一种综合多特征的句子相似度计算方法①

吴全娥　熊海灵　(西南大学计算机与信息科学学院 重庆 400715)

摘　要: 提出了一种综合多特征的句子相似度计算方法，该方法分别从句子的句法、词汇语义、词形三个方面来度量句子的相似度，最后将这三个方面加权整合计算得到句子的相似度。本方法综合考虑了句子的深层和表层信息，并对句子进行了词汇扩展，从而使句子相似度计算更加准确。

关键词: 句子相似度计算; 多特征; 树核; 权值

Method for Sentence Similarity Computation by Integrating Multi-Features

WU Quan-E, XIONG Hai-Ling

(College of Computer and Information Science, Southwest University, Chongqing 400715, China)

Abstract: A method for sentence similarity computation by integrating multi-features was proposed. According to the syntax feature, semantic feature and word feature of the sentences, the similarity was measured, respectively. Then, this paper combined the sentence similarity by endowing the above three features with different weights. Comparatively, this kind of estimation of sentence similarity is more accurate than the previous because both the deep and surface information of the sentences were taken into accounted, and the vocabulary of sentences was also extended in the process of calculation.

Keywords: sentence similarity computation; multi-features; tree kernel; weight

1　引言

在自然语言处理领域中，句子相似度计算是一个基础而核心的研究课题，它在现实中有着广泛的应用，如基于实例的机器翻译中通过句子相似度计算匹配相似的句子，找到相似的译文[1]；基于常见问题集(FAQ)的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[2]；信息检索中利用句子相似度计算找到与用户检索需求相似的句子[3]。

在句子相似度计算中，根据汉语句子的不同表现形式可以概括为三类方法：基于词特征的句子相似度计算、基于语义特征的句子相似度计算、基于句法分析特征的句子相似度计算。但是这三类方法都存在不足，如基于词特征的句子相似度计算方法统计相比较的两个句子相同词的共现频率，它只利用句子的表层信息，没有考虑词汇本身的含义，有很大的局限性；基于语义特征的方法依赖于语义词典，由于语义词典的不全面以及未登录词语义信息的缺失，也给计算带来一定的误差；而基于句法分析特征的方法需要依靠句法分析技术，目前对句子各成分之间的依存关系分析准确率还不高，使得这种方法难以取得较高的准确率。为此，考虑将这三方面特征进行综合，扬长避短，更全面地衡量句子的相似度。提出了一种综合多特征的句子相似度计算方法，综合考虑句子的词汇、语义和句法信息，使这三种特征在表达句子信息时各有侧重，同时在计算句子的语义相似度时考虑同义词的扩展，使得计算结果更为准确。

2　常用的句子相似度计算方法

根据汉语句子的不同表现形式，可以将句子的特征分为三种：词特征、词义特征、句法特征。下面分别对基于这三个特征的句子相似度计算方法进行介绍。

2.1 基于词特征的句子相似度计算

基于词特征的句子相似度计算方法就是将两个句子的有效词表示成两个向量，然后计算两个句子向量

① 收稿时间:2010-03-12;收到修改稿时间:2010-05-22

夹角的余弦，得到的结果就是两个句子的相似度[4]。例如两个句子 S_1 和 S_2，它们所有有效词构成的向量空间是 $V=\{X_1,X_2,X_3,\cdots\cdots X_n\}$，其中 Xi 为有效词。句子 S_1 的向量是 $V1=\{\omega_1,\omega_2,\omega_3,\cdots\cdots,\omega_n\}$，$\omega_i$ 是词 X_i 在句子 S_1 中出现的次数，$V_2=\{\phi_1,\phi_2,\phi_3,\cdots\cdots,\phi n\}$，$\phi_i$ 是词 X_i 在句子 S_2 中出现的次数。则两个句子的相似度为：

$$similarity(S_1,S_2)=\vec{V}_1\cdot\vec{V}_2=\sum_{i=1}^{n}w_i\cdot j_i/(\sqrt{\sum_{i=1}^{n}w_i^2}\cdot\sqrt{\sum_{i=1}^{n}j_i^2}) \quad (1)$$

