

文章编号: 1003-0077(2018)02-0022-07

一种基于聚类与分类结合的汉语隐喻短语识别方法

符建辉^{1,2}, 王 石¹, 曹存根¹

(1. 中国科学院 计算技术研究所 智能信息处理实验室, 北京 100190; 2. 中国科学院大学, 北京 100190)

摘 要: 隐喻识别是自然语言处理的一个重要研究分支。目前人们越来越清楚地认识到隐喻在思维及语言中所处的重要地位。本研究在前人工作的实验和考察基础上,发现基于分类器来识别隐喻的方法存在数据稀疏的问题,即当训练语料中缺少需要识别的源域词数据时,分类的结果将不会太好。应对数据稀疏问题,该文提出了一种基于聚类与分类结合的隐喻短语获取方法。该方法将包含源域词 S 的短语进行聚类。将聚类的结果作为分类的一类特征。实验表明,使用聚类产生的特征训练出来的分类器,不仅能很好地识别训练语料中存在源域词数据的情况,也能很好地识别训练语料中缺少源域词数据的情况,具有很高的召回率。

关键词: 隐喻短语识别; 中文隐喻短语; 短语聚类

中图分类号: TP391

文献标识码: A

Chinese Metaphor Phrase Recognition via Combining the Clustering and Classification

FU Jianhui^{1,2}, WANG Shi¹, CAO Cungen^{1,2}

(1. Key Laboratory of Intelligent Information Processing, Institute of Computer Technology,
Chinese Academy of Sciences, Beijing 100190, China;

2. University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Metaphor is popular in any natural language, and metaphor recognition is one of the challenging topics in natural language processing. Existing classification based metaphor recognition methods suffer from data sparsity, which affects the performance of the classification. In this paper, we propose a metaphor phrase recognition method by combining classification and clustering methods to improve the performance. This method firstly conducts the clustering on phrases with source words S , and then uses the clustering results as the features for classification. The classifier also produces a satisfactory performance for those phrases which miss source words. Several experiments show that our methods achieve a high recall rate.

Key words: metaphor phrase recognition; Chinese metaphor phrase; phrases clustering

0 引言

隐喻处理研究是自然语言处理的一个重要分支。人们越来越清楚地认识到隐喻在思维及语言中所处的重要地位。俞士汶甚至认为隐喻是自然语言理解必须攻克的难关^[1]。隐喻识别的提升将有助于自然语言处理其他问题识别的提升。例如,在知识获取领域,如果知道“知识海洋”不是一个“海洋”而是一个隐喻名词,那么就不会错误地判断“海洋”是

“知识海洋”的下位。又例如,在机器翻译中,隐喻名词“草木皆兵”,在缺少语料情况下很可能会翻译成“Every bush and tree is an enemy”。如果能够识别“草木皆兵”是一个隐喻名词,将有助于得到正确结果“Every bush and tree looks like an enemy”。

认知语言学认为:一个概念隐喻包含两个部分,一个“源域”(source domain)和一个“目标域”(target domain)。“源域”通常是熟知的比较具体直观、容易理解的一些概念范畴,而“目标域”通常是后来才认识的抽象的、不太容易理解的概念范畴^[2]。

收稿日期: 2017-03-13 定稿日期: 2017-09-25

基金项目: 国家自然科学基金(30973713, 61035004, 61173063, 61203284, 91224006); 国家社科基金(10AYY003); 科技部项目(201303107)

这里沿用“源域”和“目标域”的说法,将能够在句子中作为“源域”出现的词称为源域词,例如“杀手”“大军”“海洋”等都可以称为源域词。

汉语短语中存在大量的隐喻现象,我们将带隐喻义的短语称为隐喻短语。例如表 1 中有两种模式的隐喻短语。本文的工作是识别句子中的短语是否是隐喻短语。

表 1 隐喻短语举例

模式	举例
A+N(形容词+名词)	她是一朵残忍的花
N+N(名词+名词)	就业大军

我们将隐喻短语的识别看作一个分类问题,即一个短语要么是隐喻要么是非隐喻。由于源域词是一个不断发展的词汇集,训练语料中很难覆盖所有源域词数据,我们将这种训练集中缺少源域词数据的情况称为源域词的数据稀疏问题。

