基于语义角色标注的汉语句子相似度算法 *

田塑，柯永红，穗志方

（北京大学 信息科学技术学院，北京市 100871）

摘要：在语义角色标注过程中，经常需要检索相似的已标注语料，以便进行参考和分析。现有方法未能充分利用动词及其支配的成份信息，无法满足语义角色标注的相似句检索需求。基于此，本文提出一种新的汉语句子相似度计算方法。该方法基于已标注好语义角色的语料资源，以动词为分析核心，通过语义角色分析、标注句型的相似匹配、标注句型间相似度计算等步骤来实现句子语义的相似度量。为达到更好的实验效果，论文还综合比较了基于知网、词向量等多种计算词语相似度的算法，通过分析与实验对比，将实验效果最好的算法应用到句子相似度计算的研究中。实验结果显示，基于语义角色标注的句子相似度计算方法相对传统方法获得了更好的测试结果。

关键词：语义角色标注;词语相似度;知网;词向量;标注句型匹配

中图分类号：TP391　　**文献标识码：**A

Chinese Sentence Similarity Computing Based on Semantic Roles Annotation

Kun Tian, Yonghong Ke, Zhifang Sui

(Peking University, Beijing, 100871, China)

Abstract: In the process of semantic roles annotation, searching for similar annotated sentences is a common way to analyze such corpus. Existing methods cannot take full advantage of verbs and related elements, so they are unable to meet the demand of searching for similar annotated sentences. The article develops a completely new method to calculating Chinese sentence similarity focused on the verbs. Based on semantic roles annotation, the algorithm finds the similar sentences by analyzing the semantic roles, matching the annotated sentences, and calculating similarity between these matched sentences. To get a better result, the article also compares several methods to compute word similarity, including algorithms based on How-net and Distributed Representation, and applies the algorithm that performs best to the algorithm through analysis and tests. The experiment result tells that the sentence similarity computing algorithm based semantic roles annotation performs better than traditional methods.

Key words: semantic roles annotation; word similarity; How-net; Distributed Representation; annotated sentence match

1 引言

目前研究句子相似度的方法有基于关键词的方法、使用语义词典的方法[1][2]、使用语义依存的方法[3][4]、计算编辑距离的方法[5][6]、基于语境框架的方法[7]、基于属性论的方法[8]以及基于统计的方法[9]等等。这些算法归结起来可概括为两类：一类是基于词层面特征的句子相似度计算，包括基于词的统计特征、词汇语义特征等；另一类则是基于句子层面的相似

* **收稿日期：**　　**定稿日期：**

基金项目：面向三元空间的互联网中文信息处理理论与方法（2014CB340504）

作者简介：田塑（1994—），女，本科生，专业为计算机科学与技术；穗志方（1970—），女，教授，博士生导师，研究方向为计算语言学。

通讯作者：柯永红（1981—），男，讲师，研究方向为计算语言学，电话：86-10-62753081，E-mail：kyh@pku.edu.cn

度计算，包括基于句法分析、语义分析等。

现有的算法基本可以满足一般所说的句子相似度：句子的整体结构和语义上的相似度。语义角色标注是对句子中的谓词相关体词性成分在谓词表达的事件框架中所扮演的语义角色进行标注。在语义角色标注过程中，经常需要检索相似的已标注语料，以便进行参考和分析。语义角色标注中句子相似度主要建立在以动词为核心的框架的相似度上，现有方法未能充分利用动词及其支配的成份信息，因而无法满足语义角色标注的相似句检索需求。本文提出了基于语义角色标注的汉语句子相似度算法，该从语义角色标注数据出发，以标注结果为基本单元，综合考虑多个标注动词语义与标签角色的相似度。

