中国科学: 信息科学 2018 年 第 48 卷 第 11 期: 1510–1520
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
社会媒体计算与自然语言处理专刊 • 论文

基于多维语义关系的谐音双关语识别模型

徐琳宏[1*], 林鸿飞[2], 祁瑞华[1], 杨亮[2]

1. 大连外国语大学软件学院, 大连 116044
2. 大连理工大学计算机系, 大连 116024
* 通信作者. E-mail: qingniao1203@163.com

收稿日期: 2018–06–13; 接受日期: 2018–09–10; 网络出版日期: 2018–11–14

国家自然科学基金重点项目 (批准号: 61632011)、国家自然科学基金 (批准号: 61772103, 61702080)、国家社会科学基金一般项目 (批准号: 15BYY028)、辽宁省自然基金 (批准号: 20170540230, 2015020017, 20170540232) 和辽宁省优秀人才项目 (批准号: LJQ2014127) 资助

摘要 谐音双关语的识别是幽默研究领域的一个重要分支, 并逐渐发展为一个新兴的研究领域. 本文提出一种基于 4 个维度特征集的谐音双关语识别模型, 其中 4 个维度包括语义透明度、语义相关度、语音扩展性和语法特征集. 语义透明度包括词项统计和语句字符长度两个特征, 语法特征集包括人名、大写、时态、词性和位置 5 个特征. 将这 4 个维度的 9 个特征加入到二叉判定树中, 使用 K-Means 聚类获取阈值, 完成双关语的识别. 本文的实验数据来自于 SemEval2017 任务 7 的语料, 取得了较好的效果, $F1$ 值高于参赛队中的第一名, 实验证明基于 4 个维度特征的二叉判定树分类方法在谐音双关语识别中是有效的, 并且在多个特征中, 语音扩展性和语法特征集的效果比较明显, 这也符合谐音双关语识别中语音作用较大的预测.

关键词 谐音双关语, 情感分析, 二叉判定树, 语义特征集, 聚类

1 引言

英文中双关语是一种比较常用的修辞手段, 一个双关句通常采用词义模糊的方式, 使句子产生一个以上的含义 [1]. 也就是说作者特意让一个词在句中包含两个或者两个以上的独立含义, 这在许多英文著作中被广泛应用, 例如, 莎士比亚的作品集中经常使用双关语修辞. 此外, 双关语通常带有一定的幽默色彩, 在人们的日常交际中有较强的交互价值, 能更快地增进双方的感情. 从类别上来说, 双关语一般分为语义双关 (homographic pun) 和谐音双关 (heterographic pun). 它们在许多领域都有重要的研究价值, 如广告语的生成. 双关语中人们不仅仅制造幽默, 而且会引发听众联想双关语中的目标词, 识别出词语的模糊性, 从而有益于文本情感识别.

谐音双关是双关语中的一个重要类别, 它是指语句中某个词在语音上与目标词相同或相近, 用目标词替换后原句产生了新的含义, 例如, “Diets are for people who are thick and tired of it all”, 其中

引用格式: 徐琳宏, 林鸿飞, 祁瑞华, 等. 基于多维语义关系的谐音双关语识别模型. 中国科学: 信息科学, 2018, 48: 1510–1520, doi: 10.1360/N112018-00151
Xu L H, Lin H F, Qi R H, et al. Heterographic pun identification model based on multi-dimensional semantic relationships (in Chinese). Sci Sin Inform, 2018, 48: 1510–1520, doi: 10.1360/N112018-00151

“thick” 为双关词, 它的目标词为 “sick”. 找到双关词对应的目标词, 才能更好地理解语句的含义, 从而体会语句的幽默色彩, 可见双关语的理解考验着接受者的智慧. 因此, 机器理解双关语更是一个难题.

双关语识别和生成是幽默理解的一个研究分支, 在人机交互、问答系统等方面有重要的应用. 对幽默的理解, 有助于提高人与机器交互的水平, 使人与机器之间的沟通变得更加接近人与人之间的沟通. 2017 年国际语义评测 (semantic evaluation) 发布两个新的任务是关于幽默和双关语的, 这也更好地推动了机器对幽默理解力的研究. 双关语作为一个重要的幽默来源, 是近几年的研究热点, 国内外有许多这方面的研究工作.

2 相关工作

谐音双关语是双关语的一种特例, 而双关语中很大一部分具有幽默的特性, 所以双关语是幽默研究的一个重要分支. 下面从幽默、双关语和谐音双关 3 个方面介绍目前国内外主要的研究工作, 三者在研究方法上既有各自的特点, 也有很多相通的方法和特性.

2.1 幽默的相关研究

很多双关语都有不同程度的幽默感, 幽默作为一种语言形式, 能够为对话双方营造轻松、愉快的气氛, 因而具有独特的交际价值, 备受人们青睐. 近些年来, 自然语言处理中也有很多关于幽默识别方面的研究工作.

