网络出版时间：2019-10-31 13:48:00
网络出版地址：http://kns.cnki.net/kcms/detail/11.1826.TP.20191031.1234.002.html

面向上下位关系预测的词嵌入投影模型

汪诚愚[1)] 何晓丰[2)] 宫学庆[1)] 周傲英[3)]

[1)](华东师范大学 软件工程学院，上海市 200062)

[2)](华东师范大学 计算机科学与技术学院，上海市 200062)

[3)](华东师范大学 数据科学与工程学院，上海市 200062)

摘 要 上下位关系是自然语言处理领域中的重要概念，用于描述概念之间的从属关系. 上下位关系的精准预测，有助于挖掘概念之间的内在层次结构，是构建大规模语义网络、知识本体、知识图谱等知识密集型系统的重要基石. 传统上下位关系的预测算法大多依赖较为固定的语言模式，因而具有低覆盖度、高人工干预等缺陷. 此外，语言模式与语言本身的特性高度相关，在中文等表述灵活的语言中预测精度较低. 随着深度学习技术在自然语言处理领域的迅猛发展，词嵌入技术被广泛应用于建模词之间的语义关系. 特别地，词嵌入投影模型学习如何将下位词的词向量投影到上位词的词向量，显式地建模了上下位关系的关系表示. 基于已有经典研究以及最新成果，本文详细论述了词嵌入投影模型的发展过程和最新研究进展，包括基于迭代学习、转导学习、对抗学习等深度学习技术在词嵌入投影模型上的改进. 在实验中，我们对多个词嵌入投影模型在中文和英文的公开数据集上进行充分详细的评测，探讨了不同的词嵌入投影模型在不同学习场景下的优缺点. 最后，在面向特定领域和长尾上下位关系抽取等问题上探讨了未来的研究展望.

关键词 上下位关系；词嵌入；词嵌入投影模型；关系预测；深度学习

中图法分类号 TP182

Word Embedding Projection Models for Hypernymy Relation Prediction

WANG Cheng-Yu[1)] HE Xiao-Feng[2)] GONG Xue-Qing[1)] ZHOU Ao-Ying[3)]

[1)](School of Software Engineering, East China Normal University, Shanghai 200062)

[2)](School of Computer Science and Technology, East China Normal University, Shanghai 200062)

[3)](School of Data Science and Engineering, East China Normal University, Shanghai 200062)

Abstract A hypernymy (“is-a”) relation is an important concept in Natural Language Processing (NLP). It is used to describe the subordination relation between two concepts. The accurate prediction of hypernymy relations is important for mining the inherent hierarchy of concepts and building large-scale semantic networks, ontologies, knowledge graphs and other knowledge-intensive systems. Most traditional hypernymy prediction algorithms rely on relatively fixed language patterns, which have drawbacks such as the low coverage and the high degree of manual intervention. In addition, the textual patterns are highly correlated with the language itself. For languages with low regularity such as Chinese, pattern-based methods are not sufficiently accurate. With the rapid development of deep learning techniques in NLP, word embeddings are frequently employed to model

本课题得到国家重点研发计划(2006YFB1000904)和国家自然科学基金(61572194)资助.汪诚愚，男，1991年生，博士研究生,主要研究领域为关系抽取、知识图谱、自然语言处理.E-mail: chywang2013@gmail.com.何晓丰（**共同通信作者**），男，1969年生，博士，教授，博士生导师，主要研究领域为数据挖掘、机器学习、信息检索.E-mail: xfhe@sei.ecnu.edu.cn. 宫学庆（**共同通信作者**），男，1974年生，博士，教授，博士生导师，主要研究领域为数据库技术、分布式数据库管理系统.E-mail: xqgong@sei.ecnu.edu.cn. 周傲英，男，1965年生，博士，教授，博士生导师，中国计算机学会 (CCF)会士，主要研究领域为大数据处理和分析等.E-mail: ayzhou@dase.ecnu.edu.cn.

semantic relations between words. Especially, word embedding projection models learn how to map the embeddings of hyponyms to those of their hypernyms, modeling the representations of hypernymy relations explicitly. In view of existing classical and latest research, this paper introduces the development process and the latest breakthrough of word embedding projection models, including the improvements of projection learning based on deep iterative, transductive and adversarial learning. In the experiments, we conduct extensive experiments of these word embedding projection models over multiple Chinese and English datasets, and compare the advantages and disadvantages of these models under different circumstances. Finally, the future research directions are discussed, which focus on domain-specific and long-tail hypernymy prediction.

Key words hypernymy relation; word embedding; word embedding projection model; relation prediction; deep learning

1 引言

1.1 研究背景

上下位关系（Hypernymy）是一种基本的语义关系，用以描述概念之间的层次隶属关系. 例如，在“哺乳动物-动物”这一对概念中，“哺乳动物”被称为“动物”的下位词（Hyponym），而“动物”是“哺乳动物”的上位词（Hypernym）. 上下位关系不但是知识分类体系、知识图谱等知识密集型系统的重要组成部分[1, 2]，而且也常作为重要的语义资源，用以提升自然语言处理（Natural Language Processing，NLP）和信息检索中诸多下游任务的准确度，例如自然语言推理[3]、个性化推荐[4]、互联网查询理解[5].

由于上下位关系具有广泛的应用价值，这一类关系的精准预测成为 NLP 研究中重要的基础性任务. 然而，上下位关系的预测具有很大的挑战性. 这由于上下位关系的语义内涵比较丰富，包括实体与概念之间的“is-instance-of”关系（例如“iPhone X-手机”）和概念与概念之间的“subclass-of”关系（例如“手机-通讯工具”）等. 这些知识一般属于人类的常识（Commonsense knowledge），对于机器而言，获取和推理这些知识的能力较弱. 在现有的研究中，上下位关系预测方法一般分为两个主要类别：模式匹配法和分布式学习法[6]. 模式匹配法最早可以追溯至 Marti Hearst 教授 1992 年的发表研究工作[7]. 她人工制定多个固定的英语语言模式（被称为“Hearst patterns”），用文本匹配法在语料库中自动抽取英语上的上下位关系. 例如，在句子“animals such as dogs and cats”中可以利用“… such as...”模式匹配出两个上下位关系实例“dog-animal”和“cat-animal”. 后续的研究工作大多在此基础上提升模式匹配的精准度和覆盖率. 一个典型的研究工作是 Nakashole 等人设计的 PATTY 系统[8]，这个系统利用词性标注、知识本体类别和通配符匹配等额外信息对固定的上下位关系模式进行扩展. 在提升精准度方面，Luu 等人[9]加入了句子的语法结构信息来限制关系抽取的过程. 值得指出的是，尽管模式匹配法在知识抽取任务中应用广泛，它仍然具有低覆盖率等诸多缺陷. 此外，这些方法与目标语言本身的特性关系密切，往往不具有高语言迁移性. 例如，由于中文表达灵活多变，语言模式不固定，Fu 等人[10]和我们的研究工作[11]都显示出，模式匹配法在中文的表现不理想，精准度和覆盖率都比较低，难以满足实际的应用需求.

