ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.6, June 2010, pp.1267−1276
doi: 10.3724/SP.J.1001.2010.03586

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于二元分类的复述搭配抽取*

赵世奇+, 赵 琳, 刘 挺, 李 生

(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

Paraphrase Collocation Extraction Based on Binary Classification

ZHAO Shi-Qi+, ZHAO Lin, LIU Ting, LI Sheng

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: E-mail: zhaosq@ir.hit.edu.cn

Zhao SQ, Zhao L, Liu T, Li S. Paraphrase collocation extraction based on binary classification. *Journal of Software*, 2010,21(6):1267−1276. http://www.jos.org.cn/1000-9825/3586.htm

Abstract: This paper addresses the problem of paraphrase collocation extraction by using "OBJ" relationship as a case study. Specifically, the proposed method recasts paraphrase collocation extraction as a binary classification problem, which combines multiple features based on translation, thesaurus, polarity words, and web mining. Experimental results show that the binary classification-based method is effective for paraphrase collocation extraction. Especially, the exploited features are all helpful for improving the extraction performance. With the proposed method, more than 280 000 pairs of paraphrase collocations are extracted, the precision of which is above 70%. Further experiments show that nearly 40% of sentences can be paraphrased by using the extracted paraphrase collocations, which demonstrates that the proposed method is useful in practice.

Key words: paraphrase collocation; binary classification; paraphrase feature

摘 要: 以动宾关系的搭配为例研究复述搭配的抽取.具体地,该方法将复述搭配抽取视作二元分类问题,并综合使用了基于翻译、词典、极性词以及网络挖掘的多种特征.实验结果表明,所采用的二元分类方法对于抽取复述搭配是行之有效的,其中使用的各种特征对于提高复述搭配抽取的效果皆有帮助.利用该方法,共抽取出 28 万余对的复述搭配,其准确率超过 70%.进一步的实验结果表明,使用抽取的复述搭配,可以为约 40%的句子实现复述生成,从而说明了该方法的实际应用价值.

关键词: 复述搭配;二元分类;复述特征
中图法分类号: TP391 文献标识码: A

复述(paraphrase)是指对相同语义的不同表达[1].复述研究在众多自然语言处理的应用领域中都有重要的意义.例如在机器翻译研究中,对翻译短语的复述可以有效扩充翻译短语表,从而解决数据稀疏问题[2];同时,对翻译候选答案的复述可以提供更多更丰富的参考翻译句,从而促进翻译结果的自动评价[3,4].在自动问答研究中,对问句和答案抽取模板的复述可以解决问句和答案句之间的词不匹配问题,提高答案抽取的召回率[5−7].在

* Supported by the National Natural Science Foundation of China under Grant Nos.60803093, 60675034 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2008AA01Z144 (国家高技术研究发展计划(863))

Received 2008-09-04; Accepted 2009-01-15

自然语言生成研究中,复述可用于词汇选择和句子改写,以生成风格恰当且富于变化的句子[8].在自动文摘的研究中,复述既可用于识别待处理文本中的同义句[9],又有助于对系统生成文摘进行自动评价[10].除此之外,复述在信息检索和信息抽取等领域也有着重要的应用[11,12].

按照粒度划分,复述可分为复述词、复述短语、复述模板以及复述句子,具体例子可见表 1.

Table 1 Examples of paraphrases at different levels

表 1 各级复述实例

Type	Example
Paraphrase words	*aggravate* vs. *exacerbate*
Paraphrase phrases	*entire nation* vs. *whole country*
Paraphrase patterns	*X solves Y* vs. *Y is solved by X*
Paraphrase sentences	*The table was set up in the carriage shed.* vs. *The table was laid under the cart-shed.*

在以往的工作中,研究者们通常致力于利用各种语料获取复述词、复述短语以及复述模板,进而将其应用到复述句子的生成和识别中[13].然而,无论是复述词、复述短语还是复述模板,在应用中都有一定的局限性.首先,研究者们通常认为复述词即是同义词,故而从诸如 WordNet[14]这样的词典中抽取同义词作为复述词使用[3].但在实际应用中,两个同义词在特定的上下文中未必能够被视为复述词,而特定上下文中的复述词也未必是同义词[15].例如,*pay* 是 *bear* 在 WordNet 中的一个同义词,但在句子 *He has to bear the blame* 中,*bear* 不能被替换成 *pay*;相反,*shoulder* 并不是 *bear* 的同义词,但在上面给定的句子中却可以被看作是 *bear* 的复述.相比而言,复述短语由于通常含有多个词,因此歧义性较小,对上下文的依赖也不那么明显.但因为受到短语长度的限制,复述短语仅能针对句子中的局部表达进行复述,无法解决远距离搭配的复述问题.至于复述模板,由于在其抽取过程中可以引入句法信息[16,17],因此可以减轻对长度和距离的限制.然而,模板的形式和应用领域往往具有局限性,比如,很多研究规定一个模板只能含有两个槽(slots,即模板中的变量),且槽中只能填充名词[5,16].

鉴于已有方法存在的缺陷,本文研究复述搭配的获取问题.这里,搭配指的是具有某种特定句法关系的两个词[18].例如在句子 *It seems extraordinary that any thing like a general system of improvements has to* ***encounter****, in the beginning, a powerful* ***opposition*** 中,*encounter* 和 *opposition* 便构成一个具有动宾关系的搭配.为方便起见,我们将此搭配表示为⟨*encounter*,OBJ,*opposition*⟩.进而,本文将复述搭配定义为字面表达不同但意思相同的两个搭配.例如⟨*deny*,OBJ,*use*⟩和⟨*disallow*,OBJ,*use*⟩便可看作是一对复述搭配.本文以具有动宾关系的搭配为例进行研究.具体地,给定名词 n 以及与 n 构成动宾关系的任意两个动词 v_1 和 v_2,本研究的任务是识别"⟨v_1,OBJ,n⟩"和⟨v_2,OBJ,n⟩是否构成一对复述搭配.尽管本文工作仅实现了具有动宾关系的复述搭配获取,但本文所提方法可以简单地移植到其他类型的复述搭配获取当中去.

