文章编号：1003-0077(2016)06-0067-08

# 限定领域口语对话系统中的商品属性抽取

叶大枢，黄沛杰，邓振鹏，黄 强

（华南农业大学 数学与信息学院，广东 广州 510642）

**摘 要**：按功能或问题域划分，商品属性抽取(product feature mining)在限定领域的对话系统中属于口语语言理解(spoken language understanding，SLU)的范畴。商品属性抽取任务只关注自然文本中描述商品属性的特定部分，它是细粒度观点抽取(fine-grained opinion mining)的一个重要的子任务。现有的商品属性抽取技术主要建立在商品的评论语料上，该文以手机导购对话系统为背景，将商品属性抽取应用到整个对话过程中，增强对话系统应答的针对性。使用基于 CBOW (continuous bag of words)语言模型的 word2vector(W2V)对词汇的语义层面建模，提出一个针对口语对话的指数型变长静态窗口特征表达框架，捕捉不同距离词语组合的重要特征，使用卷积神经网络(convolutional neural network，CNN)结合词汇的语义和上下文层面对口语对话语料中的商品属性进行抽取。词嵌入模型给出了当前词和所给定的属性类别是否存在相关性的证据，而所提出的特征表达框架则是为了解决一词多义的问题。实验结果表明，该方法取得了优于研究进展中方法的商品属性识别效果。

**关键词**：商品属性抽取；词向量；卷积神经网络；特征表达；口语对话系统

**中图分类号**：TP391 **文献标识码**：A

## Product Feature Mining in Restricted Domain Spoken Dialogue System

YE Dashu, HUANG Peijie, DENG Zhenpeng, HUANG Qiang

(College of Mathematic and Informatics, South China Agricultural University, Guangzhou, Guangdong 510642, China)

**Abstract**: This paper applies the product feature mining on a dialogue system of a mobile phone recommendation assistant, enhancing the focus of the system during the interaction. CBOW (continuous bag of words) language model is used to represent the sematic clue. A feature framework with exponential elongate static window is introduced to capture the import features among the interactions between words of variant distance. We finally utilize convolutional neural network (CNN) to perform product feature mining task. The word embedding representing sematic clue gives the relation between current word and the product feature, while the feature framework can alleviate the word ambiguity. The experiment shows that our model outperforms the state-of the act methods on product feature mining.

**Key words**: product feature mining; word2vector; CNN; feature representation; spoken dialogue system

## 1 引言

商品属性抽取(product feature mining)是指从顾客评论语料中提取有关商品属性描述的特定部分[1]。例如，从某品牌手机的评论“手机壳超级漂亮，套上特有感觉，喜欢！”可以看出买家对“手机”这件商品的“手机壳”属性非常满意。商家就可以通过买家对商品属性的侧重点，对产品本身或者组合配件做相对应的调整满足买家们的不同需求。语料中的商品属性可分为显性属性和隐形属性[2]。显性属性就是直接体现在语料中的属性，如上面例子中提到的“手机壳”属性；隐性属性则是语料中通过描述属性的外在形态所暗指的商品属性，如“这款手机很不错，就是要一天充一次电”的评论中，“一天充一次电”实际上指的就是“电池容量”或者 “手机能耗”方面的属性。

商品的隐性属性，很大程度上是通过对整句话

收稿日期：2016-09-27 定稿日期：2016-10-11

基金项目：国家自然科学基金(71472068)；广东省大学生科技创新培育专项项目(pdjh2016b0087)

的完全理解，结合领域相关知识推断所得。就目前来看，信息抽取技术还不足以完全掌握这方面的技术，所以本文所介绍的商品属性抽取，均指商品的显性属性抽取。由于属性抽取结果形式的结构化绝大多数以(名，值)二元组的形式表现，在以非自然文本形式如表格或图表来描述商品的时候，则给买家一种直接、简洁、有效的体验。因此商品属性抽取的主要用途有：(1)对某件商品给出不同买家的感受或体验，能为新的买家对商品提供感性的认识。对普通买家而言“性价比高”会比“1 000 块”、“8 核 2.3Ghz”等具体的硬件参数标签的表现力强；(2)对商品做模糊的检索。结构化的另一个好处就是，属性能为计算机所认识，买家可以通过自身对商品的功能需求(如“运行流畅”)而不是具体的硬件参数，检索商品。