2 基于语义角色的句子相似度算法

2.1 词语相似度计算

词语相似度是本文提出的基于语义角色的句子相似度算法中衡量语义相似程度的基础。句子的相似度计算建立在对词语的相似度计算之上。下文对比了使用《同义词词林》的词语相似度算法[10]和基于 How-net 的词语相似度算法[11]，介绍了本文所使用的词向量的训练过程与模型的选择，并探讨了基于分类词典和语料库的词语相似度算法在本文的句子相似度算法中的应用。

2.1.1 基于世界知识体系的词语相似度算法

How-net 和《同义词词林》都属于根据某种世界知识或分类体系，并且有一个反映知识结构的树状层次体系，将所有的义项都组织在一棵或几棵树状的层次结构中，再利用两个结点之间有且只有的一条路径来计算义项之间的相似度。但实际上，How-net 与《同义词词林》采用的是不同的语义体系和表达方式。《同义词词林》是按照词义相似度组织树状层次结构，每个收录词在树的最底层；而 How-net 中每个义原是这个层次体系中的一个结点，词语的语义由义原构成语义表达式来定义。相比《同义词词林》，How-net 这种由多个义原及其关系对词语做的定义更为立体和准确，体现了义原相似度和义原关联度，而非单纯的词义相似度。

另外，How-net 在多义词的处理上也有优势。虽然多义词会出现在《同义词词林》中所对应的多个词义处，但这些义项的出现频率无法在词林中体现。故而在计算词语相似度时只能简单的取各义项的算术平均值。而 How-net 中有第一义原占更大的权重。

因此在后面的句子相似度计算中计算词语相似度时，本论文没有采用《同义词词林》，而是使用 How-net 与基于词向量的方法作对比。

2.1.2 基于语料库的词语相似度算法

本文选择由 Google 开发的训练工具 Word2vec[12]训练得到词向量。Word2vec 根据给定的语料库，通过优化后的训练模型快速有效地将一个词语表达成向量形式。本文实验中使用的词向量训练语料为 Giga word[13]，训练模型选择为 skip-gram[14]，维度为 50。通过利用两个句子中所有的词来构成向量空间，然后利用这两个向量夹角的余弦值作为词语相似度。

2.1.3 两类词语相似度算法的比较

基于世界知识的方法简单有效，无需用语料库进行训练，比较准确地反映了词语之间语义方面的相似性和差异，而对于词语之间的句法和语用特点考虑得比较少，得到的结果受人的主观意识影响较大，有时并不能准确反映客观事实。基于语料库的方法比较客观，综合反映了词语在句法、语义、语用等方面的相似性和差异，但比较依赖于训练所用的语料库，计算量大，且受资料稀疏和资料噪声的干扰较大。

在 How-net 的知识体系中，一个较为具体的词语是通过一系列以第一基本义原为首的、

对该词较为笼统而一般化的描述来表达的。而第一基本义原相似度在整体相似度中又占了很大的比例，因此这种词语和第一基本义原的相似度往往较高。虽然有时第一基本义原和原词语和在概念上有一定的重合，但当第一基本义原出现在文本中时，在大多数情况下其表达的意思与该词语可以表达多种意思。例如“促销”的第一基本义原为“卖”，但对于词语“卖”，其出现在句子中时用来表达“促销”这种特定义项的情况比例却并不高。这种定义和计算的方法会在处理基本语义相似但适用范围相差较大的两个动词上给出较高的相似度。而在词向量中，两个词相似度对应的是词语同时出现在一定范围内的上下文中的概率，可以在一定程度上避免这种情况的出现。更多类似的例子如下表所示：

词语 1	词语 2	基于 How-net 的相似度	基于词向量的相似度
促销	卖	0.76	0.57
打扫	消除	1.0	0.38
出发	开始	0.76	0.55
访问	看望	1.0	0.51

表 2.1 基于 How-net 和词向量的动词词语相似度比较

其中，词语 2 是词语 1 在 How-net 中的第一基本义原。

所以对于符合这种条件的动词，人们的认知往往更符合词向量给出的较低的相似度。因为词向量间的是通过词语出现在一定范围内的上下文中出现的概率给定的，更符合人们的认知。而对于其他类型的词语，形容词的第一基本义原通常为“属性值”，名词的第一义原往往是“场所”、“时间”、“人”等通用概念，因此这种表述方式和词向量相比对词语的相似度影响并不大。基于此，本文的实验部分将在考虑动词相似度时分别对这两种算法进行实验并做进一步的验证分析。