相对于复述词、复述短语和复述模板,本文所研究的复述搭配具有以下显著的优点:首先,与复述词相比,复述搭配不仅仅识别两个动词 v_1 和 v_2 是否是复述,更不是简单地利用词典获取同义词,而是考虑当 v_1 和 v_2 的宾语为 n 时,二者是否构成复述关系.在给定宾语 n 的情况下,两个互为同义词的动词也许并不是复述,而两个非同义词的动词却可能被视为复述;其次,与复述短语相比,复述搭配可以有效解决长距离搭配的复述问题.比如在前面的例子中,搭配⟨*encounter*,OBJ,*opposition*⟩之间间隔 7 个词(含标点),二者很难被包含在同一短语内,但利用本文的方法却可以很容易地将该搭配识别出来并为其抽取复述;再次,与复述模板比较,复述搭配不存在形式上的约束和应用领域的限制.因此,复述搭配可作为复述词、复述短语及复述模板的有益补充,值得深入研究.

然而截至目前,国内外对复述搭配的研究却很少.Wu 和 Zhou 率先提出同义搭配的概念(即本文定义的复述搭配),并借助翻译信息抽取同义搭配[18].该方法的基本假设是,两个同义搭配的字面表达虽然不同,但其翻译却可能相同.例如⟨*resolve*,OBJ,*problem*⟩和⟨*solve*,OBJ,*problem*⟩都可以翻译为⟨解决,OBJ,问题⟩.基于这一假设,该方法首先从一个单语语料库中统计搭配,再利用一个双语语料库为任意搭配获取中文翻译.进而,该方法将一个搭配的所有可能的翻译组成一个向量.若两个搭配的翻译向量的相似度很高,则这两个搭配将被作为同义搭配抽取出来.这种方法的弊端在于,只有当两个搭配的对应成分是 WordNet 中的同义词(如上例中的 *resolve* 和 *solve*)时才可被作为候选以计算其是否为同义搭配.也就是说,该方法限制了同义搭配的抽取范围,无法抽取出非同义

词的复述.

本文首先从一个经过句法分析的英文语料库中统计动宾搭配.进而,本文将复述搭配抽取视为二元分类问题,即给定任意搭配$\langle v_1,n\rangle$和$\langle v_2,n\rangle$(由于本文只处理动宾搭配,因此省略 OBJ,将$\langle v,\text{OBJ},n\rangle$缩写为$\langle v,n\rangle$),判定二者是否为复述(0/1).具体表示为 f:$\{(\langle v_1,n\rangle,\langle v_2,n\rangle)\}\rightarrow\{0,1\}$.本文的二元分类器使用了 4 类特征,包括翻译相似度特征、基于词典的语义相似度特征、动词极性(polarity)特征以及基于网络挖掘的上下文相似度特征.实验结果证明,本文提出的基于二元分类的方法对于复述搭配抽取问题是有效的.其中,分类器使用的 4 类特征对提高复述搭配抽取的效果均有明显帮助.利用本方法,本文共抽取出复述搭配 28 万余对,其抽样准确率达到 70.35%.在进一步的实验中,我们将抽取的复述搭配用于句子复述生成.结果表明,使用本文抽取的复述搭配,可以为约 40%的句子实现复述生成,从而证明了本方法的实际应用价值.

本文的贡献包含以下 3 点:(1) 复述搭配作为一个有价值的研究问题,前人的工作却很少涉及.本文针对该问题提出了一种有效的方法并取得了较为满意的实验结果.与 Wu 和 Zhou 的方法[18]不同,本文并不限定候选复述必须是同义词,因此抽取出的复述搭配更为灵活多样;(2) 本文率先提出将复述搭配作为二元分类问题加以解决.由于复述问题的复杂性,单一使用某一种特征往往很难解决问题.而在分类的框架下我们可以方便地融合多种特征,从而更好地解决复述搭配抽取问题;(3) 本文尝试了多种特征,既包括前人使用过的基于翻译和词典的特征,又包括本文提出的基于极性词和网络挖掘的特征.本文对多种特征加以融合、比较和分析,希望其结论对后续的研究有所裨益.

本文第 1 节介绍基于二元分类的复述搭配抽取方法,并重点介绍其中使用的分类特征.第 2 节对本文的实验设置以及实验结果进行详细描述.第 3 节为结论及对未来工作的展望.

1 基于二元分类的复述搭配抽取

1.1 搭配获取

在 Wu 和 Zhou 的方法[18]中,研究者首先利用一个单语语料库统计搭配,再利用一个双语平行语料库为统计得到的搭配获取翻译信息.本文对此稍作简化,即直接使用一个中英平行语料库的英文部分 E 统计英文搭配,再利用其中文部分 C 获取分类所需的翻译特征.本文使用的中英平行语料由 LDC 发布**.在去掉了过长(>40 词)和过短(<5 词)的句子后,该双语语料库共含 2 048 009 个中英平行句对.由于后续实验的需要,本文使用 Giza++[19]对该平行语料做了词对齐.

由于本文基于句法分析抽取搭配,因此首先对语料库的英文部分 E 进行依存句法分析.本文使用的句法分析器为 MaltParser[20].在句法分析的结果之上,本文抽取出了 E 中所有动宾搭配,在去除了出现次数为 1 的搭配之后,本文共抽取出 198 718 个不同的动宾搭配.接下来,对于具有相同宾语的任意两个搭配$\langle v_1,n\rangle$和$\langle v_2,n\rangle$,本文将其视为一对候选复述搭配.据此,本文共获得候选复述搭配 3 474 665 对.对于这些候选复述搭配,本文使用二元分类的方法从中抽取出正确的复述搭配.

1.2 特征选择

本文在实现复述搭配抽取时共使用了 4 类共 7 个特征,包括:

- 翻译相似度特征(特征 1,特征 2);
- 基于 WordNet 的语义相似度特征(特征 3,特征 4);
- 动词极性特征(特征 5);
- 基于网络挖掘的上下文相似度特征(特征 6,特征 7).

本节将对这些特征以及特征值的计算方法作详细描述.