本研究在已有工作的实验和考察基础上,应对源域词的数据稀疏问题,提出了一种新的隐喻短语识别方法,该方法能够将聚类算法和分类算法的优点相结合。该方法首先将包含源域词 S 的短语进行聚类,将聚类的结果作为分类的一类特征。在分类时,我们同时也考虑 S 所处的上下文特征和包含 S 短语的属性特征。实验表明,使用了聚类产生的特征训练出来的分类器,不仅能很好地识别训练语料中存在源域词数据的情况,也能很好地识别训练语料中缺少源域词数据的情况,具有很高的召回率。

1 相关工作

自 20 世纪 70 年代以来,各种隐喻计算模型陆续出现。Fass^[3]提出了可以处理隐喻、转喻、字面义和反常表达的隐喻理解模型 MET5 系统。Martin^[4]提出了识别和解释常规隐喻的 MIDAS 系统。Mason^[5]利用大规模语料动态提取优先参数来识别特定领域的隐喻表达。Birke 和 Sarker^[6]给出了一个识别字面义表达和非字面义表达的计算模型——TroFi(Trope Finder)系统,解决了动词的字面义和非字面义用法的识别与分类问题。Gedigian 等^[7]在 WSJ 等语料库和 PropBank 命题库以及 FrameNet 映射标注方法的基础上,利用最大熵模型给出了动

词隐喻的分类器。Shutova^[8]提出了一种通过对动词和名词的聚类来进行隐喻识别的方法。Yosef Ben Shlomo 和 Mark Last^[9]提出了一种基于分类算法的隐喻识别模型。

在汉语隐喻研究中,王治敏^[2, 10]采用最大熵模型对形如“N+N”的名词隐喻进行了识别。赵红艳^[11]利用条件随机场和最大熵模型并结合一定的语义信息对隐喻现象进行识别。李斌、于丽丽等人^[12]将最大熵模型和条件随机场模型相结合解决了“像”的明喻计算问题。黄孝喜^[13]提出了一种基于树模式匹配的隐喻识别算法。

目前隐喻知识识别的研究多采用分类器的方法,并取得了许多进展。但基于分类的方法存在源域词的数据稀疏问题,即:当训练语料中缺少源域词数据时,基于分类的方法便会失败。例如,很难利用源域词“杀手”的上下文特征来识别包含源域词“大军”的短语是否是隐喻。而隐喻是一个不断发展并时刻新增的现象,源域词也会层出不穷,我们很难构建一个包含所有源域词的训练集来保证训练的效果。

针对源域词的数据稀疏问题,我们试图让包含同一个源域词的短语集合进行聚类。希望通过短语自身的相似度比较,隐喻短语和非隐喻短语能够相互聚成不同的簇。但我们很难判断聚类后的簇的归属(是隐喻短语簇还是非隐喻短语簇)。另外,聚类方法因为没有使用训练集,也很难充分考虑隐喻短语的许多其他的特征。

本工作的创新之处在于,结合了基于聚类和分类的两种识别方法的优点,设计了一套隐喻短语识别方法。我们将包含源域词 S 的短语进行聚类。通过对聚类后的簇的分析,抽取出聚类特征,并将这些特征作为分类的一类特征。在分类时,我们同时也考虑 S 所处的上下文特征和包含 S 短语的属性特征。实验发现,在存在数据稀疏的情况下,使用聚类特征的分类结果无论在正确率上还是在召回率上都得到很大的提高,并得到较好的结果。因此,利用聚类的方法能够有效解决隐喻分类识别方法中的数据稀疏问题。

2 汉语隐喻短语的识别

前期工作中我们已积累 1 021 个源域词,部分示例如表 2 所示。

表 2 部分源域词示例

长廊	东西	天空
平台	冬瓜	平原
冠军	旅馆	开水
宫殿	豆腐	古墓
沙漠	大军	杀手
花瓶	间谍	天堂

本文工作是从句子中识别隐喻短语。针对源域词 S , 我们从语料中抽取包含 S 的 $N+N$ 和 $A+N$ 形式的短语。要判断包含 S 的短语的句子是否是

隐喻, 只需判断包含 S 的短语是否是隐喻短语。我们将隐喻短语的识别看作一个分类问题。即一个短语要么是隐喻, 要么是非隐喻。我们利用搜索引擎对每个源域词进行检索, 从包含源域词的句子中抽取包含源域词且形式是 $N+N$ 或 $A+N$ 的短语, 这些短语以及短语所处的句子构成本文工作的实验语料。

