文章编号：1000-8608(2012)05-0761-06

基于词汇范畴和语义相似的显性情感隐喻识别机制

林鸿飞*，许　侃，任　惠

（大连理工大学 计算机科学与技术学院，辽宁 大连　116024）

摘要： 隐喻是自然语言中比较常见的语言现象，在情感计算过程中有些句子的情感是由隐喻引起的，因此隐喻问题的解决将影响情感计算的结果.为此结合相关的隐喻理论，从机器学习的角度，对汉语文本中的显性隐喻的识别进行研究.以本体和喻体所属的范畴不同作为切入点，首先利用词性标注和依存句法分析提取句子中的本体词汇和喻体词汇，然后进行范畴划分和词汇语义相似度计算，最后使用支持向量机进行学习，从而对特定的隐喻句子进行识别.这一研究对后续的隐喻自动识别和隐喻理解起到了一定的作用.

关键词： 隐喻识别；范畴划分；语义相似度

中图分类号： H043；TP391.43　　**文献标志码：** A

0　引　言

分析一篇文章的情感色彩，需要对词进行着重分析，因为文章是由句子组成的，而句子是由词组成的.在文本情感计算时，常常可以发现并不是每个句子当中都含有情感词，这时计算机很难准确地给出这个句子的情感.然而情感是丰富而抽象的，往往通过隐喻给以形象的描述.因此，对这些隐喻进行识别，并且将其解释成普通的语言表达方式，有助于文本的情感计算.

例如："这律师是狐狸。"

计算机很难从字面上直接给出这句话的真实情感.但是从隐喻的角度考虑，本体是"律师"，喻体是"狐狸"，话外之音是省略掉的喻底"狡猾"、"多疑"，这句话表达的真正意思是"律师像狐狸一样狡猾."因此，隐喻的识别和理解，对于情感计算有很大的帮助.

从20世纪60年代起，西方对隐喻的研究掀起了热潮，隐喻研究者把隐喻从修辞学纳入认知语言学的范畴，人们认识到隐喻既是一种语言现象，也是一种思维方式[1].而汉语的隐喻研究也逐渐被放在一个更加重要的位置，国内语言学界对隐喻的研究给予了极大的热情，主要集中于有关汉语隐喻修辞的语言学和心理学范畴的讨论.而隐喻计算化方面的研究仍处于探索研究阶段.国内的相关学者进行了一些隐喻计算模型的探讨和试验，俞士汶从自然语言理解角度提出了多个层面的隐喻理解问题[1]，张威等在隐喻逻辑推理方面进行了研究[2-4].

王治敏[5]使用机器学习的方法识别名词短语隐喻，采用了基于实例、朴素贝叶斯、最大熵3种分类方法进行比较，得出最大熵的分类效果最好.

贾玉祥等基于现有的词典资源，对名词性隐喻进行识别[6]，并利用隐喻知识库进行了隐喻理解方面的研究[7].

本文主要针对隐喻识别进行初步探讨，以便将其应用到情感计算中.王治敏主要处理的是名词短语隐喻，而本文考虑的隐喻中的本体和喻体不限于名词短语；贾玉祥利用词典资源识别名词性隐喻，对于未登录词的处理会有限制，而本文通过依存句法分析可以弥补这一缺陷；利用隐喻知识库处理会需要大量的知识库建立工作，而本文的方法不需要用到隐喻知识库.

1　隐喻的识别

隐喻的理解过程分为两个步骤：隐喻的辨认和隐喻意义的推断[8]，分别称之为隐喻的识别和

收稿日期： 2010-12-26；　**修回日期：** 2012-06-23.

基金项目： 国家自然科学基金资助项目(60673039,60973068)；"八六三"国家高技术研究发展计划资助项目(2006AA01Z151)；高等学校博士学科点专项科研基金资助项目(20090041110002)；教育部留学回国人员科研启动基金资助项目(教外司留[2007]1108号)；大连理工大学青年教师启动基金资助项目.