2.2 语义角色标注及标签处理

本文使用的标注语料以《现代汉语谓词语义角色标注语料库规范》为准则，在文献[15]中的对谓词论旨角色的分类基础上对已完成分词的句子应用了一种用于标注句子谓词的论旨角色体系。有关语义角色标注的具体内容可以参照《现代汉语谓词语义角色标注语料库规范》。

语义角色标注的过程复杂，但由于句子相似度研究是在模拟人的判断过程，所以不必在角色的划分处理上过于精细。对于各个论旨角色，本文只考虑标注记号中的汉字部分，即忽略“+”、“&”、“?”、“@”、“VP”、“AP”等标记所带来的不同。例如，在算法中假设：“内容”和“VP 内容”是相同的角色标签。

同时，对于相似的动词，其主语和宾语部分在句中起到的作用大体相同。但由于动词不同，因而被标注的语义角色不同。所以在本文的算法中，根据论旨角色分类图，将“当事”与“施事”统一处理为相同的主语角色，宾语下面的六个角色也被认为是相同的角色。

2.3 标注句型间的相似度计算

在经过了语义角色标注的语料中，标注句型一般包含多个（动词、角色标签和论元成分），在本文中将这种结构称为语义搭配。基于语义角色标注的句子相似度算法思想是将标注句型之间的相似度转化为动词之间和两个角色标签相同的语义搭配之间的相似度。例如：

1）我 去 听 今天 下午 的 音乐会 。

[%施事 我 %] 去 [# 听 #] [%内容 今天 下午 的 音乐会 %] 。

2）你 看 过 这 部 影片 吗 ？

[%施事 你 %] [# 看 #] 过 [%内容 这 部 影片 %] 吗 ？

第一句中的（听，施事，我）和第二句中的（看，施事，你）可以组成一个语义搭配对

计算相似度；同理，（听，内容，今天 下午 的 音乐会）和（看，内容，这 部 影片）相搭配。

由于语义标注以动词为核心，本文中标注句型之间的相似度计算也围绕动词展开：在动词相似的基础上，比较相同标签下的词语相似度。对于一个含有 m 个语义角色的标注句 T，用 v 表示它的动词，$e(S)=\{e_1,e_2,\dots,e_m\}$表示 S 中所有论元成分的集合，$r(S)=\{r_1,r_2,\dots,r_m\}$表示 S 中所有角色标签的集合。则标注句型 T 可以表示为一个三元组 $\big(v,e(S),r(S)\big)$。

将标注句型T_1和T_2的相似度定义为

$$Sim(T_1,T_2)=\beta\times Sim(v_1,v_2)+(1-\beta)\times\frac{\sum_{(i,j)}Sim\big(e_i,e_j\big)}{max(m,n)}$$

其中$(i,j)\in\big\{(p,q)\mid r_p,r_q\in r(F_1)\cap r(F_2),1\le p\le m,1\le q\le n\big\}$，$m$和$n$分别为句型$T_1,T_2$中包含的标注句型数，$Sim(v_1,v_2)$为两个动词$v_1,v_2$的词语相似度，$Sim\big(e_i,e_j\big)$为论元$e_i$和$e_j$间的相似度。这里$\beta$为谓词相似度在全句中所占的权重，这里取$\beta=0.5$，即对谓词和语义角色的相似度各赋予 0.5 的权重。

在计算两个句型中各组配对论元间的总体相似度时的做法是将$Sim\big(e_i,e_j\big)$的总和除以$max(m,n)$，而非简单的求算术平均值。这是因为考虑到用于比较的两个标注句型存在复杂度不同的情况，即如果两个句型中包含的语义角色数不同，其中构成搭配对的语义角色相似度的影响将受到制约。