为了解决上述问题，分布式学习法利用两个概念的分布式表示（也称为词向量、词嵌入等），推断这两个概念是否具有上下位关系. 因为分布式表示可以从整个语料库学习到，而不是局部的语言模式，这种方法避免了模式匹配法的低覆盖率问题. 其中，上下位关系度量（Hypernymy measure）是一种主流的无监督式学习方法，用于建模两个概念具有上下位关系的程度. 典型的上下位关系度量包括 WeedsPrec[12]、SLQS[13]、HyperScore[14]等. 然而，这一方法仅利用一个指标进行判断，且不直接利用标注数据进行优化，因此他们精度相对较低.

与之不同，监督式的上下位关系分类直接利用两个概念的分布式表示作为分类模型的原始输入，预测这两个概念的语义关系（上下位或非上下位关系）. 代表性的研究工作有 Roller 等人提出的 asym 模型[15]、Turney 和 Mohammad 提出的 simDiff 模型[16]以及 Yu 等人提出的最大间隔神经网络模型[17]等. 这些方法在精度上优势明显. 然而，正如 Levy 等人

的研究[18]所指出，监督式的上下位关系分类方法存在“词汇记忆”（Lexical memorization）问题，影响了实际预测的准确度. 这个问题发生的原因在于，关系预测分类器往往学习词嵌入在不同维度的权重. 它很可能学习到哪些词是典型的上位词，而非两个概念是否具有上下位关系. 例如，当训练集中有正例“狗-动物”、“猫-动物”、“熊-动物”等，分类器极有可能会认为只要出现“动物”这一典型的上位词，输入的概念对即有上下位关系，从而错误预测“桌子-动物”也是上下位关系. 所以，如果不显式地建模两个概念之间的语义关系，当训练集和测试集中的词汇在语义上差别较大时，这一类方法的测试误差会显著增大.

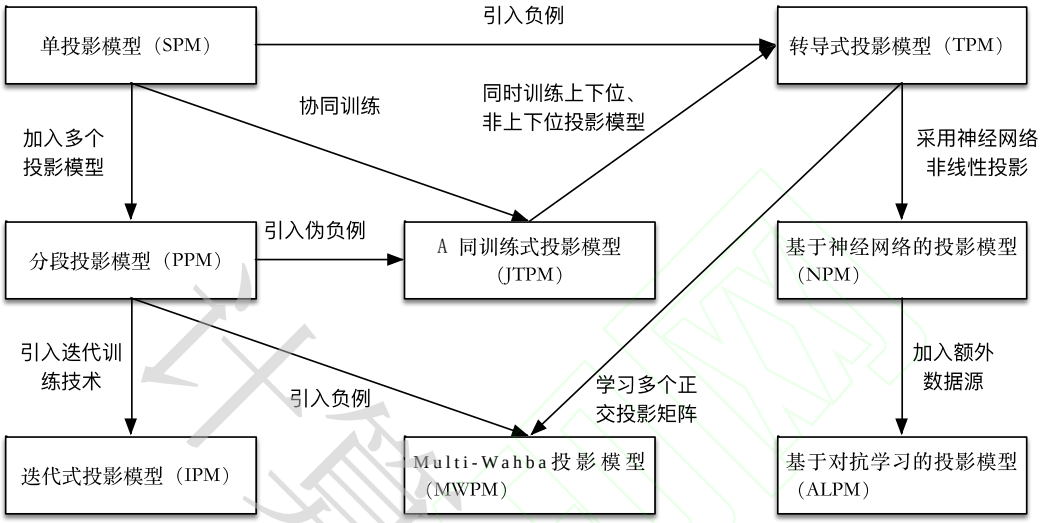


图 1　词嵌入投影模型发展框架

1.2　词嵌入投影模型概述

词嵌入投影模型是分布式学习方法的一种. 它在词嵌入模型的基础上，建模如何将下位词的词向量投影至上位词的词向量，以训练上下位关系预测分类模型. 这一思路在保持分布式方法的优点同时，通过显式地建模关系在词嵌入空间的表达，克服了“词汇记忆”问题. 这一类模型的发展框架如图 1 所示.

词嵌入投影模型的研究起源于 Fu 等人提出的线性投影模型[10]，包括单投影模型（Single Projection Model，SPM）和分段投影模型（Piecewise Projection Model，PPM）. 其主要思想为学习线性投影矩阵，使得下位词的词向量投影至其对应上位词的词向量的误差最小. 迭代式投影模型（Iterative Projection Model，IPM）[11]改进了 Fu 等人的工作[10]，考虑到不同领域中的上下位关系在词嵌入空间中往往具有不同的表示，采用迭代、半监督的方式学习多个从下位词到上位词的关系投影矩阵. 此外，IPM 考虑了中文语言的特性对算法进行优化，它利用中文上下位关系模式和同下位词关系（Co-hyponymy）模式在海量语料库中的聚合统计量来监督这一迭代学习过程，以保证迭代过程中预测精度不下降，避免半监督关系抽取中的“语义漂移”问题[19].

另一个改进方向同时利用正负例（即上下位和非上下位关系元组）训练投影模型. Yamane 等人提出协同训练式投影模型（Jointly Trained Projection Model，JTPM）[48]. 它同时学习拟合一个训练集所需的线性投影模型的数量及其相应参数. 这一模型在优化过程中采用了机器自动生成的伪负例. 我们提出的转导式投影模型（Transductive Projection Model，TPM）[20]不采用正则化的方式，而是同时学习上下位关系和非上下位关系的投影矩阵，并且考虑了训练集和测试集中概念分布的不同，引入基于 TransLP 框架[21]的转导学习正则项，来建模上下位关系映射中的非线性因子. 分段投影的技术也能和上述方法结合起来. Multi-Wahba 投影模型（Multi-Wahba Projection Model，MWPM）[22]综合考虑了这两种投影学习方法的建模优点. 它假设在上下位关系和非上下位关系中都具有K个不同的关系分量，需要分别进行投影建模. 它扩展经典的 Wahba 问题，把每个分量看成在整个数据集上的概率分布，学习模糊投影矩阵，并且在投影过程中加入投影矩阵的正交性约束，使得模型投影学习的优

化目标更合理.

随着深度神经网络技术的发展，神经网络能学习上下位关系和非上下位关系在词嵌入空间的非线性映射. 基于神经网络的投影模型（Neural Projection Model，NPM）用深度神经网络代替 TPM 中的线性投影矩阵. 由于现有分类体系（如英语的 Probase[1]、中文的 CN-Probase[23]、WikiTax[24]）中有大量上下位关系，这些资源可以用于增加现有投影模型的效果. 基于对抗学习的投影模型（Adversarial Learning based Projection Model，ALPM）[25]在 NPM 的基础上同时在已有分类体系的数据和训练集上学习两个基于神经网络的投影模型，并且通过训练数据源鉴别分类器，使得前述两个神经网络相互对抗，达到知识融合、模型互相学习增强的效果.