作者简介： 林鸿飞*(1962-)，男，博士，教授，博士生导师，E-mail：hflin@dlut.edu.cn.

隐喻的理解,其中隐喻的理解是建立在正确的隐喻识别基础之上的.因此,研究隐喻的识别对于隐喻的机器理解而言,是一项很有意义的基础性工作,对于基于文本的情感计算具有十分重要的意义[9,10].

1.1 隐喻的特点和分类

本体和喻体之间存在明显差异是隐喻的一个重要特征,完全相同的事物是不能构成隐喻的.本体和喻体不仅要有差异,也要相似,相似性也是隐喻的基本要素.而隐喻的理解就是找到本体到喻体的映射.由此可以得出影响汉语隐喻理解的四大要素:本体、喻体、相异点和相似点[2].从隐喻的表现形式、功能和效果、认知特点等角度,可以把隐喻分为显性隐喻(明喻)与隐性隐喻、根隐喻与派生隐喻、以相似性为基础的隐喻和创造相似性的隐喻等[2].本文中只讨论对显性隐喻的识别.

1.2 显性隐喻的识别

对于隐喻的识别来说,由于汉语句子本身的复杂性,不宜展开来通篇识别,应该从一个小范围入手,本文选择显性隐喻这个类别作为切入点.显性隐喻的特征是隐喻的四要素都存在于句子之中,明确说明两者是一种对比关系,汉语中的典型形式是"A像B"或"A像B一样C".如:

例1 "建筑像凝固的音乐。"

但是,并不是所有含有"像"字的句子都是隐喻句.如:

例2 "未经书面授权禁止复制或建立镜像。"

例3 "一九四二年创作的浮雕《毛泽东像》是毛泽东的侧面头像。"

这两个句子中的"像"都是名词,句子并不是隐喻句.又如:

例4 "就像邓小平同志指出的:统一战线仍然是一个重要法宝。"

例5 "像标志牌写的那样,要有志气把三峡工程建好。"

例4、5两个句子中含有动词"像",但是句子本身也不是隐喻句.

这些隐喻句是隐喻当中结构比较简单的一类,但是即便是这样,对它们的识别也不是那么容易的.在实验中,从隐喻本身的特点出发,去考虑解决显性隐喻的识别问题.从逻辑角度来看,隐喻中涉及的两个主词属于两个不同的类型,因而将它们用系词(通常是be)联结起来实际上构成了一种逻辑错误,或称"范畴置错".在显性隐喻的识别过程中就是使用隐喻中本体词汇和喻体词汇的范畴矛盾性来解决识别问题.如:

例6 "她的手像冰一样。"

例6中"手"是本体,"冰"是喻体.实际生活中,冰的温度明显低于手的温度,"手"和"冰"并不属于同一个词汇范畴,因而判定该句为隐喻句.由人对"冰"的固定认知可以推断,这个句子的真正意义是:"她的手像冰一样冷".

2 利用词性标注提取特征词

词性标注是特征词提取的基础.在词性标注阶段,利用自动分词程序对文本中的句子进行分词,然后逐词进行查找和分析.过程分为两个步骤:提取标志词汇,然后提取标志词汇的相关词汇.

在特征词的提取过程中,每个句子的结构被描述成一个三元组 $O:=\{C,T,S\}$.

其中 C 中的元素被称为标志词(character word),元素的选择范围为"像/v"和"像/n".即在分词完成的句子中,"像/v"和"像/n"需要作为标志词来提取.含有名词"像"的句子,不属于显性隐喻的识别范畴.含有动词"像"的句子将会被进一步处理.由于词性标注工具的不同,可能会产生不同的词性分类体系,本文主要考虑"像/v"和"像/n"这两类.

