北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2018.055

基于词模式嵌入的词语上下位关系分类

孙佳伟　李正华†　陈文亮　张民

苏州大学计算机科学与技术学院，苏州 215006; † 通信作者, E-mail: zhli13@suda.edu.cn

摘要　提出一种基于词模式的上下位关系分类方法，可以有效地缓解传统的基于模式的分类方法存在的稀疏问题，提高了关系分类的召回率。进一步地，通过词模式嵌入，将基于模式的方法与基于词嵌入的方法进行有效融合。为了验证方法的有效性，标注了一个包含 12000 个汉语词语对的数据集。实验结果表明，提出的词模式嵌入方法是有效的，F1 值可以达到 95.36。

关键词　上下位关系分类；模式；词嵌入；词模式嵌入

Hypernym Relation Classification Based on Word Pattern

SUN Jiawei, LI Zhenghua†, CHEN Wenliang, ZHANG Min

School of Computer Science and Technology, Soochow University, Suzhou 215000; † Corresponding author, E-mail: zhli13@suda.edu.cn

Abstract　This paper proposes a hypernym relation classification method based on word pattern, which can effectively alleviate the sparsity problem suffered by the traditional path-based method. Furthermore, this paper makes an effective combination of the path-based method and the distributional method via word pattern embedding. To demonstrate the effectiveness of the proposed approach, the authors have manually annotated a Chinese hypernym dataset containing 12000 word pairs. The experimental results show that the proposed word pattern embedding approach is effective and can achieve an F1 score of 95.36.

Key words　hypernym relation classification; word pattern; word embedding; word pattern embedding

语料库是自然语言处理任务的重要数据资源之一。上下位关系是部分语料库的基本框架，如 WordNet、HowNet。这些由人工构造的语料资源准确而且清晰，是做文本、语言研究的重要数据来源之一。然而，大型的语料库资源具有明显的缺点: 1) 维护和更新需要耗费大量的人力资源; 2) 语料库的范围和领域都非常的狭小固定[1]。因此，研究出能够自动获取上下位关系的方法是迫切需求之一。

上下位关系分类(hypernym relation classification)是对给定的名词词语对<*x*, *y*>进行自动分类的自然语言处理技术，比如将词语对<狗，动物>判断为具有上下位关系的词对，判断<花朵，蜜蜂>不具有上下位关系。上下位关系分类不仅能够支持更高层次的自然语言处理任务，如句法分析以及抽象语义表示；而且在信息处理领域也有广泛的应用价值，如在语料的属性词层次构建中可用于判断层次关系。

本文研究基于汉语的上下位关系分类的方法，并提出了一种新的词模式，构建融合词模型。研究目标在于提高基于模式方法的召回率，降低模式匹配的难度。在融合词模型中，词模式更加容易匹配，能够在短语模式基础上大幅度地提高基于模式方法的召回率，从而提高关系分类的 F1 值。在词模式保留上下文信息的基础上，本文还结合词嵌入的语

国家自然科学基金(61502325, 61673289)、江苏省高校自然科学研究重大项目(16KJA520001)资助
收稿日期: 2018–04–15; 修回日期: 2018–08–08; 网络出版时间: 2018-08-22 18:27:55

义信息，构建了词模式嵌入模型。目前没有公开的、大型的中文上下位数据库，本文提出上下位关系数据构建方法，数据构建主要根据同义词词林与NLPCC-2017 测评数据，添加部分人工构建工作。本文构建了 12000 个词语对的汉语上下位数据库。

1　相关工作

1.1　语料库构建

在英语数据中，WordNet 是目前最重要、涉及范围最广、被常用的自然语言系统数据，该数据中分为名词、副词、形容词和动词 4 组不同且互不干扰的词语网络。其中，名词网络主要通过上下位关系连接。WordNet 提供接口软件，方便查询及应用。除英文外，WordNet 还有其他语言部分，如 EuroNet，但是其他语言的数据规模远小于英语的规模，不足以进行数据研究。在荷兰语中，Sang 等[2]使用 EuroNet 的数据，但是由于数据规模太小，作者采用 Snow 等[3]的方法抽取出更多的上下位词语对。目前，比较成熟的具有上下位关系的中文数据库有 HowNet 和同义词词林等。

