●朱毅华　侯汉清　沙印亭

计算机识别汉语同义词的两种算法比较和测评

摘　要　以词素为单位进行相似度计算,在许多方面解决了字面相似度算法中不合理的问题,但汉语分词、词义分解、词素分类体系及其编码问题又亟待解决。基于语义的词素相似度算法优于基于单汉字的字面相似度算法,是今后的研究重点。表 4。参考文献 9。

关键词　中文信息处理　同义词识别　模式识别　词素　字面相似度匹配
分类号　G252.7

ABSTRACT　In this paper, the authors point out the advantages and disadvantages of the similarity computing by word elements. They think that the word-element-based similarity algorithm is better than the single-character-based similarity algorithm. 4 tabs. 9 refs.

KEY WORDS　Chinese information processing.　Recognition of synonyms.　Pattern recognition.　Word element.　Matching of word similarity.
CLASS NUMBER　252.7

随着计算机技术的飞速发展,自然语言越来越多地应用于信息检索领域,这可以说是当代检索语言发展最重要的特征[1]。然而,汉语中自然语言对概念表达灵活、自由的特点,为情报界学者提出了更多研究课题。目前,利用计算机辅助识别自然语言中存在的大量的同义词是全文检索、网络检索中亟待解决的课题之一。其中,自动识别中文同义词的测定方法主要有两种:以单汉字(即语素)为单位的字面相似度测定;以词素为单位的字面相似度测定。

1　语言学和情报学中的同义词

同义词在语言学、情报学中都存在,但其含义并不相同。汉语语言语气、语词结构的细微变化都可能造成词义的变化。因此,在语言学领域的同义词概念是比较狭窄的,可以理解为意义相同或相近的词[2]。这是严格意义上的同义词。在情报学中同义词的概念是比较宽泛的,所有能代表某一主题概念的语词都可以作为广义同义词。这里的同义词具有可替代性和模糊性,在布尔检索式中可以进行逻辑或运算,在自然语言检索中旨在提供更多的检索入口,对自然语言进行控制,实现概念检索。因此,情报学中同义词概念并不等同于语言学中和日常生活中的同义词概念。

2　基于单汉字的字面相似度算法

利用字面相似度识别同义词是将汉语的构词特征和计算机检索技术两者结合起来,是计算机辅助识别同义词的有效方法之一。字面相似度算法主要根据字面相似性原理,即汉语中绝大多数同义词、准同义词都含有相同语素这一突出特点,计算词与词之间的关联程度。王源等首先讨论了通过字面相似匹配的方法查找新术语与主题词的相关关系,并提出了在计算中同时考虑匹配字数和词汇结构两方面的因素[3]。宋明亮主要利用汉语词汇字面相似性原理进行词汇归类[4]。吴志强提出的加权相似度算法是在前面两种思想基础上又将汉语构成的"重心后移"原理加入其中,成为一种比较有代表性的算法[5]。其主要内容如下:

汉语词汇构成具有"重心后移"的特点。表达某一具体专指概念的语词,其主题中心即中心词往往在词的后半部分。在字面上语素越靠后,它在表达主题概念中所起的作用越大。基于此,对语词中各个语素表达主题概念的作用进行量化,作一加权处理:语素越靠前,作用越小,权值也越小,语素越靠后,作用越大,权值也越大。

对待匹配词甲与匹配词乙进行字面相似性分析,统计两个词中共同含有的相同语素的个数,同时对词甲和词乙中各个语素加权。根据相同语素在两词中的位置及次序,统计相同语素在各个词中所占的权值,拟定算法计算两词的相似度,确定其同义、准同义关系。设定两个词之间的相似度受两方面的影响:两个词含有相同语素的个数的影响占 60 %;相同语素在各个词中的位置关系的影响占 40 %。由此拟定相似度匹配公式:

$$xsd = 60\% \times \frac{\left[\frac{xsword}{ctrlword} + \frac{xsword}{keyword}\right]}{2} + 40\% \times dp \times \frac{\left[\sum\frac{c\_xsword(i)}{\sum ctrlword(i)} + \sum\frac{k\_xsword(i)}{\sum keyword(i)}\right]}{2}$$

其中:$xsword$ 表示两词含有相同语素即匹配字的个数;

$\sum\frac{k\_xsword(i)}{\sum keyword(i)}$ 表示匹配字在待匹配词中所处位置的权数之和;