T、S 中的元素分别被称为本体词(target word)和喻体词(source word).以 C 为中心,按照临近名词规则来确定句子中的本体词和喻体词.在处理的过程中,根据句子的复杂程度,需要把整个句子按照",,"进行断句,如果在 C 的当前分句中无法找到 T 或者 S,那么就需要扩大查找范围,在其他分句中确定 T 或 S.在本系统中,会指定 C 的下一小分句的第一个名词为要查找的本体词.如:

例7 "她/r 的/u 面庞/n 笑/v 得/u 竟/d 像/v 一/m 朵/q 盛开/v 的/u 鲜花/n"

可以得到"像/v"以及本体"面庞"和喻体"鲜花".又如:

例8 "像/v 无数/m 燃烧/v 的/u 蜡烛/n 一样/a,/wp 所有/b 老师/n 都/d 情愿/v 燃烧/v 自己/r,/wp 照亮/v 别人/r 。/wp"

按照临近名词规则,只能找到喻体"蜡烛",而对于本体"老师"临近规则失效,这时,扩大词的查找范围,指定"像/v"的下一小分句的第一个名词为要查找的本体词.从而,可以找到标志词"像/v"以及本体"老师"和喻体"蜡烛".

3 利用依存句法分析提取特征词

在利用词性标注提取本体和喻体特征词的同时,需要利用依存句法分析来加强对喻体特征词

的进一步提取. 在本文中，利用的是哈尔滨工业大学的依存句法分析系统[11]，其分析结果为特征词的提取提供了有力的支持.

系统通过依存句法分析，把句子从一个线性序列重新组织成一个结构化的依存分析树，而句子中词之间的依存关系是通过依存弧来反映的. 依存句法分析的优点是深入语言的内部结构而不再局限于表层的匹配. 为了加强词性标注提取的准确性，这里要利用依存句法分析进行本体和喻体的进一步确定，本文所用系统的无标记依存弧句法分析准确率为 80%左右，对提高词汇提取的正确性起到了一个很好的作用.

具体说来，“像”字在句子中一般充当谓语和介词，所以后面跟的一般都是动宾词和介宾词. 在依存句法分析系统中，动宾关系和介宾关系的分析标识分别为“VOB”和“POB”，而这些依存于“像”字的词汇往往就是要查找的喻体词汇. 如：

例 9　“优秀/a 大学生/n 就/d 像/v 社会主义/n 精神文明/n 建设/v 的/u 种子/n。/wp”

依存句法分析结果如图 1 所示.

本体为“大学生”，喻体为“种子”. 如果利用词性标注来判断喻体为“社会主义”，显然是错误的，但是根据依存句法分析，通过“像_种子/VOB”这一项，可以得出“种子”是“像”的动宾词，因此可以得到喻体“种子”.

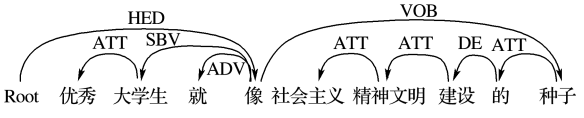

图 1　依存句法分析示例之一

Fig. 1　Example 1 of interdependence parsing analysis

4　词汇范畴判别及语义相似度计算

提取到本体和喻体特征词以后，需要对两个特征词的范畴进行判别并且计算词的相似度. 而知网中含有丰富的词汇语义知识，因此本文词汇范畴判别和相似度的计算都是基于知网(HowNet Knowledge Database)进行的.

4.1　知网的知识词典

知网是一个网状结构的知识系统，主要反映概念以及概念的属性之间的各种关系. 知网的基础文件是知识词典，在知识词典中，每一个词的概念及其描述对应一条记录，而每一条记录都主要包含词语、词语词性、词语例子以及概念定义这 4 项内容[12].

看一个具体的例子：

NO. =029814
W_C=甘
G_C=ADJ
E_C=
W_E=pleasantly
G_E=ADV
E_E=
DEF=aValue|属性值，taste|味道，sweet|甜，desired|良

在上例中，NO. 是概念编号，W_C 是汉语词语，G_C 是汉语词性，E_C 是汉语例子，W_E 是英语词语，G_E 是英语词性，E_E 是英语例子，DEF 是概念定义[12]. 其中 DEF 是最重要的. 可以认为知网中的概念定义是这个词所属的一个类别. 而这个类别，对处理隐喻，用来表示本体和喻体的范畴是很重要的.