目前，大部分有关上下位关系的研究工作都是基于英文的，其他语言的研究与语料库资源较少。汉语上下位关系的语料库规模也远远不及英文语料库。目前针对汉语上下位的研究有刘磊等[4]提出的基于“是一个”模式获取下位词概念的方法。

1.2　上下位关系挖掘方法研究

进行上下位关系分类的方法主要有两种：基于模式的方法和基于词嵌入的方法。基于模式的方法也被称为基于路径的方法(path-based method)。给定词语对<*x*, *y*>，比如<狗，动物>，在“狗是一种动物”一句中，词语对之间的短语信息为“是一种”。“是一种”是一个明显具有上下位关系的模式，也可以被称为词语对之间的路径。这种词语对之间的短语信息被称为模式(路径)。而这些利用模式进行上下位关系分类的方法就是基于模式的方法。该方法最早由 Hearst[5]提出，主要是在词语对同时出现在许多语句时，将不同语句提供的词语对之间的短语信息构成集合，通过集合来判断对应词语对的关系。Hearst[5]提供了几个明显具有上下位信息的模式，比如：是一种(is a/an)、例如(such as)、和其他(and other/ or other)等等。除短语信息以外，依存路径也可以当做模式。Snow 等[3]将词语对用他们之间的依存路径的集合来表示，在此基础上进行上下位关系分类。使用基于模式的方法会构建一个巨大的模式特征空间，而语言表达的多样性和模式的可变性使得特征空间十分的稀疏。比如，“狗是一种活泼的动物”、“猫是一种可爱的动物”，都包含“是一种”，但是不同的修饰词语导致被识别为两个不同的模式。修饰语的使用让模式更加难以匹配。为了解决这个问题，Nakashole 等[6]构建 PATTY 模型，将模式中的词语用词性来代替。

在另一种基于词嵌入的方法中，词语对被离散化地表示为向量(embedding)，通过两个词向量之间的运算来判断词语对是否为上下位关系。最早采用基于分布方法进行上下位关系判断的是 Lin[7]。Kotlerman 等[8]提出另外一种基于分布的方法，他们假设基于下位词的上下文是上位词的上下文的子集。但是，基于分布的方法有一个明显的缺点，就是训练出来的分布式向量仅仅保留了词语的语境信息，不能够包含字典等重要、准确的先验知识。

目前，在众多方法中，最优的基于模式方法的效果要比基于分布方法的效果差一些。这主要是由于基于模式的方法要求词语对出现在同一个句子中[1]，并且模式要进行匹配，限制了基于模式方法的召回率。

2　基于模式的上下位关系分类方法

本文构建一个基于短语模式与词模式的上下位关系分类系统，如图 1 所示。首先通过语料抽取词语之间的模式，然后将其应用在关系分类模型上。其关键点是在抽取短语模式的基础上，添加词模式的使用。通常情况下，与短语模式相比，词模式更容易抽取与匹配，且具有更细化的文本信息。因此，本文从一个大规模的语料中抽取并构建词语的短语模式与词模式空间，利用模式空间将词语的模式集合转化为向量，使用最大熵算法得到融合词模型(lexical-word pattern model)。

2.1　短语模式

短语模式是用于处理上下位关系分类任务的一种常见模式，能够有效地体现词语在句子中的关系。本文采用 Sang 等[2]提出的短语模式抽取方法。为了使短语模式更好地保留句子的信息，在模式抽取之前，语料需要经过分句、分词与词性标注这 3 个句法分析步骤。本文将 3 种标点(。？！)视为断句的标记。