英文幽默方面的研究工作比其他语言更多, 2009 和 2010 年, Taylor 等 [2,3] 收集幽默文本, 结合幽默经典理论, 逐句分析幽默语句的特点. 2012 年, Antonio 等 [4] 利用模糊性、情感极性和情感场景等特征识别 tweet 文本的幽默性. 2012 年, Yishay 等 [5] 在 tweet 文本中, 采用模式、词汇、形态和语音等特征识别幽默文本. 2012 和 2016 年, Pascale 等 [6~8] 标注了《生活大爆炸》和《宋飞正传》语料集, 分别使用 CNN, CRF 和 LSTM 分类方法在语音和文本两大类特征基础上识别幽默. 2014 年, Zhang 等 [9] 利用 GBRT 模型识别幽默的 tweet 文本. 以上是英文幽默的研究工作, 其他语言也有一些幽默研究的工作, 主要包括: 2007 和 2009 年, Rosso 等 [10,11] 使用 Ngram 完成意大利语幽默研究; 2016 年, Castro 等 [12] 使用 KNN 和 SVM 识别西班牙文的幽默. 除了以上幽默识别方面的工作, 也有一些关于幽默生成的研究. 2012 年, Igor 等 [13] 基于 SSTH 理论做幽默语句生成的研究, 采用人工打分的方式评测生成语句的效果. 2013 年, Alessandro 等 [14] 通过词语替换, 生成幽默文本, 采用人工评估的方式评估幽默等级.

在幽默计算方面, 国内研究者也进行了一定的探讨. 2015 年, 张冬瑜等 [15] 构建了情感隐喻语料库, 这对于幽默的识别提供可以借鉴的方法. 2016 年, 林鸿飞等 [16] 回顾幽默研究的发展历史, 详细阐述了幽默计算理论的多种基本理论和应用, 对于谐音幽默的处理也给出了相应的讨论. 其他学者对于幽默计算所涉及的语言学理论和方法进行了相应的研究. 但目前中文幽默识别还处于起步阶段, 经过深入的研究和探索, 未来还有很大的发展空间.

2.2 双关语的相关研究

双关语识别作为幽默研究的一个分支, 在幽默特性的基础上, 还包含一些句式构成和语义模糊等独有的特点. 2005 年, Ritchie 等 [17] 在对笑话语言分析的基础上, 利用语音相似、上下文融合等特征, 生成双关语. 2008 年, Hempelmann [18] 研究双关语目标词的自动识别, 以及幽默文本的自动生成问题.

2009 年, Hong 等 [19] 自动抽取语义特征和语法结构模板生成双关语. 2011 年, 华盛顿大学的 Chloe 等 [20] 将原领域的字面意义到目标领域的非字面意义的映射过程定位成一个隐喻问题, 提出委婉语和领域结构两个特征作为语义和文化背景识别双关语. 2015 年, Yang 等 [21] 在 “pun of the day” 的数据集上识别幽默和锚点, 将冲突性、模糊性、人际交互和语音风格等特征加入到随机森林和最大衰减法中, 完成双关语的识别. 2015 年, Tristan 等 [22] 在传统 LESK 算法的基础上研究双关语的词义消歧, 提出两个解绑策略: 一是利用词汇的词性和语法特征消歧, 二是将 WordNet 中词义聚类的方法. 两个方法有效地解决了 LESK 算法词义相近较多的问题. 2016 年, Tristan 采用一种计算模型检测双关语及特定的语义隔离 (the isolation of the intended meanings), 并集成 WSD-inspired 系统评估语义双关语.

2.3 谐音双关语

谐音双关语是双关语的一种类型, 其双关词和目标词之间具有发音相同和相似的特性, 因而找到发音相似的词对非常重要, 这方面的相关研究很早就已经开展. 1986 年, Arnold 等 [23] 基于双关语实例, 生成 2140 个双关语词对. 1991 年, Wlodzimierz [24] 在 Arnold 工作基础上收集了 3850 个双关语和目标词的词对. 1996 年, Kim 等 [25] 利用英语发音词典 [26], 制定发音相似的规则生成发音相似的词对, 用于双关语生成. 2003 年, Hempelmann 等 [27] 在 Wlodzimierz 研究基础上生成一种简单的语音替换代价表, 用于评价生成后的双关语的分值. 2015 年, 斯坦福大学 Kao 等 [28] 提出模糊性 (ambiguity) 和独特性 (distinctiveness) 两个特征, 以语言模型为基础识别幽默双关语. 2016 年, Jaech 等 [29] 完成谐音双关语的识别, 提出的理解模型分为音素、词汇和文本 3 个层次, 利用语言模型识别谐音双关语, 并统计出谐音双关中元音和辅音相互替换的次数. 2017 年, Samuel 等 [30] 利用 Google 的 n-gram 和 CMU 的发音词典完成谐音双关语识别. Dipankar 等 [31] 用隐 Markov 模型和循环依赖网络, 采用语法和词性特征等特征识别双管. 2018 年, Diao 等 [32] 利用 WordNet 和改进的深度学习网络识别双关语.