这种方法不需要任何对文本内容的理解，它只是利用了词的表面信息，简单易实现。当然，这个方法也存在很大的局限性。首先，它是一种统计的方法，只有当句子所包含的词语数量足够多时，相关的词重复出现，统计效果才会表现出来。其次，基于词特征的方法只是考虑了词在上下文的统计信息，而没有考虑词汇本身的语义信息，对同义词以及一词多义效果不好。

2.2 基于语义特征的句子相似度计算

基于语义特征的句子相似度计算方法是利用句子的语义信息来计算句子的相似度的[5]，这种方法需要一些语义知识资源做基础。通过计算句子之间的词语相似度，进而计算句子的相似度，例如有 2 个句子 A 和 B，A 包含的词为 $A_1,A_2,\cdots,A_m$，B 包含的词为 $B_1,B_2,\cdots\cdots,Bn$，$s(A_i,B_j)$表示 A_i 和 B_j 之间的相似度,sem_sim(A,B)表示 A,B 句子之间的单词语义相似度

$$sem\_sim(A,B)=(\sum_{i=1}^{m}a_i/m+\sum_{j=1}^{n}b_j/n)/2 \quad (2)$$

式中：$a_i=\max(s(A_i,B_1),s(A_i,B_2),...,s(A_i,B_n))$

$b_j=\max(s(B_j,A_1),s(B_j,A_2),...,s(B_j,A_m))$

这种方法考虑了词语的语义信息，对那些表面不同，深层语义相同的词语能识别出来，但由于词典的不全面和一些未登录词语义的缺失给计算带来一定的误差，此外基于语义特征的句子相似度计算方法在计算句子相似度时，采用了一种最大匹配法，没有考虑句子的结构信息，因此准确率还没有达到令人满意的程度。

2.3 基于依存树的句子相似度计算方法

基于依存树的句子相似度计算方法在算句子的相似度时考虑了被比较句子的句法结构信息和词汇信息[6]，其中，句子的整体句法结构用句子的谓语中心词及其直接支配成分来表示，将分析结果看作一棵简化了的依存树。在计算依存树之间的相似度时就是计算那些有效搭配对之间的相似度。

这种方法把句子的句法结构纳入句子的相似度计算中，对句子的理解更加充分，从而更准确地得到句子的相似度，理论上是一种较好的计算模型。但是，这种方法需要依靠句法分析技术，而现有的句法分析技术还不够成熟，相似度计算基础不牢固导致这种方法难以取得较高的准确率，实用性不强。

3 新改进的句子相似度计算方法

针对上述句子相似度计算方法的问题，提出了综合多特征的句子相似度计算方法，将句子的词特征，语义特征，句法特征加权组合起来，互为补充。

如图 1 所示，syn_sim 表示两个句子的结构相似度；sem_sim 表示单词语义相似度；word_sim 表示词形相似度。首先，将两个经过了句法分析，表示成树状结构的句子，使用树核 (Tree Kernel)算法[7]来计算句子的结构相似度 syn_sim；其次，利用《知网》进行词汇扩展，利用《同义词词林》计算句子中单词的语义相似度 sem_sim；之后，对分词和关键词抽取过后的两个句子统计单词的共现频率，得出两个句子的词语相似度；最后对句子结构相似度 syn_sim，单词语义相似度 sem_sim，词形相似度 word_sim 加权整合得出句子的相似度 sen_sim。

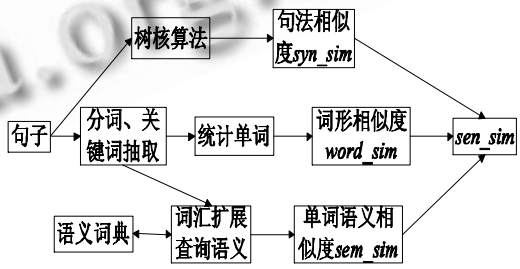

图 1　综合多特征的句子相似度计算流程

3.1 关键词抽取

汉语句子中任何句子都是由关键成分(主、谓、宾)和修饰成分(定、状、补)组成，句子的关键成分对句子起主要作用，修饰成分对句子起次要作用。进行句子相似度计算时只要考虑句子的关键成分。通常情况下，一个句子中做主语和宾语的多为名词或代词，作谓语的多为动词或形容词。为此，将句子中的所有名词、