1.3 下文内容概述及组织结构

虽然词嵌入投影模型已有不少研究，但是仍然缺乏系统化的实验分析. 例如，Biemann 等人[49]通过上位词排序检索（Hypernym Ranked Retrieval）任务来评估其算法精度；模型[20, 22]则更多考虑上下位关系和非上下位关系的分类. 本文在对现有上下位关系预测方法最新研究成果进行综述性分析的基础上，对这一系列基于词嵌入投影模型在统一的数学框架下进行概述和总结. 在实验中，我们进一步对这些方法和经典基线方法进行横向综合对比实验，探讨了不同的词嵌入投影模型在不同学习场景下的优缺点. 最后，本文探讨了面向特定领域的、以及长尾上下位关系预测等研究挑战，提出了未来的研究展望.

本文余下内容的组织结构如下所示：第 2 节概述近年来上下位关系预测的新研究进展；第 3 节建立统一的数学框架，介绍一系列基于词嵌入的投影模型；第 4 节综合比较各种模型在基准数据集上的表现；第 5 节讨论现有研究的不足和未来的研究展望，并且给出本文的结束语.

2 上下位关系预测的最新研究进展

上下位关系预测在 NLP 领域有广泛的研究，早期研究可以参考参见汇总性文献[25]. 2010 年以来的研究可参考本课题组的综述论文[6]. 不失一般性，本节从模式匹配法和分布式学习法出发，概述从海量文本中预测上下位关系的最新研究进展，并且进行深入讨论.

2.1 模式匹配法

几乎所有模式匹配法的研究都可以追溯至 1992 年提出的 Hearst patterns[7]. 即使在近几年，这些模式仍然被广泛应用于构建大规模分类体系，例如 Probase[1]、WebIsADB[26]等. 基于 Hearst patterns 的改进工作主要着眼于两个出发点：提升精准度和覆盖率. 为了提升精准度，研究者常常针对抽取出来的上下位关系候选元组设计置信度评分，以过滤低置信度的关系候选元组. 例如在 Probase 中[1]，作者采用似然性比率从候选概念集中来筛选出最有可能的上位词和下位词. Luu 等人[9]考虑了句子中的语法结构信息进行过滤，从而减少关系抽取的错误. 研究者也常常利用分类器对抽取出的候选关系元组进行进一步判别. 例如，Bansal 等人[27]使用一个概念对中两个词本身的特征，以及他们在维基百科语料库中的统计信息作为特征，预测这两个词具有上下位关系的可能性.

由于 Hearst patterns 具有高度的固定性，更多的研究工作旨在提高模式匹配的覆盖率. 一种常见的方法称为“模式泛化”，即采用更为泛化的语言模式来取代固定的 Hearst patterns. Navigli 和 Velardi[28]提出“Star pattern”的概念，利用通配符来取代模式中最频繁的词. 在 PATTY 系统中[8]，作者加入了更多额外信息，实现模式泛化，包括词性标注、知识本体类别等. 莫媛媛等人[29]考虑了在部分上下位关系的表述中，上位词和下位词距离较远的境况，这些情况很难被现有算法捕获，设计了基于层叠条件随机场（Cascaded Conditional Random Field，CCRF）的模型，解决长距离依赖问题. 陈金栋和肖仰华[30]使用了约束较低的弱模板和概念信息，构建语义模板，同时结合强句法模板，在精准抽取上下位关系的同时，保持了抽取的高召回率. 另一种重要的思路为迭代式抽取，即以少量“种子”作为系统输入，交替挖掘出新的关系元组及语言模式. Kozareva 和 Hovy[31]设计了半监督学习系统，从网络数据中自动抽取上下位关系. Carlson 等人[19]在迭代过程中利用多视图学习算法避免了“语义漂移”问题.

此外，概念本身的语言模式也为上下位关系抽取提出了额外的信息. 例如，从“哺乳动物”的中心词“动物”，我们可以推断出“哺乳动物”和“动物”具有上下位关系. 这些概念本身的语言模式在很多分类体系构建系统中频繁使用，例如 Taxify 系统[32]，以及 Gupta 等人的研究工作[33]. 我们先前基

于维基百科类别的关系抽取研究[34]也利用了上述规则，来提高上下位关系抽取的召回率.

值得注意的是，由于中文语言模式灵活多变，简单的模式匹配方法不能取得满意的效果[10][11]. 因此，常常综合利用模式匹配的各种改进技巧，以取得较好的抽取效果. 程韵如[35]综合利用模式匹配、基于依存句法分析和语义角色特征的条件随机场，以及子句间的并列关系，设计新型混合上下位关系抽取算法. 王长有和常增春[36]叠加了 CCRF 和支持向量机模型，实现了基于句子结构特征的上下位关系抽取分类模型. 与针对英语语境下上下位研究工作相对比，中文的研究工作还需要进一步深入.

2.2 分布式学习法

模式匹配法的最大缺陷在于，只有当两个概念在一句句子中同时出现时，相关的上下位关系才能被抽取出来. 为了解决概念之间的共现稀疏性问题，分布式学习法直接利用两个概念的分布式表示来推断这两个概念是否具有上下位关系.

根据学习方法的不同，分布式学习法可分为非监督的上下位关系度量和监督的上下位关系分类器. 对于任意一个概念对，上下位关系度量计算出一个评分，用于衡量它们有上下位关系可能性. 经典的关系度量包括 WeedsPrec[12]、SLQS[13]等. 由于篇幅限制，本文不再列举经典的研究工作. 读者可参考 Santus 等人[37]对这些度量的评测工作. 实验结果表明，在这些经典的度量中，并没有一个在所有的数据集中比其他方法全都取得更好的效果.

近年来，新的研究工作一般致力于学习上下位关系嵌入（Hypernymy embedding），即设计特别的词嵌入学习算法，使得概念的表示更加有利于上下位关系的预测. HyperScore[14]度量采用基于负采样的上下位关系嵌入学习模型 HyperVec. Chang 等人[38]提出了分布式包含嵌入模型（Distributional inclusion embedding）. 这一工作考虑到如下现象：上位词的上下文一般在语义上包含下位词的上下文. 由于上下位关系一般具有传递性，可以组织成层次化的结构. 诸多研究工作旨在把概念嵌入在双曲嵌入空间中（Hyperbolic embedding space）. 例如 Nickel 和 Kiela[39]提出利用 Lorentz 模型将层次化的概念结构嵌入在双曲嵌入空间中. 另一个类似的工作由 Ganea 等人[40]提出. 由于上位词的语义“蕴含”下位词的语义，他们提出了双曲蕴含圆锥（Hyperbolic entailment cone）的概念，用于同时学习上位词和下位词的词嵌入表示. 在中文领域，刘燊[41]利用中文语义词典《大词林》中大量上下位关系，提出基于字信息的词嵌入学习模型，并且以此为基础，学习上下位关系的向量表示.