** LDC: http://www.ldc.upenn.edu/.本文使用的 LDC 语料包括:LDC2000T46,LDC2000T47,LDC2002E18,LDC2002T01,LDC2003E07,LDC2003E14,LDC2003T17,LDC2004E12,LDC2004T07,LDC2004T08,LDC2005E83,LDC2005T06,LDC2005T10,LDC2006E24,LDC2006E34,LDC2006E85,LDC2006E92,LDC2006T04,LDC2007T02,LDC2007T09.

1.2.1 翻译相似度特征

Wu 和 Zhou 的研究已证明了使用翻译信息有助于识别复述搭配[18].其根本原因在于,两个复述搭配可能具有相似甚至相同的翻译.此外,Bannard 和 Callison-Burch 也验证了翻译信息在复述抽取中的作用[21].因此,本文考虑的第一类特征即为两个搭配的翻译相似度特征,具体包括:

特征 **1**(搭配的翻译相似度特征(**collocation translation similarity**,简称 $\boldsymbol{F_{CTS}}$)). F_{CTS} 的计算方法与 Wu 和 Zhou 的方法[18]相似.由于本文使用的双语语料已经过词对齐,因此很容易从语料中找到每个搭配对应的所有可能的中文翻译.设英文搭配 e_{col} 的中文翻译向量为 $V_{CT}(e_{col})=\langle(c_1,w_1),(c_2,w_2),\ldots,(c_n,w_n)\rangle$.其中,$c_i$ 为 e_{col} 的第 i 个翻译,w_i 为 c_i 的权重.这里的权重 w_i 基于以下公式计算:

$$w_i = tf_i \times \log\frac{N}{n_i} \tag{1}$$

其中:tf_i 表示 c_i 在 e_{col} 的翻译中出现的次数;n_i 表示中文翻译包含 c_i 的英文搭配的个数;N 为常数,定义为 $\max_i n_i$,即所有 n_i 的最大值.可见,该公式与通常使用的 tf.idf 原理相似,即当 tf_i 越大而 n_i 越小时,其权重 w_i 就越大.设 e_{col}^1 和 e_{col}^2 为两个搭配,则 $F_{CTS}(e_{col}^1,e_{col}^2)=\cos(V_{CT}(e_{col}^1),V_{CT}(e_{col}^2))$,其中,两个向量的余弦相似度定义为

$$\cos(V_1,V_2)=\frac{V_1\cdot V_2}{|V_1|\times|V_2|} \tag{2}$$

特征 **2**(动词的翻译相似度特征(**verb translation similarity**,简称 $\boldsymbol{F_{VTS}}$)). 特征 1 通过计算两个搭配的翻译相似度来度量搭配本身的相似性.由于双语语料库规模的限制,使用特征 1 可能受到数据稀疏的影响.作为对特征 1 的补充,本文引入了特征 2.由于每对候选复述搭配$\langle v_1,n\rangle$和$\langle v_2,n\rangle$中的名词 n 是相同的,因此特征 2 忽略 n,直接在双语语料库中抽取 v_1 和 v_2 的翻译向量.虽然忽略 n 的约束会影响准确性,但由于保留了动词更多的翻译,因此可以一定程度地缓解上面特征 1 的数据稀疏问题.具体地,搭配 e_{col}^1 和 e_{col}^2 的动词翻译相似度特征被定义为 $F_{VTS}(e_{col}^1,e_{col}^2)=\cos(V_{VT}(v_1),V_{VT}(v_2))$.这里,$T_{VT}(v_1)$和 $T_{VT}(v_2)$分别为 v_1 和 v_2 的翻译向量.其中,每个翻译权重的计算方法与特征 1 类似,可参照公式(1).向量余弦相似度的计算方法亦与前面相同,可参照公式(2).

1.2.2 基于 WordNet 的语义相似度特征

许多学者基于 WordNet 等类义词典计算词的语义相似度.WordNet 将词组织成树状结构,因此,最简单的一类计算语义相似度的方法便是通过计算两个词在该树状结构上的距离.简单地讲,两个词之间的距离越短,则相似度越大[22].在此基础上,有研究者从信息论的角度出发,对上述方法作了改进[23].另外,也有人使用两个词在 WordNet 中注释的相似度来度量词的语义相似度[24].本文基于 WordNet 定义了以下两个特征:

特征 **3**(基于 **Lin** 方法的语义相似度特征(**Lin-Based semantic similarity feature**,简称 $\boldsymbol{F_{LSS}}$)). 特征 3 基于 Lin 提出的语义相似度计算方法[23].在该方法中,两个词的相似度取决于二者所含的公共信息量的多少.具体地,该方法基于以下公式计算两个词的语义相似度:

$$sim_{Lin}(x_1,x_2)=\frac{2\times\log P(C_0)}{\log P(C_1)+\log P(C_2)} \tag{3}$$

这里,x_1,x_2 为待计算的两个词,C_0,C_1,C_2 均为 WordNet 中的同义词类(synsets),其中满足 $x_1\in C_1$,$x_2\in C_2$,且 C_0 为 C_1 和 C_2 在 WordNet 中的最近公共祖先.$P(C_i)$可理解为 C_i 类的词的出现概率,需要从一个大规模的语料库中统计得到(本文使用的语料库为 the British national corpus (world edition):http://www.natcorp.ox.ac.uk/).当 x_1 或 x_2 属于多个同义词类时,我们取二者的最大相似度.这里,定义 $F_{LSS}(e_{col}^1,e_{col}^2)=sim_{Lin}(v_1,v_2)$.

特征 **4**(基于 **Vector** 方法的语义相似度特征(**vector-based semantic similarity feature**,简称 $\boldsymbol{F_{VSS}}$)). 特征 4 基于 Patwardhan 提出的方法[24]计算得到.该方法通过计算两个词在 WordNet 中注释的相似度来度量两个词的语义相似度.设 $V_G(v_1)$和 $V_G(v_2)$分别是 v_1 和 v_2 在 WordNet 注释中的词组成的向量,则基于 Vector 方法的语义相似度特征定义为 $F_{VSS}(e_{col}^1,e_{col}^2)=sim_{Vec}(v_1,v_2)=\cos(V_G(v_1),V_G(v_2))$.向量余弦相似度计算方法见公式(2).在本文中,特征 3 和特征 4 使用开源工具 WordNet::Similarity(http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi)计算得到.

