ISSN 1000-1239/CN 11-1777/TP
50(8): 1728-1736, 2013

基于框架语义分析的汉语句子相似度计算

李 茹[1,2] 王智强[1] 李双红[1] 梁吉业[2] Collin Baker[3]
[1](山西大学计算机与信息技术学院 太原 030006)
[2](计算智能与中文信息处理教育部重点实验室(山西大学) 太原 030006)
[3](International Computer Science Institute, Berkeley, California, 94704)
(liru@sxu.edu.cn)

Chinese Sentence Similarity Computing Based on Frame Semantic Parsing

Li Ru[1,2], Wang Zhiqiang[1], Li Shuanghong[1], Liang Jiye[2], and Collin Baker[3]
[1](*School of Computer & Information Technology, Shanxi University, Taiyuan* 030006)
[2](*Key Laboratory of Computation Intelligence & Chinese Information Processing (Shanxi University), Ministry of Education, Taiyuan* 030006)
[3](*International Computer Science Institute, Berkeley, California* 94704)

Abstract Sentence similarity computing plays an important role in many tasks of natural language processing. Recent approaches to sentence similarity computing have focused on word-level information without considering the semantic structural information; these methods based on the sentence structure are not generally desirable as they are severely affected by the incomplete description of sentence semantic. Hence, similarity computing isn't able to get better results. To solve this problem, this paper proposes a novel similarity computing approach based on Chinese FrameNet. The approach implements to measure the sentences' semantics similarity by multi-frame semantic parsing, importance measure of frames, similar match of frames, similarity computing between frames and so on. From the frame perspective, the multi-frame semantic parsing comprehensively describes sentences' semantics by identifying all the target words, choosing corresponding frames and labeling the frame elements. On that basis, the similarity result can be more accurate by distinguishing the different frames' importance in accordance with the semantic coverage area of the frame. In addition, by means of extracting the semantic core words of the frame element, the approach improves the precision of similarity among the frames of chunk form. The sentences which contain multiple target words are chosen as the corpus of the experiments. In contrast with traditional approaches, the results show that the proposed approach could achieve better similarity results.

Key words Chinese FrameNet; multi-frame semantic parsing; sentence semantic similarity; frame similarity; frame importance measure

摘 要 句子相似度计算在自然语言处理的许多领域中发挥着重要作用.已有的汉语句子相似度计算方法由于考虑句子的语义不全面,使得相似度计算结果不够准确,为此提出一种新的汉语句子相似度计算方法.该方法基于汉语框架网语义资源,通过多框架语义分析、框架的重要度度量、框架的相似匹配、框

收稿日期:2011-12-26;**修回日期**:2012-07-03
基金项目:国家自然科学基金项目(60970053);国家语委"十二五"科研规划项目(YB125-19);国家"八六三"高技术研究发展计划基金项目(2006AA01Z142);山西省国际科技合作项目(2010081044);山西省重点实验室开放基金项目(2009011059-4)

架间相似度计算等关键步骤来实现句子语义的相似度量.其中多框架语义分析是从框架角度对句子中的所有目标词进行识别、框架选择及框架元素标注,从而达到全面刻画句子语义的目的;在此基础上根据句子中框架的语义覆盖范围对不同框架的重要度进行区分,能够使得相似度结果更准确.在包含多目标词的句子集上的实验结果显示,基于多框架语义分析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　汉语框架网;多框架语义分析;句子语义相似度;框架相似度;框架重要度

中图法分类号　TP391

句子相似度计算是对句子间的相似性给出一个度量,它在自然语言处理的许多领域中都发挥着重要的作用[1].例如,在基于实例的机器翻译系统中[2],需要通过句子的相似度计算来返回最优的翻译结果;在自动问答系统、专家系统、信息检索[3-5]等应用领域,一方面用户所提出的问题与系统问题库之间需要用问题相似度来获得问题匹配,另一方面对于答案抽取模块也需要利用问题与候选答案间的相似度来返回;在自动文摘系统中[3-6]需要用到句子相似度计算来去除冗余信息等.

目前针对汉语句子相似度计算,从句子的不同分析形式来看主要分为两类: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结构层面的相似度计算,包括基于句法分析、语义分析等.

基于词统计特征的方法主要是通过考虑词频、词性等信息来度量句子间的相似度,如向量空间模型的方法,它将语料库中的句子通过每个词的 TF-IDF 表示为向量,相似度用向量的夹角余弦表示.该方法只考虑句子的表层信息,并未考虑句子结构、语义等深入特征,只有当语料具有一定的规模时这种统计的效果才会体现出来,具有一定的局限性.基于词汇语义特征的方法主要依赖 HowNet[7],WordNet[8]及同义词词林[9]等词汇语义词典.李素建[10]基于 HowNet 和同义词词林提出了语句相关度的定量计算模型,该方法充分考虑了词的语义特征,能够将表面信息不同而语义信息相同的词挖掘出来.然而该方法较为依赖语义词典的完整性,词典的不全面和一些未登录词的语义代码的缺失将直接影响计算的准确性.因此,基于词汇语义特征的方法也存在一定的局限性.