由上文举的两个例句可知，e_i,e_j可能是单个词语，也可能是一个词块。当e_i,e_j均是词时，计算方法与$Sim(v_1,v_2)$相同；而当e_i,e_j中至少有一个是词块时，采用类似前一节中基于语义的相似度算法：

将e_i和e_j看作两个词集合，分别包含 M 和 N 个元素。设e_i中第 m 个词和e_j中第 n 个词之间的相似度为s_{mn}，可以得到相似度矩阵：

$$\begin{bmatrix}s_{11}&\cdots&s_{1N}\\ \vdots&\ddots&\vdots\\ s_{M1}&\cdots&s_{MN}\end{bmatrix}.$$

则这两个词集合的相似度为

$$Sim\big(e_i,e_j\big)=\frac{1}{2}*\left(\frac{\sum_{i=1}^{M}s_m}{M}+\frac{\sum_{i=1}^{N}s_n}{N}\right)$$

在上式中：

$s_m=max(s_{mi}|1\le i\le N)$　　$s_n=max(s_{ni}|1\le i\le M)$

其中词语相似度S采用基于词向量或 How-net 的方法计算。

受基于 How-net 的词语相似度算法[11]的改进方式的启发，主要部分的相似度值应该对于次要部分的相似度值起到制约作用，在本算法中，即谓词的相似度也影响到其支配的语义角色的相似度。如果两个标注句型的谓词相似度比较低，那么谓词所支配的其他语义角色的相似度对于整体相似度所起到的作用也要降低。因此将标记句型相似度公式改为

$$Sim(T_1,T_2)=Sim(v_1,v_2)\times\left[\beta+(1-\beta)\times\frac{\sum_{(i,j)}Sim\big(e_i,e_j\big)}{max(m,n)}\right]$$

2.4 标注句型的相似匹配

标注句型匹配就是将句子间具有相似语义的标注句型进行配对。一种容易想到的方法是直接先对两个句子的所有标注句型两两之间进行相似度计算，然后从计算结果中获得标注句型的相似匹配结果。但这种算法在处理复杂的多谓词长句时计算复杂度很大。

一个标注句型是有谓词及其所支配的各个论元成分构成，其中谓词决定了这个句型的结

构，同时又是整个句子的动作承担者，是语义标注的核心。因此动词在度量标注句型间的相似度中是最重要的一个因素。虽然动词间的相似度不能完全代替句型之间的相似度，但已能在很大的程度上区分标注句型间是否具有一定的相似性。因此，本文通过对动词的相似匹配来实现标注句型的相似匹配。

设句子S_1中第i个谓词和S_2中第j个谓词之间的相似度为Sim_{ij}，可以得到谓词之间的相似度矩阵

$$A = \begin{bmatrix} Sim_{11} & \cdots & Sim_{1n} \\ \vdots & \ddots & \vdots \\ Sim_{m1} & \cdots & Sim_{mn} \end{bmatrix}$$

其中m和n分别为两个句子中的谓词个数，因而也是这两个句子的标注句型数。假设$m < n$，执行下述算法：

1）找到矩阵 A 中最大的元素$Sim_{pq} = max(Sim_{ij} \mid 1 \le i \le m, 1 \le j \le n)$，得到$S_1$中第$p$个谓词与$S_2$中第$q$个目标词构成的一种谓词匹配对；