本文方法分以下两个步骤:

步骤 1 隐喻短语的聚类识别

如图 1 所示, 对于语料中的每一个源域词 i , 将包含源域词 i 的短语 P_{i1}, \dots, P_{im} 进行聚类。聚类后得到簇 C_{i1}, \dots, C_{im} 。再从这些簇中抽取每一个短语 P 的聚类特征。

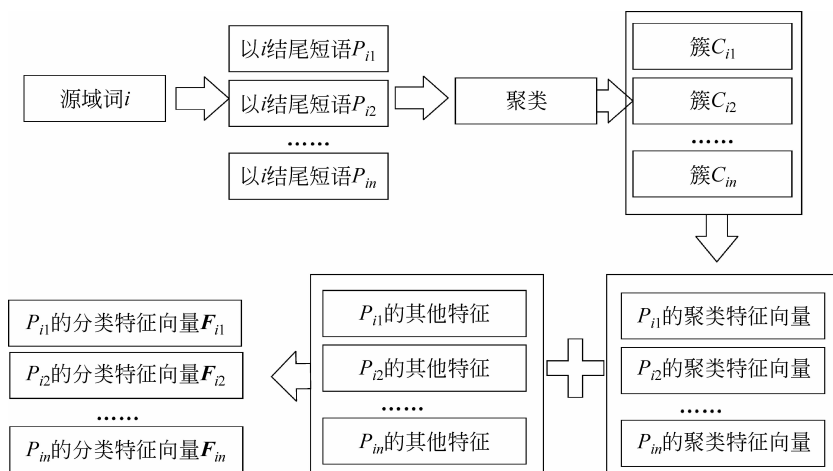


图 1 汉语短语特征的生成流程图

步骤 2 隐喻短语的分类识别

将步骤 1 生成的聚类特征结合其他特征组成 P_{ij} 最终的特征向量 $F_{i1} \dots F_{im}$ 。再将所有汉语短语生成的特征向量利用分类器进行训练和分类。

2.1 隐喻短语的聚类识别

通过对汉语隐喻短语的分析发现, 一个隐喻短语的最后一个词通常为该短语的源域词。例如, “心灵<沙漠>”“知识<海洋>”“就业<大军>”等。当然也有出现在短语首部的情况, 例如, “<花>样年华”。我们从语料中抽出了 300 个隐喻短语, 其中仅有 10 个隐喻短语的源域词是出现在前面。本文重点考查源域词出现在隐喻短语的末尾的情况。下面我们用源域词“大军”为例来说明本方法的思想。“刘邓大军”不是隐喻, 因为“刘邓大军”的上位是“大军”, 进行军事战斗的部队。而“就业大军”并不是真正意义上的“大军”, 它实际指就业人员像大军一样拥挤。在语料中考查“刘邓大军”和“蒙古大军”发

现, 如果源域词以字面义出现在短语中, 其上下文经常和“战争”“敌人”“厮杀”等字面义相关词出现。而作为隐喻义时, 往往不会出现这些相关上下文或只出现少量字面义相关词, 例如,

“36 万‘就业大军’今年步入职场, 你如何能脱颖而出……。”

在上文中更多出现的是和“就业”相关的词汇。也就是说, 源域词在汉语短语中不表现出隐喻义时, 该汉语短语常与其字面义相关词集共现频率较高, 其上下文存在一定的共性。我们利用搜索引擎抽取扩展汉语短语 P 的上下文信息。将包含源域词 S 的短语 P 利用搜索引擎检索, 抽取搜索引擎前 100 项检索到的网页片断, 这些片断都是包含检索项的一两句话。同时, 利用一个停用词表将一些词(如“网页快照”“图片”“网页”等)过滤掉。我们将这 100 项网页片断合成, 构成一篇文档 D 。短语 P_i 和 P_j 的相似度用 D_i 和 D_j 的相似度来表示。在计算 D_i 和 D_j 的相似度时, 我们采用常用的 cosine 余弦距

离来计算。具体计算方法如式(1)所示。

$$\begin{aligned} \text{Sim}(P_i, P_j) &= \text{SimDis}(D_i, D_j) \\ &= \text{COS}(D_i, D_j) \\ &= \frac{\sum_{k=0}^m (\omega(t_i^{(k)}) \cdot \omega(t_j^{(k)}))}{\sqrt{\sum_{k=0}^m \omega(t_i^{(k)})^2} \cdot \sqrt{\sum_{k=0}^m \omega(t_j^{(k)})^2}} \end{aligned} \quad (1)$$