在口语对话系统方面，对话信息以槽(slot)的形式保存下来，Chen 等[3]提出一种基于槽间关系的无监督自然语言处理模块，将基于槽信息的表达结构构建语义(semantic)层面上的关系图和基于词的在词汇(lexical)层面上的关系图结合起来，使得槽所记录的信息更为完备，提升对话系统中口语语言理解(spoken language understanding，SLU)模块对自动语音识别(automatic speech recognition，ASR)模块转换的文本的阅读能力。本文主要研究的属性抽取属于 SLU 的任务范畴。SLU 功能模块将文本中相关属性给出多个识别结果并以[名，值]的格式输出，在本文所应用的手机导购对话系统中，当用户输入为“我要三星品牌的”，系统的 SLU 模块将抽取其中的“三星”属性，并处理成结构化的[品牌，三星]表达形式。

本文将商品属性抽取技术应用于对话系统中，首先从微博、新闻等和口语对话语料存在一定相似性的外部语料与本文所收集到的对话语料一起，无监督的为每一个词训练得到一个向量表示。采用指数型变长静态窗口的特征表达框架，结合卷积神经网络(convolutional neural network，CNN)对候选词进行商品属性分类。相比已有研究，本文的主要贡献包括：

(1) 给出一种针对文本语料中数字字符串的预处理方法，能在降低数字之间的差异性的同时，又能在一定程度上保留数值的有序性。和情感分析等任务不同，数字字符串或者数字和字母的混合字符串，如手机商品的“价格”、“手机型号”等属性。而不对价格做处理将引起数据的稀疏性问题。

(2) 针对口语对话语料的特点，给出一个指数型变长静态窗口的特征表达框架，能有效捕捉不同距离范围内的词对商品属性词的影响。

(3) 经过对中文手机导购领域的对话语料的预处理和观察，针对热门的手机商品属性类别，对口语对话上下文建模，以满足对话系统在一般情况下的应答能力。

本文的后续部分安排如下：第二节介绍相关工作；第三节介绍本文使用的商品属性抽取框架；第四节给出对比实验结果；最后为总结本文工作并展望未来工作。

## 2 相关工作

Hu 和 Liu[1]基于模板实例化[4-5]和文本抽取[6-7]的研究，经过对商品评论语料的细致观察，做出了商品属性总是与对应的评价一起出现的假设前提，首次提出一种关联的规则，将商品属性和商品评论相互关联，对评论语料中的商品属性进行提取。Popescu 和 Etzioni[8]设计一些基于语法的模板用以识别文本中商品属性并使用 PMI(pointwise mutual information)对提取到的商品属性去噪。Zhuang 等[9]从电影语料中发掘各种不同的语法规则用以提取电影相关属性。Qiu 等[10]也在属性和对应的观点词总是一起出现[1]的假设前提上，提出八条语法规则抽取文本中商品属性和情感词。Zhang 等[11]对 Qiu 等[10]所提出的语法规则数目进一步扩充，并且使用了 HITSalgorithm[12]对扩展的候选词集进行排序。Wu 等[13]提出短语层面上的语法分析技术用以抽取观点和商品属性。Zhao 等[14]提出两条启发式规则对句子的语法树作简化和利用语法树的子树做特征，企图缓解数据的稀疏性，引入卷积树核(convolution tree kernel)对候选词作商品属性分类。Xu 等[15]尝试着以一种考虑词和相关语法共现层面的半监督分类器(semi-supervised classifier)抽取不常见商品属性，仍旧是无法解决稀疏性的问题[16]。Xu 等[16]用词向量模型对当前词语的词汇层面建模，用卷积神经网络对词汇的上下文层面建模，最后使用标签传播算法将两者结合，半监督地对当前词的是否为商品属性词做判定，使得商品属性的识别效果有较为显著的提升。

口语对话的语料和商品的评论语料的表现形式存在一定的差异性，使得构建在商品评论语料中的属性抽取技术难以直接应用到口语对话系统中，存

在相当程度上的挑战性，其差异主要体现在：

（1）商品的评论能以一句话表达用户所关注的商品属性的状态，也能在一句话中表达商品不同方面的属性。由于对话系统中显式和隐式的确认机制，对话中的很多时候则需要几轮（一轮对话是一次问答）对话来确定用户所描述属性。

（2）系统的应答语料，主要来自于自然语言生成（natural language generation，NLG）模块带有一定程度的结构化信息，使得某些系统应答大量重复出现，例如，开场白“您好，欢迎光临”，使得基于词频统计的方法[15]失去作用。

