基于潜在语义特性的语义双关语检测及双关词定位*

刁宇峰[1,2]，杨亮[1]，樊小超[1,3]，吴迪[1]，徐博[1]，许侃[1]，林鸿飞[1]

（1.大连理工大学 辽宁省 大连市 116024；2.内蒙古民族大学 内蒙古自治区 通辽市 028000；

3.新疆师范大学 新疆自治区 乌鲁木齐市 830054）

摘 要：语义双关语是幽默、笑话和喜剧等作品的来源之一，在人类写作的发展进程中具有重要的历史地位。由于语义双关语存在歧义难懂的特点，因此难以挖掘语义双关语的潜在语义信息，故目前语义双关语的检测和双关词的定位是自然语言处理任务中的一项困难和挑战。本文在语义双关语的理论基础上，挖掘了一系列的潜在语义特性，并构建了对应每个特性的特征集，用以检测语义双关语；同时从潜在语义特性出发，本文提出了一种基于词向量和同义词融合的语义相似度匹配算法实现双关词的定位。在SemEval 2017 Task7和Pun of the Day数据集上均取得了较好的实验结果，验证了本文所提出的检测算法和定位算法。

关键词：语义双关句；潜在语义特性；双关词定位；词向量；同义词;

中图分类号：TP391　　**文献标识码**：A

Homographic Puns Detection and Puns Location Based on Latent Semantic Characteristics

Abstract: Homographic puns, being a common source of humor in jokes and other comedic word, which have a long history in human writing. However, due to the ambiguity of homographic puns, it is hard to mine the latent semantic information. Therefore, there are still many difficulties and challenges in detection and location of homographic puns. Base on the theory, we design a series of latent semantic characteristics and corresponding features for each characteristic to detect homographic puns. Then, a semantic similarity matching algorithm is proposed to locate puns based on the fusion of Word Embedding and Sysnet. Experiment results on SemEval2017 task7 and Pun of the Day show that our methods are effective for homographic puns detection and location.

Keywords: Latent Semantic Characteristics; Homographic Puns Detection; Puns Location; Word Embedding; Sysnet;

1 引言

双关语是一种文字游戏，利用一词多义或者语音相似来达到多个含义的一种修辞方式[1]。在文学、演讲和广告语中，双关语也是标准的修辞手段。例如，莎士比亚因为他的双关语而闻名世界[2]，甚至在非喜剧作品中也广泛存在。众所周知，双关语作为一个广泛研究的有趣课题，能够洞察文字游戏和双重含义的本质性质。

双关语分类任务在 NLP 领域中有重要的意义。例如，Redfern[3]将双关语划分为语义双关语和谐音双关语，前者主要解决同义词的问题，后者主要解决同音词的问题。这两种双关语都有其自身的特点，不能用同一种模式来区分两种类型的双关语。因其常用性和数据集的原因，本文的研究主要集中于对语义双关语的研究。然而，目前对语义双关语的工作

* **收稿日期**：2018-06-10　　**定稿日期**：2018-07-25

基金项目：本课题得到国家自然科学基金资助项目（编号：61632011，61572102，61702080，61602078），中央高校基本科研业务费专项资金资助（DUT18ZD102，DUT17RC(3)016）

作者简介：刁宇峰（1987—），博士研究生，主要研究领域为文本情感计算。杨亮（1986—），博士，讲师，主要研究方向为文本情感计算；林鸿飞(1962—)， 博士，教授，主要研究领域为文本挖掘、情感计算和信息检索

中未从双关语理论的基础上进行系统的推导和解释。

本文的贡献主要有三点：第一，在双关语的理论基础上，针对语义双关语，本文挖掘出不一致、模糊、情感因素和语言学四种潜在语义特性，并设计每个结构下的特征集，提出一种有效的语义双关语检测模型；第二，在语义双关语潜在语义特性的基础之上，考虑到低维分布语义空间和同义词信息，本文提出一种基于词向量和同义词融合的语义相似度算法，能够有效的定位双关词；最后，在 SemEval2017 task7 和 Pun of the Day 两个数据集上，本文提出的方法在语义双关语检测和双关词定位两个任务上均取得了较好的实验性能。