2）删除矩阵 A 中Sim_{pq}所在的行与列；

3）循环执行前两步直到矩阵 A 中的行数或列数为 0。

由此得到m组谓词的搭配对，对应这些谓词所在的标注句型。然后对这些句型配对应用 3.3.2 节中的标注句型间的相似度算法。

2.5 句子整体相似度计算

对于包含p个谓词的句子S_1和包含q个谓词的句子S_2，分别拥有包含p和q个标注句型。则 S_1 的标注句型集合为 $T(S_1) = \{T_{11}, T_{12}, \ldots, T_{1p}\}$，$S_2$ 的标注句型集合为 $T(S_2) = \{T_{21}, T_{22}, \ldots, T_{2q}\}$。两个句子的相似度计算公式为

$$Sim(S_1, S_2) = \frac{\sum_{(i,j)} Sim(T_{1i}, T_{2j})}{max(p, q)}$$

其中(T_{1i}, T_{2j})为标注句型的匹配对。在这里暂且认为所有框架都是一样重要的，即赋予它们相同的权重。

与 3.2.2 节中的处理方法类似，考虑到句型间的相似度要受到句型总数的制约，在转化为句子相似度时将句型的相似度总和除以$max(p, q)$. 对于拥有不同数量的谓词的句子对，单个句型对间相似度在计算句子整体相似度时的作用被减小。

3 实验结果与分析

3.1 实验语料准备

目前，国内和国际上都没有关于汉语句子相似度计算的公共测试集，所以针对汉语句子相似度计算的测试语料一般只能通过人工构建来完成。本文所用语料全部是来自 973 项目中所搭建的语料标注平台，对相似句子分组搜集，最筛选获得。

本实验共进行了四轮测试。每轮测试的具体做法如下：

1）从《同义词词林》中选取 5-6 类常见、高频、意义较为明确的动词义项。这些义项类均在《同义词词林》中的前三层就已分支，且如果在第三层才分支则保证在这些义项类第三层中距离为 10 以上。同时，同类中的义项保证在第五层之前不出现分支，且在出现分支的层距离不超过 3；

2）在语料库中找出包含这五大类动词中至少一类里的某个动词的句子，句子的其他部分不做要求。这样得到的句子集作为测试集；

3）在测试集中选取 10 句作为标准集，人工为标准集中每条语句在测试集中选取若干条最为相似的句子作为“标准相似句”，数量控制在 5 条左右，最少不少于 3 条，最多不超过

7 条。其余句子均为“噪声句”。

这样得到的测试集可以保证每条标准句都有一定数量的相似句对应，且噪声句与标准句间也有一定的相似度，达到“噪声”的干扰效果。

最终四轮测试构建的测试语料集总共分别包含 181、207、236、265 条语句，大多数条语句包含词汇在 10 个以上，且语料中所有句子均是已经过分词和语义角色标注的句子。

3.2 实验对比的相关方法

为更好地体现本文所提出的算法计算句子相似度的效果，本文将此方法与几种一般常见的计算相似度算法来进行实验对比，同时通过实验对本文算法中用于词语相似度的方法进行比较和定量分析。实验对比的相关方法包括：

1）基于词特征的句子相似度算法[16]

设句子S_1和S_2共包含n个不同的词，这些词构成的向量空间$V=\{X_1,X_2,\dots,X_n\}$。句子S_1的向量$V_1=\{\omega_1,\omega_2,\omega_3,,\dots,\omega_m\}$，其中$\omega_i$为$S_1$中第 i 个词的词向量。句子$S_2$的向量$V_2=\{\varphi_1,\varphi_2,\varphi_3,,\dots,\varphi_n\}$，其中$\varphi_i$为$S_2$中第 i 个词的词向量。则两个句子的相似度为：

$$Sim(S_1,S_2)=\frac{\sum_{i=1}^{n}\omega_i\cdot\varphi_i}{\sqrt{\sum_{i=1}^{n}\omega_i^2}*\sqrt{\sum_{i=1}^{n}\varphi_i^2}}$$

2）基于词语语义特征的句子相似度算法[17]

设两个句子 A 和 B 所包含的词分别为A_1，A_2，$\cdots$，A_m和B_1，B_2，$\cdots$，B_n，则词$A_i(1\le i\le m)$和$B_j(1\le j\le n)$之间基于 How-net 的相似度可用$s(A_i,B_j)$来表示。这样就得到两个句子中任意两个个词的相似度，A，B 句子之间的语义相似度

$$Sim(A,B)=\frac{1}{2}*\left(\frac{\sum_{i=1}^{m}a_i}{m}+\frac{\sum_{i=1}^{n}b_i}{n}\right)$$

