息资源管理学院，北京　100872；3. 北京理工大学管理与经济学院，北京　100081）

2011 情报学报

研究中中文专利技术关键词或术语抽取方法研究很少见。C-value方法不需要前提条件，主要采用以下统计特征值：

1. 候选词语在语料库中出现的频次；
2. 候选词语的嵌套词在语料库中出现的频次（即作为其他更长词语的一部分出现的频次）；
3. 嵌套词语的数量；
4. 候选词语的长度，即包含多少单词。

文中提出的PC-value方法去掉了长术语中的较短词语的频率，改进后的公式为：

$$PC\text{-}value(a) = \begin{cases} \log_2^{|a|} \cdot f(a) + 2^{|a|-2} \cdot g(a) & \text{当 } a \text{ 没有被嵌套} \\ \log_2^{|a|}\left(f(a) - \frac{1}{|T_a|}\sum_{b \in T_a} f(b)\right) + 2^{|a|-2} \cdot g(a) & \text{其他情况} \end{cases}$$

数据为从专利系统中下载的燃料电池技术领域的专利数据。结果表明：PC-value方法不但能够较好地抽取多字术语，而且在抽取精度上也好于C-value方法。

优点：抽取精度有所提高； 未来工作：可以考虑采用术语伴随词来提高术语抽取的精度。

使用Bert搭建关系抽取模型进行别名抽取：

https://www.modb.pro/db/183015

数据及代码

基于条件随机场的方志古籍别名自动抽取模型构建

基于条件随机场的方志古籍别名自动抽取模型构建

李 娜

（南京林业大学 人文社会科学学院，江苏 南京 210037）

中文古籍的命名实体研究多集中在文学作品等较多规范的古籍文本中。

文中基于条件随机场理论，构建物产别名的自动识别模型，实现物产别名的自动抽取。在构建CRF模型时需加入别名长度和一元边界词等特征来提高模型性能。

数据集为《方志物产》山西分卷，采用人工手动标注别名的方式。模型结果如下：

表 4　物产别名自动识别模型的测试结果

编号	训练语料	测试语料	精确率 P/%	召回率 R/%	调和平均数 F/%
1	2—10	1	90.48	80.81	84.52
2	1,3—10	2	93.77	78.85	85.46
3	1—2,4—10	3	91.50	76.66	82.66
4	1—3,5—10	4	93.01	81.76	86.77
5	1—4,6—10	5	93.47	81.24	86.85
6	1—5,7—10	6	95.56	82.03	88.02
7	1—6,8—10	7	94.01	85.56	89.48
8	1—7,9—10	8	93.02	82.18	87.07
9	1—8,10	9	95.03	77.72	85.12
10	1—9	10	95.36	79.53	86.28
平均值			93.52	80.63	86.22