监督的上下位关系分类器直接利用两个概念的分布式表示作为分类模型的原始输入，训练二分类的关系分类器. 经典的算法直接利用这两个概念的词向量，或者对其进行简单的算术运算作为分类器的特征[42]. 近年来，Roller 等人提出 asym 模型[15]，同时利用词向量的差值和平方差值作为分类器的特征，更好地利用了上下位关系的反对称性. Turney 和 Mohammad[16]提出了 simDiff 模型，分别计算两个概念与其他词之间的语义相似度，然后把两个语义相似度向量的差值当成特征. Yu 等人[17]利用 Probase[1]中的海量上下位关系数据源，对一个概念分别学习它作为上位词和下位词时不同的分布式表示，并且提出一个最大间隔神经网络模型，用于上下位关系分类. 此外，前述由上下位关系嵌入模型生成的词向量，也常常用于上下位关系分类器的输入[14]，此处不再赘述. 如第 1 节所述，这一类方法往往存在“词汇记忆”问题[18]，限制了其进一步广泛应用. 因此，如何综合利用模式匹配法和分布式学习法，并进行改进成了一大研究焦点.

2.3 更多讨论

从上述概述可以看出，模式匹配法和分布式学习法各有优劣. 学术界对于哪种方式对上下位关系的预测最有效并无明确定论[6].

值得指出的是，这两大类方法的分类不是非黑即白的关系，而是互相结合、互相吸收. 其中，Shwartz 等人[43]采用长短期记忆网络（Long Short Term Network，LSTM）来学习上下位关系的语言模式，同时加入两个概念的词向量作为特征，协同训练混合神经网络进行分类. 马晓军等人[44]将基于词向量的分类方法与基于模式的迭代学习方法进行结合，用于领域上下位关系的获取. 孙佳伟等人[45]提出一个新的词模式嵌入模型，将词嵌入模型和语言模式有效结合，用于解决语言模式稀疏性问题.

此外，语言模式也能和上下位关系度量结合. Roller 等人[46]计算某概念对在海量语料库中与 Hearst patterns 的匹配统计量，并以此提出基于 Hearst patterns 的上下位度量，其精度超过了多个经典上下位度量. Le 等人[47]进一步改进了 Roller 等人[46]的工作，将基于 Hearst patterns 的统计量与双曲空间的词向量结合起来. 上述这些方法都可以把模式匹配法和分布式学习法融合到同一个算法中.

为了在分布式学习框架下解决“词汇记忆”的问题，另一类方法为基于投影的学习方法. Fu 等人提出的线性投影模型[10]是这个方向的开创性工作. 这一方法学习多个线性模型，学习如何将下位词的词向量投影至上位词的词向量. 在模型预测阶段，对于具有上下位关系的概念对，该模型可以较为精准地通过一个概念的词向量推算出另一个概念的词向量；反之，对于不具有上下位关系的概念对，该模型的预测偏差则较大. 所以，上述投影方法可以实现上下位关系分类. Yamane 等人[48]改进了上述方法，同时学习需要线性投影模型的数量及其参数. Biemann 等人[49]利用负采样，在学习上下位关系投影的同时，建模了非上下位关系的语义，从而使关系分类的决策边界更为清晰. 我们的研究工作[11,20,22,25]也属于这一方向，分别从不同角度提升投影学习的效果. 随着词嵌入学习技术的飞速发展，它在上下位关系预测中起的作用将越来越大.

表 1 词嵌入投影模型在多个维度上的比较

模型	投影模型数量	是否采用负样本	是否学习非线性投影	是否考虑中文语言特性	是否利用额外知识源
SPM	1	否	否	否	否
PPM	K	否	否	否	否
IPM	K	否	否	是	是
JTPM	K	*仅采用伪负样本	否	否	否
TPM	2	是	是	是	是
MWPM	2K	是	否	否	否
NPM	2	是	是	否	否
ALPM	4	是	是	否	是

3 词嵌入投影模型

本节介绍用于上下位关系预测的投影模型. 本文首先给出问题定义，之后详细介绍这一系列模型.

3.1 问题定义

记一个概念对为(x_i, y_i)，$\vec{\boldsymbol{x}}_i$和$\vec{\boldsymbol{y}}_i$分别为x_i和y_i的词嵌入表示. D^P、D^N和D^U为三个概念对数据集，分别表示上下位关系集合（训练集正例）、非上下位关系集合（训练集负例）和关系未知的概念对集合. 本文关注的上下位关系预测任务的定义如下所示：

定义 1. 上下位关系预测. 给定两个训练集D^P和D^N作为输入，上下位关系预测的目标是训练模型f，使得f在测试集D^U上做出准确预测.

下文详述词嵌入投影模型定义及其优化. 表 1 给出了各模型在多个维度上的对比.

3.2 单投影模型（SPM）

SPM 是 Fu 等人[10]最早使用的基线投影模型，它采用基本的线性模型假设，即通过线性投影，下位词的词向量可以投影至上位词的词向量. 令$\boldsymbol{M}$为投影矩阵，SPM 的优化目标J_{SPM}定义为：

$$J_{SPM} = \frac{1}{|D^P|}\sum_{i=1}^{|D^P|}\|\boldsymbol{M}\vec{\boldsymbol{x}}_i - \vec{\boldsymbol{y}}_i\|^2$$

3.3 分段投影模型（PPM）

Fu 等人[10]进一步发现，涉及不同主题的上下位关系的语义表示往往不同，例如“狗-动物”和“沙发-家具”的对应的投影矩阵差别较大. PPM 假设存在K个潜在语义关系分量，每个分量对应不同的投影矩阵. 在 PPM 中，先将训练集D^P中的概念对按特征$\vec{\boldsymbol{x}}_i - \vec{\boldsymbol{y}}_i$，利用 K-Means 聚成$K$个簇. 对于每个簇，有$\boldsymbol{M}_k\vec{\boldsymbol{x}}_i \approx \vec{\boldsymbol{y}}_i$,其中，$\boldsymbol{M}_k$为第$k$个投影矩阵. 设第$k$个簇为$C_k$，PPM 的整体优化目标$J_{PPM}$定义为：

$$J_{PPM} = \frac{1}{K}\sum_{k=1}^{K}\sum_{(x_i,y_i)\in C_k}\|\boldsymbol{M}_k\vec{\boldsymbol{x}}_i - \vec{\boldsymbol{y}}_i\|^2$$

3.4 迭代式投影模型（IPM）

IPM 的优化目标与J_{PPM}相似，它额外加入了$\vec{\boldsymbol{b}}_k$作为偏置向量，定义为：

$$J_{IPM} = \frac{1}{K}\sum_{k=1}^{K}\sum_{(x_i,y_i)\in C_k}\left\|\boldsymbol{M}_k\vec{\boldsymbol{x}}_i + \vec{\boldsymbol{b}}_k - \vec{\boldsymbol{y}}_i\right\|^2$$

根据定义 1，上下位关系预测的目标在于对D^U

的关系类别进行准确预测. 然而，当D^U中概念对的语义分布与D^P显著不同时，预测精度将会明显下降. IPM 采用迭代式、半监督学习方法渐进式更新参数的值，其算法思想如图 2 所示：