动词、形容词和副词作为关键词。在计算句子的相似度时先对句子进行分词和词性标注，根据词性标注提取出句子的关键词，句子的相似度计算只考虑这些关键词。这样进行相似度计算要更准确一些。

3.2 词汇扩展

汉语中有很多词汇词形不同但却表达同一个意思，为了提高这类词汇相似度计算的准确度，需要对分词后抽取的关键词汇进行同义词扩展。这里使用《知网》语义词典作为词汇扩展的资源。知网中同义词为具有相同的英语译文(W_E)和语义定义(DEF)的词汇。例如“我”和“俺”，其简化词条如下：

NO.=085498	NO.=000701
W_C=我	W_C=俺
W_E=I	W_E=I
DEF=firstPerson|我	DEF=firstPerson|我

可见，“我”和“俺”具有相同的英语译文(W_E)“I”和语义定义(DEF)“firstPerson|我”，是一对同义词。

3.3 句法相似度计算

为了能够利用句子结构信息更全面的计算两个句子的相似度，本文对句子进行深入分析，将两个句子表示成树状结构，用树核算法来计算两个句子的句法结构相似度。例如汉语例句“我是老师的学生”和句子“他是学生”经过句法分析后的结果如下图 2 所示：

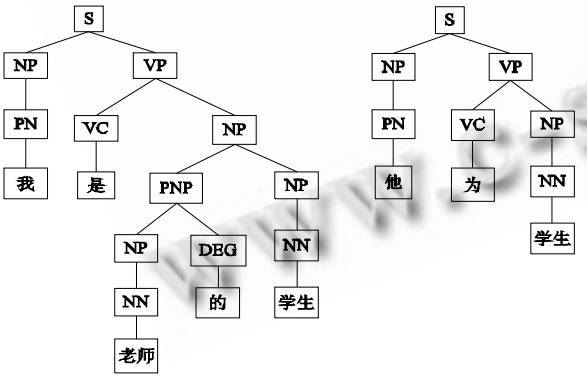

图 2　句法分析的树状结构

对于汉语句子的嵌套结构，最直观的表现形式是树状结构，这样更能体现句子的信息[8]，另外，在比较两个句子时，两种结构的相似度不仅体现在单个分支的句法结构，也体现在句子的整体结构上。

比较两个树状结构时，可将树 T 表示成全部子树类型的向量形式：$h(T)=(h_1(T),h_2(T),\cdots h_i(T)\cdots h_m(T))$，其中 $h_i(T)$是第 i 个子树类型(Substree)在 T 中出现的次数，其中图 2 中句法分析的树状结构出现的部分子树如图 3 所示。

将树表示成向量形式后，树 T_1 和树 T_2 之间的相似度可以用如公式(3)所示，用两个向量 $h_1(T)$和 $h_2(T)$之间的点积来计算。

$$tree\_sim(T_1,T_2)=h(T_1)\cdot h(T_2)=\sum_i h_i(T_1)h_i(T_2) \quad (3)$$

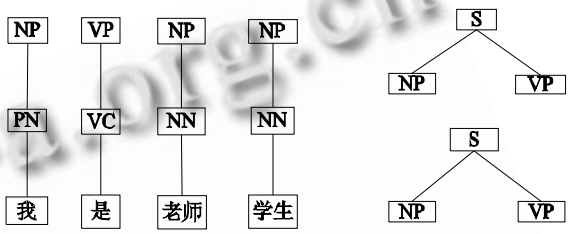

图 3　树状结构中的部分子树

由于向量 h(T)的维数与树 T 的大小呈指数关系，计算的复杂度高，因此 Collins 和 Duffy 提出如下的 Kernel 核函数[9]来隐含地计算两个向量之间的点积：

$$\begin{aligned} h(T_1)\cdot h(T_2)&=\sum_i h_i(T_1)h_i(T_2)\\ &=\sum_{n_1\in N_{T_1}}\sum_{n_2\in N_{T_2}}\sum_i I_i(n_1)I_i(n_2)=\sum_{n_1\in N_{T_1}}\sum_{n_2\in N_{T_2}}C(n_1,n_2)\end{aligned} \quad (4)$$