优点：识别的准确率较高；

缺点：召回率较低；原因有：

1. 别名单独出现；
2. 别名与物产名相同；
3. 别名重复出现；
4. 别名与其他信息混淆；
5. 别名的长度判断错误；
6. 别名连续并列出现；

未来工作：扩大语料规模；完善特征模板。

旅游场景下的实体别名抽取联合模型

旅游场景下的实体别名抽取联合模型*

杨一帆，陈文亮
（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

2020 中文信息学报

一些主题的描述文本不包含其本身，存在文本主体缺失问题。

文中提出将主体特征信息嵌入句中进行编码，明确目标别名，并以联合方式建模，实现针对主体的别名抽取。联合抽取模型如下图所示(编码层采用ELMo预训练模型)：

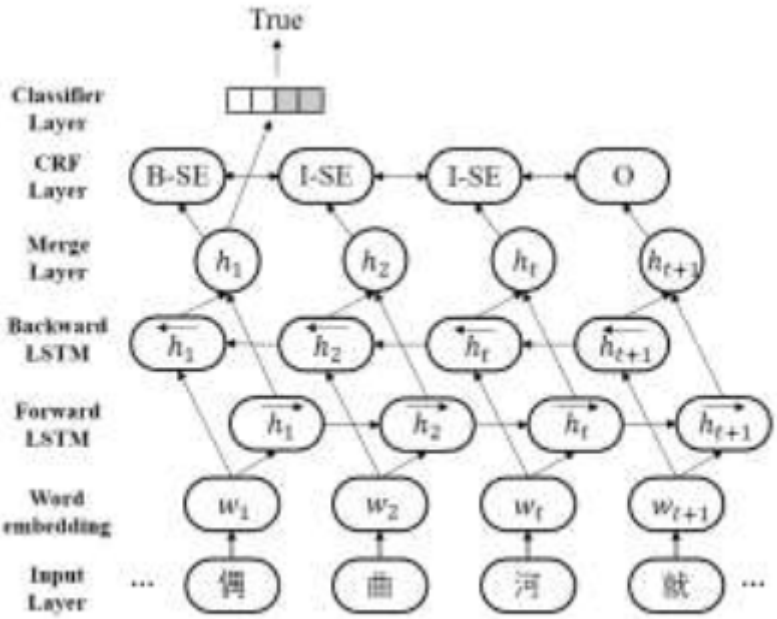

图4 联合模型框架示意

数据集为爬取的旅游景点描述文本，采用人工标注出目标实体。提供数据和代码。对比结果为：

表5 不同模型实验结果对比

Methods	Precision/%	Recall/%	F1/%
CRF+ME	66.95	46.85	55.12
BiLSTM-CRF+CNN	65.67	59.16	62.24
LSTM-LSTM-Bias	64.61	61.42	62.97
BiLSTM-CRF+C&C	69.14	**66.67**	67.88
LS&AP+BiLSTM-CRF+C&C	75.60	63.01	68.73
LS&AP+BiLSTM-CRF+C&C+ELMo	**75.90**	64.98	**70.02**

优点：能够自动识别特征，准确率有所提高；

缺点：识别错误的情况：

1. 上下文表述形式特殊；
2. 语言种类的差异性；
3. 实体命名生僻；
4. 实体枚举过多；

未来工作：考虑加入语义特征或其他语言模型（如BERT），进一步提高识别的效果。

表1 常见的别名类别

别名类别	主体	示例	实体类型
指向性昵称	故宫	故宫又叫做【紫禁城】…	AE
新旧昵称	中国国家馆	…后来更名为【中华艺术宫】	AE
	上海动物园	…原名为【西郊公园】	AE
誉名	苏州	…有【“人间天堂”】的美誉	AE
实体缩写	苏州大学	【苏大】校园内风景优美…	SE
添加或省略位置	国家图书馆	【中国国家图书馆】是中国…	SE
字符重叠	王府井天主堂	【王府井教堂】位于…	SE

Word Embedding Approach for Synonym Extraction of Multi-Word Term

Word Embedding Approach for Synonym Extraction of Multi-Word Terms

Amir Hazem and Béatrice Daille
Laboratoire des Sciences du Numérique de Nantes (LS2N)
Université de Nantes, 44322 Nantes Cedex 3, France
Amir.Hazem@univ-nantes.fr,Beatrice.Daille@univ-nantes.fr

Hazem A, Daille B. Word embedding approach for synonym extraction of multi-word terms[C]//Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.

多词同义词的抽取工作上的研究较少。目前有两类方法：组合的方式(Compositional Approach)和半组合的方式(Semi-Compositional Approach)。组合的方式依赖于规则和同义词词典，这种方法过多依赖于领域的资源；半组合的方式对多词同义词中的单个词使用概率分布的方式发现其同义词，然后组合得到同义词，这种方法难以处理不同长度的多词同义词抽取问题。