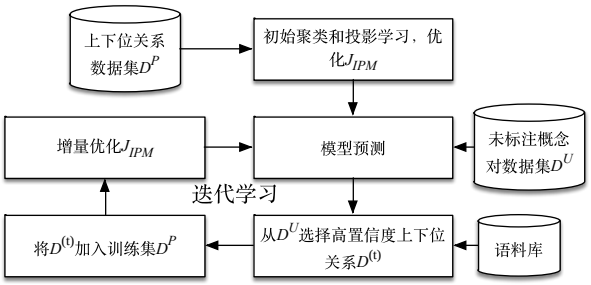


图 2　迭代式投影模型算法框架

设t为迭代序数，迭代学习过程的参数初始化值通过梯度下降优化J_{IPM}得到. 当$t>1$时，给定已训练的模型，对D^U中的未标注概念对进行预测. 同时，利用中文语料库，计算这些概念对匹配相关语言模式的统计量，选择出最为可信的上下位关系概念对作为增量的训练数据. 在第t个迭代中，令选出的增量数据集为$D^{(t)}$，训练数据集相应的更新规则为：$D^P \leftarrow D^P \cup D^{(t)}$. 更新数据集$D^P$后，IPM 也增量更新其模型参数. 设$\vec{\boldsymbol{c}}_k^{(t)}$为第$t$个迭代第$k$个簇的中心，其更新规则如下：

$$\vec{\boldsymbol{c}}_k^{(t+1)} \leftarrow \vec{\boldsymbol{c}}_k^{(t)} + \lambda \frac{1}{\left|D_k^{(t)}\right|} \sum_{(x_i,y_i)\in D_k^{(t)}} \left(\vec{\boldsymbol{x}}_i - \vec{\boldsymbol{y}}_i - \vec{\boldsymbol{c}}_k^{(t)}\right)$$

其中$D_k^{(t)}$为$D^{(t)}$中属于第k个簇的概念对集合，λ为更新率. 根据更新后簇中心，重新分配聚类结果，并优化J_{IPM}以求得$\boldsymbol{M}_k$和$\vec{\boldsymbol{b}}_k$的更新值. 当迭代结束后，可得到适应D^U的线性投影模型.

表 2　IPM 中采用的中文语言模式示例

模式类别	示例
Is-A 模式	x_i是一种y_i
*针对一个概念对(x_i, y_i)	x_i是y_i之一
Such-As 模式	y，例如x_i、x_j
*针对多个概念对(x_i, y)、(x_j, y)	y，特别是x_i、x_j
	x_i、x_j等y
同位词关系模式	x_i、x_j等
*针对一个概念对(x_i, x_j)	x_i和x_j
	x_i以及x_j

在此研究工作中，我们针对中文语言的特性进行定制. 尽管中文语言具有高度灵活性，但仍有部分语言模式可以为选择高置信度的概念对提供有效的统计量. 表 2 列举了一部分中文语言模式，其中 Is-A 和 Such-As 模式为两个概念对存在上下位关系提供了正面证据；同下位词关系模式则相反，表征这两个概念可能都是某概念的下位词，因而他们之间极有可能不存在上下位关系. 对于数据集D^U中概念对(x_i, y_i)，根据上述模式在语料库的统计量，计算正向得分$PS(x_i, y_i)$和负向得分$NS(x_i, y_i)$. 从D^U中选择$D^{(t)}$的这一关系选择的过程被建模成带成本的最大覆盖问题（Budgeted maximum coverage problem）[50] . 这一问题可以通过贪心算法进行近似优化. 算法的实现细节请参阅文献[11].

3.5　协同训练式投影模型（JTPM）

由于 PPM 和 IPM 都采用启发式的方法对上下位关系元组进行聚类，学习多个投影矩阵，这两个步骤分别计算，不能保证全局最优. Yamane 等人[48]提出的 JTPM 方法动态地对这两个步骤进行迭代计算. 其中，一个上下位关系元组(x_i, y_i)与簇C_k的相似度采用内积计算，定义为：

$$sim_k(x_i, y_i) = \sigma(\boldsymbol{M}_k \vec{\boldsymbol{x}}_i \cdot \vec{\boldsymbol{y}}_i + \vec{\boldsymbol{b}}_k)$$

其中，$\sigma(\cdot)$是 Sigmoid 函数. 此外，JTPM 采用机器自动生成的伪负例来改进投影学习的结果. 其目标函数J_{JTPM}如下所示：

$$J_{JTPM} = \frac{1}{K}\sum_{k=1}^{K} \sum_{(x_i,y_i)\in C_k} (\log sim_k(x_i, y_i) + \sum_{m=1}^{M} \log(1 - sim_k\left(x_i, y_i^{(m)}\right)))$$

其中，对于每个上下位关系对(x_i, y_i)，算法都自动生成M个伪负例$\left(x_i, y_i^{(1)}\right)$、……、$\left(x_i, y_i^{(M)}\right)$. 这一目标在优化正例投影学习的同时，增加了正例和负例的区分度.

3.6　转导式投影模型（TPM）

虽然前述模型刻画了上下位关系在词嵌入空间的投影，但是它的投影仅限于线性，非线性投影分量没有得到有效建模. 例如，如果已知“美国-国家”是上下位关系，由于“美国”和“加拿大”语义高度相似，在预测时，模型应该也可推断“加拿大-国家”是上下位关系. 在这个例子中，模型需要将“美国”和“加拿大”的这两个不同但相似的词向量都映射到“国家”的词向量. 因此，在现有模

型中引入非线性投影分量是极度必要的.

TPM 方法采用了转导学习的框架，建模了上下位关系的线性投影和非线性映射关系，并且允许加入专家制定的语言规则. 其算法框架图如图 3 所示.

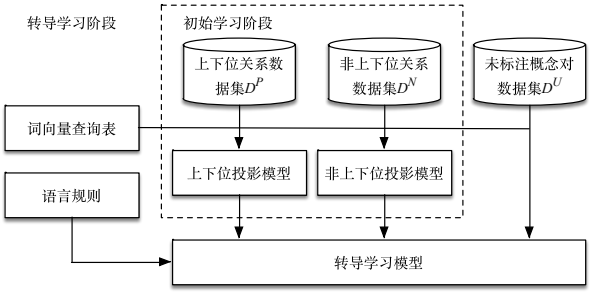


图 3 转导式投影模型算法框架

在初始学习阶段，分别学习两个线性投影矩阵. 对于$(x_i,y_i)\in D^P$，线性投影模型为$M^P\vec{\boldsymbol{x}}_i\approx\vec{\boldsymbol{y}}_i$；对于$(x_i,y_i)\in D^N$，类似的，我们有$M^N\vec{\boldsymbol{x}}_i\approx\vec{\boldsymbol{y}}_i$. 根据学习到的两个投影矩阵$M^P$和$M^N$，我们对未标注概念对$(x_i,y_i)\in D^U$进行预测评分$s_i\in(-1,1)$：

$$s_i=\tanh(\|M^N\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i\|-\|M^P\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i\|)$$

其中，s_i越高，$\|M^N\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i\|$越大，$\|M^P\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i\|$越小。这表示$x_i$和$y_i$具有上下位关系的可能性越大.