英文方面, 关于幽默和双关语的研究大致分为 3 种: 一是通过具体双关语实例, 建立双关语和幽默的理论; 二是利用机器学习方法 SVM, KNN 和深度学习等有监督方法识别双关语; 三是完成双关语或幽默文本的生成. 目前在无训练集条件下识别谐音双关语的研究较少.

本文的主要贡献如下: (1) 提出谐音双关语识别中的 4 个维度的特征集; (2) 在无训练集的条件下, 采用二叉判定树识别双关语. 文中第 2 节介绍了幽默和双关语识别方面的研究工作; 第 3 节给出我们的识别模型, 将 4 个维度的特征集加入到二叉判定树中识别谐音双关语; 第 4 节介绍了数据集和相关语言模型; 第 5 节列出了识别模型在 SemEval2017 语料中的实验结果; 第 6 节总结了本文工作, 并提出今后工作的设想.

3 谐音双关语的识别模型

3.1 谐音识别中的 4 个维度的特征集

3.1.1 语义透明度

所谓语义透明度是指整个句子的语义可以根据合成语句的多个词汇含义来推知的程度 [33,34], 其中包括单词的使用频率, 搭配的熟悉程度和典型性等. 如 Mok [35] 认为词频越高, 其语义透明度则越高. Pollatsek 等 [36] 以词频和语义透明度为变量, 分析失语症患者语言产出的正确率. 在双关语识别中, 语义透明度也是一个重要的特征, 例如 “Psychiatrists like Kentucky Freud Chicken” 和 “Hotel owners usually have suite dreams” 两句中 “Freud Chicken” 和 “suite dreams” 的搭配频率较低, 更熟悉的搭

配是 “fried chicken” 和 “sweet dreams”. 为了解决上述问题, 本文使用语句中 Unigram 值和语句长度 (senlength) 两个特征来表征语句的语义透明度.

Unigram 值代表语句中词汇的使用频率, 使用频率越低, 则语义的透明度越低. 本文使用语言模型 (language model) 计算语句中词语及搭配的使用频率和熟悉程度.

语句长度值取语句中包含的字符个数, 而不是单词的个数, 因为单纯使用单词的个数, 不能体现那些单词较长, 但单词数较低的语句的复杂程度. 字符个数越大, 语句的透明度越低.

3.1.2 语义相关度

语义相关度是计算候选词在原句中的语义流畅程度, 即候选词在原句中替换后语义越流畅, 则该候选词与句子的相关程度越大, 选择相关度最大的作为整个双关句的目标词. 候选词与原句的语义相关度是通过 WordNet [37] 的 Similarity 接口计算, 先得到候选词与原句中每个词的相似度, 然后选择最大的相似度作为该词的语义相关度. 假设 C 为语音扩展词集合, 由 $c_1, c_2, c_3, \ldots, c_m$ 组成, 则通过式 (1) 计 c_i 的语义相关性:

$$\mathrm{Relate}(c_i) = \mathrm{MAX}_{k=1}^{n}(\mathrm{Sim}(c_i, w_k)), \tag{1}$$

其中 w_k 为原句中第 k 个单词. 那么整个语句 s 的语义相关度计算公式如式 (2) 所示:

$$\mathrm{Relate}(s_i) = \mathrm{MAX}_{i=1}^{m}(\mathrm{Relate}(c_i)), \quad c_i \in C, \tag{2}$$

其中 C 为语音扩展后的候选词列表, 如果没有语音扩展词, 则该句也没有语义相关度. 一个语句的语义相关度越大, 说明找到的词越可能是目标词, 那么该句是双关语的可能性就更大.

3.1.3 语音扩展性

CMU 发音字典 (CMU pronouncing dictionary) 是卡内基梅隆大学 (Carnegie Mellon University) 研发, 为语音合成器设计的[1)]. 最近发布的版本中包含 134000 个词条, 每个词条的发音采用英文音素集 Arpabet.

发音词典中一个词可以包含多个音标, 本文在做语音扩展时使用一个词条的所有发音, 与任何一个发音匹配都作为该词的语音扩展词. 语音扩展词分为两类: 发音相同和发音相似. 发音相同的词是直接在 CMU 发音字典中查找与该词条任一发音相同的词汇, 而寻找发音相似的词汇, 需要统计双关语中原词与目标词之间的发音替换规律, 我们采用 Jaech 等 [29] 给出的元音与元音以及辅音与辅音之间的替换矩阵. 经过音素替换后, 如果在词典中能找到该发音, 则作为原词的发音扩展词.

3.1.4 语法特征集

本文采用以下 5 个特征作为谐音双关语判断的语法特征:

(1) 人名. 创建人名字典, 判断语句中的名词是否为人名, 英语双关语中存在很多人名, 尤其是 “Tom” 出现频率较高, 用于指代某个人.