短语模式指在句子中，两个指定名词之间的词

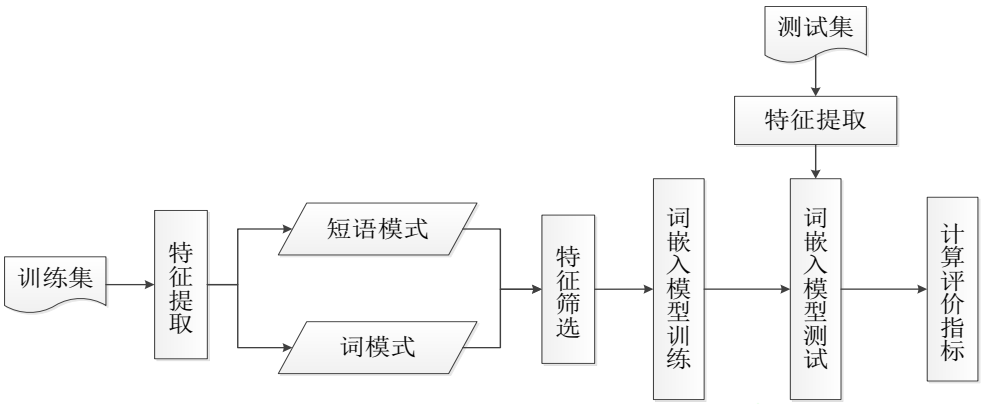

图 1　上下位关系分类系统流程图

Fig. 1　Flow chart of hypernym relation classification

长度不超过 5 的短语[3]。在上述句子中，词语对<水果，苹果>之间的短语“这 一 种”是词长度为 3 的短语模式。当句子不同时包含词语对中的两个词时，词语对无法从该句子中抽取短语模式。

基于短语模式的分类方法需要构建模式空间。抽取出词语对在语料中的短语模式集合，由所有的短语集合构成模式空间。下面给出两个词语对的样例，其中例 1 为具有上下位关系的词语对<水果，苹果>的句子，短语模式集合为{“这 一 种”，“是 一 种”，“中”}；例 2 为具有非上下位词语对<水果，人>句子，模式短语集合为{“对”，“不 爱 吃”}，虽然例 2(3)中包含“水果”和“人”两个词语，但是分词之后并没有单独的名词“人”出现，所以在这句话中没有词语对<水果，人>的短语模式。由此，上下位关系分类可以根据上下文信息进行估计与推理。

例 1　1) 我 只 喜欢 **苹果** 这 一 种 **水果** 。
2) **苹果** 是 一 种 **水果** 。
3) **水果** 中 **苹果** 产量 最高 。

例 2　1) 多 吃 **水果** 对 **人** 的 身体 有 好 处 。
2) 很多 **人** 不 爱 吃 **水果** 。
3) 一 种 **水果** 叫 **人**参果 。

2.2　词模式

基于模式的上下位关系识别是将模式转化为特征向量的关系分类任务。目前，基于模式的识别方法的召回率普遍较低[7]。研究表明，使用短语模式作为特征，能有效地进行上下位关系分类，但是存在明显的不足[2]：1) 语料规模决定短语模式空间的大小，对实验产生较大影响；2) 受修饰语、定语和语言习惯的影响，相同含义的短语模式能够相似，但很难相同。如果想要提高基于模式方法的召回率，需要扩大模式空间、降低模式匹配难度。

本文提出一种优化的词模式。该模式将短语模式中的每一个词都单独作为一个特征。根据短语模式不同但相似这一特点，本文采用词模式作为优化模式。例 3 中，上下位词语对<苹果，水果>和<动物，狗>的短语模式分别为“是 一 种 健康 的”以及“是 一 种 活泼 的”。

例 3　1) **苹果** 是 一 种 健康 的 **水果** 。
2) **狗** 是 一 种 活泼 的 **动物** 。

例 3 中，两个短语模式表达了相同的意思，但是不同的修饰词使得短语模式无法匹配。虽然这两种模式整体无法匹配，但是模式内的词语大部分是一样的。基于上述例子，本文尝试以词作为模式的基本单位，能够提高模式的匹配度。本文提出词模式的概念。词模式是将短语模式中的每一个词都独立出来，作为一种模式。不同于短语模式体现词语在句子中关系这一观点，本文认为词语之间的每一个中间词都分别体现了词语之间的关系。