2 相关工作

双关语自古以来就在修辞学和文学批判等方面中被广泛使用和讨论，近年来日益成为一个值得研究的课题。但是，在计算语言学和自然语言处理领域中类似的研究工作并不多[3]。在本节中，主要回顾了与双关语相关的前人工作。

对于语义双关语的检测方面，Kao 等人[4]提出了一种双关语中幽默的计算模型，主要从模糊性和特殊性两个维度检测语义双关句。Miller 和 Gurevych[5]提出了针对语义双关语的多个含义进行词义消歧的语义双关语识别算法。Huang 等人[6]介绍了一种新的框架，主要考虑句子中的位置信息作为检测语义双关句的重要指标。然而，上述的语义双关语检测工作没有从理论的角度出发，对语义双关语的本质进行系统的推导和合理的解释。

对于双关词的定位方面，Doogan 等人提出的 Idiom Savant 系统[7]主要基于 n-gram 和词向量，计算关联度和候选双关词的得分用以定位双关词。Vechtomova[8]等人引入互信息等传统特征，运用排序学习算法得到双关词。Indurthi 和 Oota[9]提出的 Fermi 系统主要计算句子中任意词对的同义词之间的相似度。然而，现有的双关词定位任务没有从双关的理论和本质出发，未充分考虑低维稠密语义空间和一词多义的联系。

双关语在幽默中也有很广泛的应用。Taylor 和 Mazlack[10]提出了一种基于固定句法上下文的 n-gram 识别算法来说明双关语在英文笑话中的幽默效果。林鸿飞等人[11]详细阐述了幽默的多种基本理论和实际应用，对于语义幽默的理解也给出了相应的讨论。但目前双关语的识别还处于起步阶段，未来还有很大的发展空间。尽管双关语经常在许多场合中使用，但由于歧义性和复杂性，现有的成果不能很好的进行处理和分析。

3 语义双关语的潜在语义特性及其特征

本文将语义双关语的检测归结为一个传统的文本分类问题。首先，本文根据语义双关语的相关理论，从四个方面制定了语义双关语的潜在语义特性，分别是：（1）不一致特性；（2）模糊特性；（3）情感因素特性；（4）语言学特性。然后，针对每个潜在语义特性，本文设计了一系列的特征来有效的检测语义双关句。

3.1 不一致特性

语言学家 Wales[12]指出，人们在说话时使用双关语，其主要目的在于使用不同意思以达到不同的奇妙的反应和效果。 不一致特性由语境中的冲突和语义上的不连贯导致。双关产生于两种或以上不协调不合适的状态下，一种复杂的组合方式。因此，不一致特性是双关语产生这种语言学现象的重要原因之一。

Eg1. Money doesn’t grow on tree. But it blossoms at our #branches#.

例 1. 钱不能长在树上。但是它可以在我们的#银行#上开花。

该句为双关语，[branches]为双关词，一般是[树枝]的意思，在该语境下意思为[银行]。

[Money doesn't grow on tree]和[But it blossoms at our branches]产生一种与语境的冲突，是一种不一致特性，从而达到语义双关语的效果。

针对不一致特性，首先本节定义了两类特征，间隔性和重复性，用以衡量一个句子中任意两个词对之间的语义距离。其次，不一致特性是一种语义上的不连贯，本节定义了语义连贯性来衡量语义双关语中的语义距离。本节使用 Word Embedding 和 N-gram 语言模型来计算不一致特性。Word Embedding 能够充分展现低维稠密空间下的语义信息，能够更好的表示语义双关语的潜在语义信息，这里使用 Word2Vec[1]工具。同时，N-gram 语言模型是一种能够发现语句中词与词之间关联性的规律信息，本文使用 KenLM 工具来训练 N-gram 语言模型，使用的外部语料来自开源的新闻语料(Brown 语料集)。

- 间隔性：衡量句子中任意两个词之间的最大语义距离，使用 Word2Vec 词向量计算词与词之间的余弦相似度。
- 重复性：衡量句子中任意两个词之间的最小语义距离。
- 语义连贯性：衡量句子的语义连贯性，使用 KenLM 工具对 n-gram 语言模型打分。

3.2 模糊特性

模糊特性是指句子中的一个词具有多个含义[13]，起到模糊歧义的作用，是很多语义双关语的关键成分[5]。双关语可以使一个词关系到不同的方面和角度，其双关词具有本身的字面意思，由于受到上下文语境的影响，一个词的多个可能的含义能够让人们产生不同的理解，以达到模糊的效果和突出的目的。

Eg2. Before he sold Christmas trees, he got himself #spruced# up.

