术语自动抽取方法研究综述

袁劲松 张小明 李舟军
(北京航空航天大学 北京 100191)

摘 要 术语抽取是文本处理领域的一项基础性研究工作,好的术语自动抽取方法能够提高本体构建的质量和语义检索的精度。首先,对术语的定义、特性以及术语抽取效果的评价方法进行了概述。然后,在分析和总结近20年术语自动抽取相关文献的基础上,对术语自动抽取的各种方法进行了详细的综述。介绍了这些方法的研究进展,分析了其优缺点,并详细描述了部分经典算法。最后,对术语自动抽取未来研究的趋势进行了展望。

关键词 术语抽取,文本处理,评价方法,自动抽取方法
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.8.002

Survey of Automatic Terminology Extraction Methodologies

YUAN Jin-song ZHANG Xiao-ming LI Zhou-jun
(Beijing University of Aeronautics & Astronautics, Beijing 100191, China)

Abstract Terminology extraction is a fundamental research work for text processing domain. The quality of ontology and accuracy of sematic retrieval can be improved by using a better automatic terminology extraction method. Firstly, the definition and characteristic of terminology, as well as the evaluation of terminology extraction were briefly introduced. Secondly, through a thorough analysis and summarization of literatures about automatic terminology extraction in recent twenty years, a comprehensive survey of state-of-the-art automatic terminology extraction methodologies was conducted, which includes domestic and international current research, their advantages and disadvantages and detailed descriptions of some classical algorithms. Finally, the trend of future study was discussed.

Keywords Terminology extraction, Text processing, Evaluation measures, Automatic extraction methodologies

1 引言

随着互联网的快速发展及大数据时代的来临,传统互联网的知识组织和知识共享方式越来越难以满足人们的需求。近年来,语义网的快速发展、"智能"搜索引擎的出现、知识库的构建等等无不标示着互联网正在进行一场变革,互联网将步入一个"语义时代"。作为沟通人和机器之间的重要桥梁,文本处理与分析技术在语义时代变得更加重要。

术语抽取是文本处理领域中一个基础而又重要的研究方向,可以应用在本体构建[1]、机器翻译[2,3]和语义检索[4-6]等诸多研究领域。然而术语抽取是一项复杂而困难的工作,如果只靠人工进行抽取,不仅费时耗力,而且需要一定的语言学知识以及相关的领域背景。因此,研究出一套自动、高效和高可移植性的术语抽取方法具有十分重要的意义。

为了方便理解术语抽取相关工作,本文首先对术语的定义、术语的特性以及术语抽取效果的评价方法进行概述,在此基础上对国内外术语自动抽取常用的方法进行详细的总结、分析和比较,最后指出了术语自动抽取进一步的研究方向。

2 概述

术语是随着人类对各个研究领域的不断探索和研究而逐步形成的,用来记录或标记在此过程中积累沉淀的专业知识概念。在日常生活中,术语经常被人们称为"专有名词",但其并不仅仅是名词以及复合名词,由动词、形容词、介词和名词构成的词或词组(又称短语)都有可能是术语。例如,在财经领域,动词"跌停"、动名词短语"通货膨胀"等都是术语。另一方面,随着社会的发展,许多由某个领域专有的词语逐渐到被广泛使用,这就很难明确这些词语是不是术语。例如,"指数"最初是出现在数学领域,表示数的次方,后来"指数"的含义被扩大,如经济学中的"股票指数"、"消费者价格指数"等等。因此,术语是一种结构复杂并且边界难以确定的语言单元。

2.1 术语的定义

"术语"目前还没有统一明确的定义,但是中外不少学者从语言学或者术语学角度给出了自己的观点。本文列举一些国内外比较通用的术语定义。

Sager[7]认为"terms are the linguistic representation of concepts",即"术语是概念的语言表征"。

到稿日期:2014-09-21 返修日期:2014-12-15 本文受国家自然科学基金(61170189,61370126,61202239),教育部博士点基金(20111102130003)资助。

袁劲松(1990-),男,硕士生,主要研究方向为数据挖掘,E-mail:yuanjinsong1990@gmail.com;张小明(1980-),男,博士生,主要研究方向为数据挖掘与文本挖掘;李舟军(1963-),男,教授,博士生导师,主要研究方向为数据挖掘与文本挖掘、网络与信息安全。