词语对<水果，苹果>和<动物，狗>在例 3 中的短语模式与词模式如表 1 所示。在短语模式情况下，两个词语对分别只有一个短语模式，并且并不相同。然而，在词模式情况下，两个词语对的模式集合长度为 5，且主要词模式都是相同的，两种集合的重叠程度有 4/5。

词模式不仅能够提高模式匹配度，提高方法召回率，也能够扩大模式空间，减少语料规模对实验结果的影响。所以，本文在短语模式的基础上添加

表 1　词语对短语模式与词模式

Table 1　Lexical patterns and word patterns for word pairs

词语对	短语模式	词模式
<水果, 苹果>	是 一 种 健康 的	是、一、种、健康、的
<动物, 狗>	是 一 种 活泼 的	是、一、种、活泼、的

了词模式。

本文利用给定模式构建特征空间，并利用特征空间将词语对的模式集合向量化，将词语对对应的特征向量输入分类器，对给定词语对进行关系分类。这种模式的优化方法降低模式匹配的难度，从而提高了方法的召回率，使得 F1 值也有明显的提升。

3　基于词模式嵌入表示的上下位关系分类方法

在 2.2 节提出的词模式基础上，本文构建了基于词模式嵌入表示的上下位关系分类系统。首先，根据大规模语料获取词嵌入[9]，在该语料基础上获取词语对的词模式。然后，根据词语的词嵌入以及所有词模式的词嵌入，使用简单前馈神经网络构造词融合词模式嵌入模型(word pattern embedding model, WP-EMB)。在此情况下，本文提出的 WP-EMB 模型不仅充分利用了语句的上下文信息，也利用了词语的语义信息。

3.1　词嵌入

基于模式的方法利用上下文信息对词语对进行上下位关系分类，存在较明显的缺点：1) 需要词语对作为独立名词，同时出现在同一个句子中；2) 当词语对比较少见时，词语对甚至可能没有上下文信息，无法进行基于模式的分类；3) 仅利用上下文信息，却无法使用词的语义信息[1]。这些缺点限制了基于模式方法的召回率。

与基于模式方法不同，基于词嵌入表示的上下位关系分类方法利用词嵌入来表示词语的信息[8,10]。词嵌入表示既包含词语的上下文信息，也在很大程度上表示词的语义。这一特点使得基于词嵌入的方法有较高的召回率。目前，在基于词嵌入的方法中，词语对<*x*, *y*>通常用特征向量表示。Roller 等[11]与 Baroni 等[12]分别用 $\vec{y}-\vec{x}$ 和 $\vec{x}\oplus\vec{y}$ 来表示词语对。

由于短语模式具有长度较长、包含多个词语等特点，目前还难以将短语模式整体转变成词嵌入形式，基于模式的方法还无法利用语义信息。

3.2　词模式的嵌入

本文提出在基于词模式方法的基础上添加词嵌入，该方法结合词模式的上下文信息，也利用了词语的语义信息。通常情况下，基于模式的方法难以添加词嵌入，2.2 节中提出的词模式是添加词嵌入的基础。

首先，采用与 Baroni 等[12]相同的 $\vec{x}\oplus\vec{y}$ 组合方法来表示词语对本身的含义。除此之外，通过词模式嵌入来进一步利用上下文信息。不同词语对的词模式集合$\{\overrightarrow{p_1},\overrightarrow{p_2},\dots,\overrightarrow{p_n}\}$大小也不一致，采用 $\overrightarrow{p_1}\oplus\overrightarrow{p_2}\oplus\dots\oplus\overrightarrow{p_n}$ 的向量拼接方法产生的词模式特征向量维度不一致，其中，$\overrightarrow{p_i}$ 为词语对的词模式 *i* 的词嵌入表示。