基于句子结构特征的有李彬等人[11]提出的基于语义依存的句子相似度,该方法考虑了依存句法树中词之间的语义依存,句子的相似度由依存树中有效语义搭配对之间的相似度来度量(有效语义搭配是指句子中核心词与其直接依存成分).由于该方法仅考虑句子的核心语义依存结构,未将全部语义依存信息考虑在内,因此句子相似度度量结果不够准确.Li 等人[12]提出了一种基于框架语义依存图的句子相似度计算模型,通过计算句子核心框架间的各种参与者及外围语义成分来获得句子整体的语义相似度.该方法从框架语义的角度能够有效地刻画句子的核心语义,然而只考虑核心框架,外围语义成分处理以词集合的形式进行处理,使得包含多框架的句子语义缺失,影响最终相似度的准确性.

目前基于词层面特征的相似度计算方法未考虑句子的结构信息;基于句子结构特征的相似度计算方法未能全面考虑句子语义.本文将从框架角度出发,首先通过多框架语义分析全面考虑句子的语义;其次对句子中的不同框架进行重要度度量;最终以框架为基本单元,结合不同框架的重要度来度量句子间的相似度.

1　句子的框架语义分析

汉语句子的框架语义分析是以汉语框架网(Chinese FrameNet, CFN)[13]为基础,通过 CFN 中定义的框架来对句子进行语义分析.国际上针对英文的框架语义分析任务于 2007 年举行过评测[14],具体任务有目标词识别、框架选择、框架元素标注.汉语的框架语义分析也同样包括此 3 项任务,其中涉及的框架、框架元素、词元以及目标词等相关概念如文献[13,15-16]所述.

1.1　单框架语义分析

单框架语义分析是通过单一框架来实现句子的语义分析.首先通过目标词识别,确定句子中一个可以激起框架的词;其次通过框架选择,为目标词选取相应的框架;最终通过框架元素标注,对句子中目标词所支配的框架元素进行识别与分类.

例句 1. 美国总统决定前往北京.

若句中识别出一个目标词“前往”,框架选择结果为“到达”,最终例句 1 的单框架语义分析结果如下:

〈thm 美国/ns 总统/n〉 决定/v 〈tgt=到达 前往/v〉 〈goal 北京/ns〉./w,

其中,“〈tgt=到达 前往/v〉”表示目标词“前往”所属框架为“到达”,“〈thm 美国/ns 总统/n〉”、“〈goal 北京/ns〉”表示目标词“前往”所支配的框架元素,其中“thm”,“goal”为框架元素类型标记,分别指“转移体”与“目的地”.

1.2 多框架语义分析

例句 1 的单框架语义分析仅刻画了句子的一个框架“到达”,而未考虑“决定”框架的语义.像这样包含多目标词的句子若只进行单框架语义分析则其他框架的语义很难体现,甚至会缺失句子的核心语义,在此基础上将难以准确度量句子间的相似度.

全面刻画句子的语义是准确度量句子相似度的关键,要从框架角度准确度量句子的相似度,则必须通过多框架语义分析来全面刻画句子的语义.

如例句 1,既要进行“到达”框架的语义分析,又要进行“决定”框架的语义分析:

1)〈cog 美国/ns 总统/n〉〈tgt=决定 决定/v〉〈dec 前往/v 北京/ns〉./w;

2)〈thm 美国/ns 总统/n〉 决定/v 〈tgt=到达 前往/v〉 〈goal 北京/ns〉./w.

此时例句 1 中“决定”与“前往”两个框架的语义全部体现出来,将这种对句子进行多框架的语义分析称为多框架语义分析.只考虑句子的核心框架,如例 1 只考虑“决定”框架时称为核心框架语义分析.

1.3 框架的重要度度量

当句子包含多个框架时不同框架的重要性并不一定相同,要准确度量句子间的相似度,则必须在考虑框架本身的同时考虑其重要性.然而度量句子中框架的重要度并非易事,因为依据不同的重要度度量标准,度量的结果并非一成不变.例如“妈妈喊你回家吃饭”中的“喊”与“吃”,若依据口语中的语气来衡量,“喊”语气更重时,则“喊”更为重要,反之“吃”更为重要.因此框架重要度度量标准选择是框架重要度度量的关键.

针对本文的句子语义相似度计算问题,选择从框架对句子的覆盖范围来衡量框架的重要度:一个框架对于句子的覆盖范围越大则认为此框架对于句子越重要(注意这里的框架重要度只针对本文的句子语义相似度计算问题),例如:

例句 1 分别以“决定”、“前往”为目标词的多框架语义分析结果中:

1)中“决定”框架涵盖整个句子,而 2)中“到达”框架未涵盖整个句子,即只涵盖了句子的一部分;2)的目标词“前往”嵌入于 1)的框架元素“〈dec 前往/v 北京/ns〉”中,即 2)的“到达”框架在 1)中仅作为“决定”框架的一个框架元素.此时,认为 1)中“决定”框架相对于 2)中“前往”框架更为重要.