其中, m 表示 D_i 和 D_j 中词的个数; $\omega(t_i^{(k)})$ 表示 D_i 的第 k 个词的权重, 权重计算方法使用了 tf/idf 计算方法。

基于上面的两个汉语短语的相似度计算, 我们采用层次聚类对所有包含源域词 S 的短语 P 进行聚类。具体聚类算法如算法 1 所示。

算法 1: 短语的层次聚类算法

输入: $D_1, D_2, \dots, D_i, \dots, D_j, \dots, D_n$; 阈值 λ

输出: m 个簇 C_1, \dots, C_m

- (1) Do Begin
- (2) 设置每个 D_i 为一个簇 C_i ;
- (3) repeat
- (4) 计算每两个簇 C_i 和 C_j 的距离;
 $\text{Dist}(C_i, C_j) = \min_{D_i \in C_i, D_j \in C_j} \{ \text{SimDis}(D_i, D_j) \}$
- (5) 找到 Dist 值最小的 Dist_{\min} ,
 假设 $\text{Dist}_{\min} = \text{Dist}(C_x, C_y)$ 。
 if $\text{Dist}_{\min} > \lambda$, $C_{\text{new}} = \text{merge}(C_x, C_y)$ 。
- (6) Until $\text{Dist}_{\min} < \lambda$
- (7) End

聚类后, 根据 P_i 所在簇的信息, 我们生成以下关于 P_i 的聚类特征:

- (1) P_i 所在簇的大小;
- (2) P_i 所在簇是否包含源域词 S (在聚类过程中, 我们将源域词 S 本身也参与聚类);
- (3) P_i 所在簇中低频率短语的比例 F_1 ;
 $F_1 = \text{簇中低频词数量} / \text{簇大小}$;
- (4) P_i 所在簇中“A+N”短语所占比例 F_2 ; $F_2 = \text{“A+N”短语数量} / \text{簇的大小}$ 。

2.2 隐喻短语的分类识别

在构建分类器时, 除上面短语聚类后生成的特征外, 还考虑两类特征: 短语上下文特征、短语的属性特征。

2.2.1 短语上下文特征

源域词 S_i 的上下文定义为:

$\text{Sent} = \langle W_{i-p} \dots W_{i-2}, W_{i-1}, W_i, W_{i+1}, W_{i+2}, \dots, W_{i+q} \rangle$

其中, Sent 代表源域词 W_i 所处的句子。句子 Sent 是从搜索引擎检索 S_i 获得的网页摘要中抽取,

Sent 包含 S_i 且 S_i 处在 N+N 或 A+N 形式的短语中, 同时还要过滤掉停用词, 本文考虑上下文特征, 短语自身特征及说明如表 3 所示。

表 3 短语自身特征及说明

特征类型	说明
W_{i-1} 及词性	出现在源域词左边的第一个词
W_{i-2} 及词性	出现在源域词左边的第二个词
W_{i+1} 及词性	出现在源域词右边的第一个词
W_{i+2} 及词性	出现在源域词右边的第二个词
W_i 所处短语是否间隔“的”“之”字	例如, “知识的海洋、生活的天堂、希望的原野”等
W_{i-1} 是否是姓氏	例如, “李海洋”, 带有姓氏的短语很可能是姓名
W_i 后是否有标点符号	在隐喻短语后面往往跟着标点符号
包含 W_i 的短语是否是低频词	考查发现, 当用搜索引擎检索词汇低于 10 000 条记录数时一般为低频词

2.2.2 短语的属性特征

属性规则是利用源域词本身的属性来判断一个词归属的一种方法。一般认为, 在上下位关系中, 下位共享着上位的大部分的属性。隐喻词汇因为不是源域词的下位, 所以其并不具有或者很少具有源域词本身的属性。例如, “沙漠”的属性有“面积”“温度”等。“撒哈拉大沙漠”是“沙漠”的下位, 将“撒哈拉大沙漠”和沙漠的属性词绑在一起, 并利用搜索引擎检索。我们的搜索串是: “撒哈拉大沙漠的面积”“撒哈拉大沙漠的温度”, 检索出来的词频分别是 1 030 条和 647 条。“爱情沙漠”是关于“沙漠”的一个隐喻词, 并不具备“面积”“温度”这些属性。我们用查询串“爱情沙漠的面积”“爱情沙漠的温度”来进行查询, 所得结果都为 0。

