基于语义依存的汉语句子相似度计算*

李　彬, 刘　挺, 秦　兵, 李　生

(哈尔滨工业大学 计算机科学与技术学院 智能内容管理实验室, 黑龙江 哈尔滨 150001)

摘　要: 句子间相似度的计算在自然语言处理的各个领域都占有很重要的地位, 在多文档自动文摘技术中, 句子间相似度的计算是一个关键的问题。由于汉语句子的表达形式是多种多样的, 要准确地刻画一个句子所表达的意思, 必须深入到语义一级并结合语法结构信息, 由此提出了一种基于语义依存的汉语句子相似度计算的方法, 该方法取得了令人满意的实验效果。

关键词: 相似度计算; 语义; 依存结构; 自然语言处理; 多文档文摘

中图法分类号: TP301.6　　文献标识码: A　　文章编号: 1001-3695(2003)12-0015-03

Chinese Sentence Similarity Computing Based on Semantic Dependency Relationship Analysis

LI Bin, LIU Ting, QIN Bing, LI Sheng

(*Intelligent Content Management Laboratory, College of Computer Science & Technology, Harbin Institute of Technology, Harbin Heilongjiang 150001, China*)

Abstract: Sentence similarity computation is very important in all the fields of Natural Language Processing. In Multi-document Summarization Technology, sentence similarity computation is a key problem. As we know, a Chinese sentence can be presented by many kinds of style, if we want to describe what a sentence means, we should dip into the semantic level and consider about the dependency structure. In this paper, we applied a method that based on semantic dependency relationship analysis to compute sentence similarity, and the experiment result of this method is satisfied.

Key words: Similarity Computation; Semantics; Dependency Structure; Natural Language Processing; Multi-document Summariation

1　引言

相似度是一个很复杂的概念, 在语义学、哲学和信息理论中被广泛地讨论。在不同的具体应用中, 其含义有所不同。例如, 在基于实例的机器翻译中, 相似度主要用于衡量文本中词语的可替换程度; 而在信息检索中, 相似度更多的是反映文本与用户查询在意义上的符合程度, 在自动问答中, 相似度反映的是问题与答案的匹配程度; 而在多文档文摘系统中, 相似度可以反映出局部主题信息的拟合程度。我们把句子间的相似度定义为一个在(0, 1)之间的数值, 0 代表两个句子不相似, 1 代表两个句子完全相似, 两个句子之间的相似度的值越大表示它们就越相似。现在国内外有很多学者在研究句子间相似度的计算, 如哥伦比亚大学的 Goldsdein 等人通过最大边缘相关的方法(Maximal Marginal Relevance)进行相似度计算[1], 学者 Chis H.Q.Ding 等人采用了隐含语义索引(Latent Semantic Indexing)的方法[2], 国内有学者利用骨架依存的方法计算汉语句子间相似度[3]。

句子间相似度的计算在自然语言处理的各个领域都占有很重要的地位, 在基于实例的机器翻译、自动问答和多文档文摘系统中, 语句相似度的计算是一个关键问题, 而语句相似度的衡量机制与对语句的分析深度是密切相关的。在相似度计算中, 按照对语句的分析深度来看, 主要存在两种方法: ①基于向量空间模型的方法。该方法把句子看成词的线性序列, 不对语句进行语法结构分析, 相应的语句相似度衡量机制只能利用句子的表层信息, 即组成句子的词的词频、词性等信息。由于不加任何结构分析, 该方法在计算语句之间的相似度时不能考虑句子整体结构的相似性。②对语句进行完全的句法与语义分析, 这是一种深层结构分析法, 对被比较的两个句子进行深层的句法分析, 找出依存关系, 并在依存分析结果的基础上进行相似度计算。本文采用的基于语义依存的方法也是属于第二种。

2　基于语义依存相似度计算方法

TF-IDF 是信息检索领域常用的方法, 一般来说能够产生较好的效果。但由于 TF-IDF 方法只考虑了词在上下文中的统计特性, 而没有考虑词本身的语义信息。如"我爱吃土豆"和"我爱吃马铃薯"这两个句子所表达的意思应该是完全相同, 因为"土豆"和"马铃薯"在语义上是等价的, 由于 TF-IDF 没有考虑到这种语义信息, 因此它具有一定的局限性。并且汉语句子的表达形式是多种多样的, 如果要准确地刻画一个句子所表达的意思, 还应该结合语法结构信息。由此我们引入了基于语义