在转导学习阶段，我们同时考虑训练集和测试集中概念对的相似度. 设$p_i=(x_i,y_i)$和$p_j=(x_j,y_j)$为两个概念对，定义其上下位关系的相似度为：

$$sim(p_i,p_j)=\begin{cases}\cos(\vec{\boldsymbol{x}}_i,\vec{\boldsymbol{x}}_j),y_i=y_j\\0,y_i\neq y_j\end{cases}$$

由此可见，$sim(p_i,p_j)$可以建模“美国-国家”和“加拿大-国家”在关系预测中类别的内在关联性.

设$\boldsymbol{F}$为训练集D^P、D^N和测试集D^U中所有概念对中存在上下位关系的最终预测向量，维度为$|D^P|+|D^N|+|D^U|$，每个元素的取值范围为$[-1,1]$. $\boldsymbol{S}$是初始预测值向量，对于任一概念对$(x_i,y_i)\in D^P\cup D^N\cup D^U$，根据训练数据的人工标注结果和模型初始预测，我们设定如下：

$$s_i=\begin{cases}1,(x_i,y_i)\in D^P\\-1,(x_i,y_i)\in D^N\\\tanh(\|M^N\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i\|-\|M^P\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i\|),(x_i,y_i)\in D^U\end{cases}$$

其中，$\boldsymbol{W}$用于调节训练数据和测试数据的相对权重，$\boldsymbol{R}$是基于语言学规则的预测值向量（详述见下文）. TPM 的整体优化目标是最小化如下函数的值：

$$J_{TPM}=\|\boldsymbol{W}(\boldsymbol{F}-\boldsymbol{S})\|_2+\|\boldsymbol{F}-\boldsymbol{R}\|_2+\frac{\mu_1}{2}\boldsymbol{F}^T\boldsymbol{\Phi}^{-1}\boldsymbol{F}+\frac{\mu_2}{2}\|\boldsymbol{F}\|_2$$

在上式中，μ_1和μ_2为预定义的正则化参数. $\boldsymbol{\Phi}$为基于 TransLP 框架[21]的非线性正则化矩阵，其中$\Phi_{i,j}=sim(p_i,p_i)$. 这一框架假设$\boldsymbol{F}$满足协方差为$\boldsymbol{\Phi}$的多维高斯分布. 因此，正则项$\boldsymbol{F}^T\boldsymbol{\Phi}^{-1}\boldsymbol{F}$以转导学习的方式加入了上下位关系的非线性映射，在不知道数据集D^U中概念对标签的情况下，也能实现模型的推理.

此外，在 TPM 中，$\|\boldsymbol{F}-\boldsymbol{R}\|_2$的存在可以允许这个模型加入任意语言学规则. 例如，对于需要预测得分的概念对“哺乳动物-动物”，假设当前模型预测评分$f_i=0.7$. 同时，这一概念对匹配了人为制定的规则“如果一个概念的中心词和另一个概念相同，则有上下位关系”，其置信得分为$r_i=0.95$（一般可以在训练集中学到，或由专家制定）. 优化$\|\boldsymbol{F}-\boldsymbol{R}\|_2$会使最终评分更接近0.95. 如果测试集中某一概念对不匹配任何语言规则，只要锁定$r_i=f_i$，这一项就不必优化. 详情可以参考文献[20].

3.7 Multi-Wahba投影模型（MWPM）

MWPM 结合了分段投影的技术和 TPM 对于正负例的投影建模方法，是经典 Wahba 问题（Wahba's problem）的拓展[52]. 它作出如下假设：上下位关系和非上下位关系中都存在多个不同的关系分量，需要进行投影建模. 与 IPM 中的采用的“硬聚类”不同，在 MWPM 中，训练集中的每个概念对对于每个分量具有不同的权重.

以学习上下位关系的投影为例. 首先使用 K-Means 算法对D^P的概念对聚类. 令$\vec{\boldsymbol{c}}_k$为第k个簇的中心. 概念对$(x_i,y_i)\in D^P$属于第k个簇的权重为：

$$a_{i,k}^P=\frac{\cos(\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i,\vec{\boldsymbol{c}}_k)}{\sum_{(x_j,y_j)\in D^P}\cos(\vec{\boldsymbol{x}}_j-\vec{\boldsymbol{y}}_j,\vec{\boldsymbol{c}}_k)}$$

之后，MWPM 学习K个模糊投影矩阵$\boldsymbol{M}_1^P,\cdots,\boldsymbol{M}_K^P$，用于上下位关系投影. 此外，当使用的词向量归一化后，如果投影矩阵为正交矩阵，投影之后的向量也是归一化的[51]. 所以，在本研究工作中，我们加入投影矩阵正交性约束. MWPM 对于上下位关系投影学习的目标函数J_{MWPM}^P可以定义为：

$$J_{MWPM}^P=\frac{1}{2}\sum_{k=1}^{K}\sum_{(x_i,y_i)\in D^P}a_{i,k}^P\left\|\boldsymbol{M}_k^P\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i\right\|^2$$

$$s.t.\boldsymbol{M}_k^{P\,T}\boldsymbol{M}_k^P=\boldsymbol{I}\ (k=1,\cdots K)$$

其中，$\boldsymbol{I}$表示单位矩阵. 值得注意的是，J_{MWPM}^P的优化结果对于不同的k是完全独立的. 所以，我们只需

要分别计算每个投影矩阵$\boldsymbol{M}_k^P\ (k=1,\cdots K)$的值即可. 在应用数学领域，这个问题可以看成三维Wahba问题（Wahba's problem）在高维空间的拓展[52]. 我们拓展文献[53]中基于SVD的闭式解，如算法1所示：

算法 1. MWPM投影矩阵计算算法.

输入：所有$(x_i,y_i)\in D^P$的词向量$(\vec{\boldsymbol{x}}_i,\vec{\boldsymbol{y}}_i)$，权重$a_{i,k}^P$

输出：投影矩阵$\boldsymbol{M}_k^P$

1. $\boldsymbol{B}_k=\sum_{(x_i,y_i)\in D^P} a_{i,k}^P\vec{\boldsymbol{y}}_i\vec{\boldsymbol{x}}_i^{\ T}$
2. $SVD(\boldsymbol{B}_k)=\boldsymbol{U}_k\boldsymbol{\Sigma}_k{\boldsymbol{V}_k}^T$
3. $\boldsymbol{R}_k=\mathrm{diag}(1,\cdots,1,\det(\boldsymbol{U}_k)\det(\boldsymbol{V}_k))$
4. $\boldsymbol{M}_k^P=\boldsymbol{U}_k\boldsymbol{R}_k{\boldsymbol{V}_k}^T$

这一算法的正确性证明详见文献[22].

