基于改进编辑距离和依存结构的句子相似度计算

刘宝艳，林鸿飞，杨志豪

（大连理工大学计算机科学与工程系，大连 116024 ）

摘 要： 句子相似度计算在中文自然语言处理领域有着广泛的应用背景。要准确的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，本文提出了一种基于改进编辑距离和依存结构的句子相似度计算方法。依存算法考虑到词语之间的相互作用关系和句子内部的结构，而编辑距离由于《同义词词林》的应用可以兼顾同义词之间的替换，因此该方法与其他方法相比，描述句子的信息更加全面，实验结果表明该方法是有效的。

关键词： 相似度计算；依存结构；改进编辑距离；自然语言处理

Sentence Similarity Computing Based on Improved Edit-distance and Dependency Structure

Liu Baoyan, Lin Hongfei, Yang Zhihao

(Department of Computer Science and Engineering, Dalian University of Technology , Dalian 116024)

Abstract: Sentence similarity computing has been widely used in the field of natural language processing. As we know, 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 way that based on improved Edit-distance and dependency structure to compute sentence similarity . Dependency structure can open out the relationship of words and the structure of sentence , and improved Edit-distance can take account of the substitution of synonyms by the use of Cilin. Comparing to other sentence similarity computing methods, the method can fully describe the features of the sentence. Finally some examples are presented to show that the results are satisfied.

key words: Similarity Computation; Dependency Structure; Improved Edit-distance; Natural Language Processing

1 引言

句子相似度计算在自然语言处理领域具有非常广泛的应用背景，例如：在问答系统中通过句子相似度计算找到与问题相匹配的答案；在自动文摘系统中通过句子相似度计算去除冗余信息，抽取文摘句；在信息检索系统中通过句子相似度计算找到与用户需求相似的句子；在基于实例机器翻译中通过句子相似度计算匹配相似的句子，得到需要的译文等等。因此长期以来，句子相似度计算问题，一直为人们所热衷。

目前研究句子相似度的方法有基于相同词汇的方法，使用语义依存的方法[1]，计算编辑距离的方法[2]，基于关键词的方法[3]，使用语义词典的方法，基于语境框架的方法[4]，基于属性论的方法以及基于统计的方法等等。

基金资助：国家自然科学基金（60373095）

作者简介：刘宝艳（1980），女，辽宁，硕士研究生 lbysmile@126.com.

林鸿飞（1962），男，辽宁，教授，博士，hflin@dlut.edu.cn .

其中，基于相同词汇的方法有很明显的局限性，对于同义词之间的替换则无能为力。而使用语义词典的方法，可以很好地解决这一问题，但是单纯的使用语义词典的方法，并没有考虑到句子内部的结构和词语之间的相互作用关系，准确率不高。编辑距离通常被用于句子的快速模糊匹配领域，但是其规定的编辑操作不够灵活，也没有考虑词语的同义替换。最后基于统计的方法，需要构造大量的训练语料，工作量是十分巨大的，而且还存在着数据稀疏的问题。

一个句子的信息的完整表达，不但依赖于组成句子的词汇，而且还依赖词汇之间的关系。本论文将从句子结构信息和句子的词汇信息进行研究，结合目前研究句子相似度的方法，提出了一种基于改进编辑距离和依存结构相融合的计算句子相似度的方法。

2 汉语句子相似度概念及其计算方法的研究

2.1 句子相似度概念

句子相似度是指两个句子的匹配符合程度，周 昉[5]把句子间的相似度定义为一个在[0,1]之间的实数。0 代表两个句子完全不相似，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。

2.2 句子相似度计算方法

2.2.1 基于依存的句子相似度计算方法

依存句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。依存句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，主张句子中动词是支配其他成分的中心成分，而它本身却不受其他任何成分的支配，所有受支配成分都以某种依存关系从属于支配者[6]。二十世纪七十年代，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，中国学者又提出了依存关系的第五条公理[7]：

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。

在利用依存算法进行相似度计算时，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它的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。

相似度计算公式[1]如下：

$$SIM(Sen_1, Sen_2) = \frac{\sum_{i=1}^{n} W_i}{Max\{PairCount_1, PairCount_2\}} \qquad (1)$$

$\sum_{i=1}^{n} W_i$ 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。

2.2.2 基于编辑距离的句子相似度计算方法

编辑距离的算法首先由俄国科学家Levenshtein提出的，故又叫Levenshtein Distance。 编辑距离就是用来计算从原串（s）转换到目标串(t)所需要的最少的编辑操作数目，编辑操作有“插入”、“删除”和“替换”三种。在计算汉语句子相似度的时候，传统的计算编辑距离是以字为单位，如图 1 显示了“爱吃苹果”与“喜欢吃香蕉”之间的编辑距离为4 。