通过已有的工作, 我们积累了大量的属性词和属性值词^[14-15], 对于包含源域词 S 的短语 p , 利用已有的源域词 S 的属性词或属性值词 c 构造两种查询串: “ p 的 c ”和“ p 的 * c ”, 再利用搜索引擎检索, 并设定阈值 T , 如果检索到的记录条数高于该阈值, 就认为 p 具备属性 c 。表 4 给出了源域词及其属性相关词集示例。

表 4 源域词及其属性相关词集示例

源域	属性词	属性值词
暴雨	降水量, 降雨量, 积水	雷电, 雷雨, 大风, 强风, 雷击
海洋	面积, 水深, 深度, 宽度, 物种多样性, 生物多样性, 水温, 位置, 地理位置	公里, 群岛, 海峡, 米, 浪

续表

源域	属性词	属性值词
大军	战法,战场,敌人,人数,战斗力,伤亡人数,人马,任务,使命,意图	激战,师部,中原
杀手	动机,武器,任务,长相,身高,特征	残忍,谋杀,死亡,杀人
沙漠	面积,气候,降雨量,分布,植被,主要植被,生态,生态特征,生态环境,物种多样性,生物多样性	荒漠,适合,绿洲,公里
天空	云层,变化,种类,关系,监测,特征,预报	湛蓝,雨雪,雨滴,阴天,雪天,乌云,晚霞,晴空,蓝天,积雪
上帝	<无法获取>	<无法获取>
天堂	<无法获取>	<无法获取>

实验证明,当 T 取 75 时,结果最好,如式(2)所示。

$$f(c_i, p) = \begin{cases} 1 & \text{检索 } c_i + p \text{ 条目} > T \\ 0 & \text{else} \end{cases} \quad (2)$$

当 $f(c_i, p) = 1$ 时,表示短语 p 具备属性 c_i 。我们的属性特征表示如下:

$$f(c_i, p) = \begin{cases} 1 & \text{检索 } c_i + p \text{ 条目} > T \\ 0 & \text{else} \end{cases} \quad (3)$$

$$f = \begin{cases} \frac{\sum_i f(c_i, p)}{\text{num}(c)} \text{num}(c) > 0 \\ -1 \text{num}(c) = 0 \end{cases} \quad (4)$$

f 表示短语 p 具备源域词 S 属性的程度, $\text{num}(c)$ 表示源域词 S 的属性及属性值词的大小。 $\text{num}(c) = 0$ 时表示无法获取源域词 S 的属性词或属性值词,此时无法获知 p 具备 S 属性的程度,所以用 $f = -1$ 来代替。

从表 4 可看出,有些源域词本身很难从语料中自动获取属性词或属性值词。这些词一般是一些带有强烈隐喻义的词汇,它们在句子中更多地以隐喻出现,而其字面义出现的情况反而很少。我们将无法获取属性或属性值词也作为分类特征的一种。即,如果一个源域词 S ,无法抽取其属性词或属性值词,很有可能这个词在语料中倾向于作为隐喻出现。

2.2.3 分类器选择及分类分法

2.2.3.1 分类器的选择

在分类器的选择上,我们采用以下分类器: Naïve Bayes、CRF、最大熵和 SVM(高斯核函数)。同时我们对每种分类器都使用 AdaBoost 算法进行迭代提升。

2.2.3.2 分类预处理

在分类之前,需要遍历每个源域词 S ,将包含 S 的短语进行聚类,并抽取聚类特征。此时我们还需要判断包含 S 的短语数量是否足够多。当短语数量超过 10 时才考虑使用聚类来得到短语的上下文共性特征。因为实验发现,当数量小于 10 个时,聚类的结果并不理想。

2.2.3.3 分类后处理

为充分利用聚类出来的簇的信息,针对源域词 S 的短语集合,我们循环对每一个短语 i 进行分类判断是否是隐喻之后,再利用簇的信息再次进行结果的校正,具体校正规则如下:当短语 i 所在簇的元素数量大于 1,并且簇中非短语个数大于短语个数时,认为短语 i 的归属为非隐喻。即,默认为聚类后的簇中元素,或者都为隐喻,或者都为非隐喻。