(2) 全大写. 判断语句中是否有单词是全大写的, 大写单词一般表示强调或者缩写, 表达语句的重要含义, 大写的缩写词也容易产生语音扩展.

1) https://en.wikipedia.org/wiki/CMU_Pronouncing_Dictionary.

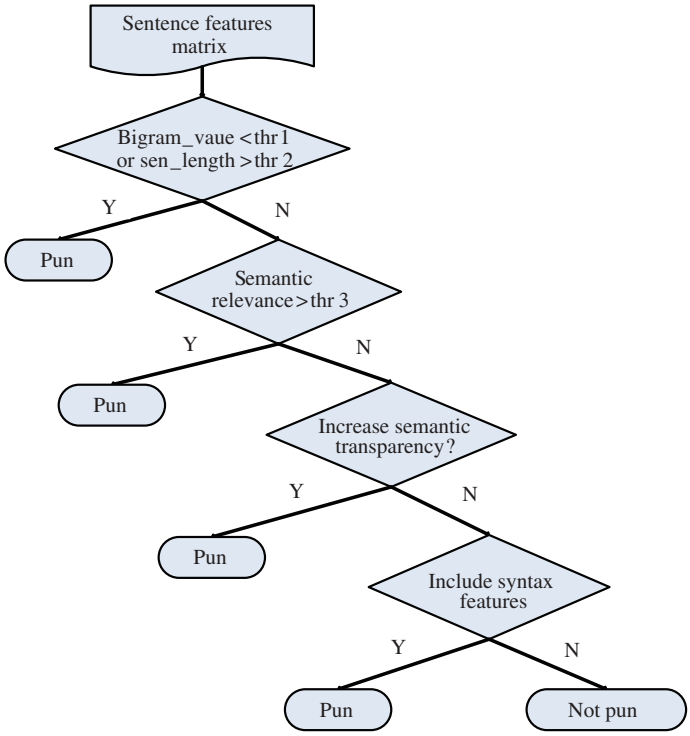

图 1 (网络版彩图) 二叉判定树识别双关语算法

Figure 1 (Color online) Puns recognition algrithm based BDT

(3) 过去时态. 判断一个语句是否包含过去时态. 一个句子可能有多个分句, 如果有一个分句使用了过去时态, 则判定整个语句包含过去时. 过去时态一般代表过去发生的动作或状态, 而双关语中有一大部分是对过去看见和听见的事情叙述时产生的.

(4) 词性特征. 在判断人名和大写单词时, 选择名词、动词、形容词和副词作为双关词的候选词.

(5) 位置特征. 一般双关句的双关词都出现在语句的后部, 所以完成语音扩展和语义计算时为语句后半部分的单词分配更高的权重.

前 3 个为布尔类型, 直接用于识别双关语, 后两个是为语音扩展、人名和时态服务的. 5 个特征都是从谐音双关语的语法特点出发, 找到用词、造句以及语法上的一些特征, 各种语法特征都可以帮助识别谐音双关语.

3.2 谐识别算法

本文利用语义透明度、语义相关度、语音扩展性和语法特征集 4 个维度的特征, 采用二叉判定树算法识别谐音双关语, 使用 SemEval2017 任务 7 的语料, 该任务主要是完成双关语的识别. 语料中只提供测试集, 没有训练集, 有监督分类方法不适用. 所以我们采用二叉判定树算法[2)]识别语句是否为双关语. 判定树中的每个非叶子结点都包含一个条件, 因此对应着一次识别. 每个叶子结点对应着种类, 因为本文是双关语的二义分类, 所以叶子结点分为两类 “双关句” 和 “非双关句”. 从上述 4 个维度出发, 每个维度对应一个非叶子结点, 识别过程如图 1.

2) https://en.wikipedia.org/wiki/Binary_decision_diagram.

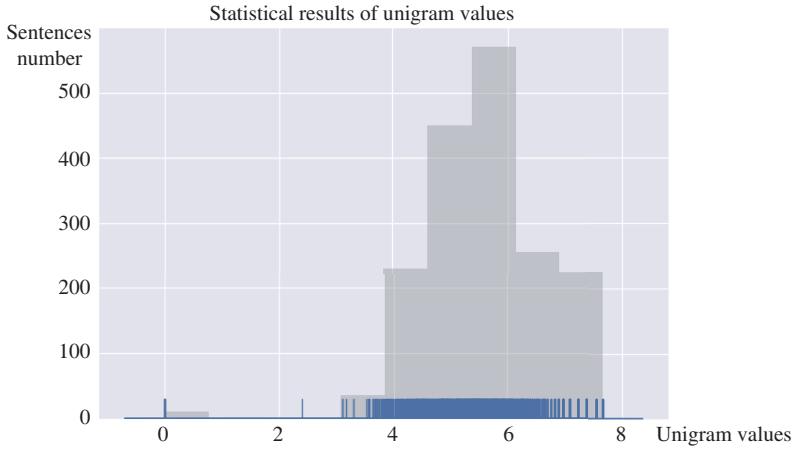

图 2 (网络版彩图) Unigram 值的语句密度

Figure 2 (Color online) Sentence density on unigram values

算法的输入是句子 S_i, 首先根据语句 S_i 中各单词计算特征矩阵 M_{ij}, i 为语句编号, j 代表上述 4 个维度中的所有特征, 然后将特征矩阵 M_{ij} 输入到二叉判定树中, 每个非叶子结点根据不同的阈值标准判断当前句是否为双关句, 如果是, 算法结束. 如果不是, 则继续到下个结点判断, 依次类推. 如果所有的特征都不满足, 说明不具备成为双关语的任何特征, 那么判定为非双关句.