（3）对商品的评论常常以“商品属性”、“观点”一起出现良好结构形式，而对话中用户直接向系统表达所关注的属性。

（4）一句用户输入（utterance）比商品的评论长度更短、信息更为不全面、且口语化程度更高。因此基于语法结构的商品属性抽取方法，都受到数据的稀疏性等问题的困扰，很快遇到性能瓶颈。

本文则从外部语料、数字平滑机制、以及引入词向量的特征表达这三方面缓解数据稀疏问题。实验结果表明，本文提出的方法比基于语法结构的方法的抽取效果有明显的提升。

## 3　口语对话系统中商品属性抽取方法

### 3.1　商品属性抽取的技术流程

图 1 是本文提出的属性抽取的技术框架。

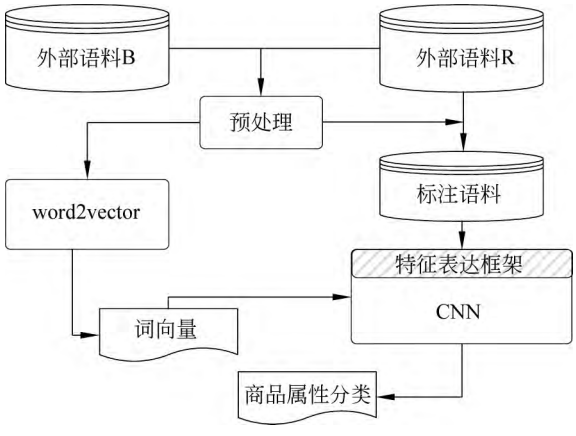


图 1　商品属性抽取技术框架

本文将通过词义和上下文层面对给定的词汇建模，语义信息主要通过词向量来体现，而上下文信息则是通过指数型变长静态窗口获取综合考虑长距离和短距离的特征表达，通过 CNN 结合上述两个层面的信息对其进行细粒度的商品属性类别的判定。一个词会有多种意思，如“三星”既可以指的是评价等级，也可以表示手机品牌“三星”，所以要结合上下文语境，对其商品属性的归属性做判断。

在图 1 的流程中，首先将收集到背景语料和对话语料经过数据预处理和分词的操作合并在一起作为 word2vec 模型的输入。利用 word2vector 模型无监督的给每一个词语训练得到一个向量表示。对对话语料进行标注，形成训练样本和测试样本，最后将训练得到的词向量与已标注的语料相结合，构造训练集作为 CNN 的输入，经过基于指数型变长窗口的特征表达框架，利用 softmax[17]方法将多分类问题转化为二分类问题，从而得到每个词的属性归属度。

### 3.2　语料库的构建

#### 3.2.1　外部语料知识库的构建

外部语料也称为背景语料，用以学习词汇词义层面的信息，本次实验收集了 100 万条微博和两万篇网络新闻作为训练词义的知识库，使用 word2vector 模型无监督的给每一个词建立一个向量表示，在向量空间中用余弦相似度衡量两个词之间的相似性。将收集到的外部语料去除标点符号等，再经过 jieba 分词①。使用微博作为主要的外部语料是出于微博也是一种口语化短文本的考虑，与口语对话系统的交互存在某些方面的共性。

#### 3.2.2　口语对话语料库的构建

将所收集到的领域内部语料共有有效对话 1 533 段，8 678 个词，同样使用 jieba 分词进行分词处理。在情感分析的任务中，语料中的数字对整句话的情感取向几乎没有影响，所以一般情况下，不将其纳入句子表达特征的考虑范围内。在细粒度的商品属性抽取任务中，当数字以价格或型号的形态出现时，则变成任务的抽取目标，有着不可忽视的地位。这里给出了对数字或数字和字母的混合字符串，做了两个处理：

（1）用 $num 表示语料中出现的纯数字字符串，用 $numeng 表示数字和英文字母混合形态的字符串。

（2）在 $num 和 $numeng 后面紧跟着一个数字，表示当前字符串的长度。

这样的处理有以下几方面的优点：

（1）将数字字符串作为一个特征的同时不会因

① https://github.com/fxsjy/jieba

此带来数据稀疏问题。相比不做任何处理数据的最大稀疏度(这里的稀疏度是指对数据进行 one-hot 编码后所增加的维数)从原来的 $O(M)$ 降低到 $O(\log_{10}M)$,其中 $M$ 是语料库字符串的最大数值。假设语料中最大的数值为 1 000,不做出任何处理,会带来最多 1 000 稀疏度(最坏假设是 0－1 000 的数值都在语料中出现)。而经过上面的处理,则最多引入 3 的稀疏度。