1.2.3 动词极性特征

极性词的概念出自意见挖掘和情感分析等研究领域.它是指那些能够表达支持/反对或者喜欢/厌恶等意见或情感倾向的词.通常,极性词包含正(即支持/喜欢)和负(即反对/厌恶)两种极性.以动词极性词为例,*enjoy*,*promote*,*win* 等词的极性为正;而 *hate*,*disallow*,*lose* 等词的极性为负.我们通过观察发现,对于候选复述搭配$\langle v_1,n\rangle$和$\langle v_2,n\rangle$,若 v_1 和 v_2 的极性不同,则二者很难互为复述(如$\langle$*worsen*,*problem*$\rangle$vs.$\langle$*solve*, *problem*$\rangle$,*worsen* 极性为负,*solve* 极性为正);而若 v_1 和 v_2 的极性相同,则二者互为复述的可能性就相对较大(如$\langle$*support*,*payment*$\rangle$vs.$\langle$*warrant*,*payment*$\rangle$,*support* 和 *warrant* 极性均为正).因此,本文在识别复述搭配时引入动词极性特征,其具体描述如下:

特征 **5**(动词极性特征(**verb polarity feature**,简称 $\boldsymbol{F_{VP}}$)). 本文在计算动词极性特征时使用了本实验室人工标注的英文极性词表.该表将词标注为 3 类,即正极性、负极性和非极性词(如非极性动词 *eat*,*use* 等).由于本文中所有待计算的动词均为及物动词(出现在动宾搭配中),因此,我们使用的是该极性词表中及物动词的极性标注信息.具体地,本部分包含 2 423 个原形动词的极性标注,其中正极性词 226 个,负极性词 443 个,无极性词 1 754 个.由于本文研究的动宾搭配中的动词不仅包含原形动词,还包含过去式、过去分词、现在分词、第三人称单数等形态变化,因此我们对上述极性标注信息作了扩充,即如果一个原形动词的极性标注为 *L*,则其所有形态变化的极性也均为 *L*.进而,我们基于以下公式计算两个搭配的动词极性特征:

$$F_{VP}(e_{col}^1,e_{col}^2)=\begin{cases}1, & \text{如果}v_1,v_2\text{均为极性词且极性相同}\\ 0.5, & \text{如果}v_1,v_2\text{均为非极性词}\\ 0, & \text{如果}v_1,v_2\text{极性不同或其中之一为非极性词}\end{cases} \tag{4}$$

1.2.4 基于网络挖掘的上下文相似度特征

有研究者基于词的分布假设实现同义词自动聚类[25].基本思想是,若两个词经常出现在相似的上下文中,则这两个词的意思也相似.本文将分布假设应用于复述搭配抽取,即认为若两个搭配的上下文相似,则这两个搭配便可能互为复述.这样,问题转化为如何获取任意搭配 e_{col} 的上下文信息.按照传统的方法,我们可以利用一个大规模的单语语料库,从中抽取包含 e_{col} 的句子作为其上下文.然而,利用通常的语料库很难做到为任意搭配都能获取足够的上下文.因此,数据稀疏问题往往比较严重.鉴于此,本文提出用网络挖掘的方法获取搭配的上下文.换言之,本文将整个互联网作为一个大规模语料库.

给定搭配 $e_{col}=\langle v,n\rangle$,网络挖掘的实现步骤是:(1) 构建查询 Q="v n"(引号""不包含在 Q 内);(2) 利用搜索引擎检索 Q;(3) 返回检索结果的前 N 个摘要(snippets);(4) 从摘要中抽取所有包含 Q 且 v 和 n 满足 OBJ 关系的句子作为 e_{col} 的上下文句子,我们将 e_{col} 的上下文句子集合记为 $ctx(e_{col})$.本文使用的搜索引擎为 Live Search (http://www.live.com/),在实验中取 N=500.在此基础上,本文提出以下两个特征:

特征 **6**(基于句法的上下文相似度特征(**syntax based context similarity feature**,简称 $\boldsymbol{F_{SCS}}$)). 与 Lin 的方法[25]类似,特征 6 只使用 $ctx(e_{col})$中与 v 和 n 有句法关系的上下文词作为 e_{col} 的上下文特征.因此,本文首先使用 MaltParser 对 $ctx(e_{col})$中的句子进行句法分析,并抽取出与 v 或 n 有句法关系的词组成其上下文特征向量 $V_{SC}(e_{col})$.这里,向量 $V_{SC}(e_{col})$的第 i 个元素为(c_i,rel_i),其中,c_i 为特征词,rel_i 为该特征词与 v 或 n 的句法关系.(c_i,rel_i)的权重 w_i 基于公式(1)计算得到.其中:tf_i 表示(c_i,rel_i)在 e_{col} 的上下文中的出现次数;n_i 表示上下文中出现了(c_i,rel_i)的搭配个数;N 为常数,定义为 $\max_i n_i$.进而,本文定义 $F_{SCS}(e_{col}^1,e_{col}^2)=\cos(V_{SC}(e_{col}^1),V_{SC}(e_{col}^2))$.

特征 **7**(基于词袋的上下文相似度特征(**bag-of-words based context similarity feature**,简称 $\boldsymbol{F_{BCS}}$)). 与特征 6 不同,特征 7 考虑搭配 e_{col} 的所有上下文词,即使用 $ctx(e_{col})$中除 v 和 n 之外的所有词构成 e_{col} 的上下文特征向量 $V_{BC}(e_{col})$.该向量中的第 i 个元素即特征词 c_i 的权重 w_i 同样基于公式(1)计算得到.其中:tf_i 表示 c_i 在 e_{col} 的上下文中的出现次数;n_i 表示上下文中出现了 c_i 的搭配个数;N 为常数,定义为 $\max_i n_i$.该特征被定义为