例 2. 在他卖圣诞树之前，他将自己#打扮的整齐漂亮#。

该例句为语义双关语，[spruced]为双关词，该词有[云杉树]的意思，还有[使自己或事物看上去整齐和漂亮]的意思。

本节使用词汇资源 WordNet[2]来计算句子的模糊特性。双关语主要由句子中的内容词（名词、动词、形容词和副词）构成[1]，称之为候选双关词。本文使用 NLTK 词性标注工具来识别候选双关词，用以体现双关语中的模糊性[1]。本文结合词性信息来计算一个词的语义分散度 *PSD*，如公式(1)所示。

$$PSD = \frac{1}{P(|S_{pos}|,2)} \sum_{S_i, S_j \in S_{pos}} d(S_i, S_j) \qquad (1)$$

S_{pos} 表示句子中具有相同词性的词的同义词(sysnet)集合(s_0,s_1,...s_n)，$P(|S_{pos}|,2)$ 表示两个同义词集合中任意组合的个数，$d(s_i,s_j)$表示同义词 s_i 和 s_j 在 WordNet 之间的上位距离。

- 最远语义距离：根据句子中相应的词性集合，计算任一词义的最远语义距离。
- 平均语义距离：计算句子中任一词义的平均语义距离。
- 最近语义距离：计算句子中任一词义的最近语义距离。

[1] https://code.goole.com/p/word2vec
[2] http://www.nltk.org/howto/wordnet.html

3.3 情感因素特性

双关语能够产生委婉、含蓄以及幽默的语言效果。Van Mulken[14]发现，经常使用双关语以达到幽默的效果，会让观众对广告中的产品增加更正面的看法和积极的认同感。因此，语义双关语与情感因素有密切的关系。

Eg3. The two guys caught drinking battery acid will soon be #charged#.

例 3. 这两个喝电池酸液的家伙很快就会被#起诉#。

该句为语义双关语，[charged]为双关词，具有强烈的情感色彩。语义双关语中的双关词表现出一定的情感色彩，所关联的正面或者负面的倾向性都是人们在情感上的真实反映。

本节使用开放资源 SenticNet[15] 识别词级别的情感。该资源提供情感极性(polarity)和情感学(sentics)，可以充分的衡量词汇的主观性和情感信息。

- 情感极性：计算所有词的情感极性分值的总和、情感极性分值的平均值、情感极性分值绝对值的总和以及情感极性分值绝对值的平均值。
- 情感学：从情感学的总分、平均分、绝对值总分和绝对值平均分四个维度来表示情感学特征。

3.4 语言学特性

本节主要采用语言学特性进行分析，从词性、位置、句子长度和语义信息四个方面设计了有效的特征。

（1）词性信息

Eg4. Boyle said he was under too much #pressure#.

例 4. 波义耳说他承受的#压力#太大了。

该句为语义双关句，[pressure]是双关词，为名词。根据词性信息，可以影响语义双关语检测。具体的特征如下所示。

- 候选双关词数量：计算句子中的各类候选双关词的数量。
- 候选双关词占比：计算各类候选语义双关词在句子中的占比。

（2）位置信息

Miller 和 Turković 认为[1]，大多数双关语的双关词处在语句中的后半部分。因此，候选双关词的位置能够影响对双关语检测的判断。

Eg5. Here is how the track meet is going to #run#.