冯志伟[8]是我国较早开始从事术语学研究的学者，他将术语定义为“通过语音或文字来表达或限定专业概念的约定性符号”。

《术语工作原则与方法》[9]中写到“术语是专业领域中概念的语言指称”。

以上几种术语的定义虽然角度和内容各有不同，但我们不难看出，术语与特定领域中的概念之间存在着紧密的联系。

2.2 术语的特性

GB/T 10112-1999[9]中对术语的特性有明确的要求：

(1)单名单义性。在创立新术语之前应先检查有无同义词，并在已有的几个同义词之间选择一个能较好满足其他要求的术语。

(2)顾名思义性，又称透明性。这里的“义”是指定义，术语应能准确扼要地表达定义的要旨。

(3)简明性。要求术语尽可能简明，以提高效率。

(4)派生性，又称能产性。术语应便于构词，特别是组合成词组使用的基本术语更应如此。基本术语越简短，构词能力越强。

(5)稳定性。使用频率较高、范围较广，已经约定俗成的术语，没有重要原因，即使是有不理想之处，也不宜轻易变更。

(6)合乎本族语言习惯。术语要适合本族语言习惯，用字遣词，务求不引起歧义，不要带有褒贬等感情色彩的意蕴。

以上要求较为详细全面，但是太过抽象，无法具体度量。Kageura[10]将术语的特性归结为两个便于度量的特性：

(1)单元性(Unithood)，指一个词或词组是否可以表达一个独立、完整的语言含义，并具有稳定的结构。

(2)术语度或领域性(Termhood)，是指一个词或词组与特定领域的相关性，用于度量该词或词组对领域知识的表达能力。

以上两个特性既能较好地反映上述几种定义的涵义，又能够进行度量。因此，目前很多术语自动抽取工作中，大多是以这两点或其中一点作为判断候选术语是否为术语的标准。但仅仅使用这两个度量标准，很难保证术语的单义性和无歧义性。

2.3 术语抽取效果的评价

目前，尚不存在统一的术语抽取效果评价方法。常见的术语抽取结果评价方法[11]有3个：准确率(Precision)、召回率(Recall)和F值(F-Measure或F-Score)。在某个领域语料上，术语抽取结果统计信息如表1所列。

表1 术语抽取结果统计表

	术语	非术语
被抽取出的	a	b
没有被抽取出的	c	

以上结果统计时需要参照一个标准术语表(golden standard)[12]。设标准术语表为 ST，提取出的术语集合为 T，则被抽取出的术语个数为：

$$a=|ST\cap T| \tag{1}$$

准确率是衡量术语抽取的准确程度，计算公式为：

$$P=\frac{a}{a+b}=\frac{|ST\cap T|}{|T|} \tag{2}$$

召回率是衡量术语抽取的全面程度，计算公式为：

$$R=\frac{a}{a+c}=\frac{|ST\cap T|}{|ST|} \tag{3}$$

F值综合考虑了准确率和召回率，计算公式为：

$$F=\frac{(\alpha^2+1)PR}{\alpha^2P+R} \tag{4}$$

其中，α是可调节参数，常使用的值为0.5,1,2。当$\alpha=1$时，准确率和召回率的权重相同；当$\alpha=2$时，准确率的权重较高；当$\alpha=0.5$时，召回率的权重较高。

另外，Korkontzelos[13]提出了一种在候选术语有序的情况下，能够对术语抽取方法进行更全面评价的方法。

设 L 为将 T 以某种度量排好序后的有序表，标准术语表为 ST，在 L 中前 $n(1,2,\cdots,N)$ 个术语中，真正的术语(Correct Terms)$CT(n)$为：

$$CT(n)=ST\cap\{L[1],L[2],\cdots,L[n]\} \tag{5}$$

L 中前 n 个术语的准确率 $P(n)$ 定义为：

$$P(n)=\frac{|CT(n)|}{n} \tag{6}$$

召回率 $R(n)$ 表示在提取出的真正术语中，排在前 n 个的真正术语的比率：

$$R(n)=\frac{|CT(n)|}{|ST\cap T|} \tag{7}$$

以上几种评价方法各有特点，在术语抽取工作中，根据术语抽取方法的特点以及不同的应用场景会选取不同的评价方法。

3 术语自动抽取方法

术语自动抽取的研究已长达20多年，20世纪90年代国外就有了一批术语自动抽取系统[14]。国内的研究则集中在近10年，主要是在国外研究基础上对已有方法进行改进。

中文术语自动抽取的研究还具有自身的特殊性。英文语料较中文语料有天然的优势，英文的词之间是用空格分隔开的，而中文语料中的词和词之间是没有分隔符的。因此，一般在中文术语自动抽取过程中，首先需要对语料进行分词。而在早期分词效果不是很理想的时候，很难进行相关研究。随着分词效果的提升和分词工具的日益完善，相关研究工作才得以顺利进行。目前最常用的中文分词工具是NLPIR(又名ICTCLAS2013)，其分词精度可以达到98.45%，另外还支持词性标注[15]。