其中，N_{T_1} 和 N_{T_2} 是 T_1 和 T_2 中的节点集合，我们定义 $C(n_1,n_2)=\sum_i I_i(n_1)I_i(n_2)$ 是以节点 n_1 和 n_2 为根的公共子树数量。若子树 i 存在于以 n 为根结点的树中则定义符号函数 $I_i(n)$为 1，否则为 0。这样，使得点积计算一方面从复杂度上降为 $O(|N_{T_1}||N_{T_2}|)$，另一方面使求解的过程更加清晰。$C(n_1,n_2)$的递归计算形式如下：

(1) 若在 n_1 和 n_2 上的产生式(Production)不同，则 $C(n_1,n_2)=0$；

(2) 若在 n_1 和 n_2 上的产生式相同，且 n_1 和 n_2 的孩子只有叶子节点，则 $C(n_1,n_2)=1$；

(3) 若在 n_1 和 n_2 上的产生式相同，且 n_1 和 n_2 的孩子没有叶子节点，则

$$C(n_1,n_2)=\prod_{j=1}^{nc(n_1)}(1+C(c_{n_1}^j,c_{n_2}^j)) \quad (5)$$

其中，$nc(n_1)$是 n_1 的孩子数量，c_n^j 节点 n 的第 j 个孩

子， $nc(n_1)=nc(n_2)$ 。节点上的产生式是指节点到全部孩子的产生表述。

3.4 单词语义相似度计算

句子结构相似度是利用句子的句法信息进行计算的，而仅仅考虑句子的句法信息是不够的，句子的语义信息一个重要成分是单个词语的语义信息，单词语义信息是由其构成的单词同义关系决定的。

计算单词语义相似度是用《同义词词林》作为系统的语义知识资源。其思想是利用《同义词词林》中对每个单词提供的语义编码进行两个单词之间的语义距离计算。我们用的《同义词词林扩展版》将单词的词义逐级划分为 5 层，描述一个由上到下，由宽泛概念到具体词义的语义分类体系，并将所有的单词按义原分门别类组织在其中。每一个词汇都按照其语义赋予了一个或多个 5 位的语义代码。

对于 w_1,w_2 两个单词之间的语义距离，我们首先查到它们的语义编码，然后利用如下的公式进行计算：

$$Dist(w_1,w_2)=2\times(7-n) \tag{6}$$

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。用公式(6)可知 Dist(苹果，香蕉)=2，Dist(喜欢，爱)=0。

单词语义相似度是指两个句子的关键词语的语义相似度，使用《同义词词林》进行词语语义相似度计算时根据关键词抽取过程中的词性标注，对两个句子中相同词性词进行匹配，得到每个匹配对，然后利用公式(6)计算每个匹配对的语义距离，得到每个匹配对的语义距离值后，然后使用公式(7)将其转化为两个单词的相似度值

$$sword(w_1,w_2)=a/(Dist(w_1,w_2)+a) \tag{7}$$

即认为单词间的相似度与单词的语义距离成反比。a 为调节参数。通过此公式将两个单词间的语义距离转化为相似度值。

计算得到两个句子中的任意两个词之间的相似度后，计算两个句子的单词语义相似度的方法就是利用上述公式(2)来计算。

3.5 词形相似度

设句子 X 的长度为句子中关键词的个数，记为 len(X)。same(A,B)表示句子 A,B 中相同关键词的个数，当一个单词在 A,B 中出现的次数不同时，以出现次数少的计数。则句子 A,B 的词形相似度计算如下：

$$word\_sim(A,B)=2\times\frac{same(A,B)}{len(A)+len(B)} \tag{8}$$

3.6 特征加权计算

将句子的句法特征、词语语义特征和词形特征加权整合，就得到了句子与句子之间的相似度，相似度计算公式如下：

$$\begin{aligned}sen\_sim(S_1,S_2)&=a\times syn\_sim(S_1,S_2)+\\ &b\times sem\_sim(S_1,S_2)+g\times word\_sim(S_1,S_2)\end{aligned} \tag{9}$$

α、β、γ分别表示句法相似度，单词语义相似度和词形相似度的权值，且满足 α + β + γ =1。

4 实验结果及数据分析

对于相似度计算结果的评价，最好放在一个实际的系统中(如自动问答系统)，观察不同的相似度计算方法对系统性能的影响。这需要一个完整的系统，在条件不具备的情况下，本文采用人工判别的方式，用本文提出的方法和其它方法分别计算句子相似度，并对它们的计算结果进行比较。本文对比了 3 种句子相似度计算方法：