为了便于计算,记例句 1 框架语义分析结果 1)中框架“决定”为第 1 层,2)中框架“到达”为第 2 层.容易看出,框架的层数越多语义重要度越小.

根据以上分析,设 F_{ij} 为句子 S 第 i 层的第 j 个框架,将其重要度函数定义为

$$\eta(F_{ij})=\frac{\alpha}{i-1+\alpha}, \tag{1}$$

其中 α 为(0,1)之间的可调节参数,$0<\eta(F_{ij})\leqslant 1$.

要注意有时句子中框架之间并不具有相互嵌入的关系,此时认为它们属于同一层,即语义重要度相同.

2 基于框架语义分析的句子相似度计算

句子的多框架语义分析利用重要度不同的多个框架来对句子进行刻画.依据整体与部分的思想,整体相似建立在部分相似基础上,本文句子的相似度能够分解为各个框架之间的相似度.

在比较两个整体间的相似度时,只有将扮演相同角色的部分进行比较才有意义.文献[17]作实词间的相似度时提到一个较为形象的例子:比较两个人长相是否相似,一般总是比较他们的脸型、眼睛、鼻子等相同部分是否相似,而不会拿眼睛去和鼻子作比较.因此,在进行句子相似度比较时,需要对两个句子中的各个框架进行相似匹配,从而实现扮演相同角色部分的比较.

以下首先介绍框架间的相似度度量,其次进行框架的相似匹配,最后给出句子之间的相似度度量方法.注意,每一节中所用到的字母及下标都只在本节中起作用.

2.1 框架间的相似度计算

框架一般包含多个“目标词→语义类型→框架元素”结构,这种结构称为语义搭配.将框架之间的相似度转化成语义搭配对之间的相似度,所谓语义搭配对[11]是指两个框架中语义类型相同的两个语义搭配,例如:

例句 2. 〈tot 阳光/n 〉〈tgt=包含 包含/v 〉〈par 各种/r 横波/n 〉./w;

例句 3. 〈tot 阳光/n 〉由/p〈par 不同/a 的/u 横波/n 〉〈tgt=包含 组成/n〉./w.

例句 2 中的"包含→tot→阳光"与例句 3 中的"组成→tot→阳光"以及例句 2 中的"包含→par→各种 横波"与例句 3 中的"组成→par→不同 的 横波"为语义搭配对.

对于一个含有 m 个框架元素的框架 F,用 a 表示它的目标词,$e(F)=\{e_1,e_2,\cdots,e_m\}$表示 F 中所有框架元素的集合,$r(F)=\{r_1,r_2,\cdots,r_m\}$表示 F 中所有语义类型的集合,框架 F 可以表示为一个三元组 $(a,e(F),r(F))$.

将框架 F_1 与 F_2 的相似度定义为

$$Sim(F_1,F_2)=\frac{\sum\limits_{(i,j)}\frac{1}{2}(Sim(e_i,e_j)+Sim(a_1,a_2))}{\max(m,n)},\tag{2}$$

其中:$(i,j)\in\{(p,q)\mid r_p,r_q\in r(F_1)\bigcap r(F_2),1\leqslant p\leqslant m,1\leqslant q\leqslant n\}$,$m$ 与 n 分别为框架 F_1 与 F_2 中所含有的框架元素个数;$Sim(a_1,a_2)$为两个目标词 a_1 与 a_2 之间的相似度,采用基于知网的词汇语义相似度计算[17](本文所有涉及词汇之间的相似度,均采用此方法实现);$Sim(e_i,e_j)$代表框架元素 e_i 与 e_j 之间的相似度,其中当 e_i,e_j 均是词时,计算方法与 $Sim(a_1,a_2)$相同.

当 e_i,e_j 至少有一个是词块时,如,e_i="我 的爸爸",e_j="爸爸 的 爱好",此时希望比较的是"爸爸"和"爱好"之间的相似性,如果不进行一定预处理,难以通过直接计算来得到合理的相似度结果.本文在计算词块间的相似度时对其进行了核心语义分析,即剔除修饰成分,提取语义核心词.采用课题组已构建的语义核心词提取规则集[18]来提取.

在提取核心词后,e_i 或 e_j 还有可能是词块形式,如"以 村委会 选举 的 方式"提取核心词为"选举 方式".此时,$Sim(e_i,e_j)$的计算方法如下[19]:

将 e_i 和 e_j 看作两个词集合,分别包含 M,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},得到相似度矩阵为

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

取 $s_m=\max(s_{m1},s_{m2},\cdots,s_{mN})$,$s_n=\max(s_{n1},s_{n2},\cdots,s_{nM})$,词集合之间的相似度为

$$Sim(e_i,e_j)=\frac{1}{2}\times\left(\frac{\sum\limits_{m=1}^{M}s_m}{M}+\frac{\sum\limits_{n=1}^{N}s_n}{N}\right).\tag{3}$$

2.2　框架的相似匹配

框架相似匹配就是将句子间具有相似语义的框架进行配对,可以直接先对两个句子的所有框架两两之间进行相似度计算,然后从计算结果中获得框架相似匹配结果.但从 2.1 节可知,当对所有框架两两之间都进行相似度计算时复杂度很大.