(2) 保留字符长度的信息,保持了字符串数值在数量级上的良序。如 $num1<$num2。

(3) 本文提出数字特征的平滑方法本身具有商品属性的偏向性。如 $num1 如果是一个商品属性则很可能代表"屏幕尺寸"或者"评价等级"等,再如 $numeng5 很可能是一个手机型号。

经过预处理后得到的语料如表 1 所示。

表 1 领域内语料存储格式

| 用户 | 对话内容 |
|---|---|
| 系统 | 您好 想 买 什么 手机 |
| 用户 | 华为 的 |
| 系统 | 请问 您 想要 多少 钱 的 |
| 用户 | $num4 左右 的 |
| 系统 | 主屏 尺寸 要 多 大 的 呢 |
| 用户 | $num1 寸 的 |
| 系统 | 您 看看 华为 $numeng5 这款 怎样 |
| 用户 | 它 的 像素 是 多少 |
| 系统 | 它 是 $num3 万 像素 |
| 用户 | 还有 其他 牌子 吗 |
| 系统 | 我 帮 您 查查 |

在经过预处理后需要对词语进行商品属性的标注。经过对语料分类整理,针对手机导购的应用场景特点,选取了后置摄像头像素、颜色、品牌、价格、主屏尺寸、型号等六种商品属性和一个非商品属性的标记进行标注。

### 3.3 词向量的特征表示法

#### 3.3.1 语言模型和词向量表示法

基于统计的语言模型能够表示成一个已出现的词和当前词的条件概率的一个极大似然表示,如式(1)所示。

$$\hat{P}(w_1^T)=\prod_{t=1}^{T}\hat{P}(w_t\mid context(w_t)) \tag{1}$$

其中 $context(w_t)$ 表示词语 $w_t$ 的上下文,$context(w_t)=w_i^j=(w_i,w_{i+1},\cdots,w_j)$ 表示从句子下标 $i$ 到 $j$ 的子串,$w_t$ 表示句子中第 $t$ 个词,则 $w_1^{t-1}$ 表示句子从第 1 个词到 $t-1$ 个词的句子前缀。这样的统计语言模型在自然语言处理如模式识别、机器翻译和信息抽取等领域都得到相当成功的应用[18]。出于计算复杂度的考虑,很自然的一个想法就是用 $w_{t-n+1}^{t-1}$,$w_t$ 的前 $n-1$ 个词近似表达其前缀,得到 n-gram 语言模型如式(2)所示。

$$\hat{P}(w_t\mid w_1^{t-1})\approx\hat{P}(w_t\mid w_{t-n+1}^{t-1}) \tag{2}$$

Bengio 等人[19]提出神经网络语言模型(neural network language model,NNLM),给每一个词一个向量表达,设计一个一层神经网络直接对 n-gram 语言模型直接建模,如式(3)所示。

$$f(w_t,w_{t-1},\cdots,w_{t-n+1})=\hat{P}(w_t\mid w_1^{t-1}) \tag{3}$$

#### 3.3.2 word2vector 模型

本文将使用 CBOW(continuous bag of word)语言模型结合 Hierarchical softmax 计算方法,收集到的语料中训练词向量。

CBOW 语言模型不仅限于前 $n-1$ 个词作为 $w_t$ 的上下文,而是考虑了句子中距离当前词为 $n$ 以内的词都看作是当前词的上下文环境,如式(4)所示。

$$C(context(w_t))=\sum_{0<|i-t|<n}C(w_i) \tag{4}$$

$C(i)$ 表示词 $i$ 的向量表达。当 $n=3$ 时,将位置 $w_{t-2}$ 到 $w_{t+2}$ (不包括 $w_t$)上下文窗口为 5 的向量相加形成 $context(w_t)$,对于每个构建好的上下文环境,神经网络语言模型都要对词库中的每一个词作参数更新,如式(5)、式(6)所示。

$$\hat{EB}=\underset{C(w_t)}{\operatorname{argmax}}\frac{1}{|V|}\sum_{t=1}^{|V|}\log p(w_t\mid context(w_t);c(w_t)) \tag{5}$$

$$p(w_t\mid context(w_t);c(w_t))=\frac{e^{y_{I(w_t)}}}{\sum_{w\in V}e^{y_{I(w)}}} \tag{6}$$

其中 V 是词库大小,$I(w)$ 表示 w 在词库中的下标。$y_i$ 表示神经网络的非归一化输出。

Hierarchicalsoftmax[17]则先将整个词库构建成一棵哈夫曼树,每个词位于树的叶子节点,中间节点表示其叶子节点的某种组合关系表示。因此每次只需要更新二叉树上从根节点出发达到 $w_t$ 的叶子节点路径上的所经过的全部节点,将复杂度从原来的 $O(|V|)$ 降低到 $O(\log|V|)$。