为了获取词模式集合的特征向量，并保持特征向量的维度一致，本文采用对所有词模式嵌入池化结果做拼接的方法。计算过程如式(1)~(4)所示。

$$\overrightarrow{p_{max}}=\max pooling(\overrightarrow{p_1},\overrightarrow{p_2},\dots,\overrightarrow{p_n}) \tag{1}$$

$$\overrightarrow{p_{min}}=\min pooling(\overrightarrow{p_1},\overrightarrow{p_2},\dots,\overrightarrow{p_n}) \tag{2}$$

$$\overrightarrow{p_{ave}}=\text{average}\,pooling(\overrightarrow{p_1},\overrightarrow{p_2},\dots,\overrightarrow{p_n}) \tag{3}$$

$$\overrightarrow{P_{xy}}=\overrightarrow{p_{max}}\oplus\overrightarrow{p_{min}}\oplus\overrightarrow{p_{ave}} \tag{4}$$

其中，$\overrightarrow{p_i}$ 表示词语对的词模式 *i* 的词嵌入，$\overrightarrow{P_{xy}}$表示给定词语对的词模式的特征向量。

基于以上得到的词模式特征向量表示，本文将词语对的词嵌入和词模式特征向量拼接，利用最终拼接的特征向量做为词语对的表示。此方法能够同时利用词语的语义信息与上下文信息进行上下位关系分类。

$$\overrightarrow{V_{xy}}=\vec{x}\oplus\vec{y}\oplus\overrightarrow{P_{xy}} \tag{5}$$

其中，$\overrightarrow{V_{xy}}$表示词语对的最终特征向量。

本文提出的词语对的最终特征向量既保留了模式的上下文信息，也获取了词语本身的语义信息。在此基础上，本文将最终特征向量输入一个简单前馈神经网络，构建一个词融合词模式嵌入模型 WP-EMB。