$$F_{BCS}(e_{col}^1,e_{col}^2)=\cos(V_{BC}(e_{col}^1),V_{BC}(e_{col}^2)).$$

1.3 PAUM分类器

目前,可用于二元分类的机器学习算法和工具很多,其中主要包括基于感知器(perceptron)、支持向量机(SVM)、最大熵(ME)等模型的分类器.各种分类器之间没有绝对的优劣之分,其性能的好坏往往取决于待处理的具体问题以及待分类数据的特点.本文问题的特殊性在于,我们通过标注发现,在待分类数据中,正例(正确的复述搭配)与负例(错误的复述搭配)的比例约为 1:13.也就是说,待分类数据中的正负例很不平衡.对于这样的数据,一般的分类器很难取得令人满意的分类结果.因此,本文选择 PAUM(perceptron algorithm with uneven margins)[26]分类器来实现复述搭配的抽取.PAUM 分类器主要用于解决数据不平衡的分类问题.由于篇幅的限制,这里只能粗略地介绍一下 PAUM 的原理.PAUM 对基本的感知器算法进行改进,对正负两类分别定义不同的边界,即对正负两类样例分别处理,从而解决两类数据不平衡的问题.本文在后续实验中对 PAUM 分类器与 SVM 分类器进行了比较,结果发现,PAUM 分类器在复述搭配抽取这一具体问题上要明显优于 SVM 分类器.

2 实验与分析

本文的实验包括 3 部分:(1) 评价本文提出的基于二元分类的复述搭配抽取方法,特别是验证分类特征与分类器的有效性;(2) 利用本文的方法实现大规模复述搭配抽取,并对抽取结果进行评价;(3) 将抽取得到的复述搭配应用于句子复述生成,以验证本方法的实际应用价值.

2.1 对基于二元分类的复述搭配抽取方法的评价

2.1.1 训练数据标注

由于二元分类的方法需要训练语料,而目前国内外并没有公开发布的相关语料,因此,本实验首先通过人工标注的方法构建一个训练集.我们从全部候选复述搭配中随机抽取出 9 140 对,将其交由两名标注者和一名仲裁者进行标注.其标注流程为:(1) 由两名标注者分别对抽取出的数据进行独立标注,每一对候选复述搭配被标注为正例(复述)或负例(非复述);(2) 计算两名标注者的标注一致性,我们通过计算得到两组标注结果的 Kappa 值为 0.934,这说明两名标注者的一致性很高;(3) 由仲裁者对两名标注者意见不同的数据进行重新标注,并将其标注作为最终标注结果.依照上述过程,我们共从 9 140 对候选复述搭配中标注得到正例 659 对,负例 8 481 对.

2.1.2 对分类特征的评价

首先利用上述标注数据对本文提出的分类特征进行评价.由于标注数据数量较少,因此采用 5-fold 交叉验证的方法进行本实验.具体地,我们将标注数据平均分为 5 份,每组实验进行 5 轮,每轮选取其中 4 份用于训练,剩余 1 份用于测试.这里采用的评价指标为准确率 P,召回率 R 以及 F-值 F.具体定义为 $P=|A\cap B|/|A|$,$R=|A\cap B|/|B|$,$F=2PR/(P+R)$.其中,A 表示分类器识别为正例的数据集合,B 表示人工标注为正例的数据集合.由于我们采用 5-fold 交叉验证的方法,因此将每组实验的 5 轮平均准确率、召回率和 F-值作为这组实验的性能指标.

为考察本文使用的 4 类特征是否对复述搭配识别都有作用,我们进行了 4 组实验,每组实验依次加入基于翻译的特征(特征 1、特征 2),基于词典的特征(特征 3、特征 4),基于极性词的特征(特征 5)以及基于网络挖掘的特征(特征 6、特征 7).其实验结果见表 2 ***.从表 2 中我们可以看到,随着加入每一类特征,分类的 F-值都有明显提高.尤其是当使用全部 4 类特征时,分类准确率、召回率和 F-值均达到最高.这说明本文所采用的 4 类特征对于提高二元分类的性能都是有帮助的.也就是说,全部 4 类特征均有助于复述搭配的识别.这里需要说明的是,本文并没有实现 Wu 和 Zhou 的方法[18]并与之进行直接比较.然而,由于 Wu 和 Zhou 的方法仅使用了搭配的翻译相似度特征,因此本文特征 1 和特征 2 的组合即可看作是对 Wu 和 Zhou 的方法的近似.从表 2 的实验结果可以看出,仅使用翻译相似度特征与使用全部特征相比,准确率、召回率和 F-值均有明显差距.由此也证明了本文使用的其余各种特征的贡献.

*** 需要说明的是,我们的目的是比较各个特征组合能够达到的最优性能,因此对于每种特征组合都通过调整分类器的参数使分类性能达到最优.其中最主要的两个参数是 PAUM 分类器的正边界 $M+$和负边界 $M-$.参数调整方法见文献[27].

Table 2 Contributions of the 4 kinds of features

表 **2** 4 类特征的贡献

Feature combination	*P* (%)	*R* (%)	*F* (%)
Features 1-2	56.97	57.51	57.12
Features 1-4	62.07	56.76	59.14
Features 1-5	69.77	57.51	62.97
Features 1-7	72.60	61.30	66.46

Table 3 Contribution of each feature

表 **3** 每个特征的贡献

Feature combination	*P* (%)	*R* (%)	*F* (%)
Removing feature 1	70.45	60.55	65.08
Removing feature 2	65.32	54.03	59.09
Removing feature 3	64.01	55.69	59.49
Removing feature 4	69.03	64.49	66.65
Removing feature 5	60.06	60.70	60.27
Removing feature 6	72.95	61.61	66.76
Removing feature 7	67.05	62.98	64.91