### 3.4　CNN 与指数型变长静态窗口的特征表达框架

#### 3.4.1　卷积神经网络

在原始的全连接的神经网络中，如果第 $l$ 层有 $n^l$ 个隐层节点，第$(l-1)$层有 $n^{l-1}$ 个隐层节点，则从第$(l-1)$层到第 l 层有 $n^l \times n^{l-1}$ 个参数，也表明有 $n^l \times n^{l-1}$ 条边将两层连接。当 $n^l$ 和 $n^{l-1}$ 都很大的时候，参数空间很大，训练的速度会非常慢。因此，用卷积运算来替代全连接。在卷积过程中第 l 层的每一个神经元在每一步(一步产生一个神经元)都只和第$(l-1)$的某一个局部产生全连接，如式(7)所示。

$$a^{\langle l\rangle} = f(w^{\langle l\rangle} \otimes a^{l-1} + b^{(1)}) \tag{7}$$

其中 $a^l$ 表示神经网络中第 l 层的输出，$w^{\langle l\rangle}$ 是一个长度固定 m 的但小于 $a^{l-1}$ 的维数的滤波器，特别地，当 $w^{\langle l\rangle} = \left[\frac{1}{m},\cdots,\frac{1}{m}\right]$ 时，是一个均值滤波器，$\otimes$ 表示卷积运算。从公式可以看出，卷积神经网络的每一步都共享同一个滤波器。参数的空间大小为 $m+1$。$f(\cdot)$ 是激活函数，也叫特征映射。

卷积的计算方法相比于全连接参数个数显著减少，但是经过特征映射后神经元的个数相比而言，没有明显变化。此时做一个 softmax 全连接的分类器其参数空间仍然相当庞大。因此通常会在卷积操作之后，特征映射之前加上一个采样的操作，也称为池化(Pooling)。池化操作不仅能进一步的减少参数的个数，还能降低特征的维度，从而避免过拟合。如式(8)所示。

$$a^{\langle l\rangle} = f(w^{\langle l\rangle} \otimes down(a^{l-1}) + b^{(1)}) \tag{8}$$