从图 1 算法可以看出, 语句透明度和语义相关度需要根据阈值识别双关句, 较小的阈值会导致召回率较高, 从而错误地将所有的样本分为正例. 为了以更合理的方式计算阈值, 本文分别将 Unigram 值、语句长度和语义相关度 3 个特征矩阵, 用 K-Means 方法聚类. K-Means 算法是按照最邻近原则把待分类样本点分到各个簇, 质心 (centroid) 是指各个类别的中心位置, 质心的维数等同于单条数据的维数. 针对每个特征矩阵, 我们聚类为双关语和非双关语两类, 得到左右两个质心. 统一选择右质心作为阈值, 因为左质心是非双关语的中心, 右质心是双关语的中心, 这样可以有效防止召回率太高, 以致错误地将所有的样本分为正例. 如果选择双关语类别的质心作为阈值, 即使一个特征将正例分为负例, 后面还有特征可以矫正分类结果. 本文中 3 类特征都是一维的, 得到的质心是一个数值, 可以直接作为阈值. 图 2 以 Unigram 值为例, 统计了特征值与同一特征值下语句数量的关系. 横坐标为所有可能的 Unigram 取值, 纵坐标为语句数量. 3 个特征矩阵经过 K-Means 聚类后, 选择双关语的质心作为该特征的阈值, 因此 Unigram 特征、语句长度和语义相关度的阈值分别为: 6.5, 100 和 0.3. 根据上述阈值, 采用二叉判定树算法, 逐层增加双关语的召回率.

4 数据集及语言模型

4.1 数据集

本文的实验语料来自 SemEval2017 子任务 7 的谐音双关语 (heterographic pun) [3]. 子任务 7 包含两类语料, 语义双关语 (homographic puns) 和谐音双关语 (heterographic pun). 我们的实验选择谐音双关语的语料, 完成子任务一, 即识别语句是否为双关语. 该语料只有测试集, 没有训练集, 因而不能采用有监督的分类方法完成. 数据集采用 XML 格式, 每个语句用一个 text 元素表示, 为方便任务二

3) http://alt.qcri.org/semeval2017/task7.

和任务三, 语句中每个单词用 word 元素表示. 无论语句还是单词, 都有一个唯一的标识. 谐音双关语料中, 共有 1780 个语句, 其中双关语 1271 条, 非双关句 509 条. XML 文件具体格式如下:

```
<text id="het_5">
    <word id="het_5_1">Are</word>
    <word id="het_5_2">evil</word>
    <word id="het_5_3">wildebeests</word>
    <word id="het_5_4">bad</word>
    <word id="het_5_5">gnus</word>
    <word id="het_5_6">?</word>
</text>
```

4.2 评价指标

本文采用准确率、召回率、精确率和 $F1$ 值 4 项评估双关语的识别效果. 如果把正例预测为正例的个数记作 TP, 负例预测为正例的个数记作 FP, 正例预测为负例的个数记作 FN, 负例预测为负例的个数记作 TN. 那么

Precision = (TP + TN) ÷ (TP + FP + TN + FN),

Recall = TP ÷ (TP + FN),

Accurary = TP ÷ (TP + FP),

F1 = 2 × (Accurary × Recall) ÷ (Accurary + Recall).

4.3 语言模型

计算语义透明度主要是通过单词的 Unigram 值确定, 下面简单介绍计算 Unigram 用到的语言模型 (language model). 语言模型是自然语言处理中被广泛应用的统计模型, 目前主要采用 n 元语法模型 (ngram model). 这种模型简单, 相关的平滑技术也比较成熟. 语言模型主要是构建一个字符串 s 的概率分布 $p(s)$, 反应一个字符串 s 出现的频率. 假设 s 由单词 $w_1, w_2, w_3, \ldots, w_n$ 组成, $P(s)$ 的公式如下:

$$P(w_1, w_2, w_3, \ldots, w_n) = \prod_{i=1}^{n} p(w_i|w_1, \ldots, w_{i-1}). \tag{3}$$

上述公式中如果句子太长, 则概率会接近 0, 由于参数过多, 而不能正确训练. 所以实际应用中多采用 n 元语法, $n = 1$ 时, 即一元文法 (unigram), 认为每个词都是孤立存在的. $n = 2$ 时, 认为第 i 个词 w_i 仅与它前面的一个词相关, 成为二元文法 (bigram), 依次类推. 下面以二元文法为例, 给出 $P(s)$ 的计算方法, 如下:

$$P(w_1, w_2, w_3, \ldots, w_n) = \prod_{i=1}^{n} p(w_i|w_{i-1}). \tag{4}$$

本文集成 KenLM Toolkit [38] 工具包, 训练 ngram 语言模型. 其中训练语料来自布朗数据集的新闻语料 (newswire sections of the Brown corpus) [39], 语料规模为 6G.