针对上述问题，文中提出两点改进方法：

1. 使用词向量模型来代替半组合方式中概率分布的方法来发现语义相似的词；
2. 依赖于词向量的语义可加性，可以使用词向量来直接发现变长多词同义词。

数据集：French/English 风能语料和French/English 乳腺癌语料。结果如下：

Method	French	English
Hamon&Nazarenko	0.25	3.63
Mikolov	4.56	6.78
Semi-Comp (MI-COS)	27.4	32.6
Semi-Comp (LO-COS)	26.8	27.2
Semi-Comp (LLR-JAC)	31.4	36.1
Semi-Comp (SG50)	30.9	50.3
Semi-Comp (SG100)	**34.9**	55.9
Semi-Comp (SG200)	34.8	52.7
Semi-Comp (CBOW50)	23.0	49.0
Semi-Comp (CBOW100)	23.7	49.4
Semi-Comp (CBOW200)	23.8	49.4
Full-Comp (SG100)	27.3	57.8
Full-Comp (SG200)	28.9	**58.4**
Full-Comp (SG300)	28.5	55.3
Full-Comp (CBOW50)	22.6	47.0
Full-Comp (CBOW100)	20.1	45.1
Full-Comp (CBOW200)	21.6	44.5

Table 2: Results (MAP%) on the wind energy corpus.

Method	French	English
Hamon&Nazarenko	4.92	7.03
Mikolov	8.37	9.12
Semi-Comp (MI-COS)	19.9	12.6
Semi-Comp (LO-COS)	27.1	11.0
Semi-Comp (LLR-JAC)	13.9	13.3
Semi-Comp (SG50)	32.1	15.0
Semi-Comp (SG100)	**32.2**	15.2
Semi-Comp (SG300)	27.9	9.60
Semi-Comp (CBOW50)	29.1	15.1
Semi-Comp (CBOW100)	29.2	15.3
Semi-Comp (CBOW300)	29.4	15.8
Full-Comp (SG100)	25.6	17.4
Full-Comp (SG200)	28.0	**18.9**
Full-Comp (SG300)	30.5	16.0
Full-Comp (CBOW100)	24.9	10.6
Full-Comp (CBOW200)	24.9	11.6
Full-Comp (CBOW300)	25.0	10.5

Table 3: Results (MAP%) on the breast cancer corpus.

优点：引入词向量模型可以显著提高组合和半组合方式抽取同义词模型的性能；

缺点：候选词中会出现重复词而降低模型性能；

未来工作：设计特定的过滤方法去掉重复的词；考虑不同长度的同义词和语法规则的使用。

数据及代码

A Measure of Similarity between Graph Vertices: Applications to Synonym Extraction and Web Searching

A Measure of Similarity between Graph Vertices: Applications to Synonym Extraction and Web Searching*

Vincent D. Blondel†
Anahí Gajardo‡
Maureen Heymans§
Pierre Senellart¶
Paul Van Dooren†

Blondel V D, Gajardo A, Heymans M, et al. A measure of similarity between graph vertices: Applications to synonym extraction and web searching[J]. SIAM review, 2004, 46(4): 647-666.

文中从单语言的词典出发，根据词典构建一个有向图，对于出现在一个词的定义中的另一个词，存在一条指向该词的边。然后，对于一个给定的单词，可以从上述图中抽取出一个子图，子图以该单词为中心。根据这个子图，迭代计算相似度矩阵，选取矩阵中相似度较高的顶点作为该给定词的同义词。

数据集为Online Plain Text English Dictionary，结果如下图：

	Distance	Our method	ArcRank	WordNet
1	vanish	vanish	epidemic	vanish
2	wear	pass	disappearing	go away
3	die	die	port	end
4	sail	wear	dissipate	finish
5	faint	faint	cease	terminate
6	light	fade	eat	cease
7	port	sail	gradually	
8	absorb	light	instrumental	
9	appear	dissipate	darkness	
10	cease	cease	efface	
Mark	3.6	6.3	1.2	7.5
Std dev.	1.8	1.7	1.2	1.4

缺点：对于具有弱连通的图不太适用。

A Minimally Supervised Approach for Synonym Extraction with Word Embeddings

A Minimally Supervised Approach for Synonym Extraction with Word Embeddings

Artuur Leeuwenberg[a], Mihaela Vela[b], Jon Dehdari[bc], Josef van Genabith[bc]

[a] KU Leuven – University of Leuven, Belgium
[b] Saarland University, Germany
[c] DFKI, German Research Center for Artificial Intelligence, Germany

Leeuwenberg A, Vela M, Dehdari J, et al. A minimally supervised approach for synonym extraction with word embeddings[J]. The Prague Bulletin of Mathematical Linguistics, 2016, 105(1): 111.