其中的 $down(\cdot)$ 表示对特征的映射结果作下采样。一般选择最大池化(max-pooling)作为下采样的函数，如式(9)所示。

$$down_{max}(a) = \max_i a_i \tag{9}$$

#### 3.4.2　指数型变长静态窗口的特征表达框架

图 2 展示了是本文所采用的商品属性抽取方法，

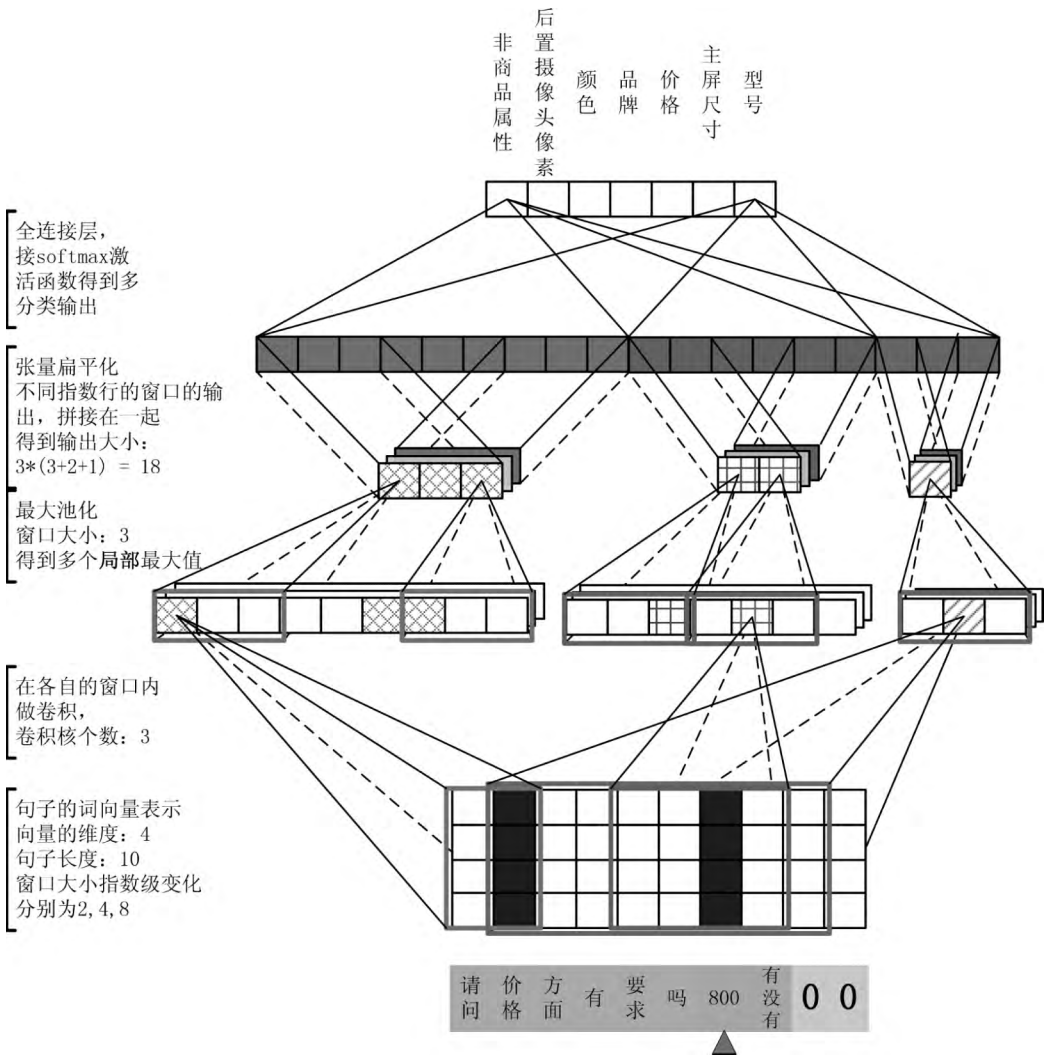


图 2　指数型变长静态窗口特征表达

用户的当前输入是“800 有没有”，则分别需要对“800”和“有没有”两个词做商品属性的判别。如果还是像商品评论语料中属性抽取任务一样的以句子为上下文，则当前词“800”既有可能是手机属性中“后置像头像素”，或者是“价格”。因此，需要将更大范围的上下文也考虑进来，如用户输入之前的系统输出。CNN 的输入格式需要固定的长度，对于输入小于预设长度 $L$ 的时候，本文采用句子末尾补 0 的方法。这里的 $L$ 为模型的超参(hype-parameter)。

对于预设定的 $L$，分别设定相对应的指数型的变长窗口以捕捉不同距离范围内的词对当前词的影响，如式(10)所示。

$$set_{(\cdot)}=\{2^l \mid l\in N^+, 2^l<L\} \quad (10)$$

$set(\cdot)$ 表示设定的不同长度的窗口的集合。$w^l \in set(\cdot)$ 表示 $set(\cdot)$ 中长度为 $l$ 的窗口，$N^+$ 表示正整数集合。图中 $set_{(\cdot)}=\{2,4,8\}$。不同的窗口分别在各自的长度范围内做卷积，max-pooling 和特征映射。然后将不同窗口的映射结果扁平化后拼接在一起，最后全连接的 softmax 做多分类输出。

指数型变长静态窗口特征表达框架，具有以下两种优良特性：

(1) 短的窗口能够捕捉近距离的词语本身在句子中的重要性，如图距离为 2 的窗口经过 max-pooling 以后，能够捕捉重要特征“价格”和“800”。而长距离的窗口则能对短距离学习出来的重要特征加以修正，最终可以抽取到[价格，800]。

(2) 给定有序元组 $TI=[1,2,\cdots,2^i,\cdots,2^j,\cdots,2^m]$，$j>i$，$L\approx 2^m$。则经过卷积计算相对应的所得到的窗口长度为 $TO=[L-1+1,\cdots,_{(}L-2^i+1_{)},\cdots,_{(}L-2^j+1_{)},\cdots,(L-2^m+1)]$，将其中的第 $i$ 项和第 $j$ 项相比有式(11)。

$$R=\frac{TO_i}{TO_j}=\frac{L-2^i+1}{L-2^j+1}\approx\frac{2^m-2^i+1}{2^m-2^j+1} \quad (11)$$

当 $j-i$ 趋向于 $m$ 时，有 $i\to 0$ 且 $j\to m$，因此得到式(12)。

$$\lim_{(j-i)\to m} R=\frac{2^m-1+1}{2^m-2^m+1}=2^m \quad (12)$$

这表明随着静态的输入窗口指数型增长，其相应的特征表达所占有的份额也指数型降低。因此“800”和“有没有”即使有着几乎相同的上下文，由于模型更为看重短窗口的所映射出特征的表达能力，因此对两者具有很好的区分度。