5 实验结果

数据集中谐音双关语共 1780 条数据, 其中双关语 1271 条, 非双关语 509 条. 本文共选择 4 个维度的 9 个特征, 实验结果如表 1 和 2 所示.

表 1 特征迭加的双关语识别效果

Table 1 The pun recognition results of superimposed all features

Features	Precision (%)	Recall (%)	Accuracy (%)	$F1$ (%)
Semantic transparency	46.57	28.80	88.83	43.49
Semantic transparency + semantic relevance	48.88	33.02	89.85	47.22
Semantic transparency + semantic relevance + phonetic expansibility	62.87	57.12	86.22	68.72
Semantic transparency + semantic relevance + phonetic expansibility + syntax features	78.48	82.93	86.39	84.62
The number one of 2017SemEval [30]	78.37	81.90	87.04	84.39

表 2 分别使用各维度特征的识别效果

Table 2 The recognition results of each dimension features

Features	Precision (%)	Recall (%)	Accuracy (%)	$F1$ (%)
Semantic relevance	32.02	4.88	98.41	9.30
Semantic transparency	46.57	28.80	88.83	43.49
Phonetic expansibility	52.58	39.89	86.37	54.57
Syntax features	69.61	63.49	91.29	74.90

从表 1 中可以看出 4 个维度的特征是逐个添加到系统中的, 单纯的语句透明度虽然准确率较高, 但召回率较低. 加入语义相关度后, 召回率和 $F1$ 值均提高了 3% 左右. 再加入语音扩展性后, $F1$ 值大幅度提高了 21%, 这是因为本次任务是识别谐音双关语, 语句中是否存在语音相同或相似的扩展词是双关语识别的一个重要标志, 如果找到语义上关联的语音扩展词, 那么这个句子就更可能是谐音双关语. 最后加入多个语法特征后, 在准确率没有下降的前提下, $F1$ 值又提高 15.5%. 这说明谐音双关语在语法和语用上有一定的特点, 例如习惯使用过去式等. 4 个维度的特征都加入后, 分类的 $F1$ 值达到 84.62%, 比 SemEval 评测公布结果的第一名高出 0.3%, 说明以上 4 个维度的特征在谐音双关语识别中有较好的效果. 为了定量分析每个特征的作用, 我们分别使用每个特征在数据集中分类, 结果如表 2.

从表 2 中可以看出, 4 个维度的精确率都比较高, 主要是召回率差别较大, 语义相关度的召回率最低, 但精确率最好, 达到 98.41%. 召回率较低是因为只有部分谐音双关语具有语义关联性, 精确率较高说明具有这个维度特征的语句基本都是双关语. 召回率最高的是语法特征集, 达到 63.49%, 是语义相关度的 13 倍. 该维度中包含多个特征, 每个特征都贡献了部分召回率. 语句透明度和语音扩展性召回率相差 10% 左右, 精确率基本一致. 从实验结果可以看出 4 个维度的特征, 精确率都比较高, 如果单一特征的精确率较低, 二叉判定树的方法会把语句错分为双关句, 没有修正的机会. 而精确率较高, 保证每次划分大部分是正确的. 单个特征的召回率都比较低, 比较适合用二叉判定树的方法, 使多个特征结合, 每个分支结点都能增加部分召回率, 多个特征叠加, 使最终的召回率达到合理水平.

表 2 中数据也说明语法特征集的精确率和召回率都比较高, 为细化语法特征集中各个特征的作用, 我们只使用语法特征集中的 3 个特征识别双关语, 结果如表 3 所示. 该表中给出语法特征集中 3 个特征逐个用于识别双关语的效果, 其中过去时态特征召回率和精确率都较高, 所以 $F1$ 值最高, 而人名特征具有最高的精确率. 可见, 时态特征是双关语识别的一个重要特性.

表 3　3 个语法特征的识别效果

Table 3　The recognition results of three syntactical structure features

Features	Precision (%)	Recall (%)	Accuracy (%)	$F1$ (%)
Names	42.75	20.61	96.32	33.96
Capitalization	30.56	3.93	76.92	7.49
Tense	67.08	58.22	93.03	71.64

除了 K-Means 聚类方法获取阈值, 本文也尝试使用其他阈值, 以 Unigram 值为例, 分别使用 4.5, 5.5, 6.5 和 7.5 为阈值, $F1$ 值依次为 86.09%, 86.61%, 84.62% 和 84.07%. 阈值的变化对实验结果有一定的影响, 大约有 2% 左右的波动. 此外, 本文也尝试使用了有监督的机器学习方法, 利用支撑向量机为基础的分类模型, 使用上述无监督的特征进行双关语分类, 十则交叉验证下的分 $F1$ 值为 73.35%. 比无监督方法下降了 10% 左右, 这可能与分类特征数量较少有关.