例 5. 这里是赛道如何#运行#的说明。

该句为语义双关句，其中 run 为双关词，且位置在句子的后端。特征如下所示。

- 最大位置：计算候选双关词集合在句子中的最大位置。
- 最小位置：计算候选双关词集合在句子中的最小位置。
- 平均位置：计算候选双关词集合在句子中的位置的平均值。

（3）句子长度

Barbieri 和 Saggion[16]提出句子的结构信息能够有效的衡量不同实体之间的差异。因此，句子的不同长度会影响语义双关句的检测。

- 句子长度：计算句子的句子长度。

- 长度之差：计算当前句子长度与句子平均长度的差值。

（4）语义信息

本节定义句子间的搭配关系是同词性候选双关词之间的语义关系，使用 WordNet 来计算候选双关词间的语义相似度。

Eg6. I used to be a banker but I lost #interest#.

例 7. 我过去是一个银行家但是我失去了#利益#。

该句为语义双关语，[interest]为双关词，具有[利益]和[兴趣]的含义，这里是[利益]的意思。本节通过计算[used]和[lost]，[banker]和[interest]之间的语义相似度来检测出该句是否为语义双关语。

同样，本节衡量候选词之间的反义关系。如词[fall]在 WordNet 中的反义词有：[ascent]，[rise]，[ascend]和[increase]。特征如下所示。

- 最大语义相似度：通过 WordNet 计算名词与名词、动词与动词、形容词与形容词、副词与副词之间的路径相似度。
- 是否存在反义词：计算句子中的候选双关词在 WordNet 中是否具有反义词。
- 最大反义词数量：计算候选双关词在 WordNet 中的反义词个数的最大值。
- 平均反义词数量：计算候选双关词在 WordNet 中的反义词个数的平均值。

4 基于词向量和同义词融合的双关词定位

每一条语义双关语都包含一个双关词，本节需要给出线索并定位到哪个词是双关词。本节将双关词定位归结为一个无监督问题，提出 LOCATION_PUN 相似度匹配算法，如表 1 所示。该算法的输入为每一条语义双关语，输出为具体的双关词。

表 1：LOCATION_PUN 算法

算法 1　LOCATION_PUN 算法
1、输入：每一条语义双关句
2、输出：双关词
3、初始化：MaxSimilarity = 0
4、步骤一：对输入的语义双关句进行词干化。
5、步骤二：去除标点和停用词。
6、步骤三：保留语义双关句的名词、动词、形容词和副词，构建候选双关词集合。
7、步骤四：去除在 WordNet 中少于两个同义词的词。
8、步骤五：去除句子中位置在前面的词。
9、if 候选双关词集合只有一个 then 返回(双关词)
10、for 候选双关词集合中的任一词 w_i do
11、　步骤六：定义词 w_i 对应的相似度得分 s_i
12、　for 候选双关词集合中的其他词 w_j do
13、　步骤七：在预训练的 GloVe 词向量中查找 w_i 和 w_j 对应的向量 e_i 和 e_j
14、　步骤八：在 WordNet 中查找 w_i 和 w_j 的同义词集合，分别取平均词向量 c_i 和 c_j
15、　步骤九：分别将上述两个向量进行拼接，$d_i=[e_i; c_i]$　$d_j=[e_j; c_j]$
16、　步骤十：计算 d_i 和 d_j 的余弦相似度，并加入到 s_i 中
17、　if MaxSimilarity <= s_i then MaxSimilarity = s_i，并记录词 w_i end if
18、　end for
19、end for
20、返回(双关词)

Eg7. Getting rid of your boat for another could cause a whole #raft# of problems.

例 7. 把你的#船#换成另一艘船可能会造成很多的问题。

该例句为语义双关语，其中[raft]为双关词，有[一批]和[船]的意思。

通过语义双关语的潜在语义特性可知，(1)语言学特性的位置信息：双关词通常出现在语义双关语的句尾；(2)语言学特性的词性信息：候选双关词通常为名词、动词、形容词和副词；(3)模糊特性：双关词在 WordNet 中有至少两个词义；(4)不一致特性的间隔性和重复性：双关词与非双关词之间在低维稠密空间的语义关联性不大。