4.2　词汇范畴的判别

根据知网知识词典的语言描述，可以利用其中的概念定义来判别一个词所属的类别. 如果提取的本体和喻体特征词属于同一个类别或者相近的类别，就可以判断这个句子不是隐喻；反之，如果这两个特征词汇不属于同一个类别，那么就构成隐喻.

例 10　“这些地区都应像京郊灾区一样，振奋精神，精打细收，力争抗灾夺丰收。”

提取特征词为“地区”、“灾区”，这两个词汇在知网中的概念定义均为“地区”，因此可以判断它们属于同一个概念范畴，那么就不构成隐喻.

例 11　“远处的她像一束飘动的红绸子。”

提取特征词为“她”、“绸子”，这两个词汇在知网中的概念定义分别为“他”、“材料”，因此可以判断它们不属于同一个概念范畴，那么就构成隐喻.

对于词汇相近范畴的识别，需要支持向量机的学习来进行.

4.3　语义相似度计算

如果两个词语相互替换而不改变替换文本的句法语义结构，那么这两个词是相似的. 而它们的相似度在于改变替换文本句法语义结构的程度，改变得越小，相似度越高，改变得越大，相似度越低.

计算两个词语的语义相似度，对于隐喻的识别有重要的意义，本体和喻体的相似度可以作为隐喻识别的一个重要特征. 根据知网提供的词汇义原以及知网本身的网状结构，可以计算出词之间的距离[13].

首先，应该计算两个词之间相似度的大小. 对于两个汉语词语 W_1 和 W_2，如果 W_1 有 n 个义项

(概念):S_{11},S_{12},…,S_{1n},W_2 有 m 个义项(概念):S_{21},S_{22},…,S_{2m},W_1 和 W_2 的语义相似度即是各个概念的相似度之最大值[12]. 即

$$sim(W_1,W_2)=\max_{i=1,\cdots,n,j=1,\cdots,m} sim(S_{1i},S_{2j})$$

这样可以根据两个概念之间的相似度来计算两个词之间的相似度.

然后,根据计算得到的相似度大小,设定一个阈值来界定两个词的范畴,从而判断它们是不是属于同一个类别. 这样,就可以把概念之间的相似度问题归结成两个词之间的范畴问题,最终达到是否属于一个类别判断的目的.

在系统中,相似度的取值范围规定在 0～1. 针对实验语料,在 0 ～ 1 定义一个阈值 d,相似度大于 d 的词被视为同一范畴,小于等于 d 的则被视为不同范畴,即构成隐喻. 阈值 d 的计算公式为

$$d=\frac{\sum_{k=1}^{m} S_k^{y}}{\left(\sum_{i=1}^{m} S_i^{y}/m+\sum_{j=1}^{n} S_j^{n}/n\right)}$$

式中:S_i^{y} 为训练语料中正例相似度的值;m 为正例个数;S_j^{n} 为训练语料中负例相似度的值;n 为负例个数. 通过计算得到划分相似度的阈值为 0.1. 阈值 d 计算公式的设计主要考虑训练语料中正负例的非均衡问题,由于语料有限,语料中存在明显的正负例非均衡问题,公式主要通过求平均数的思想来解决这一问题.

例 12 “头发白得像腊月里的雪。”

通过计算得到,本体词“头发”和喻体词“雪”的相似度为 0.1. 根据阈值判断,它们的相似度比较低,本体和喻体不属于同一个范畴,因此构成隐喻.

5 基于支持向量机的隐喻识别

支持向量机根据特征空间的数据是否线性可分,分为线性支持向量机和非线性支持向量机. 非线性支持向量机的原理是把输入向量从低维特征空间映射到高维特征空间,从而把非线性问题转换成线性问题. 而这种变换是通过核函数 $k(\boldsymbol{x},x_i)$来实现的[14].