$\sum\frac{c\_xsword(i)}{\sum ctrlword(i)}$ 表示匹配字在待匹配词中所处位置的权数之和;

dp 表示位置系数,其值为被匹配词与待匹配词语素总数之比,如果被匹配词语素总数大于待匹配词语素总数,dp = keyword/ ctrlword,反之则为 dp = ctrlword/ keyword。

该相似度计算公式是通过计算匹配字串与被匹配词和待匹配词的比例的算术平均数,以及匹配字串在被匹配词和待匹配词的位置次序关系的权数之和的算术平均数,然后分别乘以两个影响的加权数,最终得出两词的相似度。例如,"经济信息管理"与"商业信息管理"的相似度:

xsword = 4; ctrlword = 6; keyword = 6;

$$\sum\frac{c\_xsword(i)}{\sum ctrlword(i)} = \frac{3+4+5+6}{1+2+3+\cdots+6} = \frac{6}{7}$$

$$\sum\frac{c\_xsword(i)}{\sum keyword(i)} = \frac{3+4+5+6}{1+2+3+\cdots+6} = \frac{6}{7}$$

$$dp = \frac{ctrlword}{keyword} = \frac{6}{6} = 1$$

$$xsd = 60\% \times \left[\frac{4}{6} + \frac{4}{6}\right] \div 2 + 40\% \times 1 \times \left[\frac{6}{7} + \frac{6}{7}\right] \div 2 = 74.3\%$$

两词的相似度为 74.3%;其他例子见表 1。

表 1　基于单个汉字的字面相似度算法计算实例

词对	A 股	TNT	西红柿	中华人民共和国	流动偏好	财会制度
	B 股	梯恩梯	番茄	中国	灵活偏好	财务制度
相似度	0.57	0	0	0.46	0.9	0.77

3　基于语义的词素相似度识别算法

字面相似度算法只适用于识别由纯汉字构成的词汇,不适用于识别纯粹由非汉字组成的词汇。因此开始有人致力于研究以词素为单位识别同义词的方法。汉语词素分析首先应用在"《军用主题词表》应用管理系统"中。后来查贵庭也做了一些分析与应用[6]。朱毅华则系统论述了这种以语素为单位的基于语义的同义词识别算法[7]。

词素相似度识别算法的主要思想是:

首先,建立常用词素的语义词典,对识别词进行切分,在此基础上以词素为单位,以相似性原理为依据,将词素的字面形式转换为语义代码进行相似度判别,在考虑词组的结构关系的前提下进行同义词的识别。其中引入了表达度这一概念,表示词的部分对整体的涵义所起的作用大小,据此进行加权。

公式的成立首先假设以下条件为已知:

待匹配词 ctrlword 的信息量总和为 A;

匹配词 keyword 的信息量总和为 B;

两词中表示相同语义的信息量为 C_1,C_2;

共同部分 C_1 对 A 的表达度为 x,C_2 对 B 的表达度为 y。

根据这些条件可得:

$$x = \frac{C_1}{A},\ y = \frac{C_2}{B} \qquad (C_1 = C_2 = C)$$

则相似度:

$$xsd = \frac{2}{\frac{1}{x} + \frac{1}{y}} \qquad (x, y \text{ 不为 } 0)$$

例如计算"经济信息管理"与"商业信息管理"两词的相似度:

(1)词素切分。分别将两词切成词素,"经济信息管理"切分为"经济"、"信息"、"管理";"商业信息管理"切分为"商业"、"信息"、"管理"。

(2)找到相同部分为"信息"与"管理"。

(3)权重计算:仍然使用传统算法中的重心后移原理分配权重,但单位换为词素,则"信息"在"经济信息管理"中的权重为 2,"管理"为 3;

"信息"在"商业信息管理"中的权重也为 2,"管理"为 3;

$$x = \sum\frac{c\_xsword(i)}{\sum ctrlword(i)} = \frac{2+3}{1+2+3} = \frac{5}{6}$$

$$y = \sum\frac{c\_xsword(i)}{\sum keyword(i)} = \frac{2+3}{1+2+3} = \frac{5}{6}$$

$$xsd = \frac{2}{\frac{1}{x} + \frac{1}{y}} = \frac{2}{\frac{6}{5} + \frac{6}{5}} = \frac{5}{6} = 83.33\%$$