3.3　词模式嵌入模型

本文构建的词模式嵌入的上下位关系分类模式采用双层前向神经网络。结构如图 2 所示。

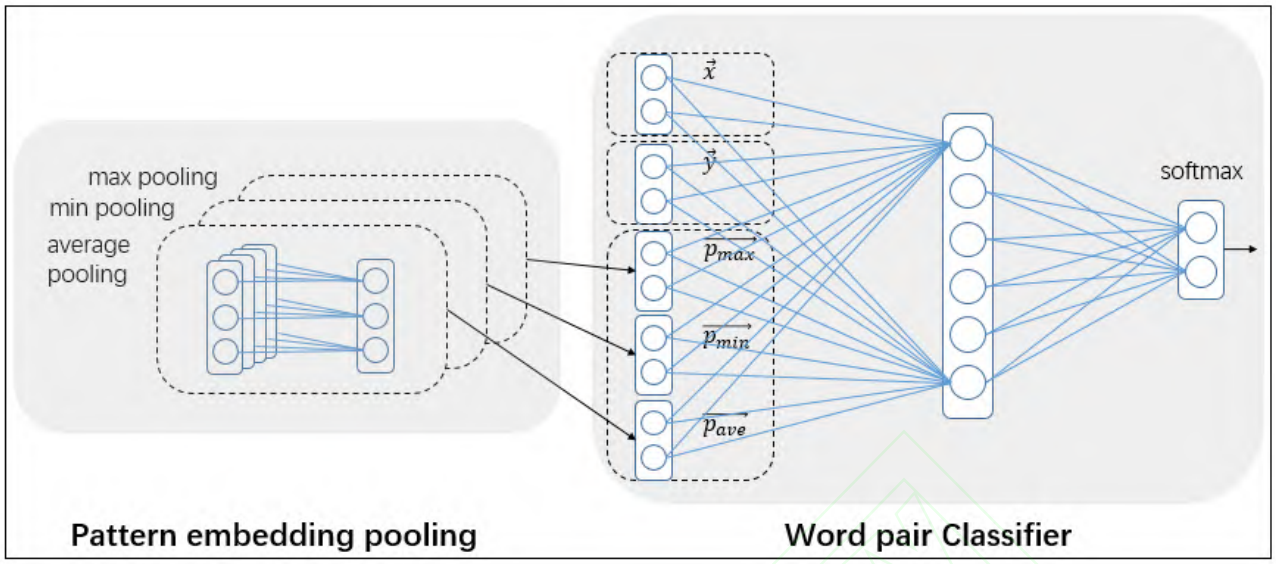

图 2　基于词模式嵌入的神经网络模型结构

Fig. 2　Architecture of model based on word pattern embedding

模型使用 softmax 计算类别概率。

$$\vec{c} = softmax(W \cdot \overrightarrow{V_{xy}}) \tag{6}$$

其中，$\vec{c}$ 是一个两维的概率向量。词语对仅在$\vec{c}[1] > 0.5$ 时被分类为上下位关系词语对。

在输入层，词语对被映射为最终特征向量$\overrightarrow{V_{xy}}$，作为神经网络的输入。在隐层，神经元数量为 100，采用 sigmoid 激活函数提高神经网络的非线性建模能力。在输出层，采用 softmax 计算词语对被分类为上下位词语与非上下词语对的分值。

4　汉语上下位数据标注

由于目前没有公开、大型的上下位数据库资源，为了完成实验，本文构建一个由约 4200 个上下位词对、8500 个非上下位词与对构成的数据库。

Snow 等[3]提供了一种主流的方法来构建词库。根据现有数据训练出分类器，然后从语料中爬取出现在同一个句子中的名词语对，作为测试集。用之前训练好的分类器对测试集进行分类，分类之后再加以人工检查，检查后的数据可以继续加入词库。这个方法较难用于中文语料，因为中文没有英文的词根变化，所以在词性标注这一步，很多修饰词或前缀词会被标注为名词，会出现一句话中有超过 5 个名词。所以本文并没有采取这种方法。

本文构建的上下位词语对约 4200 个，有 3 个来源。

1) 从同义词词林和 HowNet 中爬取可能具有上下位关系的 20000 对词语对，然后通过人工检查，最终保留 2300 对上下位词语对。

2) 人工构建词类，包括动物、植物、家电、国家、城市、建筑、方言、乐器、节日、木材、学科、职业等等，上位词为如上概括的词语，下位词为常见的该类事物。比如，节日的下位词为清明、春节、端午、中秋节等，乐器的下位词为钢琴、小提琴、竖笛、二胡等。

3) NLPCC-2017 的测评任务涉及包括上下位在内的 4 种关系，主办方公开了用于评价的 2000 对词语对，其中有 500 对为上下位词语对。

非下位词有 8500 个，主要有 2 个来源。

1) NLPCC-2017 公开评价数据的 500 个部分整体的名词语对。

2) 在上述第二条中，采用大类错开和并列下位词构造的词语对。比如，所有节日的下位词和乐器这一词构成非上下位词语对，如<端午，乐器>、<中秋节，乐器>；所有具体大类的下位词的并列，如节日的两个下位词<端午，中秋节>，国家的两个下位词<中国，古巴>等等。再对构造出的词语对进行人工检查，保留需要数目的非上下位词语对即可。

5　实验与结果分析

5.1　实验语料

为了进行对比，本文将 Sang 等[2]的基于短语模

式的方法在汉语的语料与数据上进行实验。本文需要大量的无标注语料进行模式抽取与词嵌入训练，无标注语料主要为百度百科和 BCC 语料库。百度百科语料包含目前百度百科中的所有词条信息和对应的网页数据。BCC 语料是由北京语言大学构建的语料平台，可以根据需要进行语料查询和保留。BCC 语料包括报刊、文学、科技和古汉语等多领域语料，是可以全面反映当今社会语言生活的大规模语料库。各语料的基本情况如表 2。

表 2　数据规模及语料信息

Table 2　Details of corpus for hypernym datasets

数据	词语对数量	语料来源	句子数
上下位词语对	4274	百度百科	6150000
		BCC 语料库	90000
非上下位词语对	8500	百度百科	850000
		BCC 语料库	50000

本文主要依据 BCC 语料来抽取短语模式与词模式。部分上下位和非上下位词语对在 BCC 语料中没有匹配的语句，这些没有相关语句的词语对会直接被分类为非上下位，影响模型的分类效果。在此情况下，本文也基于百度百科的语料进行了同样的实验。本文采用精确率(P)、召回率(R)以及 F1 值作为评价模型分类效果的指标。