收稿日期: 2002-10-18

基金项目: 国家自然科学基金资助项目(60203020); 哈尔滨工业大学校基金资助项目(HIT.2000.50)

依存结构的相似度计算方法。

2.1 知网简介

计算语义的相似度，需要一定的语义知识资源作为基础。在汉语中，人们常用董振东和董强先生创建的知网(HowNet)作为语义知识资源。知网是一个以概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库，它是一个网状的有机的知识系统[4]。

语义词典是知网的基础文件，在这个文件中每一个词语的概念及其描述形成一个记录，每一个记录都主要包含四项内容。其中每一项都由两部分组成，中间以“=”分隔。每一个“=”的左侧是数据的域名，右侧是数据的值。它们排列如下：

NO. =词或短语序号　W_X=词或短语

G_X=词或短语的词性　E_X=词或短语的例子

DEF=概念定义

其中，W_X，G_X，E_X构成每种语言的记录，X用于描述记录所代表的语种，X为C则为汉语，为E则为英语。每个词语由DEF来描述其概念定义，DEF的值由若干个义原及其与主干词之间的语义关系描述组成，义原是知网中最基本的、不易于再分割的意义的最小单位。

2.2 语义消歧

为了计算句子之间的语义相似度，首先要确定句子中的词在这个句子中所表达的语义。如“打毛衣”中的“打”作为“编织”的意思，而“打酱油”中的“打”作为“买”的意思，所以我们需要确定“打”这个词在不同的句子中的不同含义，这一步工作称为语义消歧。语义消歧(WSD)是解决如何在给定上下文语境中确定多义词的义项(Sense)的问题。我们用的是哈尔滨工业大学计算机科学与技术学院智能内容管理实验室所做的语义消歧系统。目前该系统在开放测试下准确率能够达到91.89%，封闭测试准确率能够达到98.67%。该系统能够对经过分词和词性标注后的句子进行语义消歧，并在每个词后面标注上相应的语义号。例如对于句子：“哈尔滨/nd 在/p 什么/r 地方/ng？/wj”，经过语义消歧后变为：“哈尔滨/17 在1 269 什么/468 地方/17？/-1”。

每个语义号都对应知网中的一个义原。例如，17对应的义原为“place|地方”，1 269对应的义原为“{location}”，468对应的义原为“aValue|属性值，kind|类型”，-1表示在知网中找不到这个词(如“公转”)或者这个词是没有价值的语义信息(如标点符号)。

对于上面所说的“打酱油”中的“打”，语义号为348(buy|买)，“打毛衣”中的“打”语义号为525(weave|编织)。由此可以看出，语义消歧能够挖掘出一个词在上下文中确切的含义，而不是仅仅停留在词的表面。

2.3 依存文法的定义

依存文法是由法国语言学家L Tesniere在其著作《结构句法基础》(1959年)中提出，对语言学的发展产生了深远的影响，特别是在计算语言学界备受推崇。依存语法通过分析语言单位内成分之间的依存关系揭示其句法结构，主张句子中核心动词是支配其它成分的中心成分，而它本身却不受其它任何成分的支配，所有受支配成分都以某种依存关系从属于支配者[5]。在20世纪70年代，Robinson提出依存语法中关于依存关系的四条公理，在处理中文信息的研究中，中国学者提出了依存关系的第五条公理[6]：①一个句子中只有一个成分是独立的；②其它成分直接依存于某一成分；③任何一个成分都不能依存于两个或两个以上的成分；④如果A成分直接依存于B成分，而C成分在句中位于A和B之间，那么C或者直接依存于B，或者直接依存处于A和B之间的某一成分；⑤中心成分左右两边的其它成分相互不发生关系。

句子成分间相互支配与被支配、依存与被依存的现象普遍存在于汉语的词汇(合成语)、短语、单句、复合直到句群的各级能够独立运用的语言单位之中，这一特点为依存关系的普遍性[7]。依存句法分析可以反映出句子中各成分之间的语义修饰关系，它可以获得长距离的搭配信息，并与句子成分的物理位置无关。

2.4 句子依存结构的建立

利用依存结构计算句子间的相似度，关键的一步是如何获得句子各成分间的依存关系信息。在此，采用了哈尔滨工业大学计算机科学与技术学院智能内容管理实验室所做的依存句法分析器。目前该分析器对依存弧的标记准确率能达到86%以上。通过该依存句法分析器的分析，句子各成分之间的依存关系如图1所示。