早期的术语自动抽取中使用的大多是基于语言学知识。后来，随着统计自然语言处理技术的快速发展，术语抽取系统中逐步引入了一种或多种统计策略。而随着隐马尔可夫模型(Hidden Markov Model)[16]、条件随机场(Conditional Random Fields)[17]等机器学习算法在词性标注、命名实体识别等领域的使用[18]，结合机器学习算法的方法也被引入到术语的抽取研究中[19,20]。

总的来说，目前术语自动抽取方法主要分为3大类：1)基于规则的方法；2)基于统计的方法；3)多种方法相结合。

3.1 基于规则的方法

基于规则的方法主要使用术语的词语词性以及词法模式等语言知识，利用这些知识可以从语料中自动抽取出术语。

国外，基于语言学知识的术语自动抽取研究主要集中在20世纪90年代[14]。

1994年Jacquemin研究出的FASTER系统首先从已知的术语库中归纳出术语构成规则，并结合73条扩展操作规则，从法语医药领域语料中进行术语抽取，最终准确率为

86.7%，召回率为74.9%。

1995年Lingsoft公司推出的NODALIDA-95使用结构化的方法组织语言知识，以NPtool[21]为核心。其首先判断句子的边界、句子中的成语、复合形式等等，然后通过词形分析进行消歧处理。该系统的术语抽取效果非常好，在英文宇宙学技术语料中术语抽取的准确率达到95%～98%，召回率达到98.5%～100%。

1995年Justeson & Katz提出的Terms系统首先对文档进行标注和消歧处理，并提供一个语法模式列表。利用这些模式，可以抽取出文档中的单词或多个词语搭配作为候选术语，然后根据词频进行排序。在英文的光谱分析领域语料上该系统的准确率达到96%。

国内由于起步较晚，很少见到单纯使用语言知识进行术语自动抽取的相关研究。

这类方法基于已有的术语集以及领域特点进行规则总结，在准确率上有一定的优势。但该类方法的可移植性很差，即不同的语言、不同的领域、不同的语料集，语言规则各不相同，需要根据具体情况制定。同时，制定规则需要较强的语言知识和领域知识背景，并且当规则较为复杂时，还需要解决多个规则间可能存在的冲突。因此，目前已经很少有仅使用基于规则的方法进行术语自动抽取的研究。

3.2 基于统计的方法

基于统计的方法以统计学理论为基础，利用术语已经在语料库中的分布统计属性来识别术语。

经常使用到的统计方法可以分为两大类：一类衡量词或词组的领域性，如词频(Frequency)[11,22]、TF-IDF值[23,24]、领域相关性(Domain Relevance)和领域共识(Domain Consensus)[25]等；另一类衡量词组的单元性，如互信息(Mutual Information)[26]、对数似然比(Log-Likehood Rato)[27,28]等。

3.2.1 领域性度量

(1)词频。术语的词频在领域语料中一般会高于普通词汇，但有许多通用词汇在语料中也会有较高词频。因此，如果只简单考虑词频信息，将会导致抽取的术语中有许多通用词汇，并且无法抽取词频较低的术语。对此，可以使用背景语料进行改进[11,22,27,29]，通过词或词组出现在领域语料和背景语料中的词频差异进行术语抽取，从而提高术语抽取的效果。

(2)TF-IDF值。TF-IDF可以评估一个词在某个语料库中对其中一份文档的重要程度。TF-IDF值的相关计算公式为：

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{8}$$

$$idf_i=\frac{|D|}{|\{j:t_i\in d_j\}|} \tag{9}$$

$$tf\text{-}idf_{i,j}=tf_{i,j}\cdot idf_i \tag{10}$$

在术语抽取中，可以用TF-IDF值评价候选术语的领域性，TF-IDF值越大则领域性越强。

(3)DR和DC值。术语的领域相关性和领域共识是指：给定n个领域文本库的集合$(D_1,D_2,\cdots,D_n)$，词t的领域相关性和领域共识为：

$$DR(t,D_i)=\frac{P(t|D_i)}{\sum_{j=1}^{n}P(t|D_j)} \tag{11}$$

$$DC(t,D_i)=\sum_{d_j\in D_i}P(t,d_j)\log\left(\frac{1}{P(t,d_j)}\right) \tag{12}$$

其中，$P(t|D_i)$的估计值是$P(t|D_i)=\frac{freq(t\text{ in }D_i)}{\sum_{j=1}^{n}freq(t\text{ in }D_j)}$，$P(t,d_j)$的估计值是$P(t,d_j)=\frac{freq(t\text{ in }d_j)}{\sum_{d_i\in D_i}freq(t\text{ in }d_j)}$。当词或词组的领域相关性大于$\alpha$且领域共识大于$\beta$时，该词或词组则为术语。

