CnKPRank：结合统计特征和图模型的半监督式中文关键短语抽取方法

谢海华[1], 陈雪飞[1], 都仪敏[1], 吕肖庆[1,2], 汤帜[1,2]

1. 数字出版技术国家重点实验室，北大方正信息产业集团有限公司，北京
2. 王选计算机研究所, 北京大学, 北京

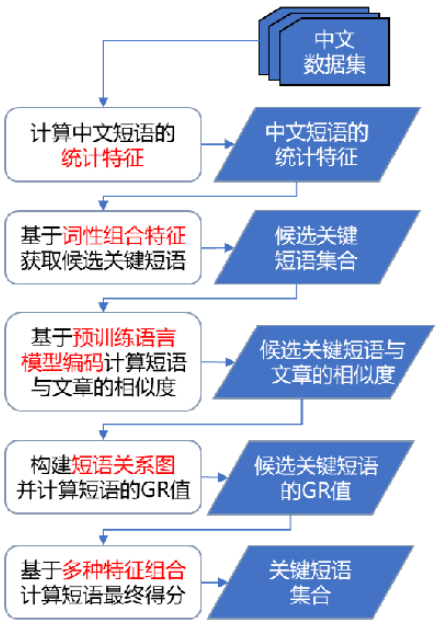
任务描述

从中文文档中抽取能够表达文档主题和内容的关键短语集合。

论文贡献

- 基于中文期刊论文数据，构建中文关键短语标注数据集，用于计算中文短语的统计特征，以及评估中文关键短语抽取算法的性能。
- 针对中文特点，提出基于词性及词性组合特征的候选关键短语获取方法。
- 为了更准确地提取关键短语，提出采用多种因素结合的方式来计算短语得分，包括：短语和文章的语义相似度、图模型排序、统计特征得分。

算法流程



详细步骤

1. 计算中文关键短语的统计特征

	类型	出现次数	概率
短语的词性组合	n+n	6873	0.1009
	n+v	2638	0.054
	vn	2130	0.0313
短语长度（字符数）	4	11095	0.4129
	3	3637	0.1354
	2	3489	0.1299
短语在文章中的位置	0-10%	16379	0.2404
	10-20%	7270	0.1067
	20-30%	6894	0.1012

2. 基于词性组合获取候选关键短语

<b>?<n|vn|v|a|b|vi>?<n|g.|nz|vn|v|ns|vi>*

3. 计算候选关键短语与文章的相似度

(1) 运用预训练语言模型，生成词语的向量表示；

(2) 构建主题词性集，并从文章中获取主题词；

(3) 将所有主题词的向量表示进行累加并求平均，得到文章的主题向量表示。

(4) 将短语中的词语的向量进行累加并求平均，得到短语的向量表示。

$$\text{□□□}(\square_p,\square_d)=\frac{\sum_{i=1}^{n}(\square_{p_i}\times\square_{d_i})}{\sqrt{\sum_{i=1}^{n}\square_{p_i}^2}\times\sqrt{\sum_{i=1}^{n}\square_{d_i}^2}}$$

4. 构建短语关系图并计算短语的GR值

如果两个候选关键短语出现在同一个句子中，而且在同一窗口内（窗口尺寸设置为5个词语），那么这两个关键短语对应的节点会有连线。连线的权重等于节点对应的短语出现在同一窗口内的次数。

$$\text{□□}(\square_i)=(1-\square)\text{□□□}(\square_i,\square)+\square\times\sum_{j\in M_i}\left(\frac{\square_{ij}}{\square(\square_j)}\times\text{□□}(\square_j)\right)$$

5. 基于多种统计特征计算短语得分

$$\square(\square_i)=\text{□□}(\square_i)+\text{□□□}(\square_i)+\text{□□□}(\square_i)+\text{□□□}(\square_i)$$

实验与分析

- 中文论文数据集：https://github.com/binggoml/Key-phrase-extraction_Chinese-corpus

Table 6. 中文关键短语抽取对比实验结果

关键短语选择方法	模型	准确率(P)	召回率(R)	F1
5个短语	TFIDF	0.1966	0.2236	0.2092
	TextRank	0.1555	0.1769	0.1655
	TopicRank	0.1318	0.1499	0.1403
	PositionRank	0.198	0.2252	0.2107
	SIFRank	0.1372	0.1559	0.146
	sim based	0.2243	0.255	0.2387
	gr-based	0.2793	0.3175	0.2972
	CnKPRank	**0.3074**	**0.3496**	**0.3272**
10个短语	TFIDF	0.1323	0.3010	0.1839
	TextRank	0.1132	0.2574	0.1572
	TopicRank	0.0888	0.2013	0.1232
	PositionRank	0.1625	0.3696	0.2258
	SIFRank	0.1181	0.2676	0.1639
	sim based	0.1828	0.4152	0.2537
	gr-based	0.188	0.4275	0.2612
	CnKPRank	0.2167	0.4927	0.301
基于阈值 $t=1.2, min=3, max=7$	CnKPRank	**0.2568**	**0.4114**	**0.3162**

未来研究方向

- 优化候选关键短语的选择。候选关键短语的选择是关键短语抽取任务的重点和难点，具有较大的提升空间和研究价值。从短语的词语构成、词性构成、语法结构等方面进行深入研究，有助于提升关键短语选择的效果。
- 抽取未在文本中出现的关键短语。目前方法抽取出的关键短语都曾出现在文本当中，而测试集中的有些关键短语并没有直接在文本中出现。抽取出不曾出现在文本当中的关键短语也是一个有价值的研究方向。

联系方式

✓ www.cndplab.com
✓ www.founderit.com
✓ www.foundersc.com
✓ cndplab@cndplab.com