3 实验与结果

3.1 短语聚类结果及分析

通过以前的工作,我们已积累源域词 1 021 个,从句子中抽取隐喻短语 10 023 个,非隐喻短语 40 097 个。具体源域词的积累工作如下:

(1) 从三千万名词短语中将最后一个词抽取出来,得到 30 056 个尾词;

(2) 人工从 30 056 个尾词中抽取可能的源域词,具体利用百度搜索引擎检索候选源域词,查看是否能发现隐喻短语,如果存在隐喻短语,则说明候选源域词是源域词。

首先针对 1 021 个源域词,对每个源域词 S ,抽取包含 S 的短语作为实验语料,然后对包含源域词 S 的短语进行聚类。源域词“大军”的聚类结果如图 2 所示。

从图 2 可看出,当源域词作为本义出现时,其对应短语倾向于聚合在一起;当源域词作为隐喻出现时,部分短语也会被聚在一起,这是因为这些隐喻词在一定程度上共用源域词的某些属性导致上下文有一定的相似性。另外,有许多隐喻或非隐喻词汇被聚散,其中大部分被聚散的是隐喻词汇。

我们采用聚类结果的纯度^[16]来评价聚类的效果。其定义如下:给定一个聚类 C 和一个类别 A ,对于每个在 C 中的簇 c ,我们计算类分布如式(5)所示。

$$p_{ca} = \frac{f(c, a)}{f(c, *)} \quad (5)$$

“大军”“蒙古大军”“成吉思汗大军”“中国大军”“李自成大军”“造反大军”“起义大军”“苏联大军”“希特勒大军”“美国大军”“联盟大军”“侵略大军”“百万大军”“记者大军”“三路大军”“林彪大军”“刘邓大军”“陈粟大军”“解放大军”“抗日大军”“贺龙大军”“彭德怀大军”“北伐大军”“帝国大军”“远征大军”“曹操大军”“刘备大军”“皇马大军”“秦国大军”	“地摊大军” “失业大军”
“生产大军”“销售大军”“科技大军”“打工大军”“学生大军”“租房大军”“考研大军”“求职大军”“毕业生大军”“就业大军”“考古大军”“扫墓大军”“旅游大军”	“恋爱大军”
“地摊大军” “失业大军”	“绿化大军”
“减肥大军” “跑步大军”	“蝗虫大军”
	“生育大军”
	“项羽大军”
	“古罗马大军”
	“化妆师大军”

图 2 源域词“大军”对应语料聚类结果

其中 a 是 A 中的一个类, $f(c, a)$ 是簇 c 中元素在类 a 中的个数。* 为通配符。

簇 c 的熵的计算如式(6)所示。

$$E_c = - \sum_{a \in A} p_{ca} \times \log(p_{ca}) \quad (6)$$

纯度计算如式(7)所示。

$$E = \sum_{c \in C} \frac{|c| \times E_c}{|C|}$$

另外,有些簇中短语个数非常少,常有出现个数为 1 的情况,这种簇无实际意义,故不加入纯度计算。本实验只考虑簇中元素个数大于 5 的情况,简称这种元素个数大于 5 的簇为大簇。否则就为小簇。部分源域词聚类结果如表 5 所示。

表 5 部分源域词聚类结果

	C	E(C)
沙漠	17	0.54
大军	19	1.21
杀手	18	1.19
海洋	12	0.15
天空	35	1.35
.....		

表 5 统计所有参与聚类的源域词,其平均 $E(C) = 0.87$ 。从这个值来看,聚类出来的簇的纯度是非常高的。也即,聚类的簇中的元素一般是隐喻短语,或者是非隐喻短语。另外,也有许多非隐喻短语没有被聚成簇,分析影响聚类效果的原因如下:

(1) 有时短语本身就含有多个义项,比如“马路杀手”,既可以指某一种对马路破坏很大的东西,也可以指某一类专门在马路上杀人的罪犯。这两种意

思都可能在语料中出现。

(2) 有些短语在语料中并不表现出词本身的意思,而常为一些公司的名称。即使是非隐喻术语也如此。比如,“东方海洋”,搜索前 60 个网页中,全部都嵌在一个公司名中。

(3) 聚类的效果与算法本身有关,因为层次聚类本身的不可逆性导致获得结果并非一定是最优的。

3.2 短语分类结果及分析

(1) 训练集和测试集的构造

为了测试训练出来的分类器对未在训练集中的源域词也有效果,我们将已有源域词分为两部分,一部分源域词及其短语作为训练集;另一部分源域词及其短语作为测试集。这样就保证了测试集中的源域词没有在训练集中出现。