3.2.2 单元性度量

(1)互信息。互信息是指两个事件集合之间的相关性。互信息在术语抽取中用来度量两个词之间的单元性。互信息的计算公式为：

$$MI=\log_2\left(\frac{p(AB)}{p(A)p(B)}\right)\approx\log_2\left(\frac{N\cdot f(AB)}{f(A)f(B)}\right) \tag{13}$$

其中，p表示概率，当语料足够充分时，公式中的概率可以用频率代替，在术语抽取中A和B分别表示词组W的最长前缀和最长后缀。

(2)对数似然比。基本思想是比较由或然表得到的概率和假设两个单词相互独立时得到的概率是否一致[30]。给定两个单词u和v，u、v的或然表如表2所列。

表2 两个词的或然表

	V=v	V≠v
U=u	a	b
U≠u	c	d

表2中a、b、c、d表示对应的搭配在语料中出现的次数。设语料中出现的词语总数为N，则对数似然比公式为：

$$LLR=-2\log\left(\frac{p_1^a(1-p_1)^b p_1^b(1-p_1)^d}{p_2^a(1-p_2)^b p_3^b(1-p_3)^d}\right) \tag{14}$$

其中，p_1表示单词u出现的概率，当语料足够充分时，$p_1\approx\frac{a+b}{N}$；p_2表示在v出现的情况下u出现的概率，$p_2\approx\frac{a}{a+c}$；p_3表示在v不出现的情况下u出现的概率，$p_3\approx\frac{b}{b+d}$。对数似然比的值越大，表示语料中u、v搭配的短语是随机出现的概率越小，即该短语是术语的可能性越大。当对数似然比大于一定阈值时，认为该短语是术语。

基于统计的方法不需要句法、语义上的信息，不局限于某一专门领域，也不依赖标注数据，通用性较强。但是，统计信息的可靠性依赖于语料的质量，因此，这类方法对语料的规模以及候选术语的分布依赖较强。

3.3 多种方法相结合

多种方法相结合是指同时使用规则和统计相结合的方法，或者在此基础上引入其他方法进行术语抽取。由于这类方法较为多样化，本文只对较为常见且有代表性的方法进行介绍。

3.3.1 结合语法规则和统计信息

结合语法规则和统计信息进行术语自动抽取的相关研究较多，其中较早且比较有代表性是C-value方法[31]，其基本思想是先用语言规则得到候选术语集，然后使用统计信息来进行过滤。

C-value方法是一种简单高效的术语抽取方法。Piao[32]在相同的语料上，对5种表现较好的术语自动抽取的方法进行了效果对比实验。由不同的领域专家对5种方法抽取出的前100个术语进行检查，结果5种方法中C-value方法的准确率在两轮检查中均为最好，并且稳定在90%左右，其他4种

方法在两轮检查中的准确率波动较大。

文献[31]中使用到的语言规则的正则表达式为：

Noun$^+$ Noun

(Adj|Noun)$^+$Noun

((Adj|Noun)$^+$((Adj|Noun)*(Noun Prep)$^?$ (Adj|Noun)*) Noun

其中，Noun 为名词，Adj 为形容词，Prep 是介词。这 3 条语言学规则中，第一条最严格，认为术语中只能是名词以及名词短语；第三条最宽松，认为术语中还可以包含形容词和介词，使用该规则可以抽取出更多的术语，但也会抽取出很多的非术语词组。

在使用语言学规则得到候选术语后，计算候选术语的 C-value 值，计算公式如式(15)所示：

$$C(a)=\begin{cases}(\log_2|t|)\cdot f(t), & t\text{未被嵌套}\\ (\log_2|t|)\cdot\left(f(t)-\frac{1}{|C_t|}\sum_{a\in C_t}f(a)\right), & \text{其他}\end{cases} \tag{15}$$

其中，t 是抽取的某个候选术语，$|t|$ 是候选术语 t 的长度，$f(t)$表示候选术语 a 在语料库中的词频，C_t 表示包含 t 的候选术语(即嵌套术语)集合。将 C-value 大于阈值 α 或按 C-value 值排序后 top$N$ 的候选术语作为术语。

C-value 方法在抽取中、高频术语方面明显优于仅仅基于词频的方法，对高频术语准确率可以达到 70.84%，较仅使用词频的方法提高近 2 个百分点。C-value 方法在低频部分准确率和基于词频的方法基本持平，总的准确率略高于仅仅基于词频的方法。