WordNet这种大型的同义词词库只适用于英语领域，人工构建这种大规模的知识库需要耗费大量的人力物力。

为了降低同义词抽取的监督依赖性，以及改善余弦相似度的度量标准和同义词抽取的性能，并使该方法能够应用到不同的语言领域，文中提出了一种使用词向量的方式，并使用相对余弦相似度作为相似度的衡量指标。

词向量参考Word2Vec的方法，使用CBOW或者SkipGram模型进行训练。相对余弦相似度计算前n个最相似单词的相似度，定义如下：

$$\mathrm{rcs}_n(w_i, w_j) = \frac{\mathrm{cosine\_similarity}(w_i, w_j)}{\sum_{w_c \in \mathrm{TOP}_n} \mathrm{cosine\_similarity}(w_i, w_c)}$$

进一步提高性能可以使用词性标注的方式，将词性作为一个特征进行相似度判断。

文中使用的数据集为the NewsCrawl corpus from the 2015 Workshop on Machine Translation，实验结果为：

Manual Evaluation	P^{-}_{syn}	P^{+}_{syn}	$P^{-}_{\neg syn}$	$P^{+}_{\neg syn}$	$P_{disagree}$	P_{uu}
English	0.55	0.59	0.15	0.21	0.16	0.05
German	0.30	0.35	0.42	0.49	0.15	0.03

Table 13. Manual evaluation of the final systems.

优点：降低了对监督的依赖性；可以适用于那些没有大规模同义词知识库的语言；降低了上下位词以及相关词对同义词抽取的影响。

缺点：准确率上不如WordNet等数据库。

未来工作：研究相似度阈值变化对模型性能的影响；整合POS标签来防止不正确的同义词泛化。

Automatic Synonym Extraction Using Word2Vec and Spectral Clustering

Automatic Synonym Extraction Using Word2Vec and Spectral Clustering

Li Zhang, Jun Li, Chao Wang

School of Information Science and Technology, University of Science and Technology of China, Hefei 230027
E-mail: zhangli1@mail.ustc.edu.cn

Zhang L, Li J, Wang C. Automatic synonym extraction using Word2Vec and spectral clustering[C]//2017 36th Chinese Control Conference (CCC). IEEE, 2017: 5629-5632.

目前已有的同义词抽取算法，如基于语义差异的、基于句法分析的和基于查询的算法，都存在计算复杂度高，时间成本高的特点。

文中使用word2vec模型来计算给定术语之间的相似度，然后这些词通过谱聚类算法进行聚类。算法描述如下：

Table 2. Multi-path Normalized Spectral Clustering Algorithm

Input: the similarity matrix S, the number of clusters K

- Construct the similarity graph, W is the adjacency matrix of the graph
- According to formula (3), calculate *Laplacian* L_{sym}
- Compute the first k eigenvectors $u_1,...,u_k$ of L_{sym}
- The eigenvectors $u_1,...,u_k$ of step 3. are arranged as columns in a new matrix $u \in R^{n\times k}$
- Compute the normalized matrix $T \in R^{n\times k}$ $T_{ij} = u_{ij} \Big/ (\sum_k u_{ik}^2)^{1/2}$
- Let $y_i \in R^k (i=1,...,n)$ be the i-th row vector of matrix T
- For data points $y_i \in R^k (i=1,...,n)$, use a specific clustering algorithm clustering into K classes: $C_1,...,C_K$