本系统采用SVMlight 4.0版本软件包[15,16]来实现分类器的训练和测试.

使用基于特征向量的机器学习方法进行隐喻识别,最重要的过程就是特征向量的构造. 选取恰当的特征能够较好地对实体进行表述,有利于学习效果的提高.

隐喻的特征选取,需要针对其结构特点,依据词汇范畴的原则来进行. 通过对训练语料进行分析,把“像/v”,“像/n”,本体、喻体词在知网中的概念定义,本体、喻体词的相似度 4 个方面作为隐喻识别的特征.

“像/v”表示“像”在句子中以谓词的成分出现,它是判断一个句子有可能是要处理的隐喻的一个基本特征;“像/n”是判断一个句子为非隐喻句的一个重要而直接的特征,是区分隐喻和非隐喻的标志. 根据这个特征可以把语料中非隐喻的句子直接排除出去.

本体、喻体词在知网中的概念定义是词汇范畴详细划分的特征,根据这个特征可以给出词在概念层次的范畴分类. 运用分词和依存句法分析对句子进行主成分分析,以“像”为中心,根据本体、喻体特征词进行提取. 根据范畴分类,可以有效地判断特征类别范畴相同的句子为非隐喻句.

本体、喻体词的相似度为本体、喻体词范畴二元划分的特征. 除了词的概念层次范畴分类以外,对于类别范畴相似或相近的词汇,需要采取相似度的计算来判断本体、喻体词的范畴. 根据此特征可以进一步地对词的类别进行划分,并且可以通过调整相似度的阈值来进行范畴的界定.

在实验过程中,把以上 4 个方面作为隐喻识别的特征加入特征向量中,进行机器学习的训练和预测.

6 实验与结果分析

6.1 实验说明

本文中利用的语料来自于国家语委语料库(http://219.238.40.213:8080/). 在网站中,通过语料库的检索系统搜索出含有“像”字的句子 4 590 句,从中抽取前 1 000 句作为本次实验的语料,然后手工整理到数据库中. 因为标注的语料数量有限,很难覆盖所有的词,而且汉语的表达方式灵活多样,因此在训练语料和测试语料划分时,其中 938 句作为训练语料,62 句作为测试语料. 实验的流程如图 2 所示.

在实验过程中,首先把语料分成训练样本和测试样本. 然后,在数据转换模块对语料进行分词、句法分析,并提取其中的本体和喻体信息,形成特征数据,生成 SVMlight 所需要的训练数据和测试数据. 其次,利用 SVMlight 对得到的数据进行训练和分类,最后对 SVMlight 得到的结果进行分析整理,得到最终的数据.

6.2 实验结果及分析

根据以上流程,实现了一个基于知网的隐喻计算程序模块. 共设计了两个对比实验.

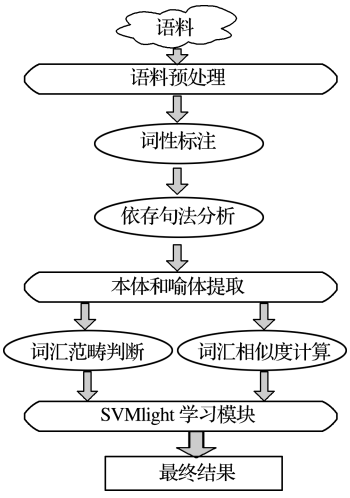

图 2　显性隐喻识别的实验流程

Fig. 2　Experiment process of dominant metaphor indentification

实验 1　仅仅利用词性标注来提取本体和喻体词，并通过知网判断词的范畴.

实验 2　利用词性标注和依存句法分析提取本体和喻体词，通过知网以及相似度计算来判断词的范畴.

实验结果如表 1 所示.

表 1　实验结果

Tab. 1　Experiment results

实验	准确率/%	召回率/%	F 值
1	54.3	100.0	0.700
2	59.2	93.5	0.725

实验 2 同实验 1 相对比，加入了依存句法分析和相似度计算，结果有了一定的提高. 但是两个实验共同的问题是：准确率相对不高，而召回率却比较高，说明实验中的噪音比较大.