C-value 方法考虑了候选术语的长度和嵌套性，在长术语抽取方面表现较好。同时，C-value 方法简单、适用性强，具有较强语言、领域无关性。因此，国内外有许多研究是基于该方法的改进[31,33]。

Frantzi[31]在原有的 C-value 方法基础上加入上下文信息，称之为 NC-value 方法。NC-value 方法分为 3 步：第一步，使用 C-value 的方法抽取出候选术语并进行排序；第二步，对每个候选术语生成一个上下文词表(list of term context words)，每个词都有一个权重；用 C-value 和上下文词表计算 NC-value 值。计算公式如式(16)所示：

$$NC(t)=0.8C(t)+0.2\sum_{a\in C_t}f_t(a)w(a) \tag{16}$$

其中，t 是候选术语，C_t 是 t 的上下文词集，$f_t(a)$是指 b 语料中作为 t 的上下文出现的次数，$w(a)$是赋予词的权重。

NC-value 在高频的术语抽取方面比 C-value 表现更好，在高频部分可以达到 75.70%，比基于词频的方法在准确率上高出 6 个百分点，在低频部分 NC-value 与 C-value 表现类似。

周浪[34]按照 C-value 方法的思路，首先对科技文档领域中的术语进行统计和分析，制定了 4 条规则，这 4 条规则较 C-value 方法中使用的语言规则宽松；然后考察了术语在语料中的词频分布情况，发现术语在文档中的词频变化比较大，而普通短语的出现较为平稳。由此参照 TF-IDF 的计算公式提出了基于词频分布变化的 termhood 计算方法。对于利用规则抽取出的候选术语 t，计算其 DV 值的公式为：

$$DV(t)=\frac{tf(t)}{df(t)}\sqrt{\frac{1}{N}\sum_{i=1}^{N+1}[tf_i(t)-\overline{tf^*(t)}]^2} \tag{17}$$

$$\overline{tf^*(t)}=\frac{tf(t)+\frac{tf(t)}{M}}{N+1}=\frac{(M+1)tf(t)}{M(N+1)} \tag{18}$$

其中，$tf(t)$表示候选术语 t 在整个测试语料中出现的总频率；$df(t)$表示候选术语 t 出现的文档频率；N 表示包含候选术语 t 的文档数；$tf_i(t)$表示候选术语 t 在第 i 篇文档中出现的频率；$\overline{tf^*(t)}$表示候选术语 t 修正后在整个语料中出现的平均频率；M 为语料中的文档数。最终抽取出的效果，较 C-value 方法准确率在 top100－top2000 上提升较为明显，有 10～20 个百分点的提升，top100 的准确率能达到 96%，召回率在 top1000－top5000 上基本与 C-value 方法持平。

结合语法规则和统计信息的术语自动抽取方法还有很多，但它们的思路大体相似，一般分为两步：首先，根据对语料以及术语做初步的统计，制定相应的语法规则，使用规则从语料中抽取出候选术语；然后，使用统计信息衡量候选术语的单元性和领域性，满足一定要求的候选术语则为术语。这类方法兼有基于语法规则和基于统计方法的优点，具有较强的语言无关性和领域无关性。

3.3.2 结合机器学习算法

隐马尔可夫模型(Hidden Markov Model，HMM)[35,36]是统计模型，用来描述一个含有隐含未知参数的马尔可夫过程，可以解决从观察序列中获得隐含参数的问题。

HMM 算法比较成熟，再加上易于训练等特点被广泛应用于自然语言处理领域。然而因为 HMM 是生成模型，尚存在许多问题：

(1)它没有直接对条件概率建模，而是采用联合概率代替，认为观察值由状态生成；

(2)HMM 独立性假设认为当前观察值仅与当前状态相关，忽略了其与之前状态和之前观察值之间的相关性。

HMM 上述缺点导致其并不是最适合用于解决序列标注问题的概率统计模型。

条件随机场(conditional random fields，CRFs)[17,37]是一种鉴别式概率模型，是随机场的一种，常用于标注或分析序列数据[44]。

相比于 HMM，CRFs 有如下优势：

(1)CRFs 能够在同一个模型中集成多个特征，特别是可加入长距离约束，能更好揭示语言学特征；

(2)CRFs 采用联合条件概率建模，避免了 HMM 的独立性假设和二元假设，从数学建模角度而言，CRFs 较 HMM 具有更可靠更合理的数学推导；

(3)HMM 是有向图模型，通过 Viterbi 算法搜索到当前对象为止的最佳路径，不考虑之后对象及其标记概率，而 CRFs 则采用无向图模型，是对整个标记序列求解联合概率，在整个序列范围内归一化，避免了因求解局部观察值概率所带来的标记偏置(Label Bias)问题。