两词的相似度为 83.33%,其他例子见表 2。

表 2　基于语义的词素相似度算法计算实例

词对	A 股	TNT	西红柿	中华人民共和国	流动偏好	财会制度
	B 股	梯恩梯	番茄	中国	灵活偏好	财务制度
相似度	1	1	1	1	1	0.83

具体做法是建立以词素为单位的语义词典,将词素按语义上的分类体系进行相似比较,再将组成语词的各个词素相似度按一定的权重计算出表达度,再通过两词的表达度计算出相似度。

语义的比较必须以义原、而不是词素作为比较单位。义原的集合根据其相互间的联系,可以组成一个语义体系。语义体系中同一分支节点或相邻节点的范畴具有相同或相似的含义,从而使得通过词素的语义代码进行比较成为可能。如"中央银行"(专有名词,不能分解)对应的义原串为"M14. 3. 1-M14. 6. 03. 2. 1-M14. 1","M14. 3. 1"指机构范畴,"M14. 6. 03. 2. 1"指金融范畴,"M14. 1"指国家范畴。

义原的比较原则上是计算类号相同节点占平均节点长度的比值。对于类号"M14. 6. 03"和"M14. 6. 04",其相关度为:相同节点数 2/平均节点长度(3 + 3) / 2 = 2/ 3 = 66. 7 %。

4　两种同义词识别算法的比较

(1)以单汉字为单位的字面相似度算法具有直观、简单、易行的特点。尤其在自动切分问题上,单汉字的切分比词素切分简单得多,而且切分速度较快。基于语义的词素相似度算法虽然在实现效率上大大优于字面相似度算法,但是基于一定的条件。如果没有一个完善的分类体系,没有一部收词丰富、编制精良的语义词典,没有一种较好的切词算法等基本条件,其出错率也会很高。

(2)自然语言中存在着大量的学名与俗称、新名与旧称、全称与简称、不同的译名,实指同一问题的反义词和否定词以及两种语言的等价词。这些词在字面相似度算法中都会因各种的变化,对计算结果产生不同影响。如"INTERNET"与"因特网"或"国际互联网",从概念角度看相似度为 100 %,而以字面相似度算法计算相似度却为 0。由于汉字数量繁多,并且一字多义现象很多,容易产生歧义,单汉字字面相似度测定必定会影响到系统的实现效率[8]。以词素为单位来计算同义词相似度则有效地弥补了以单汉字为单位带来的大部分不足。

(3)字面相似度算法不适用于识别由非纯汉字组成的词汇。然而,从《汉语主题词表(自然科学增订本)》字顺表中抽取的 12364 个同义词中,纯汉字组成的非叙词和叙词(如:波粒子-光子,标准地球-地球模型等)在所有《汉表》同义词中所占比重约为 80 %。纯汉字及非纯汉字混合组成的非叙词和叙词(如:T 触发器-计数式触发器,VA 族元素-氧族元素等)约占 10 %。另外纯罗马字母或纯数字组成的非叙词和叙词(如:1059、APDC、ADH、gA/ gV 等)数量较少一些[9]。后两部分同义词字面相似度数值很低,影响了识别率。然而使用基于语义的词素相似度算法可以根据语义把它们识别出来。

(4)字面相似度算法中主观因素较多,如计算公式中的 60 %、40 %的取值以及 dp 在公式中所起的作用都没有经过统计分析,缺乏理论依据,因此影响了同义词识别的质量。然而基于语义的词素相似度算法在数学上意义非常明确,去掉了 dp、60 %及 40 %等主观因素,计算公式相当简单。

4　系统的测试与结果分析

4. 1　实验数据来源与实验方法

(1)封闭式实验:数据来源于《社会科学检索词表》中经济类的 Y、D 项构成的同义词及《现代汉语分类词典》中的经济类同义词共 914 条。实验方法是利用朱毅华研究的同义词识别系统进行计算机识别。

(2)开放式实验:数据来源于《重庆库》光盘中下载的 F83 金融类标引关键词及部分封闭实验语料混合而成的同义词与非同义词共 651 条。实验方法是首先利用封闭实验的系统计算相似度,然后人工判别其同义词的正确识别率及误识别率。

4. 2　实验结果及分析

表 3　封闭式实验数据测试(前 500 条)