同样,我们也测试源域词在训练集中出现的情况。我们将在训练集中的源域词的部分短语抽取出来作为测试集。在训练时,我们都采用十折交叉验证。

(2) 分类器选择

采用精度 $P(\text{precision})$ 、召回率 $R(\text{recall})$ 以及 F 值($F\text{-measure}$)来评价我们的最终结果。在考虑上下文特征、聚类特征、属性特征的情况下,使用不同分类算法所得结果如表 6 所示。

表 6 不同分类算法结果

	$P/\%$	$R/\%$	$F/\%$
SVM	81.70	83.20	82.4
CRF	86.80	73.30	79.5
最大熵	79.40	71.30	75.1
Bayes	77.00	71.50	74.1

通过表 6 中数据发现 SVM 在这些特征下效果表现最好。

(3) 不同特征组合下的实验结果比较

下面我们将使用 SVM 继续考察各分类特征在分类中的作用。我们设计以下分类器。

SVM 分类器 a: 训练和分类时只考虑上下文的特征;(不加入聚类过程)

SVM 分类器 b: 训练和分类时考虑短语上下文特征和属性特征;(不加入聚类过程)

SVM 分类器 c: 训练和分类时考虑短语上下文特征、属性特征、聚类特征;(加入聚类过程和属性特征)

SVM 分类器 d: 训练和分类时考虑短语上下文特征、聚类特征;(加入聚类过程)

各分类器的分类结果如表 7 所示。

表 7 源域词 S 已出现在训练集中的分类结果

	P/%	R/%	F/%
分类器 a	71.70%	83.20%	77.0%
分类器 b	83.40%	71.70%	77.1%
分类器 c	88.40%	85.70%	87.0%
分类器 d	81.70%	73.20%	77.2%

通过表 7 可看出,因为源域词 S 已在训练集中出现,在训练集中存在源域词数据的情况下分类器 a 的效果是不错的。在加入属性特征之后,分类器 b 的结果比分类器 a 的结果明显提高。从分类器 a 和 b 可看出:如果能保证源域词的训练集大小,是可以通过分类器很好的识别隐喻现象。加入聚类特征和属性特征的分类器 c 的结果明显有所提升。说明聚类特征即使在源域词充分的情况下也有提升作用。

表 8 中考查了源域词 S 没有在训练集出现的情况。分类器 a 和分类器 b 因为缺少源域词信息,导致识别结果较差。通过加入聚类特征,分类器 c 的效果明显提升。因为分类器 d 没有考虑属性特征,所以其结果比分类器 c 差。

表 8 源域词 S 未出现在训练集中的分类结果

	P/%	R/%	F/%
分类器 a	56.70	63.20	59.8
分类器 b	67.40	75.70	71.3
分类器 c	87.50	78.60	82.8
分类器 d	75.80	71.30	73.5

4 结论和下一步工作

汉语隐喻处理在中文信息处理领域是一个新的研究方向。本文在对前人的实验进行考察的基础上,发现通过分类器来识别隐喻的方法存在严重的数据稀疏问题。为应对数据稀疏问题,本文提出了一种聚类 and 分类结合的隐喻短语识别方法。该方法将包含源域词 S 的短语进行聚类,产生基于源域词自身的聚类特征。在利用分类器训练时,将聚类特征加入。同时我们也考虑了上下文特征和属性特征。在最后的实验结果分析部分,我们重点分析了聚类特征所起的作用。实验表明,使用聚类产生的特征训练出来的分类器,不仅能很好地识别训练语

料中存在源域词数据的情况,也能很好的识别训练语料中缺少源域词数据的情况,具有很高的召回率。

另外,我们分析了目前该方法中仍存在的问题,并认为本方法还有很大的提升空间。

(1) 本方法第一步需要获取源域词,源域词的多少直接关系到本方法的结果,而源域词集合是通过人工进行抽取的。该抽取过程耗时耗力,并且新的源域词也会随着语言的发展不断增多。所以有必要增加自动获取源域词方法。后续我们将重点在这方面进行考察。