对于候选双关词与其他词之间语义相似度的计算方式，本节使用词向量和 WordNet 两种方式。对于词向量，使用 Word2Vec 和 Glove，语义距离采用余弦距离和编辑距离；对于 WordNet，使用 WordNet 中的同义词集合，利用 Path Similarity 计算候选词的不同同义词的相似度。最后，从词向量和同义词出发，融合 GloVe 和 Sysnet 的方式计算语义相似度，最终定位双关词的位置。

5 实验与分析

本节首先介绍实验方面的设置，然后验证本文提出的潜在语义特性在双关语检测任务中的性能，最后验证 LOCATION_PUN 相似度匹配算法在双关词定位任务中的表现。

5.1 实验设置

本节首先分析实验使用的数据集，然后介绍具体的评价指标和基线方法，最后给出在模型训练过程中的实现细节。

(1) 数据集

SemEval2017[3] Task7：该任务主要检测和定位语义双关句，包括语义双关句和谐音双关句两部分。本文主要关注语义双关句，子任务一是检测语义双关句，每条文本至多含有一个双关表达；子任务二是定位双关词，每条语料均为语义双关句，需要线索定位哪个词为双关词。

Pun of the Day[4]：该网站的数据用于检测语义双关句，其正例来源于日常的用户，为了构建平衡的数据集进而获取合适的负例，该数据集从以下四个网站收集负例：AP News[5]、New York Times、Yahoo!Answer[6]和 Proverb。统计分析见表 2。

表 2：数据集 SemEval-2017 Task7 和 Pun of the Day 的统计信息

数据集	子任务	总数量	总词数
SemEval2017 Task7	子任务一(语义双关语检测)	2250	24499
SemEval2017 Task7	子任务二(语义双关语定位)	1607	18998

数据集	正例	负例	平均长度	正例平均长度	负例平均长度
Task7(子任务一，语义双关语)	1607	643	13.1	13.9	10.8
Pun of the Day	2403	2403	13.5	12.2	13.8

(2) 评价指标

对于语义双关语的检测任务，本文的评价指标与 SemEval2017 Task7 任务一的评价方

[3] SemEval2017 Task7: http://alt.qcri.org/semeval2017/task7.
[4] Pun of the Day: http://www.punoftheday.com/.
[5] AP News: http://hosted.ap.org/dynamic/fronts/HOME?SITE=AP.
[6] Yahoo!Answer: http://answers.yahoo.com/.

法一致，采用准确率、召回率和 F1 值指标。

对于语义双关词定位任务，本文与 SemEval2017 Task7 子任务二的评价指标一致，采用覆盖率、准确率、召回率和 F1 值指标。

(3) 基线方法

对于语义双关语的检测任务，本文设置了如下的基线方法。

- Bag of Words(BOW)：主要捕获句子中的词序关系信息，检测是否为语义双关句。
- Language Model(LM)：在统计学的基础上，通过句子中词的概率分布计算对应的双关概率值，不需要训练集和训练语料。
- AVGWord2Vec：根据潜在语义分布表示，将句子的任意词对应的词向量相加取平均值。
- HPCF：本文将提出的四个潜在语义特性定义为语义双关语核心特征（Homographic Puns Core Features，简称 HPCF）。
- AVGWord2Vec_HPCF：本文将 HPCF 和 AVGWord2Vec 结合在一起使用，性能已超过基线方法。

对于语义双关词的定位任务，本文设置了如下的基线方法。

- Idiom Savant：该方法采用 word2vec 计算候选词的得分，使用 WordNet 提供的 gloss vector 计算关联度，在 SemEval2017 task7 任务二评测中取得了第一名的成绩。
- UWaterloo：该方法引入互信息等特征，运用得分公式进行排序，从而得到双关词，在 SemEval2017 task7 中排名第二。
- Fermi：计算词与词的同义词之间的相似度，在 SemEval2017 task7 中排名第三。
- LOCATION_PUN：本文提出的基于词向量和同义词融合的语义相似度匹配算法，用于定位语义双关词，取得了最好的性能。

(4) 实验细节

对于语义双关句的检测任务，本文采用 5 倍交叉验证来进行实验，使用 60%的数据进行训练模型，使用 20%的数据进行调参，使用 20%的数据进行预测。训练 Word2Vec 词向量维度的语料来自 Wiki，分别对比了维度 100、200、300 的维度，最终选择 300 维。本文使用 GBDT 这个基于决策树的方法作为分类算法，与文献[17]一致。