Output: Clusters $A_1,...,A_K$,with $A_i = \{j \mid y_j \in C_i\}$

训练word2vec模型的数据集为英文维基百科数据，同义词抽取的数据集为Reuters-21587。实验结果如下：

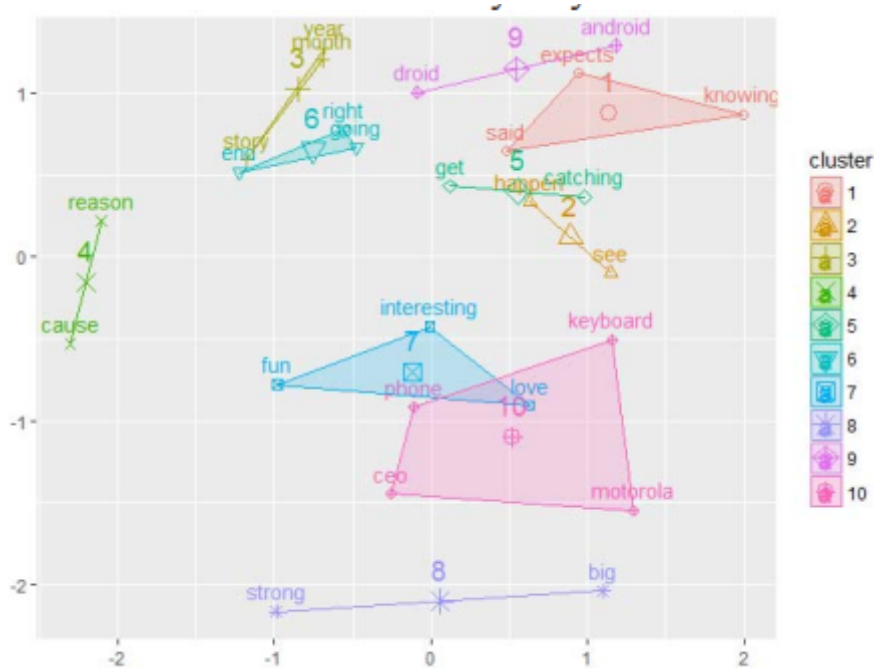

Fig. 1: Key words clustering result use spectral clustering

Synonym Extraction

Clustering Methods	Precision	Recall	F1
Spectral Clustering	0.808	0.744	0.775
K-means	0.279	0.473	0.351

优点：降低了同义词抽取的时间；提升了抽取的准确率。

Automatic Relation Extraction - Can Synonym Extraction Benefit from Antonym Knowledge?

Automatic Relation Extraction –
Can Synonym Extraction Benefit from Antonym Knowledge?

Anna Lobanova, Jennifer Spenader, Tim van de Cruys,
Tom van der Kleij, Erik Tjong Kim Sang
University of Groningen
{a.lobanova@| j.spenader@| t.van.de.cruys@| a.a.j.van.der.kleij@ai.| e.f.tjong.kim.sang@}rug.nl

Lobanova A, Spenader J, Van De Cruys T, et al. Automatic Relation Extraction − Can Synonym Extraction Benefit from Antonym Knowledge?[J]. Proceedings of WordNets and other Lexical Semantic Resources-between Lexical Semantics, Lexicography, Terminology and Formal Ontologies, Odense, Denmark, 2009: 17-20.

反义词和同义词拥有一样的上下文分布环境，因此会被误抽取。文中提出在抽取的同义词候选集中去除反义词的方法来提高抽取的同义词的质量。

反义词抽取采用基于模式的方法，先设计模式，然后抽取出满足条件的句子，从中学习新的模式后迭代学习新的模式。

数据集为Twente News Corpus(200 million words)，结果如下：

	cut-off (cosine)					
	.40	.30	.20	.18	.15	.10
Baseline (unfiltered)						
Precision	.008	.025	.053	.045	.036	.014
Recall	.003	.005	.035	.038	.048	.099
$F_{\beta=1}$	.004	.008	.042	.041	.041	.025
Filtered						
Precision	.008	.025	.055	.047	.039	.015
Recall	.003	.005	.035	.038	.048	.099
$F_{\beta=1}$	.004	.008	.042	.042	.043	.026