## 4 实验

### 4.1 数据集

本文将中文手机导购对话系统[20]收集到的对话语料的有效对话 1 533 段，经过预处理、分词、词标注等操作后，以词为单位构造数据集，数据集的标注情况如表 2 所示。从表格中可以看到不同的商品属性的数目分布都比较均匀。按每类商品属性照 7∶1 的比例划分训练集和测试集。

表 2 商品属性类别在语料库上的分布情况

| 极性 | 属性类别 | 个数 |
|---|---|---|
| 非商品属性 | 非商品属性 | 1 720 |
| 商品属性 | 后置摄像头像素 | 1 082 |
| | 颜色 | 1 272 |
| | 品牌 | 1 380 |
| | 价格 | 1 346 |
| | 主屏尺寸 | 717 |
| | 型号 | 1 161 |
| | 总数 | 8 678 |

### 4.2 实验设置

本实验首先采用收集到 100 万条微博、两万篇新闻以及 1 533 段对话语料训练 word2vector 模型，每一个词对应一个 50 维的实数向量。因为微博和对话语料本身是一种短文本，所以 CBOW 语言模型的上下文窗口设置为 5。

将用户的当前输入和对话系统的上一轮输出拼接在一起，以当前词为中心向前后截断固定的长度(本实验设置为 9)，把句子中的每个词用训练得到的词向量拼接作为 CNN 模型的输入，则静态窗口包含 2、4、8 三种不同的窗口长度，不同窗口长度的卷积核设置为 32，对卷积输出层做最大池化，池化的窗口设置为 3，将最大池化的输出扁平化处理，并首尾拼接成一列，最后接一层 softmax 全连接层。

本文的方法与对比的研究进展中的方法如下：

(1) CBOW+EXP：本文首先给每个词无监督的训练得到一个词向量表示，用词向量替换句子中词，通过指数型变长静态窗口特征表框架与 CNN 的结合体，对商品属性词进行抽取。

（2）CBOW＋CNN：Xu 等[16]首先使用 DRM（domain relevance measure）抽取商品属性种子集对语料进行标注，但该方法只适合是否为商品属性的二分类标注，因此本文采用人工标注的语料库。然后用词向量以及固定长度卷积窗口（本文实验中为 5）的 CNN 对商品评论中的候选（待定）商品词作属性词的判定。

（3）PMI：Popescu 和 Etzioni[8]提出一种基于词频统计的网页商品属性抽取方法，使用 PMI 词频统计方法无监督的抽取商品属性。$\mathrm{PMI}_{(f,d)}=\frac{Hits(d+f)}{Hits(d)*Hits(f)}$，$Hits(a)$是搜索引擎端查找关键词 $a$ 得到的网页数目。

本实验将从各个商品属性上 $F$ 的平均值（$MF$）和正确率（$accuracy$），对商品属性的抽取效果作比较，其中 $F$ 值是准确率和召回率的调和平均数。

### 4.2　实验效果对比

如表 3 所示，为本实验的商品属性抽取效果对比结果。$acc$ 表示正确率。表中的数值均以百分数的形式表示。

表 3　实验结果对比

| 抽取方法 | 非商品属性 | 后置摄像头像素 | 颜色 | 品牌 | 价格 | 主屏尺寸 | 型号 | $MF$ | $acc$ |
|---|---|---|---|---|---|---|---|---|---|
| CBOW＋CNN | 82.58 | 84.51 | 92.86 | 87.63 | 94.00 | 87.39 | 75.86 | 86.40 | 87.75 |
| PMI | 88.97 | 76.71 | 96.15 | 93.75 | 88.21 | 76.92 | 62.96 | 84.42 | 86.35 |
| CBOW＋EXP | 87.01 | 90.91 | 92.31 | 88.12 | 93.81 | 90.60 | 81.36 | 89.16 | 90.36 |

经过观察表中的本文所使用的方法在“后置摄像头像素”、“主屏尺寸”、“型号”等这三方面的属性相比于已有的做法，有大幅度的提升，而在其余属性上的抽取效果基本上与最好的相持平，因此从总体上看，$MF$ 和 $acc$ 都比研究进展的方法效果好。其中 $MF$ 相比于 CBOW＋CNN、PMI 分别提高了 3.2%和 5.6%，$acc$ 相比于 CBOW＋CNN、PMI 分别提高了 3%和 4.6%。实验结果说明，本文提出的指数型变长静态窗口的特征表达方法，对细粒度的商品属性具有更高区分度。