5.2　融合词模型

基于短语模式和词模式，本文进行实验的特征有如下几种：1) 短语模式 (lexical pattern, LP)：词语对在句子之间、词长度不超过 5 的短语；2) 词模式 (word pattern, WP)：词语对在句子之间、长度不超过 5 的短语模式中的每一个词；3) 词模式频率 (word frequency pattern, WFP)：由于词模式非常容易匹配，使得词模式出现频率较高。为了利用词模式的频率信息，采用桶方法统计词模式出现次数，将次数分入不均匀的桶。对于每一个词模式及其出现次数，构建6个桶，分别为1-5, 6-10, 11-20, 21-50, 51-100, 100。给定词语对，当词模式“是”出现次数为 23 时，“是 21-50”对应的特征值为 1，其余桶值为 0。

采用最大熵作为分类模型，分别在 BCC 语料和百科语料进行对比实验。基线实验为基于 LP 的短语模式模型，本文提出基于 LP 和 WP、基于 LP 和 WFP、基于 LP, WP 与 WFP 的 3 种融合词模型与基于 WP 的词模型。根据上述评价指标，利用不同特征情况下，各模型的正确率和召回率见表 3。

从表 3 中可以看出，在 BCC 语料中，当模式中添加词模式时，模型得到的 F1 值最高为 75.15%；在百度百科语料中，在短语模式中添加词模式时，模型得到最高的 F1 值为 82.46%。根据 McNemar 显著性检测，模型之间的差距($p < 0.05$)证明词模式为基于模式的上下位关系分类方法的性能带来可靠的提高。

与已有工作相比，文本提出的融合词模型性能明显优于 Sang 等[2]提出的基于短语模式的模型，说明模式匹配度高的重要性。词模式能够有效地降低模式匹配的难度，提高模式匹配率，从而提高基于模式方法的召回率。

5.3　词模式嵌入模型

为了确定最佳的分类模型，本文以基于单层 bilinear 的词嵌入模型为基线实验。对比实验分别为基于单层 bilinear 的词融合模式嵌入模型、基于 linear 的词模式嵌入模型与基于 linear 的词融合词模式嵌入模型。每个模型的输入与计算方式如表 4 所示。

表 3　基于模式方法对比实验结果

Table 3　Performance of different models on two corpuses

实验系统	BCC 语料			百度百科语料		
	P/%	R/%	F1/%	P/%	R/%	F1/%
短语模式	84.53	55.77	67.20	82.17	77.89	79.98
+词模式	83.19	68.53	**75.15**	81.01	**83.96**	**82.46**
+词模式频率	85.56	65.49	74.19	81.33	78.89	80.07
+词模式+词模式频率	**86.03**	64.40	73.66	**84.49**	78.74	81.51
词模式	80.62	**69.74**	72.26	79.63	83.11	81.33

表 4　模型输入与计算

Table 4　Input and calculation of models

模型	输入	计算方法
bilinear 词嵌入模型	$\vec{x}$, $\vec{y}$	$\vec{x}\cdot W\cdot\vec{y}$
bilinear 词融合词模式嵌入模型	$\vec{x}\oplus\overrightarrow{P_{xy}}$, $\vec{y}\oplus\overrightarrow{P_{xy}}$	$(\vec{x}\oplus\overrightarrow{P_{xy}})\cdot W\cdot(\vec{y}\oplus\overrightarrow{P_{xy}})$
linear 词模式嵌入模型	$\overrightarrow{P_{xy}}$	$W\cdot\overrightarrow{P_{xy}}$
linear 词融合词模式嵌入模型	$\vec{x}\oplus\vec{y}\oplus\overrightarrow{P_{xy}}$	$W\cdot(\vec{x}\oplus\vec{y}\oplus\overrightarrow{P_{xy}})$