Table 3: Effects of filtering out antonyms derived with chosen patterns from a set of 114 candidate synonyms: a small positive effect on the low-cut-off sets.

	cut-off (cosine)					
	.40	.30	.20	.18	.15	.10
Baseline (unfiltered)						
Precision	.025	.035	.077	.097	.071	.024
Recall	.017	.025	.053	.091	.120	.174
$F_{\beta=1}$	.020	.029	.063	.094	.090	.042
Filtered						
Precision	.013	.023	.070	.090	.069	.024
Recall	.004	.013	.041	.078	.107	.161
$F_{\beta=1}$	.006	.017	.051	.084	.084	.042

Table 4: Effects of filtering out antonyms derived with learned patterns from a set of 80 candidate synonyms: a large negative effect on the high-cut-off sets.

优点：在同义词抽取的质量上有较高的提升；

缺点：抽取的反义词候选集中会存在同义词，从而降低了同义词抽取的准确率；为了使反义词抽取的更准确，需要更多的数据集或者设计更多的模式，这些实现成本较高。

未来工作：考虑使用其他相关词来代替反义词。

Automatic translation of scholarly terms into patent terms using synonym extraction techniques

Automatic Translation of Scholarly Terms into Patent Terms Using Synonym Extraction Techniques

Hidetsugu Nanba[1], Toshiyuki Takezawa[1], Kiyoko Uchiyama[2], Akiko Aizawa[2]
[1] Hiroshima City University
[2] National Institute of Informatics
[1] 3-4-1 Ozukahigashi, Asaminamiku, Hiroshima 731-3194 JAPAN
[2] 2-1-2 Hitotsubashi, Chiyodaku, Tokyo 101-8430 JAPAN
E-mail: nanba@hiroshima-cu.ac.jp, takezawa@hiroshima-cu.ac.jp, kiyoko@nii.ac.jp, aizawa@nii.ac.jp

Nanba H, Takezawa T, Uchiyama K, et al. Automatic translation of scholarly terms into patent terms using synonym extraction techniques[C]//Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). 2012: 3447-3451.

在专利中使用的术语比研究报告中使用的术语更具抽象性和创造性，因此，需要一种将学术术语翻译成专利术语的方法。

文中提出了几种翻译方法：

1. 基于统计机器翻译的方法：如果两种表述的翻译是相同的，那么这两种表述可以被认为是相关的。
2. 基于分布相似度的方法：相似的词可以被用于相似的上下文环境中，两个术语之间的相似度被定义为术语之间含有的信息与上下文中含有的信息的商。步骤如下：

1. Analyze the dependency structures of all sentences in a research paper database using the Japanese dependency parser CaboCha[2].
2. Extract noun-phrase-verb (with a postpositional particle) pairs that have dependency relations from the dependency trees obtained in Step 1.
3. Count the frequencies of each noun-phrase-verb pair.
4. Collect verbs and their frequencies for each noun phrase, creating indices for each noun phrase.
5. Create indices from a patent database in the same way as a research paper database (Steps 1 to 4).
6. Calculate the similarities between two indices of noun phrases created in Steps 4 and 5 using the SMART similarity measure (Salton, 1971).
7. Obtain a noun phrase with the highest similarity score as a translation of a given scholarly term.

数据集为Patent Mining Task at the NTCIR-7 Workshop。结果如下：

Methods	Subgroup (5^{th} level)	Main Group (4^{th} level)	Subclass (3^{rd} level)
SMT_ABST	0.3786	**0.5186**	0.6691
SMT_ABST+IDF	**0.3812**	**0.5197**	0.6709
SMT_TITLE	**0.3797**	**0.5208**	0.6688
SMT_TITLE+IDF	**0.3799**	**0.5204**	0.6710
DS	**0.3793**	0.5182	0.6717
DS+IDF	**0.3794**	0.5175	**0.6744**
PAPER (baseline)	0.3792	0.5185	0.6720

Table 2: MAP scores by our methods and a baseline method.

结论：基于机器翻译模型的方法更适用于较窄领域的任务，基于统计的方法更适用于较广领域的任务。