式中：

$a_i=max(s(A_i,B_j)|1\le j\le n)$　　$b_i=max(s(B_i,A_j)|1\le j\le m)$

$s(A_i,B_j)$为A_i与B_j的基于词向量的词语相似度。

3）基于语义标注的句子相似度算法

本文算法。由 2.1 节的分析，词向量与 How-net 的相似度结果主要在比较动词时差异较大，且推测使用词向量的算法相对更符合人的心理认知。因此在比较词语相似度统一使用词向量和 How-net 计算的两种方法的基础上，将动词使用词向量、其他部分使用 How-net 的方法和前两种分别进行对比分析，验证猜想。

4）基于句法依存特征的句子相似度算法

基于句法依存的相似度算法通常需要分析句子的整体树状结构，衡量结构间的相似度。由于目前的句法分析工具得到的句子结构通常不够准确，使用基于句法依存的句子相似度算法得到的实验结果与上述三种方法相比，直观上即有明显的差距。并且之前已有研究[18]表明在和本文具有类似特征的语料库上，由于句法分析产生的错误较多，难以全面刻画句子的语义。因此本文的实验方法中暂时不包括此类方法。

3.3 实验结果评价标准

为全面评估本文算法，实验中使用召回率、准确率和 F 值三个指标来衡量。具体做法为：

1）从标准集中依序找出第$i(1\le i\le n)$条语句，与测试集中的所有句子计算相似度，对相似度数值从大到小排序，得到按相似度大小排名的相似句。

2）记实验前对第i条标准集语句预先人工设定的标准相似句数目为M_i，实际算法返回的排名前M_i的句子中包含的标准相似句数为$CorrectSen_i$，则召回率

$$R_i = \frac{CorrectSen_i}{M_i} \times 100\%$$

3）统计若要召回第i条语句的所有标准相似句需要包含的最少噪声句数目，记为Sen_i，则准确率

$$P_i = \frac{M_i}{Sen_i} \times 100\%$$

为方便统计，Sen_i超过 20 的句子统一记为 20.

4）评价函数公式：

$$F = \frac{1}{n}\sum_{i=1}^{n}\frac{2R_iP_i}{R_i + P_i}$$

其中n为标准集例句总数。

3.4 实验结果与分析

方法 1 和方法 2 分别对应 3.2 节中的两种传统的计算句子相似度的算法；后三种方法均采用基于语义角色标注的算法，其中：方法 3 中动词相似度使用词向量计算，其他词语相似度使用 How-net 计算；方法 4 中全部使用 How-net 计算；方法 5 中全部使用词向量计算。

四轮测试得到的数据如表 3.1-3.2 所示：

表 3.1 第一、二轮实验结果

方法	第一轮			第二轮		
	召回率	准确率	F 值	召回率	准确率	F 值
1	54.67%	48.89%	51.61%	52.50%	41.25%	46.20%
2	49.67%	44.17%	46.76%	55.83%	48.00%	51.62%
3	70.00%	56.67%	62.63%	78.33%	73.64%	75.91%
4	54.00%	50.00%	51.92%	55.83%	50.15%	52.84%
5	59.33%	57.79%	58.55%	80.83%	78.33%	78.56%

表 3.2 第三、四轮实验结果

方法	第三轮			第四轮		
	召回率	准确率	F 值	召回率	准确率	F 值
1	58.67%	50.33%	52.18%	53.43%	45.17%	48.95%
2	52.00%	45.61%	47.60%	66.43%	61.33%	63.78%
3	76.00%	74.19%	74.93%	77.64%	75.49%	75.65%
4	63.67%	59.52%	60.53%	57.79%	50.33%	53.80%
5	68.67%	65.60%	66.10%	77.64%	75.49%	76.13%