6　结论及下一步工作

本文提出一种包含 4 个维度、9 个特征的二叉判定树算法, 识别谐音双关语. 4 个维度的特征对双关语的识别都有较好的效果, 其中语音扩展性和语法特征集效果最好, 大幅度提高了算法的召回率, 识别出了更多谐音双关语, 这也符合我们谐音双关语中语音作用较大的猜想. 二叉判定树算法将多个特征融合, 采用分治的方法, 弥补了单个特征召回率较低的问题. 针对分支结点中关键阈值的选择, 本文尝试了将特征数据 K-Means 聚类的方法, 选取双关语类的质心作为阈值, 这样做虽然在单特征下损失了部分召回率, 但保证了精确率的水平. 而召回率可以通过多特征融合的方法弥补. 语音扩展性虽然在双关语识别中有较高的精确率, 但是因为扩展策略的限制, 能够找到语音扩展词的语句数量只有一半. 下一步工作考虑继续改进语音扩展的策略, 使更多的语句找到语音扩展词. 另外, 还要进一步验证上述特征迁移到其他语料中是否有较好的效果.

参考文献

1　Tristan M. Towards the automatic detection and identification of English puns. Eur J Humour Res, 2016, 1: 59–75

2　Taylor J M. Ontology-based view of natural language meaning: the case of humor detection. J Ambient Intell Human Comput, 2010, 1: 221–234

3　Taylor J M. Computational detection of humor: a dream or a nightmare? the ontological semantics approach. In: Proceedings of International Joint Conferences on Web Intelligence and Intelligent Agent Technology, Milano, 2009. 429–432

4　Antonio R, Paolo R, Davide B. From humor recognition to irony detection: the figurative language of social media. Data Knowl Eng, 2012, 74: 1–12

5　Yishay R. Automatic humor classification on twitter. In: Proceedings of the NAACL HLT, Montréal, 2012. 66–70

6　Dario B, Pascale F. Deep learning of audio and language features for humor prediction. In: Proceedings of International Conference on Language Resources and Evaluation, Portoroz, 2016. 496–501

7　Dario B, Pascale F. Predicting humor response in dialogues from TV sitcoms. In: Proceedings of ICASSP, Shanghai, 2016. 5780–5784

8　Dario B, Pascale F. A long short-term memory framework for predicting humor in dialogues. In: Proceedings of NAACL HLT, San diego, 2016. 130–135

9　Zhang R X, Liu N S. Recognizing humor on twitter. In: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, Shanghai, 2014. 889–898