对实验流程进行分析，发现特征词提取不准确是造成噪音比较大的一个主要原因.

在本体和喻体特征词提取过程中，由于语料中句子比较复杂，词的提取规则并不能适用于所有的句子. 如：

例 13　“城市梦对于地处偏僻山区龙华的农民来说，不再像天上的月亮可望不可即。”

人工对本句进行词信息提取，可以得到本体“城市梦”和喻体“月亮”，但是如果通过机器提取，句子成分的复杂程度已经超出了规则范围. 因此，语料的复杂程度是影响特征词提取准确与否的一个重要因素.

另一方面，依存句法分析在特征词提取过程中确实起到了一定的作用，但对于语料中某些句子，并不能用依存句法分析来提取本体词，而用来提取喻体词的准确率也不是很高. 如：

例 14　“我们决不能像无头苍蝇一样乱撞乱飞。”

依存句法分析的结果如图 3 所示.

通过例句可以看出，并不能得到本体“我们”和喻体“苍蝇”，因此这也是造成特征词提取不准确的一个原因.

另外，由于语料有限，相对于知网来说，特征词汇覆盖面比较小，造成 SVMlight 特征向量矩阵比较稀疏，不能达到一个理想的学习效果，也是影响实验结果的一个重要因素.

由于受到目前句法分析工具准确率的限制，本体和喻体的提取会影响最终实验结果的准确率和召回率，由此可见加强对本体和喻体的提取是提高隐喻识别精度的一个重要因素.

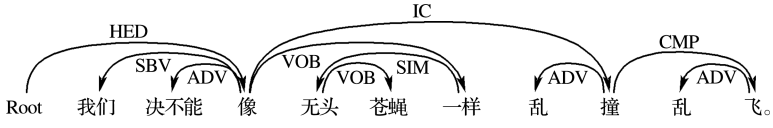

图 3　依存句法分析示例之二

Fig. 3　Example 2 of interdependence parsing analysis

7　结　语

本文提出了一种基于语义和词汇相似度的汉语显性隐喻识别方法. 仔细分析实验结果可以发现，当前实验中，由于选择的语料句子比较复杂，在特征词抽取方面的准确率还不是很高，造成词汇范畴判断和相似度计算得不到正确的结果.

在以后的研究中，除进一步提高抽取准确率外，可以考虑更多类型的隐喻，对多种类型隐喻的识别进行研究.

参考文献：

[1] 俞士汶. 语料库与综合型语言知识库的建设[C]//中文信息处理若干重要问题. 北京：科学出版社，2002.
YU Shi-wen. Construction of corpus and comprehensive language knowledge-base [C] // **Several Important Problems for Chinese Information Processing**. Beijing: Science Press, 2002. (in Chinese)

[2] 张　威，周昌乐. 汉语隐喻理解的逻辑描述初探[J]. 中文信息学报，2004，**18**(5)：23-28.
ZHANG Wei, ZHOU Chang-le. Study on logical description of Chinese metaphor comprehension [J].

Journal of Chinese Information Processing, 2004, **18**(5):23-28. (in Chinese)

[3] 杨 芸,周昌乐,王雪梅,等. 基于机器理解的汉语隐喻分类研究初步[J]. 中文信息学报, 2004, **18**(4):31-36.
YANG Yun, ZHOU Chang-le, WANG Xue-mei, *et al*. Research into machine understanding-based classification of the metaphor of Chinese [J]. **Journal of Chinese Information Processing**, 2004, **18**(4): 31-36. (in Chinese)

[4] 周昌乐. 隐喻、类比逻辑与可能世界[J]. 外国语言文学研究, 2004, **4**(2):17-22.
ZHOU Chang-le. Metaphor, analogous logic and possible worlds [J]. **Research in Foreign Language and Literature**, 2004, **4**(2):17-22. (in Chinese)