将多次实验的结果取平均值，得到图 3.1：

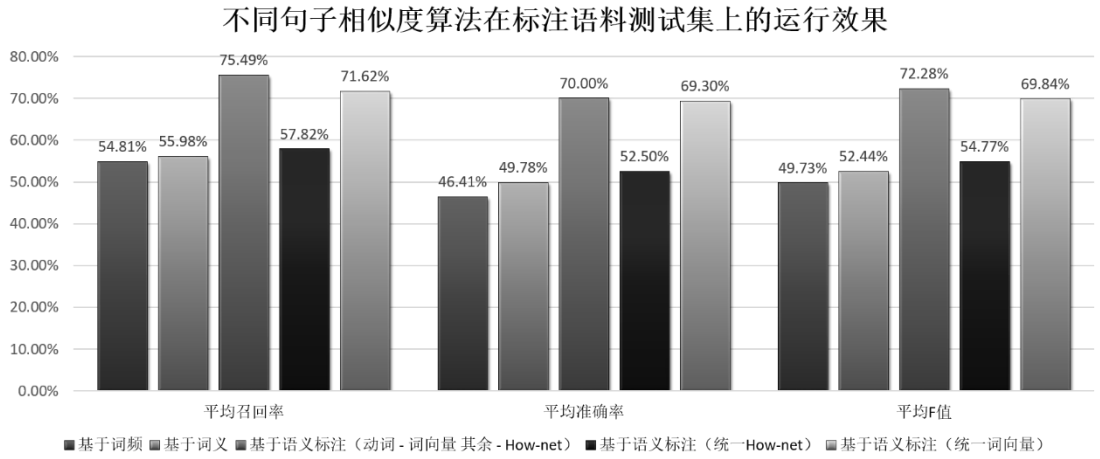

图 3.1 不同句子相似度算法在标注语料测试集上的运行效果

从表中可看到，传统的两种算法效果较为接近，F 值基本在 50%左右，而基于语义角色标注的方法效果整体要优于传统算法。分别观察方法 1 和方法 2 中的错误结果发现，方法 1 对于和相似句拥有重复词较多的标准句，测试结果非常高；但常常无法找出相同词语较少却语义相似的句子，对于这类句子其召回率通常为 0，因而整体效果不好。方法 2 考虑了词汇的语义，但由于对句子没有进行结构上的划分，所有词权重均相等，因而对于包含词汇较多的句子无法综合刻画句子的语义。

后三种方法中的方法 3 和 5，即动词使用词向量来计算相似度的方案平均 F 值最高可以达到 75%以上，而采用基于 How-net 计算相似度算法的结果则相对较低，从一定程度上验证了第一章中对词向量和 How-net 效果的比较和分析。方法 3 和 5 的效果较为接近，在不同的测试语料中这两种方法的效果小范围波动。

3.5 对算法的进一步思考

在对基于语义角色标注方法在实验中的错误结果的分析后发现，汉语中常用的许多单字节动词如“在”、“是”、“有”、“要”等，它们的意义比较宽泛，能够表达多种语义，用法也很灵活，在句中对其独立分析比较困难，因此在目前的算法中若它们被标注为动词往往会影响句子相似度的准确程度。初步想法是对这类特殊动词的义项进行细化，通过句子中的其他词汇信息判断其在某种上下文关系中的用法，对提升句子相似度结果会有所帮助。

另一方面，当前算法将所有标注句的权重统一按照相等处理，若要更准确度量句子间的相似度，更科学的方法是应该考虑标注句本身的同时考虑其重要性，即权重。一个很自然的想法是，通过对句子的句法分析将句子结构转换为树结构，越靠近根节点的词支配的部分更广，从而具有更高的相似度。但在人的阅读认知中有可能出现与此相反的情况。并且目前可用于分析汉语句子结构的工具（如 Stanford Parser）在分析较长的汉语句子时提炼出的结构不完全准确。这种权重度还有待更多的研究和实验进一步分析。