10 Reyes A, Buscaldi D, Rosso P. An analysis of the impact of ambiguity on automatic humour recognition. Lecture Notes Comput Sci, 2009, 5729: 162–169
11 Buscaldi D, Rosso P. Some experiments in humour recognition using the italian wikiquote collection. In: Proceedings of International Workshop on Fuzzy Logic and Applications: Applications of Fuzzy Sets Theory. Berlin: Springer, 2007. 464–468
12 Castro S, Cubero M, Garat D, et al. Is this a Joke? detecting humor in spanish tweets. In: Proceedings of IBERAMIA, San José, 2016. 139–150
13 Igor L, Hod L. Humor as circuits in semantic networks. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Baltimore, 2012. 150–155
14 Alessandro V, Hannu T. "Let everything turn well in your wife": generation of adult humor using lexical constraints. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofa, 2013. 243–248
15 Zhang D Y, Yang L, Zheng P Q, et al. Construction and application of affective metaphor corpus. Sci Sin Inform, 2015, 45: 1574–1587 [张冬瑜, 杨亮, 郑朴琪, 等. 情感隐喻语料库构建与应用. 中国科学: 信息科学, 2015, 45: 1574–1587]
16 Lin H F, Zhang D Y, Yang L, et al. Computational humor researches and applications. J Shandong Univ, 2016, 7: 1–10 [林鸿飞, 张冬瑜, 杨亮, 等. 幽默计算及其应用研究. 山东大学学报, 2016, 7: 1–10]
17 Ritchie G. Computational mechanisms for pun generation. In: Proceedings of the 10th European Workshop on Natural Language Generation, 2005. 8–10
18 Hempelmann C F. Computational humor: beyond the pun? the primer of humor research. Humor Res, 2008, 8: 333–360
19 Hong B A, Ong E. Automatically extracting word relationships as templates for pun generation. In: Proceedings of the NAACL HLT, Boulder, 2009. 24–31
20 Chloe K, Yuriy B. That's what she said: double entendre identification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, 2011. 89–94
21 Yang D Y, Lavie A, Dyer C, et al. Humor recognition and humor anchor extraction. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, Lisbon, 2015. 2367–2376
22 Tristan M, Iryna G. Automatic disambiguation of English puns. In: Proceedings of Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, 2015. 719–729
23 Arnold Z, Elizabeth Z. Imperfect puns, markedness, and phonological similarity: with fronds like these, who needs anemones. Folia Linguist, 1986, 20: 493–503
24 Wlodzimierz S. Metaphonology of English Paronomasic Puns. Bern: Peter Lang Pub Inc, 1991. 1–325
25 Kim B. Machine humour: an implemented model of puns. Dissertation for Ph.D. Degree. Edinburgh: University of Edinburgh, 1996
26 Robinson T. The British English example pronunciation dictionary. Cambridge: Cambridge University Press, 1996. 1–8
27 Hempelmann C F. Paronomasic puns: target recoverability towards automatic generation. Dissertation for Ph.D. Degree. West Lafayette: Purdue University, 2003
28 Kao J T, Roger L, Goodman N D. A Computational model of linguistic humor in puns. Cogn Sci, 2016, 5: 1270–1285
29 Jaech A, Koncel–Kedziorski R, Ostendorf M. Phonological pun uderstanding. In: Proceedings of NAACL HLT, San Diego, 2016. 654–663
30 Samuel D, Aniruddha G, Hanyang C. Idiom savantat semeval–2017 task7: detection and interpretation of English puns. In: Proceedings of the 11th International Workshop on Semantic Evaluations, Vancouver, 2017. 103–108
31 Dipankar D, Aniket P. JUCSE NLP at SemEval2017 Task7: Employing rules to detect and interpret English puns. In: Proceedings of the 11th International Workshop on Semantic Evaluations, Vancouver, 2017. 432–435
32 Diao Y F, Lin H F, Wu D. WECA: a wordnet-encoded collocation attention network for homographic pun. In: Proceedings of EMNLP, Brussels, 2018. 432–435
33 Wang C M, Peng D L. The roles of surface frequencies cumulative morpheme frequencies, and semantic transparencies in the processing of compound words. Acta Psychol Sin, 1999, 3: 266–273 [王春茂, 彭聃龄. 合成词加工中的词频、词素频率及语义透明度. 心理学报, 1999, 3: 266–273]
34 Cruise D A. Lexical Semantics. Cambridge: Cambridge University Press, 1991

35 Mok L W. Word-superiority effect as a function of semantic transparency of Chinese bimorphemic compound words. Lang Cogn Process, 2009, 24: 1039–1081
36 Pollatsek A, Hyönä J. The role of semantic transparency in the processing of Finnish compound words. Lang Cogn Process, 2005, 20: 261–290
37 Fellbaum C. WordNet: An Electronic Lexical Database. Cambridge: MIT Press, 1998. 1–423
38 Kenneth H, Ivan P, Jonathan H, et al. Scalable modified kneser-ney language model estimation. In: Proceedings of Annual Meeting of the Association for Computational Linguistics, Sofia, 2013. 690–696
39 Henry K, Nelson F. Computational Analysis of Present-day American English. Providence: Brown University Press, 1967. 1–424

Heterographic pun identification model based on multi-dimensional semantic relationships

Linhong XU[1*], Hongfei LIN[2], Ruihua QI[1] & Liang YANG[2]

1. *Software School, Dalian University of Foreign Languages, Dalian* 116044*, China*;
2. *Computer Department, Dalian University of Technology, Dalian* 116024*, China*
* Corresponding author. E-mail: qingniao1203@163.com

Abstract Identifying heterographic puns is an important branch of humor research, which has gradually developed into a new research area. This paper presents a heterographic pun identification mechanism based on feature sets in four dimensions, namely, semantic transparency, semantic relevance, phonetic expansibility, and syntax feature sets. The semantic transparency feature sets consist of the lexical item statistics and the character length; the syntax feature sets include names, capitalization, tense, part of speech, and location. Nine features of the above four dimensions are added to a binary decision tree to generate a threshold and complete a pun identification with the help of K-means clustering. Using the corpus of the SemEval2017 Task 7, the proposed method achieves satisfactory results, and its F1 value outscores the top one out of all participating teams. The experiment outlined in this paper proves that the taxonomic approach of the binary decision tree algorithm based on four dimensions is effective in identifying heterographic puns. The phonetic expansibility and the syntax feature sets are particularly effective among all other dimensions, which is consistent with our presumption that the phonetic feature plays a bigger role in identifying heterographic puns.

Keywords heterographic pun, sentiment analysis, binary decision tree, semantic feature set, cluster

Linhong XU was born in 1979. She is currently a lecturer at Dalian University of Foreign Languages. Her main research interests include nature language processing and sentiment analysis.

Hongfei LIN was born in 1962. He is currently a professor at Dalian University of Technology. Her main research interests include information retrieval and sentiment analysis.