例句：爱因斯坦是一位当代杰出人才。

我们把该结果形成立体结构的依存树(图2)。

有了句子的依存结构信息，就可以用它来计算句子间的相似度了。

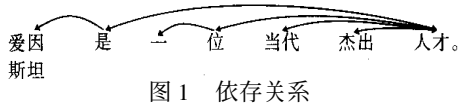

图1 依存关系

2.5 相似度计算

我们知道，依存树是一个复杂的非线性关系，如果对整个依存树进行完全匹配的话，所花费的代价是巨大的；另外，一个完整的汉语句子是由句子的关键成分和修饰成分所构成，而人们往往从关键成分就可以了解一个句子的大概意思。但由于汉语表达形式的多样性，相同的关键成分可用不同的修饰成分来修饰，如果强调修饰成分，这无疑会给句子间相似度的计算增加噪音。基于以上两点，在利用依存结构进行相似度计算时，只考虑那些有效搭配对之间的相似程度。所谓有效搭配对是指全句核心词和直接依存于其有效词组成的搭配对，这里有效词定义为动词、名词以及形容词，它是由分词后的词性标注决定的。例如以下两个句子间的比较：

例句1　事发后，伤员被及时送往就近医院救治。

例句2　晚上7时左右，所有伤员被送到了医院。

从图3和图4中可以看出，图中标记为斜体的词就可以代表各自句子的主要意思，所以句子1的有效搭配对为：送往__伤员、送往__医院、送往__救治。句子2的有效搭配对为：送到__伤员、送到__医院。我们只要比较它们之间的相似程度即可，这样一来比较算法的复杂度就大大降低了，而准确率也会得到一定程度的提高。相似度计算公式如下：

$$SIM(Sen1, Sen2)=\frac{\sum_{i=1}^{n} Wi}{Max\{PairCount1, PairCount2\}} \quad (1)$$

式(1)中, $\sum_{i=1}^{n} Wi$ 为句子 1 和句子 2 有效搭配对匹配的总权重, PairCount1 为句子 1 的有效搭配对数, PairCount2 为句子 2 的有效搭配对数。

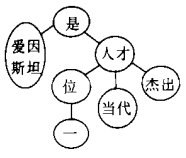

图 2　依存树

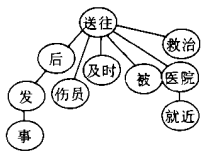

图 3　例句 1 的依存树

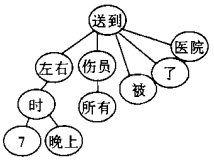

图 4　例句 2 的依存树

在此算法中, 搭配对的匹配权重被定义为:

假设有两个搭配对: ①Word1 __ Word2; ②Word1′__ Word2′。如果 Word1 = Word1′且 Word2 = Word2′则搭配对①和搭配对②的匹配权重为 1; 如果 Word1 ≠Word1′但 Word2 = Word2′或者 Word1 = Word1′但 Word2 ≠Word2′, 则搭配对①和搭配对②的匹配权重为 0.5; 否则为 0。

所以由上面的公式就可以求出例句 1 和例句 2 的相似度:

$$SIM(Sen1, Sen2) = \frac{0.5 + 0.5}{3} = 0.33$$

我们可以看到, 在上面的两个例子中, 句 1 的核心词“送往”和句 2 的核心词“送到”是同义词, 但以关键词匹配的方法并不能匹配上, 所以我们又引入了语义依存树作为补充(图 5、图 6)。

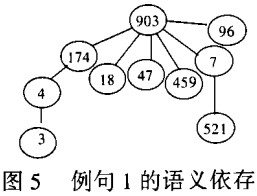

图 5　例句 1 的语义依存树

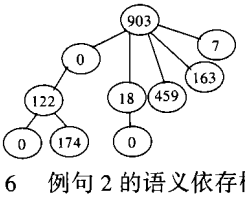

图 6　例句 2 的语义依存树

语义相似度的计算公式如下:

$$SIM'(Sen1, Sen2) = \frac{\sum_{i=1}^{n} W'i}{Max\{PairCount1', PairCount2'\}} \quad (2)$$

在式(2)中, $\sum_{i=1}^{n} W'i$ 为句子 1 和句子 2 的有效语义搭配对匹配的总权重, PairCount1′为句子 1 的有效语义搭配对数, PairCount2′为句子 2 的有效语义搭配对数。