4 结论

本文提出了一种基于语义角色标注的汉语句子语义相似度计算方法，以动词为核心，以标注句为基本单元，结合词语相似度来计算句子间的相似度。相比传统的方法，本文的方法更加适用于经过了语义角色标注的汉语句子的相似度分析，使句子相似度的度量结果更为准确。在实验所用的语料中显示，本文方法相比其他传统的句子相似度算法能够在相似度测试中获得更好的效果，更加接近于人的相似排序。但算法对汉语中一些特殊动词的处理不够细

致，对不同动词所在的标注句的重要度衡量还需进一步的研究。

下一步我们将对汉语的特殊动词、标注句重要度的衡量两方面展开研究与分析，继续提升将基于语义角色标注的句子相似度算法的准确率，最终应用于 973 项目的相似句检索功能中，为集中研究分析句子中的相似标注提供参照。

参考文献

[1] 秦兵，刘挺 等．基于常问问题集的中文问答系统研究[J]．哈尔滨工业大学学报，2003，35(10): 1179-1182.

[2] Li S.J., et al. Semantic computation in a Chinese question-answering system. Journal of Computer Science and Technology, 2002, 17 (6): 933-939.

[3] 穗志方，俞士汶．基于骨架依存树的语句相似度计算模型[A]．中文信息处理国际会议(ICCIP\'98)[C]，北京：清华大学出版社, 1998, 458-465.

[4] 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, 20(12): 15-17.

[5] 车万翔 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, 14(7): 15-20.

[6] E. Ristad and P. Yianilos, Learning String Edit Distance. IEEE Trans. PAMI, 1998, 20(5): 522-523.

[7] 晋耀红 等. 基于语境框架的文本相似度计算[J]．计算机工程与应用, 2004, 40(16): 36-39.

[8] 潘谦红, 史忠植 等. 基于属性论的文本相似度计算[J]．计算机学报, 1999, 22(6): 651-655.

[9] Chatterjee N. A Statistical approach for similarity measurement between sentences for EBMT. 1999.

[10] 田久乐，赵蔚．基于同义词词林的词语相似度计算方法[J]．吉林大学学报(信息科学版)，2010，28(6): 602-608.

[11] 刘群，李素建．基于《知网》的词汇语义相似度计算[C]// 第三届汉语词汇语义学研讨会，台北，2002.

[12] 维基百科 Word2vec 词条页面[OL]. https://zh.wikipedia.org/wiki/Word2vec.

[13] LDC(Linguistic Data Consortium)主页[OL]. https://www.ldc.upenn.edu/language-resources/data.

[14] 维基百科 Skip-gram 词条页面[OL]. https://en.wikipedia.org/wiki/N-gram#Skip-gram.

[15] Xue, Nianwen and Martha Palmer, 2009, Adding semantic roles to the Chinese Treebank[J]. Natural Language Engineering, 2008, 15(1):143-172.

[16] 秦兵，刘挺等．基于常问问题集的中文问答系统研究[J]．哈尔滨工业大学学报，2003(10): 1179-1182.

[17] Ji Wenqian, Li Zhoujun, Chao Wenhan, et al. A new method for calculating similarity between sentences and application on automatic abstracting[J]. Intelligent Information Management, 2009, 1(1): 38-45.

[18] Ru Li，Zhiqiang Wang，Shuanghong Li，Jiye Liang，Collin Baker, Chinese sentence similarity computing based on frame semantic parsing[J]. Journal of Computer Research and Development, 2013, 50(8): 1728-1736.

作者简介及照片

1. 第一作者：田堃（1994—），女，本科生，专业为计算机科学与技术

2. 通讯作者：柯永红（1981—），男，讲师，研究方向为计算语言学，电话：86-10-62753081，E-mail：kyh@pku.edu.cn

3. 第三作者：穗志方（1970—），女，教授，博士生导师，研究方向为计算语言学。