上面的实验验证了每一类特征的作用.接下来,我们通过实验考察是否每一个特征都是有用的.因此我们进行了 7 组实验,每组去掉 1 个特征而使用剩余的 6 个特征.实验结果见表 3.通过表 3 可以发现,分别去掉特征 1~特征 3、特征 5、特征 7 都会使分类结果的 F-值降低.而去掉特征 4 和特征 6 却不会对分类性能造成影响,其 F-值甚至略有升高.这里,我们简要地对每一个特征的作用加以分析.首先,特征 1 和特征 2 均为基于翻译相似度的特征,且从表 3 中可见,二者对分类性能的提高均有贡献,然而特征 1 的贡献明显小于特征 2.我们认为其原因在于,特征 1 计算两个搭配的翻译相似度,而特征 2 计算两个动词的翻译相似度.前者虽然更为准确,但统计过程中的数据稀疏问题会更为严重,因此影响了其效果.特征 3 和特征 4 为基于 WordNet 的语义相似度特征.其中,特征 3(基于 Lin 的方法[23])作用明显,而特征 4(基于 vector 的方法[24])却没有作用.我们分析认为,基于 Vector 的方法(即利用两个动词的词典注释的相似度来衡量两个动词的相似度)不十分合理.我们还在对标注数据的观察中发现了大量针对特征 4 的反例,即两个复述搭配的特征值很小,而两个非复述搭配的特征值却很大.特征 5 为基于动词极性的特征.表 2 和表 3 均证明了该特征的有效性.其中表 2 表明,增加特征 5 使得准确率显著提高;表 3 则表明,去掉特征 5 会使准确率明显降低.这一结果证明了特征 5 可以有效地过滤掉那些动词极性不同的候选复述搭配,从而提高复述搭配抽取的准确性.特征 6 和特征 7 为基于网络挖掘的上下文相似度特征.我们认为,之所以特征 6 无效而特征 7 有效,是因为特征 6 仅考虑了与指定搭配 e_{col} 存在句法关系的上下文词,而特征 7 则考虑了 e_{col} 的全部上下文词.我们可以认为,特征 7 包含了特征 6 的全部信息,因此单独去掉特征 6 不会对分类性能造成影响.换言之,该结果证明了某些与 e_{col} 没有句法关系的上下文词对于刻画和约束 e_{col} 的语义也是有作用的.

另外,我们还做了一组实验,即从全部 4 组特征中每组抽出一个相对更有效的特征并加以组合.具体地,从第 1 组特征(特征 1、特征 2)中抽出特征 2,从第 2 组特征(特征 3、特征 4)中抽出特征 3,从第 3 组特征(特征 5)中抽取特征 5,从第 4 组特征(特征 6、特征 7)中抽出特征 7.即只使用特征 2、特征 3、特征 5、特征 7.利用上述 4 个特征,按照上面的方法进行了实验.结果表明,使用上述 4 个特征所能得到的准确率、召回率和 F-值分别为 67.21%,62.52%和 64.74%.与表 2 对比可知,只使用 4 个优选特征的结果不如使用全部 7 个特征好.

2.1.3 PAUM 分类器与 SVM 分类器的比较

为了证明 PAUM 分类器在复述搭配抽取中的有效性,我们将其与 SVM 分类器进行了对比.本实验使用的 SVM 分类器为 libsvm-2.82.我们利用本文提出的 4 类特征基于 5-fold 交叉验证的方法在标注数据集上对 SVM 分类器进行了实验(这里使用了 libsvm-2.82 默认的 RBF 核函数,并通过调整参数使 SVM 的分类效果(F-值)达到最优).结果表明,SVM 分类器的准确率、召回率和 F-值分别为 83.09%,43.09%和 56.72%.以 PAUM 分类器的 F-值(见表 2 末行)为基准,SVM 分类器的 F-值低了 14.66%.该结果验证了本文第 1.3 节的结论,即 PAUM 分类器更适合解决数据不平衡的二元分类问题,从而在复述搭配抽取的问题上取得了更好的效果.

2.2 对复述搭配抽取结果的评价

在本节实验中,我们将 9 140 对人工标注数据用于训练分类器,进而基于二元分类的方法从全部 3 474 665 对候选复述搭配中抽取正确的复述搭配.这里,分类器使用的特征为上述全部 4 类特征,使用的参数为第 2.1 节中利用 5-fold 交叉验证的方法选定的最优参数.最终共抽取出 283 670 对复述搭配.

2.2.1 抽样标注与评价

我们从抽取出的复述搭配中随机抽样 2 000 对进行人工标注并计算准确率.结果表明,在 2 000 对抽样数据

中,有 1 407 对为正确的复述搭配,其准确率为 70.35%.表征两个标注者一致性的 Kappa 值为 0.819.表 4 列举了抽取出来的一些正确的复述搭配.

Table 4 Examples of the extracted paraphrase collocations

表 4 抽取出的复述搭配示例

⟨*confuse*,*activities*⟩ ⟨*dampen*,*activities*⟩	⟨*meeting*,*goal*⟩ ⟨*realizing*,*goal*⟩
⟨*assessing*,*applications*⟩ ⟨*examining*,*applications*⟩	⟨*favours*,*idea*⟩ ⟨*supports*,*idea*⟩
⟨*demands*,*approach*⟩ ⟨*entails*,*approach*⟩	⟨*combating*,*impact*⟩ ⟨*eliminating*,*impact*⟩
⟨*bear*,*blame*⟩ ⟨*shoulder*,*blame*⟩	⟨*amending*,*measures*⟩ ⟨*enhancing*,*measures*⟩
⟨*reasserts*,*commitment*⟩ ⟨*reiterates*,*commitment*⟩	⟨*repatriating*,*migrants*⟩ ⟨*returning*,*migrants*⟩
⟨*address*,*conflict*⟩ ⟨*settle*,*conflict*⟩	⟨*coordinate*,*parties*⟩ ⟨*reconcile*,*parties*⟩
⟨*drafting*,*policy*⟩ ⟨*drawing*,*policy*⟩	⟨*advocating*,*cooperation*⟩ ⟨*stimulating*,*cooperation*⟩
⟨*attack*,*crime*⟩ ⟨*suppress*,*crime*⟩	⟨*producing*,*proposals*⟩ ⟨*tabling*,*proposals*⟩
⟨*encouraged*,*development*⟩ ⟨*spurred*,*development*⟩	⟨*avoid*,*responsibility*⟩ ⟨*shed*,*responsibility*⟩
⟨*renew*,*efforts*⟩ ⟨*revitalize*,*efforts*⟩	⟨*better*,*standards*⟩ ⟨*modernize*,*standards*⟩