与之类似，MWPM对同样D^N中的概念对进行聚类，并计算每个概念对$(x_i,y_i)\in D^N$的权重$a_{i,k}^N$. 之后，它学习非上下位关系的K个模糊投影矩阵$\boldsymbol{M}_1^N,\cdots,\boldsymbol{M}_K^N$. 其优化目标$J_{MWPM}^N$定义如下：

$$J_{MWPM}^N=\frac{1}{2}\sum_{k=1}^{K}\sum_{(x_i,y_i)\in D^N}a_{i,k}^N\left\|\boldsymbol{M}_k^N\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i\right\|^2$$

$$s.t.\,{\boldsymbol{M}_k^N}^T\boldsymbol{M}_k^N=\boldsymbol{I}\ (k=1,\cdots K)$$

J_{MWPM}^N的优化过程与优化J_{MWPM}^P相同.

当$\boldsymbol{M}_1^P,\cdots,\boldsymbol{M}_K^P$和$\boldsymbol{M}_1^N,\cdots,\boldsymbol{M}_K^N$学习完成后，MWPM训练一个神经网络分类器用于关系分类. 对于任意一个概念对$(x_i,y_i)\in D^P\cup D^N$，都进行两种关系的投影计算：$\boldsymbol{M}_1^P\vec{\boldsymbol{x}}_i,\cdots,\boldsymbol{M}_K^P\vec{\boldsymbol{x}}_i$和$\boldsymbol{M}_1^N\vec{\boldsymbol{x}}_i,\cdots,\boldsymbol{M}_K^N\vec{\boldsymbol{x}}_i$，并且计算其差值$\boldsymbol{M}_1^P\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i,\cdots,\boldsymbol{M}_K^P\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i$和$\boldsymbol{M}_1^N\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i,\cdots,\boldsymbol{M}_K^N\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i$作为特征训练基于前馈神经网络的关系分类器.

根据J_{MWPM}^P和J_{MWPM}^N的定义，如果$(x_i,y_i)\in D^P$，向量$\boldsymbol{M}_1^P\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i,\cdots,\boldsymbol{M}_K^P\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i$元素的值较小；反之，如果$(x_i,y_i)\in D^N$，向量$\boldsymbol{M}_1^N\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i,\cdots,\boldsymbol{M}_K^N\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i$元素的值较小. 所以，这一特征计算算法能较好地区分上下位关系与非上下位关系. MWPM的神经网络示意图如图4.

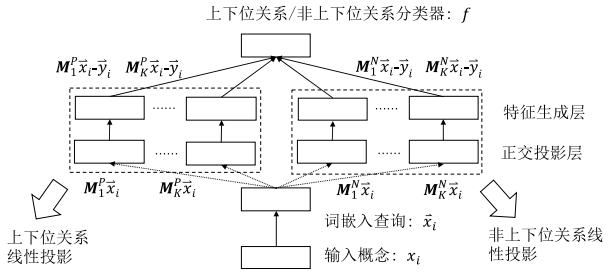


图4 模糊正交投影模型神经网络示意图

3.8 基于神经网络的投影模型（NPM）

随着深度学习的进一步发展，深度神经网络可以有效地学习上下位关系和非上下位关系的复杂非线性投影. 图5展示了文献[25]中采用了基础神经网络投影模型. 这一模型同时在训练集D^P和D^N上进行训练. 它的损失函数L_D定义为：

$$L_D=\mathbb{E}_{(x_i,y_i)\sim D^P}\left\|H(\vec{\boldsymbol{x}}_i;\theta_D^P)-\vec{\boldsymbol{y}}_i\right\|^2+\mathbb{E}_{(x_i,y_i)\sim D^N}\left\|H(\vec{\boldsymbol{x}}_i;\theta_D^N)-\vec{\boldsymbol{y}}_i\right\|^2$$

其中$\mathbb{E}$表示期望，$H(\vec{\boldsymbol{x}}_i;\theta_D^P)$和$H(\vec{\boldsymbol{x}}_i;\theta_D^N)$分别为投影模型学习的$x_i$的上位词和非上位词的预测词向量，$\theta_D^P$和$\theta_D^N$为相应模型参数. 由此可见，NPM使用多任务学习技术，同时使上下位关系和非上下位关系投影的误差最小化.

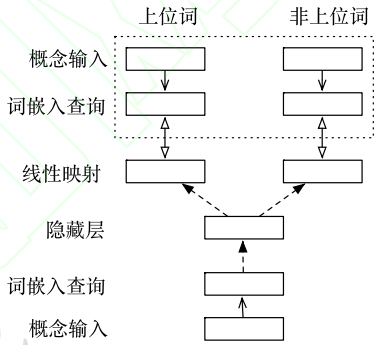


图5 神经网络投影模型示意图

最后，对于训练集D^P和D^N，根据表3所列举特征，训练用于上下位关系分类的SVM分类器，从而在测试集上完成预测.

表3 SVM采用的特征列表

词嵌入	词嵌入的范数
$\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i$	$\|\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i\|_1$、$\|\vec{\boldsymbol{x}}_i-\vec{\boldsymbol{y}}_i\|_2$
$H(\vec{\boldsymbol{x}}_i;\theta_D^P)-\vec{\boldsymbol{y}}_i$	$\|H(\vec{\boldsymbol{x}}_i;\theta_D^P)-\vec{\boldsymbol{y}}_i\|_1$、$\|H(\vec{\boldsymbol{x}}_i;\theta_D^P)-\vec{\boldsymbol{y}}_i\|_2$
$H(\vec{\boldsymbol{x}}_i;\theta_D^N)-\vec{\boldsymbol{y}}_i$	$\|H(\vec{\boldsymbol{x}}_i;\theta_D^N)-\vec{\boldsymbol{y}}_i\|_1$、$\|H(\vec{\boldsymbol{x}}_i;\theta_D^N)-\vec{\boldsymbol{y}}_i\|_2$

3.9 基于对抗学习的投影模型（ALPM）

NPM直接使用人工标注的数据作为训练数据，而自动构建的大规模分类体系中有大量上下位关系，将分类体系中的上下位关系融入投影模型，可以从知识迁移的角度提升预测精度.

ALPM采用分类体系增强的对抗学习框架[25]，它在已有分类体系的上下位关系数据和训练集上学习两个同构的基于神经网络的投影模型，并且通过训练两个数据源鉴别分类器，使得前述两个投影神经网络相互对抗，从而使在训练集上学习的投影

神经网络逐步融合分类体系中的知识. 其神经网络架构图如图 6 所示.