从图 5 和图 6 可看出, 句 1 中的“送往”和句 2 中的“送到”的语义都为 903, 这样一来, “送往”和“送到”就自然匹配上了。所以上例两句之间的语义依存相似度为:

$$SIM'(Sen1, Sen2) = \frac{1 + 1}{3} = 0.67$$

由于基于关键词和基于语义的方法有着各自的优缺点, 所以我们的算法最后用下面的公式确定句子之间的相似度:

$$S(Sen1, Sen2) = \lambda \cdot SIM(Sen1, Sen2) + (1 - \lambda) \cdot SIM'(Sen1, Sen2) \quad (3)$$

在此算法中, 取 λ = 0.5。

3　实验及结果分析

我们的测试集为 5 843 个语句, 这些句子分为两个部分: 其中有 5 800 句为噪音句子, 构成噪音集; 另外 43 个是以手工获取的句子, 构成标准集。标准集中的句子按它们两两间的相似程度可以分为 16 个类, 每个类中有 2~4 个句子不等。也就是说, 在标准集的 43 个句子中, 每个句子都存在 1~3 个相似的句子。最后我们把标准集与噪音集混杂在一起作为测试集。

测试试验是这样进行的:

对于标准集中的 43 个句子, 我们按顺序从中抽出 1 个句子, 然后计算这个句子与测试集中的句子之间的相似度, 并按照相似度的大小对测试集中句子进行排序, 输出相似度最大的前三个, 如果与该句属于同一类的其它句子都被输出, 则说明这个句子的相似度计算是成功的。我们分别用 TF-IDF 方法和本文提出的方法做了试验, 并把试验结果作了对比。试验结果计算公式:

$$正确率 = \frac{\sum CorrectSen}{\sum Sen} \cdot 100\% \quad (4)$$

在式(4)中, ∑CorrectSen 表示测试结果正确的句子总数, ∑Sen 表示被测的句子总数。其试验结果如下:

方法	测试句子(个)	结果正确的句子(个)	正确率
TF-IDF	43	20	46.5%
语义依存	43	35	81.4%

试验结果分析: 从以上试验结果可以看出, 本文采用的方法所得的正确率要远远高于 TF-IDF 方法。我们还对输出不正确的句子做了依存结构的跟踪, 发现当一个句子比较长, 而且该句有较多动词的时候, 依存分析的结果不正确, 这就导致句子的核心词找不准确, 由此得到的搭配对并不能表达整个句子的意思, 从而带来了错误的计算结果。

4　结束语

本文采用了一种基于语义依存的汉语句子相似度计算方法, 该方法把语义与依存文法分析结合起来, 有效地刻画了句子的表达意思。在计算依存树之间的相似度时, 本方法并没有匹配所有的搭配对, 而是计算那些有效搭配对之间的相似程度, 这样使计算的时间复杂度大大降低。最后我们进行了该方法与 TF-IDF 之间的对比试验, 实验结果证明该方法要优于 TF-IDF 方法。

由于本方法受依存分析的影响, 如果能进一步提高依存分析的准确率, 将得到更好的计算结果。

参考文献:

[1] Jaime Carbonell, et al. The Use of MMR, Dirversity-based Reranking for Recording Documents and Producing Summaries [C]. Proceedings of ACM-SIGIR' 98, Melbourne Australia, 1998.

[2] Chris H Q Ding. A Similarity-based Probability Model for Latent Se-mantic Indexing[C]. Proc. of 22nd ACM SIGIR Conference, 1999. 59-65.

[3] 穗志方, 俞士汶. 基于骨架依存树的语句相似度计算模型[C]. 中文信息处理国际会议(ICCIP' 98), 1998.

[4] 董振东, 等. 知网[EB/OL]. http://www.keenage.com.

[5] 刘海涛. 依存语法和机器翻译[J]. 语言文字应用, 1997, (3): 89-93.

[6] 郭艳华, 周昌乐. 一种汉语语句依存关系网协动生成方法研究[J]. 杭州电子工业学院学报, 2000, 20(4): 24-32.

[7] 车万翔, 等. 面向依存文法分析的搭配抽取方法研究[C]. 全国第六届计算语言学联合学术会议, 2001.

作者简介:

李彬(1979-), 硕士生, 研究方向为自动文摘; 刘挺(1972-), 副教授, 博士, 研究方向为智能内容管理; 秦兵(1968-), 博士生, 研究方向为自动问答和自动文摘; 李生(1943-), 教授, 研究方向为机器翻译。