对于双关词的定位任务，本文使用 GloVe[7]词向量，分别对比了维度 50、100、200 的维度，最终选择 100 维，使用 WordNet 提供的同义词集，本文对比了余弦距离和编辑距离两种相似度算法，最终选择了余弦距离算法。

5.2 语义双关语检测

本文将提出的基于潜在语义特性的检测方法与基线方法进行对比，具体结果见表 3。

表 3：语义双关语检测任务中不同方法对比结果

	SemEval2017 Task7 任务一			Pun of the Day		
	准确率	召回率	F1	准确率	召回率	F1
HPCF	0.808	0.938	0.861	0.702	0.807	0.767
Bag of Words(BOW)	0.832	0.847	0.806	0.732	0.663	0.685
Language Model(LM)	0.688	0.774	0.668	0.508	0.764	0.612
Word2Vec	0.803	0.800	0.756	0.903	0.899	0.900
BOW_HPCF	0.805	**0.954**	0.871	0.908	0.905	0.906
AVGWord2Vec_HPCF	**0.836**	0.944	**0.887**	**0.914**	**0.906**	**0.910**

(1) HPCF 主要包括不一致、模糊、情感因素和语言学共计四个潜在语义特性，在语义

[7] GloVe: https://nlp.stanford.edu/projects/glove/

双关语检测任务中，其性能优于 BOW 和 LM 方法。这充分证明了基于双关语理论提出的潜在语义特性是合理且有效的。HCPF 的结果高于 LM，表明潜在语义特性消除了领域之间的差异从而更精确的检测语义双关句。HCPF 的结果高于 BOW，表明潜在语义特性能够更合理的理解词在句子中出现的顺序。

(2) BOW_HPCF 是融合 BOW 和 HPCF 的算法，在两个数据集上的结果均高于 BOW 和 HPCF 方法。这个结果表明 BOW_HPCF 方法可以充分的表示潜在语义信息和句子中的词序顺序。但是 BOW_HPCF 的结果不如 AVGWord2Vec_HPCF，因为前者仅涉及了充足的潜在语义特性信息，但是没有考虑分布式语义信息。

(3) AVGWord2Vec_HPCF 在语义双关语检测任务中，在 Pun of the Day 数据集上，取得了 0.91 的最优 F1 值。原因在于该方法充分考虑到潜在语义特性和分布式低维稠密语义信息的关系。而在 SemEval2017 task7 的任务一中，除了 BOW_HPCF 取得了最高的召回率之外，其他结论几乎是一致的。从这些结果可以看出，本文提出的潜在语义特性能够深刻的理解语义双关语。

在 SemEval2017 task7 中最佳的系统 Fermi[9,18]，该方法同样将语义双关语检测看作一种有监督的分类问题，使用深度学习模型中的循环神经网络来训练分类器，结果达到 0.899 的 F1 值。因此，未来本文也将尝试使用现有的深度学习方法来解决这类问题。

本文针对提出的潜在语义特性进行展开实验，分析每个潜在语义特性对语义双关句检测的影响。对于上述的两个数据集，使用统一的分类器 GBDT 来验证上述不同特性的表现。为了公平性，本文使用统一的参数设置。实验结果如图 1 所示。

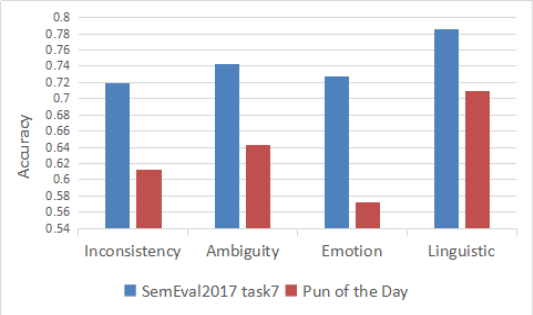

图 1：不同的潜在语义结构对语义双关语检测任务的贡献程度

(1) 根据实验结果，可以看出本文提出的潜在语义特性在两个数据集上取得了类似的性能。语言学特性在 SemEval2017 task7 任务一和 Pun of the Day 上均取得了最优的实验结果。原因在于语义双关语的检测与语言学特性中的位置信息、词性信息、搭配信息和反义信息具有相当密切的联系。