如上所述,本文抽取的复述搭配⟨v_1,n⟩和⟨v_2,n⟩并不限定 v_1 和 v_2 为同义词.因此,我们有必要考察一下在抽取出来的复述搭配中有多少为非同义词的复述.这里,我们参照的对象包括 WordNet 中定义的同义词集(记为 *Syn*-1)以及 Lin 通过自动聚类的方法构建的同义词集(记为 *Syn*-2)[25](该同义词集可从以下网址下载:http://www.cs.ualberta.ca/~lindek/downloads.htm).经过统计发现,在上面抽取的 1 407 对正确复述搭配中,动词 v_1 和 v_2 不属于 *Syn*-1 的有 1 117 对(79.39%)(若 v_1 和 v_2 为经过词形变化的动词,则我们会使用其原形判断二者是否为 WordNet 中定义的同义词),v_1 和 v_2 不属于 *Syn*-2 的有 667 对(47.41%),v_1 和 v_2 既不属于 *Syn*-1 又不属于 *Syn*-2 的有 538 对(38.24%).以上结果说明,利用本方法抽取的复述搭配中,有相当一部分为非同义词的复述,无法利用同义词词典获得.

2.2.2 错误分析

对于前面 2 000 个样例中分类错误的 593 个样例,逐一进行错误分析后发现,最主要的一类错误源自基于 WordNet 的语义相似度特征.在很多搭配对中,两个动词 v_1 和 v_2 基于 WordNet 的语义相似度很大,二者甚至为同义词,但在给定的搭配中意思却不同.例如在搭配⟨*release*,*guidelines*⟩和⟨*turn*,*guidelines*⟩中,*release* 和 *turn* 是 WordNet 中定义的同义词,其语义相似度为 1,但这两个搭配却并不是复述搭配.对于这种情况,分类器很容易将其错分为正例.另外一类主要错误源自基于网络挖掘的上下文相似度特征.这一类错误容易理解,因为两个出现在相似上下文中的搭配可能意思并不相同,甚至有可能是相反的.例如搭配⟨*eliminate*,*weapons*⟩和⟨*possess*,*weapons*⟩,二者的意思显然不同,但其上下文相似度却很大,其频数最高的前 3 个上下文词均为 *nuclear*,*destruction* 和 *mass*.此外,也有部分错误与基于翻译相似度和基于动词极性的特征相关.

2.3 复述搭配在句子复述生成中的应用

在本部分实验中,我们将抽取到的复述搭配应用于句子复述生成,以此来验证本方法在实际应用中的有效性.该任务可定义为:给定句子 *S*,倘若 *S* 中的搭配 e_{col}^1 存在复述搭配 e_{col}^2,则将 *S* 中的 e_{col}^1 替换为 e_{col}^2,从而生成 *S* 的复述句 *T*.事实上,该任务的实现过程可能比较复杂,例如当 e_{col}^1 存在多个复述搭配时,我们可以基于语言模型计算使用不同的复述搭配生成的 *T* 的分值,从而选取其中最优的复述搭配进行替换.然而,本文旨在评价抽取得到的复述搭配的质量,因此不考虑其他诸多复杂因素,而是简单地将每个句子中所有匹配的复述搭配保留下来.

本实验使用的测试集包含 119 个随机选取的英文句子,每个句子平均含有 28.5 个词(含标点).经过句法分析后发现,在 119 个测试句中有 99 个至少含有 1 个动宾搭配,测试集含有的动宾搭配的总数为 175 个.通过上述过程,我们共为 47 个句子(占全部测试句的比例为 39.50%)中的 64 个搭配(占全部动宾搭配的 36.57%)获取了共 508 个复述搭配.接下来,标注者对这些复述搭配进行了人工标注.标注者在标注过程中考虑了上下文的约束,即两个搭配只有在给定的句子中可以相互替换才视其为正确的复述搭配.标注结果表明其准确率为 63.58%.表征两个标注者一致性的 Kappa 值为 0.783.显然,由于考虑了上下文信息的约束,因此复述替换的准确率低于第 2.2.1 节中介绍的 70.35%.图 1 显示了在给定的上下文句中进行复述搭配替换的例子.

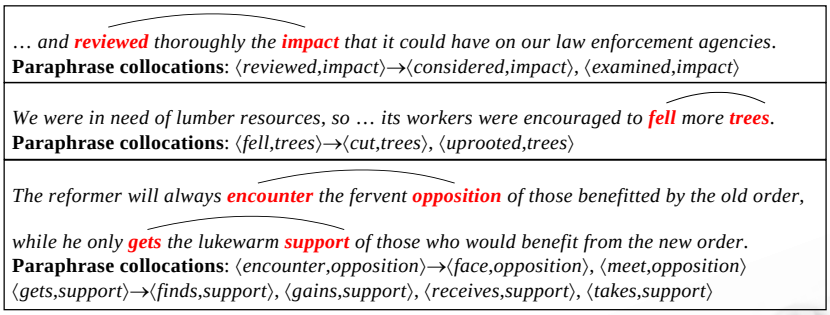

Fig.1 Examples of paraphrase collocations replacement in given sentences

图 1 给定句子中的复述搭配替换示例

3 结论与展望

本文提出了一种基于二元分类的复述搭配抽取方法.具体地,本文利用 PAUM 分类器解决数据不平衡问题,并综合使用了 4 类特征,即:(1) 翻译相似度特征;(2) 基于 WordNet 的语义相似度特征;(3) 动词极性特征;(4) 基于网络挖掘的上下文相似度特征.实验结果表明,本文使用的分类器和 4 类特征对于抽取复述搭配都是有效的.利用本方法,我们共抽取出复述搭配 28 万余对,其抽样准确率超过 70%.另外,利用本文抽取的复述搭配,我们可以为约 40%的句子实现复述生成,从而证明了本方法在实际应用中的意义.

致谢 在此,我们向对本研究工作提供帮助的老师和同学表示感谢.特别要感谢刘树伟、蓝翔等同学在实验数据标注上的工作,还要感谢车万翔老师、赵妍妍及郭宇航同学对本文初稿的审阅以及提出的宝贵意见.