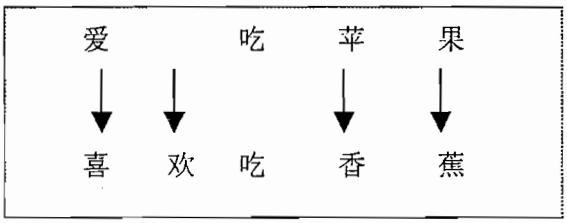

图 1 编辑距离计算

Fig.1 The computing of Edit-distance

从该计算过程可以看出，单纯使用以字为单位编辑距离的方法，计算出的语义距离和实际情况是有很大出入。首先，编辑距离算法以字为基本计算单位，而在汉语中，单个的字往往是不具备意义的。例如上面的“苹”、“果”等字，并不能反映其所合成词的意义。其次，词语之间替换操作的代价并非都是相同的。例如，“爱”被“喜欢”替换，代价不应该很大。其次，如果在被检索句子或短语中间加入为数不多的词，语义也不会有太大改变。例如“爱吃苹果”与“爱吃甜苹果”就非常相似。因此在编辑距离方法的基础上我们借鉴了车万翔[2]的利用改进编辑距离方法计算中文句子的相似度。该方法的主要思想是：以普通编辑距离算法为基础，采用词语取代单个的汉字或字符作为基本的编辑单元参与运算。同时使用了 How-net 和同义词词林两种语义资源，计算词汇之间的语义距离，加入词语的语义相似信息确定词语之间的替换代价，并且赋予不同编辑操作不同的权重。在计算相似度时，该种方法充分考虑了句子中每个词的深层信息，使表面不同，深层意义相同的词被挖掘出来，在不用经过词义消歧和句法分析的情况下，兼顾了词汇的顺序和语义等信息，使之更加符合中文句子相似度计算的要求。本文采用了哈尔滨工业大学信息检索实验室的《同义词词林扩展版》作为系统的词义知识资源。

2.2.3 同义词词林的使用

利用改进编辑距离进行句子相似度计算时用到了《同义词词林》，在车万翔[2]的论文中有《同义词词林》的使用介绍，其基本思想就是利用词林中对每个词提供的语义编码进行两个词之间的语义距离计算。我们用的《同义词词林扩展版》将词的词义逐级划分为5层，描述了一个由上到下，由宽泛概念到具体词义的语义分类体系，并将所收的词按词义分门别类组织在其中。每个词汇都按照其语义，赋予了一个或多个5位的语义代码。与此分类体系相对应的是一个词义的编码体系，描述如下：

〈词义编码〉::=〈1层〉〈2层〉〈3层〉〈4层〉〈5层〉

〈1层〉::=〈大写英文字母〉

〈2层〉::=〈小写英文字母〉

〈3层〉::=〈数字〉〈数字〉

〈4层〉::=〈大写英文字母〉

〈5层〉::=〈数字〉〈数字〉

对于 A，B 两词之间的语义距离，我们首先查到他们的语义编码，然后利如下的公式进行计算[2]：

$$Dist(A,B)=\min_{a\in P,b\in Q} dist(a,b) \tag{2}$$

其中，P，Q 分别为A，B 两词具有语义的集合。语义a，b 之间的距离为：

$$dist(a,b)=2\times(7-n) \tag{3}$$

其中，n 为它们之间的语义代码从第n 层开始不同，全部相同语义距离为0。如“苹果” Bh07A14，“香蕉” Bh07A34，“喜欢” Gb09A01，“爱” Gb09A01。用上面的公式可知$Dist$（苹果，香蕉）= 2，$Dist$（喜欢，爱）= 0。从以上的操作可以看出利用词林进行语义距离计算相似度比较方便、快捷。

3 改进编辑距离与依存算法的结合

基于依存的句子相似度计算方法体现了句子内部的结构和词语之间的相互作用关系，而编辑距离由于同义词词林的应用可以兼顾同义词之间的替换，并体现了组成句子的每个词深层的语义信息。我们的目标是将两种计算方法组合起来，扬长避短，互为补充，共同描述一个句子，从而根据这些特征计算句子和句子之间的相似度，获

得较高的准确率。这里就涉及到如何将这两种方法进行融合的问题，最普遍的方法[3]就是分别用这两种方法进行相似度的计算，然后对每种方法赋予不同的权重，并求和。

本文采用的方法是一种折中的方法，首先借鉴骨架依存树的思想[8]，仅分析出句子的整体句法结构，所谓的整体句法结构用该句的谓语中心词及其有效支配成分来表示。它的特点是把一个句子分成两个层次，第一层为句子的谓语中心词，第二层为句中谓语中心词的有效支配成分。当得到这两个层次以后，对第一层利用语义词典进行语义距离计算，第二层利用改进编辑距离的方法计算，最后将两个层次得到的结果相加。