方法 1：利用基于词特征的句子相似度计算方法，按照公式(1)计算句子的相似度。

方法 2：利用语义特征，使用《同义词词林》的句子相似度计算方法，词语的相似度计算通过调用文献[5]的计算模块，使用公式(7)完成，按照公式(2)计算句子相似度。

方法 3：本文提出的综合多特征的方法，按公式(9)计算句子相似度，选取参数 α =0.5、β =0.4、γ =0.1。其中的词语相似度计算调用文献[5]中的计算模块完成，句法结构相似度计算使用树核算法计算完成。

我们取 300 个句子为测试集，其中 200 个噪音句子构成噪音集；另外 100 个句子是手工获取的句子构成的标准集，标准集中相似的句子是已知的。对于标准集的 100 个句子，我们按顺序抽出 1 个句子，然后用上面三种方法计算这个句子与测试集中的句子之间的相似度，并按照相似度的结果进行排序，然后观察相似度最大的句子，如果与该句相似的句子都输出来了，则说明这个相似度计算的结果是正确的。我们用基于词特征的方法、基于语义特征的方法和本文的

方法进行对比试验。实验结果的计算公式为：

$$p = \frac{correctsen}{totalsen} \times 100\%$$

P 表示计算的准确率，correctsen 表示测试准确的句子数量，totalsen 表示测试的句子总数量。

表 1　基于词特征、基于语义特征和本文方法对比实验结果

方法	totalsen	correctsen	P
基于词特征	300	140	46.67%
基于语义特征	300	209	69.67%
综合多特征	300	233	77.70%

从以上试验结果可以看出，本文采用的方法所得的结果准确率要高于基于词特征和基于语义特征的方法。这说明综合句子表层和深层信息，对词汇进行扩展在句子相似度计算中发挥了作用，使得计算的准确率有了提高。

5　结论

本文提出的综合多特征的句子相似度计算方法，提高了句子相似度计算的准确率。但还有一些不足的地方比如新方法的准确率在提高的同时复杂度也相应提高了，计算效率下降了。另外，本文在三个特征的权值确定时是根据三个特征在句子相似度计算中的重要性人为确定的三个值，对方法的准确性有一定影响。找到三个特征的最佳融合点还需要用遗传算法进行训练分析得到三个特征的融合权值，这是本文下一个研究的地方。由于本文方法受句法分析以及语义词典等因素的影响，为了达到更好的效果，需要进一步提高句法分析的准确率，不断完备语义词典。

参考文献

1 尹存燕,胡国全,陈家骏,戴新宇.一种基于实例的汉英机器翻译策略.计算机工程与设计, 2005,26(4):900—903.

2 黄河燕,张亮,冯冲,陈肇雄.基于语句相似度计算的FAQ 自动回复系统设计与实现.小型微型计算机系统, 2006,27(4):720—723.

3 陶跃华.基于向量的相似度计算方案.云南师范大学学报(自然科版), 2001,21(5):17—19.

4 秦兵,刘挺,王洋,郑实福,李生.基于常问问题集的中文问答系统研究.哈尔滨建筑大学学报, 2003,35(10):1179—1192.

5 李月雷,师瑞峰,林丽冰,周一民.汉语语句语义相似度计算.计算机科学, 2008,35(4):3—4.

6 李彬,刘挺,秦兵,李生.基于语义依存的汉语句子相似度计算.计算机应用研究, 2003(12):15—17.

7 Moschitti A, Pighin D, Basili R. Engineering of Syntactic Features for Shallow Semantic Parsing.In Proceedings of the ACL05 Workshop on Feature Engineering for Machine Learning in Natural Language Processing, Ann Arbor (MI),USA, 2005(6):48—56.

8 张志昌,张宇,刘挺,李生.基于浅层语义树核的阅读理解答案句抽取.中文信息学报, 2008,22(1):80—86.

9 Collins M, Duffy N. Convolution Kernels for Natural Language. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics Table of Contents. Spain: Barcelona, 2004.