一个框架由目标词及其所支配的框架元素构成,其中目标词决定了这个语义场景的"动作"是框架的承担者,是激起框架的核心.2.1 节描述的框架间相似度是由语义搭配对之间的相似度得到,其中每计算一次语义搭配对间的相似度时就需要计算一次目标词之间的相似度.因此目标词是度量框架间相似度的一个最重要的因素,虽然目标词间的相似度不能代替框架间的相似度,但已足够区分框架间是否具有一定的相似性.本文则利用了这一点通过目标词的相似匹配来实现框架的相似匹配.

设句子 $\boldsymbol{S}_1$ 中第 i 个目标词和 $\boldsymbol{S}_2$ 中第 j 个目标词之间的相似度为 Sim_{ij},则容易得到目标词之间的相似度矩阵:

$$\boldsymbol{\Lambda}=\begin{pmatrix}Sim_{11}&\cdots&Sim_{1n}\\ \vdots&&\vdots\\ Sim_{m1}&\cdots&Sim_{mn}\end{pmatrix}.$$

m 和 n 分别为两个句子中目标词的个数,也是框架的个数,不妨设 $m<n$.循环执行以下两个步骤,直到矩阵 $\boldsymbol{\Lambda}$ 所有的行(或列)都被删除.

1) 求出矩阵 $\boldsymbol{\Lambda}$ 中最大的元素 $Sim_{pq}=\max Sim_{ij}$,$1\leqslant i\leqslant m$,$1\leqslant j\leqslant n$,则将 $\boldsymbol{S}_1$ 中第 p 个目标词与 $\boldsymbol{S}_2$ 中第 q 个目标词作为已匹配的目标词对.

2) 删除矩阵 $\boldsymbol{\Lambda}$ 中 Sim_{pq}所在的行与列,得到的新矩阵赋给 $\boldsymbol{\Lambda}$.

由此得到 m 组目标词配对,对应 m 组框架配对.此时依据 2.1 节框架间的相似度计算方法能够得到相似匹配后框架间的相似度.

2.3　句子的相似度计算

对于一个包含 m 层框架的句子 $\boldsymbol{S}$,设第 i 层的框架个数为 n_i,则 $\boldsymbol{S}$ 所包含的所有框架用集合表示为 $F(\boldsymbol{S})=\{F_{11},F_{12},\cdots,F_{1n_1},F_{21},F_{22},\cdots,F_{2n_2},\cdots,F_{m1},F_{m2},\cdots,F_{mn_m}\}$,元素个数为 $\sum\limits_{i=1}^{m}n_i$.

对于句子 $\boldsymbol{S}_1$ 与 $\boldsymbol{S}_2$,通过 2.2 节中框架间的相似匹配得到框架匹配对:

$$(F_1^1,F_1^2),(F_2^1,F_2^2),\cdots,(F_k^1,F_k^2),$$

其中 $F_l^1 \in F(\boldsymbol{S}_1), F_l^2 \in F(\boldsymbol{S}_2), l=1,\cdots,k$. 通过 2.1 节中框架间相似度的计算方法,得到框架匹配对的相似度结果为

$$Sim(F_1^1,F_1^2), Sim(F_2^1,F_2^2),\cdots,Sim(F_k^1,F_k^2).$$

不同的框架承担了句子的不同语义内容,且不同层次的框架对句子的重要度也是不相同的. 由 1.3 节可得出 F_l^1 在句子 $\boldsymbol{S}_1$ 中的重要度 $\eta(F_l^1)$ 和 F_l^2 在句子 $\boldsymbol{S}_2$ 中的重要度 $\eta(F_l^2)$,取 $\min(\eta(F_l^1),\eta(F_l^2))$ 为这个框架匹配对的重要度,则两个句子的相似度计算公式如下:

$$Sim(\boldsymbol{S}_1,\boldsymbol{S}_2)=\frac{\sum_{l=1}^{k}(Sim(F_l^1,F_l^2)\times\min(\eta(F_l^1),\eta(F_l^2)))}{\max\left(\sum_{i,j}\eta(F_{ij}^1),\sum_{i,j}\eta(F_{ij}^2)\right)}. \quad (4)$$

这种考虑了所有可匹配的框架来计算句子相似度的方法称为基于多框架语义分析的句子相似度计算.

若只考虑句子的核心框架则称为基于核心框架语义分析的句子相似度计算. 此时,只需要计算第 1 层的框架(即核心框架),$F(\boldsymbol{S})=\{F_{11},F_{12},\cdots,F_{1n_1}\}$,且 $\eta(F_{1j})=1, j=1,\cdots,n_1$. 对于两个句子 $\boldsymbol{S}_1$ 与 $\boldsymbol{S}_2$,核心框架的个数分别为 n_1 和 n_2. 得到的核心框架匹配对为 $(F_1^1,F_1^2),(F_2^1,F_2^2),\cdots,(F_k^1,F_k^2)$,其中 $F_l^1 \in F(\boldsymbol{S}_1), F_l^2 \in F(\boldsymbol{S}_2), l=1,\cdots,k$. 则式(4)变为

$$Sim(\boldsymbol{S}_1,\boldsymbol{S}_2)=\frac{\sum_{l=1}^{k}Sim(F_l^1,F_l^2)}{\max(n_1,n_2)}. \quad (5)$$

本文强调句子语义的完整性对最终相似度结果的影响,这种基于核心框架语义分析的句子相似度计算只考虑句子的核心框架,在第 3 节的实验中将会用此方法与基于多框架语义分析的句子相似度计算进行实验对比.