(2) 在两个数据集中，模糊特性取得了第二的效果，这表明大多数的语义双关语都有着良好的句式结构和多义性，具有难以理解和分析的特点。从图 1 中可以看出，在 Pun of the Day 数据集中，情感因素特性表现的性能最差，原因在于人类的情感表达，尤其是语义双关句中的情感信息是难以挖掘和分析的。

(3) 与 Pun of the Day 数据集的结论不同，在 SemEval2017 task7 任务一中，不一致特性取得了最差的结果，原因在于不协调、不和谐的含义使人们难以找到和理解丰富有用的信

息。潜在语义特性在两个数据集上有不同的表现结果，这表明语义双关句的潜在语义特性在不同数据集上会有不同的潜在语义表示信息。

5.3 语义双关词定位

本文选取 SemEval2017 task7 的任务二作为数据集，提出一种基于词向量和同义词融合的相似度匹配算法，即 LOCATION_PUN 算法来定位语义双关词，实验结果如表 4 所示。

由表 4 可知，本文的 LOCATION_PUN 算法在双关词定位任务的所有指标中均取得了最优的性能，其 F1 值超过评测的第一 Idiom Savant 近 3.4%。原因有两点：第一，本文提出的潜在语义特性对语义双关词定位任务是有效的，如语言学特性的位置、词性信息，不一致特性的间隔性和重复性，以及模糊特性；第二，在潜在语义特性的基础之上，本文从低维语义空间和传统词典提供的同义词两个角度入手，提出了词向量和同义词融合的方式，既考虑了词共现的分布式语义空间表示，又结合了 WordNet 提供的同义词信息。

表 4：语义双关词定位任务中不同方法的对比结果

	SemEval2017 task7 任务二			
方法	覆盖率	准确率	召回率	F1 值
Idiom Savant	0.9988	0.6636	0.6627	0.6631
UWaterloo	0.9994	0.6526	0.6521	0.6523
Fermi	**1.0000**	0.5215	0.5215	0.5215
LOCATION_PUN	**1.0000**	**0.6976**	**0.6976**	**0.6976**

下面，本文详细的对 LOCATION_PUN 算法进行分析，衡量维度有：语言学潜在语义特性（是否考虑位置信息和词性信息，Linguistics）、模糊特性（是否考虑同义词信息，Sysnet）和不一致特性（是否考虑 Word2Vec/GloVe 词向量），均使用余弦相似度算法进行对比，实验结果如表 5 所示。从表 5 中，可以得到以下的结论。

表 5：语义双关词定位任务中不同维度的对比结果

	SemEval2017 task7 任务二			
方法	覆盖率	准确率	召回率	F1 值
Word2Vec	**1.0000**	0.1350	0.1350	0.1350
Sysnet	**1.0000**	0.4142	0.4142	0.4142
Linguistics+Sysnet	**1.0000**	0.4798	0.4798	0.4798
Linguistics+Word2Vec	**1.0000**	0.6192	0.6192	0.6192
Linguistics+Word2Vec+Sysnet	**1.0000**	0.6733	0.6733	0.6733
Linguistics+GloVe	**1.0000**	0.6764	0.6764	0.6764
Linguistics+GloVe+Sysnet(LOCATION_PUN)	**1.0000**	**0.6976**	**0.6976**	**0.6976**

(1) Linguistics+Sysnet、Linguistics+Word2Vec 的结果要分别高于 Sysnet、Word2Vec，说明本文提出的语言学潜在语义特性的位置信息和词性信息可以帮助定位语义双关词，从而侧面表明了该特性的有效性。

(2) Linguistics+Word2Vec+Sysnet、Linguistics+GloVe+Sysnet 的结果均高于 Linguistics +Sysnet，说明不一致特性能够帮助定位双关词，同时低维分布式语义空间对双关词的定位也有很大的影响，且 GloVe 的结果均优于对应的 Word2Vec，说明了词共现信息的有效性。