HMM、CRFs 在词性标注、命名实体识别等领域表现出较好的效果。Agarwal[18]对最大熵模型(MaxEnt)、HMM、CRFs 算法在词性标注、分词和命名实体识别领域的使用进行了对比和分析，发现 CRFs 在 3 个领域中的效果均优于另外两种算法。

Zheng[19]最早使用 CRFs 模型进行术语抽取，以词语、词性以及词语的 TF-IDF 值作为特征进行术语抽取，取得了不错的成果，准确率达到 79.63%，召回率达到 73.54%，F 值达到 76.46%。

李丽双[38]使用 CRFs 方法抽取汽车领域术语。在总结了前人的工作和针对汽车领域术语特点的基础上选取了词本身、词性、词的长度、是否在词典中、窗口中词的词典特征、领域词频、背景语料词频、两类语料中的词频差以及在当前句子中与其他词的词频差的和这 9 个特征。从仅使用词本身、词性和词长度这 3 个特征开始不断添加新的特征进行术语的抽取。当使用到 6 个特征时效果最佳,精确率、召回率和 F 值分别达到 84.61%、80.50%和 82.50%,较仅使用词频、词性和词的长度特征的 CRFs 方法召回率和 F 值略有提升。

除了 CRFs 算法,Conrado[39]也对其他机器学习算法进行了尝试,比如决策树算法、朴素贝叶斯算法;并与词频、TF-IDF 和假设检验等方法进行比较,发现所用机器学习方法并没有较结合统计的方法有明显的提升。

结合机器学习的方法能够避免制定复杂的规则和公式,通用性更强。但由于需要大量的标注语料进行学习,其跨领域能力较弱。目前,该方法虽然取得了不错的效果,但是还不成熟,需要进行更多的尝试和验证。

3.3.3 基于术语部件的方法

吴云芳[40]首先提出了术语部件(term component)的概念,也就是构成术语能力较强的单词。通过考察 30000 条科技术语,提出了使用属性特征描述部件的方法,这些特征包含了术语表层和内部构成语言知识。

汤青[41]等提出基于部件扩展的本体术语抽取方法:首先,利用部件的领域聚合性和词性特征,基于词频比较的方法抽取部件;然后,综合考虑术语长度、术语词性构成以及术语内部结合度等因素,设计了基于部件的扩展规则,使用术语部件结合规则抽取候选术语;最后,利用候选术语的上下文构建候选术语的特征向量,将候选术语向量间的余弦值作为术语间的相似度,当与候选术语 t 最相近的 k 个术语的相似度都大于某个阈值 β 时,则 t 是术语。其最终效果比基于术语分布度、术语活跃度、术语主题度的多策略术语抽取方法准确率高出 2.5 个百分点。

基于术语部件扩展的方法由于无法抽取不包含术语部件的术语,对术语部件库的依赖较强。人工构建术语部件库费时费力,可移植性差;自动构建术语部件库的方法还不成熟。因此,术语部件自动获取和基于部件扩展的算法都需要进一步的研究,从而提高术语抽取的效果。

总的来说,多种方法相结合能够形成互补,兼有各种方法的优势,避免单一方法的缺陷,在研究中取得了较好的效果。因此,目前多种方法相结合的术语自动抽取研究成为趋势。

4 研究展望

总体看来,现有的方法相比于研究早期已经有了很大的进步,其中部分方法已经取得了不错的效果,有一定实际应用价值。但是,现有术语自动抽取技术还不够成熟,未来相关研究工作可以围绕以下两个方面开展。

(1)术语抽取理论体系的完善

术语抽取理论体系的完善包括但不限于以下两个方面:评价指标、语料选取和效果评价方法的完善。

目前研究中所使用的评价指标主要是本文 2.3 节所提到的准确率、召回率和 F 值等几种统计学指标。这些指标缺乏语义上的考量,如是否存在歧义、同义词等,无法对抽取出的术语质量进行全面的评价。

Piao[32]在研究中发现一些术语抽取方法在不同的领域语料上表现并不稳定,即在某领域语料上表现较好的术语抽取方法可能在其他领域语料上表现较差。而目前大多数术语自动抽取研究自成体系,使用的语料和评价的方法各不相同,研究中所取得的成果很难有说服力。