3 实验及结果分析

3.1 实验语料准备

目前国际上还没有关于汉语句子相似度计算的公共测试集,测试语料一般通过人工构建[20-21]. 由于没有统一标准,人工选取相似句子时主要依赖主观判断,因此获取高质量的相似语料并非易事. 本文所用语料是通过百度、Google 搜索平台以及北京大学 CCL(网络版)语料资源平台,由课题组 16 人进行相似句子分组搜集、相似度交叉打分、最终经相似度平均分值排序、筛选获得.

最终构建的语料大多包含多个目标词,句子一般较长,包括相似句与噪声句共 320 条,按实验需求划分为两部分:一部分为 30 的标准集;另一部分是针对标准集构建的测试集,包括 90 的匹配集和 200 的噪声集. 90 的匹配集为 30 组×3 的相似句,每组中的 3 条句子与 30 的标准集中的 1 条句子对应相似. 200 的噪声语料全部来源于 CFN 句子库,为满足噪声特点,其目标词均来自 30 组相似句中的目标词.

语料依据不同部分实验的要求,进行分词、词性标注、依存句法分析、目标词识别、框架选择、框架元素标注等预处理. 其中分词、词性标注使用了中国科学院计算技术研究所 ICTCLAS 2010 分词工具,依存句法分析使用了哈尔滨工业大学的依存句法分析器 LTP2.0[22]. 由于目前汉语句子的框架语义自动分析性能较低[23-24],特别是框架元素标注的 F 值仅为 60%左右,因此目标词识别、框架选择及框架元素标注工作均通过人工参与进行标注及矫正.

3.2 评价指标

本文的句子相似度测试结果采用正确率 P 来评价[11,21],具体做法为:从标准集中按顺序依次抽出第 $i(1\leqslant i\leqslant n)$ 条句子,与测试集中的所有句子计算相似度. 从每次的测试结果中取出相似度由大到小排名前 M 的句子,记 M 条句子中为人工设定的相似句数目为 $CorrectSen_i$,即找出的句子与人工设定的相似句一致时才认为该句子为找出的正确结果. 其中标准集大小为 30,即 $n=30$;由于在本文的实验语料中人工设定的每条标准句对应的相似句为 3 条,因此应将 M 设定为 3.

则正确率 P 为

$$P = AvgP_i = Avg\sum_{i=1}^{n}\frac{CorrectSen_i}{M}\times 100\%. \quad (6)$$

3.3 实验对比的相关方法

1) 基于向量空间模型(vector space model, VSM)的相似度计算

对于目标句子 $\boldsymbol{S}_1$ 与 $\boldsymbol{S}_2$ 计算句中每个词的 $TF\text{-}IDF$ 值,得到句子对应的向量分别为 $\boldsymbol{S}_1=(T_1,T_2,\cdots,T_n)$ 和 $\boldsymbol{S}_2=(T_1',T_2',\cdots,T_n')$,句子间的相似度利用两个向量之间的夹角余弦值来表示:

$$Sim(\boldsymbol{S}_1,\boldsymbol{S}_2)=\frac{\sum_{i=1}^{n}(T_i\times T_i')}{\sqrt{\sum_{i=1}^{n}T_i^2\times\sum_{i=1}^{n}T_i'^2}}. \quad (7)$$

2) 基于知网语义(HowNet semantic, HNS)的相似度计算[19]

对于目标句子 $\boldsymbol{S}_1$ 与 $\boldsymbol{S}_2$ 中的任意词 A_i 与 B_j,

基于 HowNet 计算其相似度 $S(A_i,B_j)$,取 $a_i=\max(S(A_i,B_1),S(A_i,B_2),\cdots,S(A_i,B_n))$,$b_i=\max(S(B_i,_{A1}),S(B_i,_{A2}),\cdots,S(B_i,_{An}))$,则目标句子 $\boldsymbol{S}_1$ 与 $\boldsymbol{S}_2$ 之间的相似度为

$$Sim(\boldsymbol{S}_1,\boldsymbol{S}_2)=\left(\frac{\sum_{i=1}^{m}a_i}{m}+\frac{\sum_{i=1}^{n}b_i}{n}\right)\Big/2. \quad (8)$$

3) 基于语义依存(semantic dependency, SD)的相似度计算[11]

首先对句子进行依存句法分析,在依存句法树上通过匹配有效搭配对,分别计算句子的句法相似度 $Sim_1(\boldsymbol{S}_1,\boldsymbol{S}_2)$与语义相似度 $Sim_2(\boldsymbol{S}_1,\boldsymbol{S}_2)$,并通过加权最终得到句子间的相似度:

$$Sim(\boldsymbol{S}_1,\boldsymbol{S}_2)=\lambda\times Sim_1(\boldsymbol{S}_1,\boldsymbol{S}_2)+(1-\lambda)\times Sim_2(\boldsymbol{S}_1,\boldsymbol{S}_2). \quad (9)$$

4) 基于核心框架语义分析(core frame semantic parsing, CFSP)的相似度依据式(5)进行计算.