(3)Linguistics+GloVe+Sysnet、Linguistics+Word2Vec+Sysnet 的结果对应高于 Linguistics +GloVe、语言学+Word2Vec，原因在于模糊特性提供的同义词信息能够合理的定位双关词。本文提出的方法（Linguistics+GloVe+Sysnet，即 LOCATION_PUN 算法）的结果最优，在

位置信息和词性信息的基础上，考虑了低维分布式语义空间和 Sysnet，充分融合了词向量和同义词的信息，能够合理高效的定位语义双关句中的双关词。

5 结论与未来工作

本文的研究工作旨在于检测语义双关语和定位双关词。基于双关语的理论基础之上，挖掘了四个潜在语义特性，针对每个特性设计了有效特征集，用以检测语义双关语。在双关词定位任务方面，本文从潜在语义特性出发，提出一种基于词向量和同义词融合的无监督语义相似度匹配算法。在两个数据集上得到的实验结果表明，本文提出的潜在语义特性具有足够的检测语义双关语的能力，能够准确的定位双关词。

在未来工作中，本文将尝试探索更高效的特征来体现语义双关的特点，并结合深度学习算法用以检测语义双关句，运用无监督方法、弱监督方法来实现定位双关词的工作。

参 考 文 献

[1] Miller T, Turković M. Towards the automatic detection and identification of English puns[J]. The European Journal of Humour Research, 2016, 4(1): 59-75.

[2] Tanaka K. The pun in advertising: A pragmatic approach[J]. Lingua, 1992, 87(1-2): 91-102.

[3] Redfern W. Puns[J]. The Scriblerian and the Kit-Cats, 1987, 19(2): 204.

[4] Kao J T, Levy R, Goodman N D. A computational model of linguistic humor in puns[J]. Cognitive science, 2016, 40(5): 1270-1285.

[5] Miller T, Gurevych I. Automatic disambiguation of English puns[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015, 1: 719-729.

[6] Huang Y H, Huang H H, Chen H H. Identification of homographic pun location for pun understanding[C]//Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017: 797-798.

[7] Doogan S, Ghosh A, Chen H, et al. Idiom Savant at Semeval-2017 Task 7: Detection and Interpretation of English Puns[C]//Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). 2017: 103-108.

[8] Vechtomova O. UWaterloo at SemEval-2017 Task 7: Locating the Pun Using Syntactic Characteristics and Corpus-based Metrics[C]//Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). 2017: 421-425.

[9] Indurthi V, Oota S R. Fermi at SemEval-2017 Task 7: Detection and Interpretation of Homographic puns in English Language[C]//Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). 2017: 457-460.

[10] Taylor J M, Mazlack L J. Computationally recognizing wordplay in jokes[C]//Proceedings of the Annual Meeting of the Cognitive Science Society. 2004, 26(26).

[11] 林鸿飞，张冬瑜，杨亮，徐博. 幽默计算及其应用研究[J].山东大学学报，2016,7(51):1-10.

[12] Wales K. A dictionary of stylistics[M]. Routledge, 2014.

[13] Bekinschtein T A, Davis M H, Rodd J M, et al. Why clowns taste funny: the relationship between humor

and semantic ambiguity[J]. Journal of Neuroscience, 2011, 31(26): 9665-9671.

[14] Van Mulken M, Van Enschot-van Dijk R, Hoeken H. Puns, relevance and appreciation in advertisements[J]. Journal of pragmatics, 2005, 37(5): 707-721.

[15] Cambria E, Hussain A. Sentic computing: Techniques, tools, and applications[M]. Springer Science & Business Media, 2012.

[16] Barbieri F, Saggion H. Modelling Irony in Twitter: Feature Analysis and Evaluation[C]. LREC. 2014: 4258-4264.

[17] Zhang R, Liu N. Recognizing Humor on Twitter[C]. ACM International Conference on Conference on Information and Knowledge Management. ACM, 2014:889-898.

[18] Miller T, Hempelmann C F, Gurevych I. SemEval-2017 Task 7: Detection and interpretation of English puns[C]. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Vancouver, BC. 2017.