方　法	阈值 ≥0. 33		阈值 ≥0. 50		阈值 ≥0. 66	
	词数	比例 %	词数	比例 %	词数	比例 %
使用词素分析法	486	97. 2	452	90. 4	376	75. 2
使用字面相似度	439	87. 8	331	66. 2	185	35

表 4　开放式实验数据测试(前 500 条)

方　法	阈值 ≥0.33				阈值 ≥0.5				阈值 ≥0.66			
	计算机识别总词数	误识别词数	误识别率%	正确识别率%	计算机识别总词数	误识别词数	误识别率%	正确识别率%	计算机识别总词数	误识别词数	误识别率%	正确识别率%
使用词素分析法	464	318	68.5	31.5	191	54	28.3	71.7	124	1	0.08	99.92
使用字面相似度	448	316	70.5	29.5	275	176	64	36	60	7	11.7	88.3

封闭式实验中,采用的数据为《社会科学检索词表》中经济类的 Y、D 项构成的同义词及《现代汉语分类词典》中部分同义词,因此数据可靠性较高,实验结果比较有说服力。

从表 3 中各项统计数据可以看出,无论是阈值定为 0.33、0.5 或是更高的 0.66,结果均是使用词素分析法时同义词正确识别量比采用字面相似度算法要高。在阈值提高到 0.66 时,词素分析法的识别性能更好。而且随着阈值的提高,两者的差别越大。当阈值等于 0.33 时,采用词素分析法识别出的同义词数为 486,而采用字面相似度算法识别出的同义词数为 439,两者相差 47 个;当阈值等于 0.5 时,两者相差 121 个;当阈值等于 0.66 时,两者则相差 191 个。这说明词素分析法确定词汇同义关系效率比字面相似度算法效率要高。

在开放式实验中,数据来源于《重庆库》光盘中下载的金融类词汇(非同义词),且混合了封闭实验中语料的部分同义词,因此作为此项实验的语料也比较具有代表性。

据表 4 中所得数据可知,使用字面相似度算法在开放性实验中的出错率较高,而且随着阈值的提高,误识别率也逐渐提高:当阈值等于 0.33 时,误识别率为 70.5%;当阈值等于 0.5 时,误识别率为 64%;当阈值等于 0.66 时,误识别率为 11.7%。虽然它是呈下降趋势的,但一直比使用词素分析法时的误识别率高。而且当阈值提高到 0.66 时,使用词素分析法的误识别率仅为 0.08%,达到很低,趋向于零。

5　结语

以词素为单位进行相似度计算确实在许多方面解决了字面相似度算法中不合理的问题,但是汉语分词、语义分解、词素分类体系及其编码问题又悄然而生,亟待解决。汉语自动分词是任何中文自然语言处理系统都难以回避的第一道"工序"。至于汉语中存在大量兼类词,影响词性标注的准确性,进而影响语义标引结果,也会造成同义词识别出错。

总而言之,基于语义的词素相似度算法明显优于基于单汉字的字面相似度算法。当然它仍然存在一些问题有待解决。今后,基于语义的词素相似度算法是研究的重点,简单易行的、基于单汉字的字面相似度算法则会作为一种辅助方法继续存在。

参考文献

1　侯汉清. 新闻信息数据库后控词表的设计和编制. 江苏图书馆学报,2000(1)
2　现代汉语词典. 北京:商务印书馆,1994
3　王源等. 后控规范的计算机处理、现代图书情报技术,1993(2)
4　宋明亮. 报纸文献机助自由标引研究及对后控制词表动态维护的思维. 张琪玉指导,硕士论文,空军政治学院,1994,6
5,8　吴志强. 经济信息检索后控制词表的研制. 侯汉清指导,硕士论文,南京农业大学,1999,6
6　查贵庭. 经济新闻自动标引系统的研究. 侯汉清指导,硕士论文,南京农业大学,2000,6
7　朱毅华. 智能搜索引擎中同义词识别算法的研究. 侯汉清指导,硕士论文,南京农业大学,2001,6
9　李朝阳,侯汉清. 汉语科技同义词字面相似度测试. 理论学术年刊,1998

朱毅华　江苏南京农业大学情报系讲师。通讯地址:江苏南京卫岗。邮编 210095。
侯汉清　江苏南京农业大学情报系教授。通讯地址同上。
沙印亭　苏州大学图书馆工作。通讯地址:江苏苏州。邮编 215000。

(来稿时间:2002-01-17)