5) 基于多框架语义分析(multi-frame semantic parsing, MFSP)的相似度依据式(4)进行计算.

3.4　实验结果及分析

实验包含以下 4 部分:

1) 基于核心框架语义分析与多框架语义分析的相似度计算结果对比;

2) 框架元素核心词块提取前后基于多框架语义分析的相似度计算结果对比;

3) 基于多框架语义分析的相似度计算与其他相似度计算方法的结果对;

4) 不同相似度计算方法分别受噪声集的影响.

3.4.1　基于 CFSP 与 MFSP 方法的结果比较

对核心框架语义分析 CFSP 与多框架语义分析 MFSP 的句子相似度方法进行对比,目的是分析当考虑句子语义不全面时会多大程度上影响相似度结果.

对于重要度函数(式(1))中的参数 α,经多次实验,$\alpha=0.9$ 时结果最优. 因此,以下实验中 α 均设定为 0.9.

在 30 的标准集与 90 的匹配集中(规模 120)的测试结果如表 1 所示:

Table 1　The Comparison Between CFSP and MFSP

表 1　CFSP 与 MFSP 方法的实验结果比较　%

Experiments	P
CFSP	50.00
MFSP	87.78

基于 MFSP 的方法明显优于基于 CFSP 的方法,正确率 P 高出 37%. 追踪 CFSP 的错误结果,其中多数错误结果出现在包含多框架的句子中,这主要由于 CFSP 只考虑句子中的单一框架框,无法全面考虑包含多框架句子的语义,影响最终相似度的度量结果(本节的实验中还未提取框架元素的核心词块,3.4.2 节将进行核心词词块提取前后的实验对比).

3.4.2　核心词块提取前后结果比较

语义核心词提取前后基于 MFSP 的相似度测试结果如表 2 所示,测试语料同 3.4.1 节.

Table 2　The Comparison Between the Experimental Results Before and After Extracting Semantic Core of the Word

表 2　语义核心词提取前后的相似度测试结果比较　%

Experiments	P
MFSP(Before extracting)	87.78
MFSP(After extracting)	91.11

在框架元素的核心词块提取后相似度测试结果有明显提升,正确率 P 提高 3%. 证明词块形式框架元素的相似度一定程度上影响句子的相似度.

3.4.3　基于 MFSP 与其他方法的结果比较

基于 MFSP 相似度计算方法与相关相似度计算方法的对比,测试语料同 3.4.1 节,测试结果如表 3 所示:

Table 3　The Experimental Comparison with Some Other Sentences Similarity Methods

表 3　本文的方法与其他句子相似度方法的实验结果比较　%

Experiments	P
VSM	88.89
HNS	72.22
SD	33.33
MFSP	91.11

基于 MFSP(提取核心词后)的方法优于其他 3 种方法. 其中 VSM 方法的结果较为接近,而基于句法语义依存 SD 方法的结果较低. 追踪 SD 方法中的错误结果发现:在包含多目标词的语料中,句法分析的错误较多;对于句法分析正确但相似度测试结果不正确的句子,SD 方法中的有效搭配对提取难以全面刻画句子的语义,但同样的例句在基于 MFSP 方法进行处理时结果会有明显提升.

3.4.4　不同相似度计算方法分别受噪声集的影响

在 3.4.1 节语料基础上,加入具有相同目标词的噪声集,噪声集规模分别为 50,100,150,200. 测试

结果如表 4 所示：

Table 4 Experimental Results on the Different Similarity Calculation Methods and Different Scale of Noise Corpus

表 4 在不同相似度方法及不同噪声语料规模下的实验结果

%

Data Sets	Experiments	*P*
120+0	VSM	88.89
	HNS	72.22
	SD	33.33
	MFSP	91.11
120+50	VSM	87.78
	HNS	70.00
	SD	31.11
	MFSP	91.11
120+100	VSM	86.67
	HNS	68.89
	SD	31.11
	MFSP	90.00
120+150	VSM	88.89
	HNS	67.78
	SD	26.67
	MFSP	90.00
120+200	VSM	90.00
	HNS	67.78
	SD	26.67
	MFSP	88.89