表 5　词模式嵌入实验结果

Table 5　Performance of pattern embedding model

模型	BCC 语料			百科语料		
	P/%	*R*/%	F1/%	*P*/%	*R*/%	F1/%
bilinear 词嵌入模型	87.55	82.02	84.69	83.68	83.48	83.58
bilinear 词融合词模式嵌入模型	73.37	90.40	81.00	49.81	**95.75**	65.53
linear 词模式嵌入模型	**95.91**	**96.81**	**96.36**	83.81	81.77	82.78
linear 词融合词模式嵌入模型	95.07	96.11	95.59	**94.90**	94.90	**95.36**

根据表 5 中基于单层 bilinear 的词嵌入模型与基于 linear 的词模式嵌入模型，可以看出模式代表的上下文信息与词嵌入代表的语义信息都是上下位关系分类的重要信息。本文设计的词融合词嵌入模型在百度百科语料上的 F1 为 95.36%，与只利用词嵌入或词模式的方法相比，得到的 F1 值高，因为综合利用了词语对的上下文信息与语义信息，提高上下位关系识别的精确率与召回率，也能够解决基于模式方法部分词语对没有模式、无法进行分类的情况。

在表 5 中，基于 BCC 语料的实验结果普遍比基于百度百科语料的结果高。因为 BCC 语料比百度百科语料规模小，部分非上下位词语对没有模式。在基于模式的模型中，不具有模式信息的非上下位词语对可以直接被分类正确，而在词模式嵌入模型中，需真实分类的词语对数量较小。词语对没有模式时，也说明词语对之间的关系较远，也是一种上下文信息。由此可见，语料的规模对实验结果也有影响。

通过比较不同模型的 F1 值，说明语义信息与上下文信息对上下位关系分类都有重要的影响。实验结果表明，本文提出的词嵌入模型能够利用上下文信息与语义信息，进而提高模型分类的性能。

6　总结与展望

本文在短语模式基础上提出词模式，降低模式匹配难度。在基于短语模式与词模式的融合词模型中，模型的召回率得到提高，模式更为简洁高效。在词模式基础上，本文结合词嵌入的语义信息，构建了词融合词模式嵌入模型。实验结果表明，本文方法能够有效地进行上下位词语对关系分类，效果超过基于短语模式的方法。

未来工作中，我们将考虑从 3 个方面进一步优化上下位关系识别的模型：在词模式向量的基础上，提高词模式特征向量的代表性；构建一个能够从文本中自动挖掘上下位关系词语对的方法；利用依存信息提高模式的有效性。

参考文献

[1] Shwartz V, Goldberg Y, Dagan I. Improving hypernymy detection with an integrated path-based and distributional method. 2016

[2] Sang E T K, Hofmann K. Lexical patterns or dependency patterns: Which is better for hypernym extraction? // Thirteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2009: 174–182

[3] Snow R, Jurafsky D, Ng A Y. Learning syntactic patterns for automatic hypernym discovery // Advances in Neural Information Processing Systems 17. 2004: 1297–1304

[4] 刘磊, 曹存根, 王海涛, 等. 一种基于“是一个”模式的下位概念获取方法. 计算机科学, 2006, 33(9): 146–151

[5] Hearst, Marti A. Automatic Acquisition of hyponyms from large text corpora // Proc. of the International Conference on Computational Linguistics. 1992: 539–545

[6] Nakashole N, Weikum G, Suchanek F. PATTY: a taxonomy of relational patterns with semantic types // Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012: 1135–1145

[7] Lin D. An information-theoretic definition of similarity // International Conference on Machine Learning. 1998: 296–304

[8] Kotlerman L, Dagan I, Szpektor I, et al. Directional distributional similarity for lexical inference. Natural Language Engineering, 2010, 16(4): 359–389

[9] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems, 2013, 26: 3111–3119

[10] Weeds J, Weir D. A general framework for distributional similarity // Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2003: 81–88

[11] Roller S, Erk K, Boleda G. Inclusive yet selective: supervised distributional hypernymy detection // COLING. 2014: 1025–1036

[12] Baroni M, Lenci A. How we BLESSed distributional semantic evaluation // Gems 2011 Workshop on Geometrical MODELS of Natural Language Semantics. 2011: 1–10