[5] 王治敏. 汉语名词短语隐喻识别研究[D]. 北京:北京大学, 2006.
WANG Zhi-min. Chinese noun phrase metaphor recognition [D]. Beijing: Peking University, 2006. (in Chinese)

[6] 贾玉祥,俞士汶. 基于词典的名词性隐喻识别[J]. 中文信息学报, 2011, **25**(2):99-104.
JIA Yu-xiang, YU Shi-wen. Nominal metaphor recognition based on lexicons [J]. **Journal of Chinese Information Processing**, 2011, **25**(2): 99-104. (in Chinese)

[7] 贾玉祥,俞士汶. 基于实例的隐喻理解与生成[J]. 计算机科学, 2009, **36**(3):138-141.
JIA Yu-xiang, YU Shi-wen. Instance-based metaphor comprehension and generation [J]. **Computer Science**, 2009, **36**(3):138-141. (in Chinese)

[8] 束定芳. 隐喻学研究[M]. 上海:上海外语教育出版社, 2000.
SHU Ding-fang. **Studies in Metaphor** [M]. Shanghai: Shanghai Foreign Language Education Press, 2000. (in Chinese)

[9] 陈建美,林鸿飞. 中文情感常识知识库的构建[J]. 情报学报, 2009, **28**(4):492-498.
CHEN Jian-mei, LIN Hong-fei. Constructing the affective commonsense knowledgebase [J]. **Journal of the China Society for Scientific and Technical Information**, 2009, **28**(4):492-498. (in Chinese)

[10] 徐琳宏,林鸿飞,赵 晶. 情感语料库的构建和分析[J]. 中文信息学报, 2008, **22**(1):116-122.
XU Lin-hong, LIN Hong-fei, ZHAO Jing. Construction and analysis of emotional corpus [J]. **Journal of Chinese Information Processing**, 2008, **22**(1):116-122. (in Chinese)

[11] 刘 挺,车万翔,李正华. 语言技术平台[J]. 中文信息学报, 2011, **25**(6):53-62.
LIU Ting, CHE Wan-xiang, LI Zheng-hua. Language technology platform [J]. **Journal of Chinese Information Processing**, 2011, **25**(6):53-62. (in Chinese)

[12] 董振东,董 强. 知网[DB/OL]. http://www.keenage.com.
DONG Zhen-dong, DONG Qiang. HowNet [DB/OL]. http://www.keenage.com. (in Chinese)

[13] 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:[s n], 2002.
LIU Qun, LI Su-jian. Word meaning similarity computation based on HowNet [C] // **The 3th Chinese Lexical Semantics Workshop**. Taipei:[s n], 2002. (in Chinese)

[14] Cristianini N, Shawe-Taylor J. **An Introduction to Support Vector Machines** [M]. Cambridge: Cambridge University Press, 2000.

[15] Canny J. A computational approach to edge detection [J]. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 1986, **8**(6):679-698.

[16] Joachims T. Making large-scale SVM learning practical [M] // Schölkopf B, Burges C, Smpla A. **Advances in Kernel Methods — Support Vector Learning**. Cambridge: MIT Press, 1999.

Mechanism of dominant sentimental metaphor identification based on lexical domain and semantic similarity

LIN Hong-fei*, XU Kan, REN Hui

(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)

Abstract: Metaphor is a natural phenomenon of the universal language. In the process of affective computing, the feelings of some sentences are aroused by the metaphor. Therefore, sentimental metaphor will affect the outcome of affective computing. Based on the theory of metaphor, from the perspective of machine learning, dominant metaphor identification of Chinese text is studied. The difference of areas between target domain and source domain is regarded as breakthrough point. Firstly, target domain and source domain are got using part of speech tagging and the interdependence parsing analysis; then, areas classifying and word similarity calculation are given; lastly, support vector machine is used so as to identify metaphor of sentence. This study will play a certain role in the automatic identification of metaphor and metaphorical understanding.

Key words: metaphor identification; domain division; semantic similarity