因此,为了术语抽取相关研究能够更好地发展,必须完善术语抽取的基础理论体系。

(2)术语抽取方法的研究

无论早期的从基于规则的方法到基于统计的方法,还是后来从基于规则、统计相结合到机器学习算法的引入,术语自动抽取技术都有很大的进步,特别是融入机器学习算法的研究已经取得了一定的成果。但是,现有术语自动抽取方法的通用性以及效果还不够理想,有很大的上升空间,对此还需要进一步的探索和研究。以提高术语抽取自动化程度和术语抽取的质量为目标,未来抽取方法相关研究工作主要有:在现有的方法基础上进行改进;借鉴其他领域的成功经验,不断探索和研究新的方法;针对具体的应用领域或场景进行针对性的研究。

结束语 随着大数据时代的来临,文本处理相关技术迅速发展,作为文本处理中基础性工作的术语抽取研究变得极为迫切。因此,深入研究术语的自动抽取方法具有重大意义。本文在简单介绍了术语的定义、术语的特性和术语抽取效果的评价方法的基础上,将已有的术语自动抽取方法分成 3 类进行了介绍、分析和比较,并对其中较为经典的公式和算法进行了详细的说明。

参 考 文 献

[1] Brewster C A, Iria J, Zhang Z, et al. Dynamic Iterative Ontology Learning[C]// Recent Advances in Natural Language Processing (RANLP'07). 2007

[2] Wolf P, Bernardi U, Federmann C, et al. From Statistical Term Extraction to Hybrid Machine Translation[C]// 15th International Conference of the European Association for Machine Translation. 2011:225

[3] Liang Y H, Li J, Ye L, et al. The Chinese Unknown Term Translation Mining with Supervised Candidate Term Extraction Strategy[J]. Procedia Engineering, 2011, 15:1388-1392

[4] Pavlopoulos J, Androutsopoulos I. Aspect Term Extraction for Sentiment Analysis: New Datasets, New Evaluation Measures and an Improved Unsupervised Method[C]// Proceedings of the 5th Workshop on Language Analysis for Social Media(LASM). 2014:44-52

[5] Bhagdev R, Butters J, Chakravarthy A, et al. Doris: Managing Document-based Knowledge in Large Organisations via Semantic Web Technologies[C]// Semantic Web Challenge. 2007

[6] Kozakov L, Park Y, Fin T, et al. Glossary extraction and utilization in the information search and delivery system for IBM Technical Support[J]. IBM Systems Journal, 2004, 43(3):546-563

[7] Sager J C, Dungworth D, McDonald P F. English special languages: principles and practice in science and technology[M]. Wiesbaden: Brandstetter, 1980

[8] 冯志伟. 现代术语学引论[M]. 北京:语文出版社, 1997
Feng Zhi-wei. An introduction to modern terminology[M]. Beijing: Language and Literature Press, 1997

[9] 术语工作原则与方法:GB/T 10112-1999[S]. 北京:中国标准出版社,2000
Terminology work principles and methods: GB/T 10112-1999 [S]. Beijing: Standards Press of China, 2000

[10] Kageura K, Umino B. Methods of automatic term recognition: A review[J]. Terminology, 1996, 3(2): 259-289

[11] Vivaldi J, Rodríguez H. Evaluation of terms and term extraction systems: A practical approach[J]. Terminology, 2007, 13(2): 225-248

[12] Zheng Y, Dou W, Wu G, et al. Automated Chinese domain ontology construction from text documents[M]// Bio-Inspired Computational Intelligence and Applications. Springer Berlin Heidelberg, 2007: 639-648

[13] Korkontzelos I, Klapaftis I P, Manandhar S. Reviewing and evaluating automatic term recognition techniques[M]// Advances in Natural Language Processing. Springer Berlin Heidelberg, 2008: 248-259

[14] Castellví M T C, Bagot R E, Palatresi J V. Automatic term detection: A review of current systems[M]// Recent advances in computational terminology. 2001: 53-88

[15] NLPIR 汉语分词系统[EB/OL]. http://ictclas.nlpir.org, 2014
NLPIR Chinese segmentation system[EB/OL]. http://ictclas.nlpir.org, 2014

[16] Eddy S R. Hidden markov models[J]. Current opinion in structural biology, 1996, 6(3): 361-365

[17] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]. 2001

[18] Agarwal M, Goutam R, Jain A, et al. Comparative Analysis of the Performance of CRF, HMM and MaxEnt for Part-of-Speech Tagging, Chunking and Named Entity Recognition for a Morphologically rich language[C]// Proceedings of the Pacific Association For Computational Lingustics(PACLING2011). 2011

[19] Zheng D, Zhao T, Yang J. Research on domain term extraction based on conditional random fields[M]// Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy: ICCPOL. Springer Berlin Heidelberg, 2009: 290-296