(2) 有些词本身就有二义性。比如“少女杀手”,该词既可以表示专杀少女的杀手,也可以表示获得少女芳心的情场高手。这种词的存在造成区分界线不明显,对结果带来一定的影响。另外,在测试集中存在着一些姓名和商标名,比如李海洋、赵大军等这种词。因为这些词本身不是隐喻短语,但源域词在其中又不作为本义出现。所以,用本方法对它们进行识别,常得出错误的结果。

(3) 属性词作用有限,有些词汇虽然有某种属性,但并不一定在语料中和该属性词同时出现。例如,“中国沙漠”虽然具备沙漠的属性,但“中国沙漠的面积”“中国沙漠的温度”的检索结果都为 0。类似这样的词汇有很多,例如,“西方大军”“东方海洋”等。

以上问题都是我们今后所要研究和解决的重点。

参考文献

- [1] 徐波, 孙茂松, 靳光瑾. 中文信息处理若干重要问题[M], 北京: 科学出版社, 2003: 55-56.
- [2] 王治敏. 名词隐喻相似度及推理识别研究[J]. 中文信息学报, 2008, 22(3): 37-43.
- [3] Fass D. met *: A method for discriminating metonymy and metaphor by computer[J]. Computational Linguistics, 1991, 17(1): 49-90.
- [4] Martin J H. A computational model of metaphor interpretation[M]. San Diego, CA, USA: Academic Press Professional Inc, 1990.
- [5] Mason Z J. CorMet: A computational, corpus-based conventional metaphor extraction system[J]. Computational Linguistics, 2004, 30(1): 23-44.
- [6] Birke J, Sarkar A. A clustering approach for nearly unsupervised recognition of nonliteral language[C]// Proceedings of the European Chapter of the Association for Computational Linguistics. Trento, Italy, 2006: 329-336.

(下转第 49 页)



王东升(1982—), 博士, 讲师, 主要研究领域为自然语言处理、知识工程、语义网等。

E-mail: wds_ict@163.com



王石(1981—), 博士, 副研究员, 主要研究领域为文本处理、问答系统、知识工程等。

E-mail: wangshi@ict.ac.cn



王卫民(1977—), 博士, 讲师, 主要研究领域为自然语言处理、知识管理、问答系统等。

E-mail: wangweimin@google.com

(上接第 28 页)

- [7] Gedigian M, Bryant J, Narayanan S, et al. Catching metaphors[C]//Proceedings of the Third Workshop on Scalable Natural Language Understanding, New York, 2006: 41-48.
- [8] Shutova E, Korhonen A. Metaphor identification using verb and noun clustering[C]//Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, 2010: 1002-1010.
- [9] Yosef B S, Mark L. MIL: Automatic metaphor identification by statistical learning[C]//Proceedings of the Workshop on Interactions Between Data Mining and Natural Language Processing, Porto, Portugal, 2015: 19-29.
- [10] 王治敏. 汉语名词短语隐喻识别研究[M]. 北京: 北京语言大学出版社, 2010: 1-19.
- [11] 赵红艳, 曲维光, 张芬, 等. 基于机器学习与语义知识的动词隐喻识别[J]. 南京师范大学学报(工程技术版), 2011, 11(3): 59-64.
- [12] 李斌, 于丽丽, 石民, 等. “像”的明喻计算[J]. 中文信息学报, 2008, 22(6): 27-32.
- [13] 黄孝喜. 隐喻机器理解的若干关键问题研究[D]. 杭州: 浙江大学博士学位论文, 2009.
- [14] 汪平仄. 面向 Web 语料的概念属性和属性值获取方法研究[D]. 北京: 中国科学院大学博士学位论文, 2014.
- [15] 汪平仄, 曹存根, 王石. 一种迭代式的概念属性名称自动获取方法[J]. 中文信息学报, 2014, 28(4): 58-67.
- [16] Steinbach, M., G. Karypis, V. Kumar. A Comparison of Document Clustering Techniques [C]//proceedings of KDD Workshop Text Mining, Boston, MA, USA, 2000: 1-20.



符建辉(1985—), 硕士, 工程师, 主要研究领域为知识获取、数据挖掘。

E-mail: fujianhui@ict.ac.cn



王石(1981—), 博士, 副研究员, 主要研究领域为知识的获取、表示与推理, 机器学习。

E-mail: wangshi@ict.ac.cn



曹存根(1964—), 博士, 研究员, 主要研究领域为大规模知识获取与管理。

E-mail: cgcao@ict.ac.cn