相似度测试结果受噪声集的影响如图 1 所示：

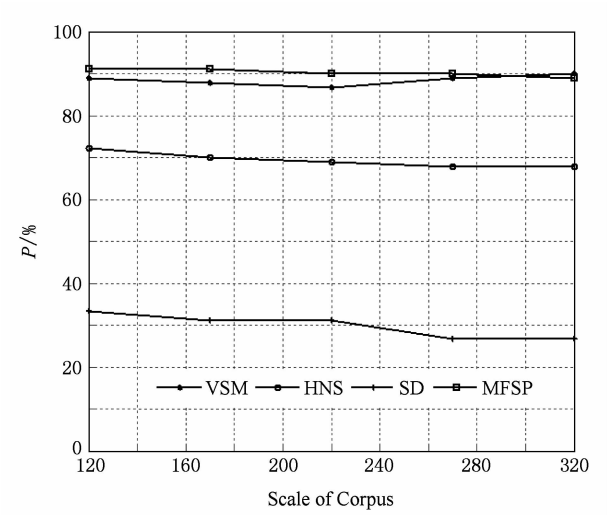


Fig. 1 The accuracy on the different scale of noise corpus.

图 1 在不同规模噪声语料上的正确率

随着噪声集的增加，每种方法的测试结果都有下降趋势. 其中 HNS，SD 下降较为明显，而 VSM 与 MFSP 较为稳定，且 MFSP 能够保持较高的正确率.

4 结论与展望

本文提出了一种基于框架语义分析的汉语句子语义相似度计算方法，首先通过对句子进行多框架的语义分析来刻画句子的语义；其次以框架为基本单元，结合框架的重要度来计算句子间的语义相似度. 本文方法对句子的语义考虑更全面，区分不同框架的重要度使句子相似度度量结果更准确. 在包含多目标词的语料中显示，本文方法相比其他方法能够在相似度测试中获得更好的结果. 但由于目前 CFN 语义资源中框架的覆盖率低，使得基于框架语义分析的句子相似度计算方法局限于处理资源库中已有框架的句子集数据.

下一步将在资源方面不断提高现有 CFN 的覆盖率，将基于框架语义分析的相似度计算方法从包含多目标词的句子集数据上拓展至大规模真实语料上，同时也将研究疑问句之间的相似度，并逐渐向在线互动平台、山西旅游问答等应用领域拓展.

致谢 本文实验用到中国科学院计算技术研究所的 ICTCLAS2010 分词工具；哈尔滨工业大学信息检索研究中心的依存句法分析器 LTP2.0；知网平台提供的词汇语义相似度计算工具，在此表示感谢！

参 考 文 献

[1] Lee M C. A novel sentence similarity measure for semantic-based expert systems [J]. Expert Systems with Applications, 2011, 38(5): 6392-6399

[2] Sui Zhifang, Yu Shiwen. The skeletal-dependency-tree-based computational model for the sentence similarity [C] //Proc of Int Conf on Chinese Information Processing. Beijing: Tsinghua University Press, 1998: 458-465 (in Chinese)
(穗志方，俞士汶. 基于骨架依存树的语句相似度模型[C] //中文信息处理国际会议录. 北京：清华大学出版社，1998：458-465)

[3] Aliguliyev R M. A new sentence similarity measure and sentence based extractive technique for automatic text summarization [J]. Expert Systems with Applications, 2009, 36(4): 7764-7772

[4] Zhao Jun, Jin Qianli, Xu Bo. Semantic computation for text retrieval [J]. Chinese Journal of Computers, 2005, 28(12): 2068-2078 (in Chinese)
(赵军，金千里，徐波. 面向文本检索的语义计算[J]. 计算机学报，2005，28(12)：2068-2078)

[5] Che Wanxiang, Liu Ting, Qin Bing, et al. Chinese sentence similarity computing for bilingual sentence pair retrieval [C] //Proc of the 7th Joint National Conference on Computational Linguistics, Beijing: Tsinghua University Press, 2003 (in Chinese)
(车万翔，刘挺，秦兵，等. 面向双语句对检索的汉语句子相似度[C] //全国第七届计算语言学联合学术会议录. 北京：清华大学出版社，2003)

[6] Zhang Qi, Huang Xuanjing, Wu Lide. A new method for calculating similarity between sentencees and application on automatic text summarization [J]. Journal of Chinese Information Processing, 2005, 19(2): 93-99 (in Chinese)
(张奇，黄萱菁，吴立德. 一种新的句子相似度及其在文本自动摘要中的应用[J]. 中文信息学报，2005，19(2)：93-99)

[7] Dong Zhendong, Dong Qiang. "HowNet" [EB/OL]. 1999 [2011-08-20]. http://www.keenage.com (in Chinese)
(董振东，董强. "知网". 1999 [2011-08-20]. http://www.keenage.com)

[8] Miller G A, Beckwith R, Fellbaum C D, et al. WordNet: An online lexical database [J]. Int Journal of Lexicography, 1990, 3(4): 235-244

[9] Mei Jiaju, Zhu Yiming, Gao Yunqi, et al. Synonyms Cilin [M]. Shanghai: Shanghai Lexicographical Publisher, 1983 (in Chinese)
(梅家驹，竺一鸣，高蕴琦，等. 同义词词林[M]. 上海：上海辞书出版社，1983)