References:

[1] Barzilay R, McKeown KR. Extracting paraphrases from a parallel corpus. In: Proc. of the ACL/EACL. 2001. 50–57. http://aclweb.org/anthology-new/P/P01/P01-1008.pdf

[2] Callison-Burch C, Koehn P, Osborne M. Improved statistical machine translation using paraphrases. In: Proc. of the HLT-NAACL. 2006. 17–24. http://aclweb.org/anthology-new/N/N06/N06-1003.pdf

[3] Kauchak D, Baizilay R. Paraphrasing for automatic evaluation. In: Proc. of the HLT-NAACL. 2006. 455–462. http://aclweb.org/anthology-new/N/N06/N06-1058.pdf

[4] Zhou L, Lin CY, Hovy E. Re-Evaluating machine translation results with paraphrase support. In: Proc. of the EMNLP. 2006. 77–84. http://www.isi.edu/natural-language/people/hovy/papers/06EMNLP-MTeval-paraphrases.pdf

[5] Ravichandran D, Hovy E. Learning surface text patterns for a question answering system. In: Proc. of the ACL. 2002. 41–47. http://aclweb.org/anthology-new/P/P02/P02-1006.pdf

[6] Rinaldi F, Dowdall J, Molla D. Exploiting paraphrases in a question answering system. In: Proc. of the IWP. 2003. 25–32. http://acl.ldc.upenn.edu/W/W03/W03-1604.pdf

[7] Zhao SQ, Zhou M, Liu T. Learning question paraphrases for QA from encarta logs. In: Proc. of the IJCAI. 2007. 1795–1800. http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-290.pdf

[8] Stede M. Lexical paraphrases in multilingual sentence generation. Machine Translation, 1996,11:75–107. http://www.springerlink.com/content/lp75vw78277v2789/

[9] Mckeown KR, Barzilay R, Evans D, Hatzivassiloglou V, Klavans JL, Nenkova A, Sable C, Schiffman B, Sigelman S. Tracking and summarizing news on a daily basis with Columbia'S Newsblaster. In: Proc. of the HLT. 2002. 280–285. http://www1.cs.columbia.edu/~sable/research/hlt-blaster.pdf

[10] Zhou L, Lin CY, Munteanu DS, Hovy E. ParaEval: Using paraphrases to evaluate summaries automatically. In: Proc. of the HLT-NAACL. 2006. 447–454. http://aclweb.org/anthology-new/N/N06/N06-1057.pdf

[11] Zukerman I, Raskutti B. Lexical query paraphrasing for document retrieval. In: Proc. of the COLING. 2002. 1–7. http://aclweb.org/anthology-new/C/C02/C02-1161.pdf

[12] Shinyama Y, Sekine S, Sudo K. Automatic paraphrase acquisition from news articles. In: Proc. of the HLT. 2002. 40–46. http://nlp.cs.nyu.edu/pubs/papers/shinyama-hlt02.pdf

[13] Zhao SQ, Niu C, Zhou M, Liu T, Li S. Combining multiple resources to improve SMT-based paraphrasing model. In: Proc. of the ACL 2008: HLT. 2008. 1021−1029. http://aclweb.org/anthology-new/P/P08/P08-1116.pdf

[14] Fellbaum C. WordNet: An Electronic Lexical Database. Cambridge: MIT Press, 1998.

[15] Zhao SQ, Liu T, Yuan XC, Li S, Zhang Y. Automatic acquisition of context-specific lexical paraphrases. In: Proc. of the IJCAI. 2007. 1789−1794. http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-289.pdf

[16] Lin DK, Pantel P. Discovery of inference rules for question answering. Natural Language Engineering, 2001,7(4):343−360. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.2538&rep=rep1&type=pdf

[17] Zhao SQ, Wang HF, Liu T, Li S. Pivot approach for extracting paraphrase patterns from bilingual corpora. In: Proc. of the ACL 2008: HLT. 2008. 780−788. http://aclweb.org/anthology-new/P/P08/P08-1089.pdf

[18] Wu H, Zhou M. Synonymous collocation extraction using translation information. In: Proc. of the ACL. 2003. 120−127. http://aclweb.org/anthology-new/P/P03/P03-1016.pdf

[19] Och FJ, Ney H. Improved statistical alignment models. In: Proc. of the ACL. 2000. 440−447.

[20] Nivre J, Hall J, Nilsson J, Chanev A, Eryigit G, Kubler S, Marinov S, Marsi E. MaltParser: A language-independent system for data-driven dependency parsing. Natural Language Engineering, 2007,13(2):95−135. http://citeseerx.ist.psu.edu/viewdoc/download?[doi=10.1.1.105.7483&rep=rep1&type=pdf]

[21] Bannard C, Callison-Burch C. Paraphrasing with bilingual parallel corpora. In: Proc. of the ACL. 2005. 597−604. http://aclweb.org/anthology-new/P/P05/P05-1074.pdf

[22] Wu Z, Palmer M. Verb semantics and lexical selection. In: Proc. of the ACL. 1994. 133−138. http://aclweb.org/anthology-new/P/P94/P94-1019.pdf

[23] Lin DK. An information-theoretic definition of similarity. In: Proc. of the ICML. 1998. 296−304. http://webdocs.cs.ualberta.ca/~lindek/papers/sim.pdf

[24] Patwardhan S. Incorporating dictionary and corpus information into a vector measure of semantic relatedness [MS. Thesis]. Duluth: University of Minnesota, 2003.

[25] Lin DK. Automatic retrieval and clustering of similar words. In: Proc. of the COLING/ACL. 1998. 768−774. http://aclweb.org/anthology-new/P/P98/P98-2127.pdf

[26] Li Y, Zaragoza H, Herbrich R, Shawe-Taylor J, Kandola J. The perceptron algorithm with uneven margins. In: Proc. of the ICML. 2002. 379−386. research.microsoft.com/pubs/66862/129.ps

[27] Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, 1992. 412−420.

赵世奇(1981－),男,辽宁抚顺人,博士生,CCF 学生会员,主要研究领域为复述,自然语言处理.

赵琳(1983－),男,硕士生,CCF 学生会员,主要研究领域为复述,自然语言处理.

刘挺(1972－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,信息检索.

李生(1943－),男,博士,教授,博士生导师,CCF 会员,主要研究领域为自然语言处理,信息检索,机器翻译.