本文在计算句子 S_1 和 S_2 相似度时，首先，用哈工大的依存算法分析器析出句子的谓语中心词，即句子的第一层，然后再利用依存算法分析器的分词和词性标注功能分别得到两个句子的第二层的 m 个和 n 个有效成分序列：$w_{11},w_{12}...w_{1m}$ 和 $w_{21},w_{22}..w_{2n}$，得到这两个层次以后就可以对两个句子的相似度进行计算。计算公式如下：

$$Dis(S_1,S_2)=\alpha\times dis_1(S_1,S_2)+\beta\times dis_2(S_1,S_2) \tag{4}$$

$$SIM(S_1,S_2)=\frac{Dis(S_1,S_2)}{Max(m,n)} \tag{5}$$

其中 $Dis(S_1,S_2)$ 为两个句子的编辑距离，$dis_1(S_1,S_2)$ 和 $dis_2(S_1,S_2)$ 分别为两个层次的距离，并对不同层赋予不同的权重。m 和 n 分别为两句子第二层有效成分的个数。

4 实验结果与分析

本论文实验所用的测试集为300个语句，这些句子分成两个部分：其中有249句为噪音句子，构成噪音集；另外51个是我们手工获取的句子，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17个类，每个类中有3个句子。也就是说，在标准集的51个句子中，每个句子都存2个我们人为觉得相似的句子。在噪音集里也有一小部分是手工获取的句子，这些句子多是与标准集中句子相似，但是相关度比较小（人为的观测），目的是减少构造标准集的人为性，使测试结果更公平，更能够体现理论的普遍性。最后把标准集与噪音集混杂在一起作为论文的测试集。

论文的测试实验是这样进行的，对于标准集中的51个句子，按顺序从中抽出1 个句子，也就是说从17类标准集每一类选一个句子，每个句子要从299个句子里找到与之相似的句子，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，然后人为地观测输出结果，如果与该句属于同一类的其他句子都被输出，则说明这个句子的相似度计算是成功的。本论文有两组实验结果，第一组对每个句子输出2句相似的句子，第二组每个句子输出3句相似的句子，每组实验结果与编辑距离的算法进行比较，比较结果如表1和表2所示。

表 1　返回2个句子的比较

Tab.1 The comparison of returning two sentences

方法	测试句子	结果正确句子	正确率
编辑距离	34	23	67.6%
结合算法	34	24	70.6%

表 2　返回3个句子的比较

Tab.2 The comparison of returning three sentences

方法	测试句子	结果正确句子	正确率
编辑距离	34	30	88.2%
结合算法	34	32	94.1%

实验结果用正确率进行评估，计算公式[3]如下：

$$正确率=\frac{\sum 测试结果正确的句子}{\sum 被测句子}\times 100\% \quad (6)$$

公式中的被测句子为标准集中去掉用来查询的句子的个数，在本试验里被测句子数为34。

从试验结果可以看出在改进编辑距离的基础上结合句子的依存结构能够提高实验结果的正确率。

5 结束语

本文采用了一种基于改进编辑距离和依存相结合的汉语句子相似度计算方法，该方法把语义同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，最后我们将该方法与只用改进编辑距离方法进行了比较，实验结果证明该方法要优于单单用改进编辑距离的方法。从实验的结果我们可以分析影响本算法正确率提高的原因有两个，首先由于本方法受依存分析的影响，要使本方法的正确率得到进一步的提高，提高依存分析的正确率是一个关键的问题，其次由于《同义词词林扩展版》收录的词是有限的，例如一些专有名词就没有收录到其中，如果将该方法用于特殊领域，可以增加一些专业领域的词典，这样也可以提高正确率。

参考文献：

[1] 李彬等. 基于语义依存的汉语句子相似度计算[J].计算机应用研究.2003年第12期.

[2] 车万翔 刘挺 秦兵等.基于改进编辑距离的中文相似句子检索[J].高技术通讯, 2004. 7 , 15-19.

[3] 赵妍妍 秦兵 刘挺等. 基于多特征融合的句子相似度计算

[4] 晋耀红.基于语境框架的文本相似度计算[J] .计算机工程与应用，2004，36-39

[5] 周 航. 汉语句子相似度计算方法及其应用的研究[D].河南大学，2005 年 5 月

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

[7] 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.

[8] 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(ICCIP98)论文 集，1998:458～465