[10] Li Sujian. The research of relevancy between sentences based on semantic computation [J]. Computer Engineering and Applications, 2002, 38(7): 75-76, 83 (in Chinese)
(李素建. 基于语义计算的语句相关度研究[J]. 计算机工程与应用，2002，38(7)：75-76,83)

[11] Li Bin, Liu Ting, Qin Bing, et al. Chinese sentence similarity computing based on semantic dependency relationship analysis [J]. Application Research of Computers, 2003, 20(12): 15-17 (in Chinese)
(李彬，刘挺，秦兵，等. 基于语义依存的汉语句子相似度[J]. 计算机应用研究，2003，20(12)：15-17)

[12] Li Ru, Li Shuanghong, Zhang Zezheng. The semantic computing model of sentence similarity based on Chinese FrameNet [C] //Proc of Web Intelligence/IAT Workshops. Los Alamitos, CA: IEEE Computer Society, 2009: 255-258

[13] Hao Xiaoyan, Li Ru, Liu Kaiying. Description systems of the Chinese FrameNet database and software tools [J]. Journal of Chinese Information Processing, 2007, 5(21): 96-100, 138 (in Chinese)
(郝晓燕，李茹，刘开瑛. 汉语框架语义知识库及软件描述体系[J]. 中文信息学报，2007，21(5)：96-100，138)

[14] Baker C, Ellsworth M, Erk K. SemEval-2007 Task 19: Frame semantic structure extraction [C] //Proc of the 4th Int Workshop on Semantic Evaluations. Quebec, USA: ACL, 2007: 99-104

[15] Fillmore C J. Frame semantics [C] //Proc of Linguistics in the Morning Calm. Seoul: Hanshin Publishing Co., 1982: 111-137

[16] Baker C F, Fillmore C J, Lowe J B. The Berkeley FrameNet project [C] //Proc of COLING/ACL. Quebec, USA: ACL, 1998: 86-90

[17] Liu Qun, Li Sujian. Word similarity computing based on Hownet [J]. Computational Linguistics and Chinese Language Processing, 2002, 7(2): 59-76 (in Chinese)
(刘群，李素建. 基于《知网》的词汇语义相似度[J]. 中文计算语言学，2002，7(2)：59-76)

[18] Li Shuanghong, Li Ru, Zhong Lijun, et al. Multi-word chunking based automatic identification of the semantic core word of the frame element [J]. Journal of Chinese Information Processing, 2010, 24(1): 30-36 (in Chinese)
(李双红，李茹，钟立军，等. 基于多词块的框架元素语义核心词自动识别研究[J]. 中文信息学报，2010，24(1)：30-36)

[19] 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 (in Chinese)
(纪文倩，李舟军，巢文涵，等. 句子相似度计算及其在自动文摘系统中的应用[J]. 智能信息管理，2009，1(1)：38-45)

[20] Li Yuhua, McLean D, Bandar Z A, et al. Sentence similarity based on semantic nets and corpus statistics [J]. IEEE Trans on Knowledge and Data Engineering, 2006, 18(8): 1138-1150

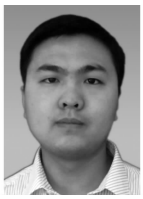
[21] Yang Sichun. An improved model for sentence similarity computing [J]. Journal of University of Electronic Science and Technology of China. 2006, 35(6): 956-959 (in Chinese)
(杨思春. 一种改进的句子相似度计算模型[J]. 电子科技大学学报，2006，35(6)：956-959)

[22] Research Center for Information Retrieval of Harbin Institute of Technology [OL] Language Technology Platform, 2011 [2011-08-20]. http://ir.hit.edu.cn/demo/ltp/
(哈尔滨工业大学信息检索研究中心[OL]. 语言技术平台，2011 [2011-08-20]. http://ir.hit.edu.cn/demo/ltp/)

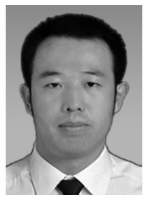
[23] Li Ru, Liu Haijing, Li Shuanghong. Chinese frame identification using T-CRF Model [C] //Proc of Int Conf on Computational Linguistics. Beijing: Tsinghua University Press, 2010: 674-682

[24] Li Jihong, Wang Ruibo, Wang Weilin, et al. Automatic labeling of semantic roles on Chinese FrameNet [J]. Journal of Software, 2010, 21(4): 597-611 (in Chinese)
(李济洪, 王瑞波, 王蔚林, 等. 汉语框架语义角色自动标注 [J]. 软件学报, 2010, 21(4): 597-611)

Li Ru, born in 1963. PhD and professor. Member of China Computer Federation. Her main research interests include Chinese information processing and informantion retrieval.



Wang Zhiqiang, born in 1987. PhD candidate. His main research interests include Chinese information processing (zhiq. wang@163. com).



Li Shuanghong, born in 1984. Master. His main research interests include Chinese information processing.

Liang Jiye, born in 1962. Professor and PhD supervisor. Senior member of China Computer Federation. His main research interests include rough set theory, data mining, artificial intelligence, etc.

Collin Baker, PhD. FrameNet project manager. His main research interests include computational linguistics.