[20] Li L S, Dang Y Z, Zhang J, et al. Domain Term Extraction Based on Conditional Random Fields Combined with Active Learning Strategy[J]. Journal of Information & Computational Science, 2012, 9(7): 1931-1940

[21] Voutilainen A. NPtool, a detector of English noun phrases[C]// Proceedings of the Workshop on Very Large Corpora Columbus. Ohio: Ohio State University, June 1993

[22] Park Y, Byrd R J, Boguraev B K. Automatic glossary extraction: beyond terminology identification[C]// Proceedings of the 19th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2002: 1-7

[23] Evans D A, Lefferts R G. Clarit-trec experiments[J]. Information processing & management, 1995, 31(3): 385-395

[24] Bolshakova E, Loukachevitch N, Nokel M. Topic models can improve domain term extraction[M]// Advances in Information Retrieval. Springer Berlin Heidelberg, 2013: 684-687

[25] Velardi P, Missikoff M, Basili R. Identification of relevant terms to support the construction of domain ontologies[C]// Proceedings of the workshop on Human Language Technology and Knowledge Management-Volume 2001. Association for Computational Linguistics, 2001: 5

[26] Daille B. Study and implementation of combined techniques for automatic extraction of terminology[M]// The balancing act: Combining symbolic and statistical approaches to language. MIT Press, Cambridge, 1996, 1: 49-66

[27] Gelbukh A, Sidorov G, Lavin-Villa E, et al. Automatic term extraction using log-likelihood based comparison with general reference corpus[M]// Natural Language Processing and Information Systems. Springer Berlin Heidelberg, 2010: 248-255

[28] Cohen J D. Highlights: Language-and domain-independent automatic indexing terms for abstracting[J]. Journal of the American Society for Information Science, 1995, 46(3): 162-174

[29] Drouin P. Term extraction using non-technical corpora as a point of leverage[J]. Terminology, 2003, 9(1): 99-115

[30] Dunning T. Accurate methods for the statistics of surprise and coincidence[J]. Computational linguistics, 1993, 19(1): 61-74

[31] Frantzi K, Ananiadou S. The C-value/NC-value domain independent method for multi-word term extraction[J]. Journal of Natural Language Processing, 1999, 6(3): 20-27

[32] Piao S, Forth J, Gacitua R, et al. Evaluating tools for automatic concept extraction: A case study from the musicology domain [C]// Proceedings of Digital Futures. 2010

[33] Ventura J A L, Jonquet C, Roche M, et al. Combining C-value and Keyword Extraction Methods for Biomedical Terms Extraction[C]// International Symposium on Languages in Biology and Medicine(LBM'2013). 2013: 45-49

[34] 周浪,张亮,冯冲,等. 基于词频分布变化统计的术语抽取方法[J]. 计算机科学,2009,36(5):177-180
Zhou Lang, Zhang Liang, Feng Chong, et al. Terminology Extraction Based on Statistical Word Frequency Distribution Variety [J]. Computer Science, 2009, 36(5): 177-180

[35] Eddy S R. Hidden markov models[J]. Current opinion in structural biology, 1996, 6(3): 361-365

[36] Wikipedia. Hidden_Markov_model[EB/OL]. http://en.wikipedia.org/wiki/Hidden_Markov_model, 2014

[37] Wikipedia. Conditional_random_field[EB/OL]. http://en.wikipedia.org/wiki/Conditional_random_field, 2014

[38] 李丽双,党延忠,张婧,等. 基于条件随机场的汽车领域术语抽取[J]. 大连理工大学学报,2013,53(2):267-272
Li Li-shuang, Dang Yan-zhong, Zhang Jing, et al. Automotive Term Extraction Based on Conditional Random Fields[J]. Joural of Dalian University of Technology, 2013, 53(2): 267-272

[39] da Silva Conrado M, Pardo T, Rezende S O. A Machine Learning Approach to Automatic Term Extraction using a Rich Feature Set[C]// HLT-NAACL. 2013: 16-23

[40] 吴云芳,穗志方,邱利坤,等. 信息科学与技术领域术语部件描述[J]. 语言文字应用,2003(4):34-39
Wu Yun-fang, Sui Zhi-fang, Qiu Li-kun, et al. The Approaches and Strategies to Describe the Term Component in Information Science and Technology[J]. Applied Linguistics, 2003(4): 34-39

[41] 汤青,吕学强,李卓,等. 领域本体术语抽取研究[J]. 现代图书情报技术,2014(1):43-50
Tang Qing, Lv Xue-qiang, Li Zhuo, et al. Research on Domain Ontology Term Extraction[J]. New Technology of Library and Information Service, 2014(1): 43-50