## 5　结束语

本文通过对比分析已有商品属性抽取方法的优缺点，结合人机交互过程中，口语对话的特点，首先利用词嵌入对词汇本身的词义建模，给出当前词是否与给定属性类别存在相关性的初步证据。然后保留相对完整的对话上下文，通过指数型变长静态窗口特征表达框架，解决一词多意的问题。实验结果表明，该特征框架对具有几乎相同上下文的相邻两个词也能具有很好区分能力，比原有的商品属性抽取方法的识别效果提升明显。然而对话过程中否定词的特征表达，仍然是口语对话系统的属性抽取的一大挑战。此外，对话系统测试与应用的广度和深度也影响着属性抽取效果的进一步提升。因此未来的工作，将加大系统的测试规模收集质量更高的对话语料，分析本文提出方法的误分案例，并探索更好的特征框架，对句子的构建更为完备的语义表达式。

## 参考文献


[1] Hu M, Liu B. Mining opinion features in customer reviews[C]//Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI 2004), 2004: 755-760.

[2] Yi J, Niblack W. Sentiment mining in Web Fountain [C]//Proceedings of the 21st IEEE Conference on Data Engineering (ICDE 2005), 2005: 1073-1083.

[3] Chen Y N, Wang W Y, Rudnicky A I. Jointly modeling inter-slot relations by random walk on knowledge graphs for unsupervised spoken language understanding[C]//Proceedings of the 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2015), 2015: 619-629.

[4] DeJong G. An overview of the FRUMP system[M]. Strategies for Natural Language Processing, 1982: 113.

[5] Radev D R, McKeown K R. Generating natural language summaries from multiple on-line sources[J]. Computational Linguistics, 1998, 24(3): 470-500.

[6] Paice C D. Constructing literature abstracts by computer: techniques and prospects[J]. Information Processing & Management, 1990, 26(1): 171-186.

[7] Hovy E, Lin C Y. Automated text summarization and the SUMMARIST system[C]//Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization, 1998: 197-214.

[8] Popescu A M, Etzioni O. Extracting product features and opinions from reviews[M]. Natural Language Processing and Text Mining. Springer London, 2007: 9-28.

[9] Zhuang L, Jing F, Zhu X Y. Movie review mining and summarization[C]//Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM 2006), 2006: 43-50.

[10] Qiu G, Liu B, Bu J, et al. Expanding Domain Sentiment Lexicon through Double Propagation[C]//Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009), 2009: 1199-1204.

[11] Zhang L, Liu B, Lim S H, et al. Extracting and ranking product features in opinion documents[C]//Proceedings of the 23rd international conference on computational linguistics (COLING 2010), 2010: 1462-1470.

[12] Kleinberg J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM, 1999, 46(5): 604-632.

[13] Wu Y, Zhang Q, Huang X, et al. Phrase dependency parsing for opinion mining[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009), 2009: 1533-1541.

[14] Zhao Y, Qin B, Hu S, et al. Generalizing syntactic structures for product attribute candidate extraction[C]//Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2010), 2010: 377-380.

[15] Xu L, Liu K, Lai S, et al. Mining opinion words and opinion targets in a two-stage framework[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), 2013: 1764-1773.

[16] Xu L, Liu K, Lai S, et al. Product feature mining: semantic clues versus syntactic constituents[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2013), 2014: 336-346.

[17] Morin F, Bengio Y. Hierarchical probabilistic neural network language model[C]//Proceedings of the International Workshop on Artificial Intelligence and Statistics (AISTATS 2005), 2005: 246-252.

[18] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, (3): 1137-1155.

[19] Bengio Y, Schwenk H, Senécal J S, et al. Neural probabilistic language models[M]. Innovations in Machine Learning. Springer Berlin Heidelberg, 2006: 137-186.

[20] Huang P J, Lin X M, Lian Z Q, et al. Ch2R: a Chinese chatter robot for online shopping guide[C]//Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2014), 2014: 26-34.

**叶大枢**(1993—),本科,主要研究领域为自然语言处理。

E-mail: yedashu2011@163.com

**邓振鹏**(1995—),本科,主要研究领域为机器学习。

E-mail: yy4f5da2@hotmail.com

**黄沛杰**(1980—),通信作者,博士,副教授,主要研究领域为人工智能、自然语言处理、口语对话系统。

E-mail: pjhuang@scau.edu.cn