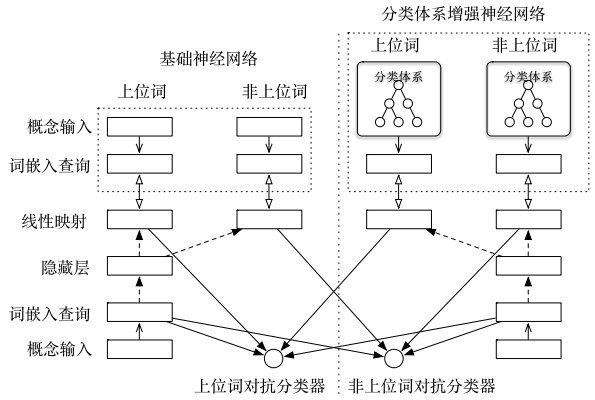


图 6 基于分类体系增强的对抗学习神经网络

具体而言，这一神经网络模型包括了两个子网络：基础神经网络和分类体系增强的神经网络. 基础神经网络同 NPM. 与此同时，通过对分类体系中的数据进行采样，获得上下位关系和非上下位关系训练集T^P和T^N. 分类体系增强的神经网络的损失函数定义为:

$$L_T = \mathbb{E}_{(x_i,y_i)\sim T^P}\|H(\vec{\boldsymbol{x}}_i;\theta_T^P)-\vec{\boldsymbol{y}}_i\|^2 + \mathbb{E}_{(x_i,y_i)\sim T^N}\|H(\vec{\boldsymbol{x}}_i;\theta_T^N)-\vec{\boldsymbol{y}}_i\|^2$$

其中，θ_T^P和θ_T^N为模型在T^P和T^N上学习的参数.

为了实现对抗学习的效果，模型同时训练两个对抗分类器. 给定一个概念的词嵌入作为输入，分类器的目标是区分一个实体的上位词或者非上位词的词嵌入是通过哪一个子网络产生的. 这两个分类的训练思想类似条件对抗生成网络[54]. 在下式中，L_P描述了上位词对抗分类器的二分类损失函数，用于区分给定$\vec{\boldsymbol{x}}_i$生成的上位词词嵌入的来源(基础神经网络或分类体系增强神经网络)；L_N为非上位词对抗分类器的损失函数：

$$L_P = \mathbb{E}_{(x_i,y_i)\sim D^P} log(1-\delta(H(\vec{\boldsymbol{x}}_i;\theta_D^P),\vec{\boldsymbol{x}}_i)) + \mathbb{E}_{(x_i,y_i)\sim T^P} log\,\delta(H(\vec{\boldsymbol{x}}_i;\theta_T^P),\vec{\boldsymbol{x}}_i)$$

$$L_N = \mathbb{E}_{(x_i,y_i)\sim D^N} log(1-\delta(H(\vec{\boldsymbol{x}}_i;\theta_D^N),\vec{\boldsymbol{x}}_i)) + \mathbb{E}_{(x_i,y_i)\sim T^N} log\,\delta(H(\vec{\boldsymbol{x}}_i;\theta_T^N),\vec{\boldsymbol{x}}_i)$$

上述投影神经网络和对抗分类器迭代训练，直到这两个对抗分类器无法区分词嵌入来自于哪个投影神经网络为止. 当这一模型训练终止后，在D^P和D^N上训练 SVM 分类器，这一步骤与 NPM 相同，此处不再赘述.

4 综合实验评测与分析

在本节中，我们对词嵌入投影模型在多个数据集上进行综合横向评测，并且把这些方法与诸多基线方法对比. 此外，我们旨在探索不同词嵌入投影模型在不同学习场景下的优劣性.

4.1 数据集与实验设置

由于中文维基百科数据集比较小，所训练的词嵌入模型质量较低，所以我们从百度百科中获得更大的中文语料库. 我们首先爬取中文百度百科的词条正文，一共获得包括分词后约 10 亿词的中文语料库. 我们用此语料库训练中文词级别的 Word2Vec 的 Skip-Gram 模型[55]，词向量的维度为 100. 由于英语语言的构词法对学习英语的词嵌入有很大帮助，我们在英语维基百科语料库上，训练 fastText 词嵌入模型[56]，其维度为 300.

本研究使用公开的数据集对上下位关系预测模型的效果进行评测. 其中，英语使用两个通用领域的数据集 BLESS[57]和 ENTAILMENT[58]、三个特定领域的数据集：ANIMAL、PLANT 和 VEHICLE. 这三个数据集来自 Velardi 等人构建的领域分类体系[59]，由 Luu 等人采样的得到包括正负例的数据集[9]。中文使用两个公开数据集：FD 和 BK，分别由 Fu 等人[10]和我们的研究[11]所构建. 这些数据集的统计数据汇总信息见表 4.

表 4 评测数据集统计信息汇总

数据集	#上下位关系	#非上下位关系
英语数据集（通用领域）		
BLESS	1337	13210
ENTAILMENT	1385	1385
英语数据集（特定领域）		
ANIMAL	4169	8471
PLANT	2266	4520
VEHICLE	283	586
中文数据集		
FD	1391	4294
BK	3870	3582

4.2 英语通用数据集上的实验结果与分析

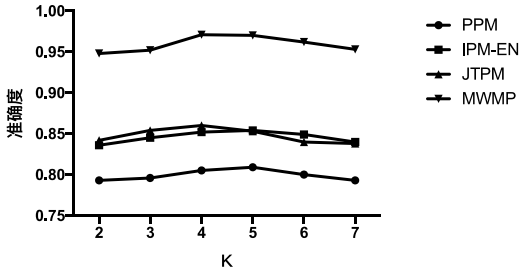
由于 IPM 和 TPM 两个模型为中文语言进行了定制，为了在英语数据集上进行评测，我们去除了 IPM 中的中文模式统计计算模块和 TPM 的优化项$\|\boldsymbol{F}-\boldsymbol{R}\|_2$，把他们用于英语数据集的评测. 我们把

这两种模型记为IPM-EN和TPM-EN.我们在BLESS和ENTAILMENT这两个英语基准数据集上用相同的词向量作为输入，评测所有词投影模型的实验效果. 同时，我们也将Mikolov等人[57]、Yu等人[17]、Luu等人[9]及Nguyen等人[14]提出的算法作为基线算法. 基线方法的具体的实验设置参见文献[9，14]. 实验采用Leave-One-Out的实验方法进行评测，其评测指标是准确度（Accuracy），具体步骤与文献[14]等中的方法保持一致.

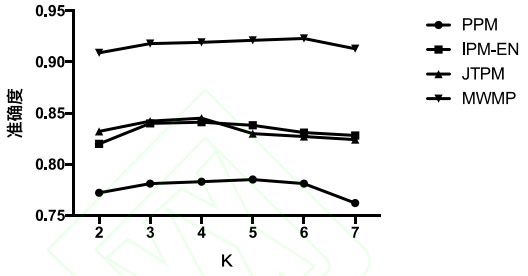
PPM、IPM-EN、JTPM和MWMP需要设置参数K的值，我们设置默认值$K=4$，并且在下文中汇报参数变化对实验效果的影响. IPM默认运行10个迭代，JTPM的默认设置为$M=5$. 我们采用Adam算法训练MWPM、NPM和ALPM中神经网络，在NPM和ALPM的神经网络中只使用一层隐藏层，SVM分类器采用RBF核函数. 读者可以在参考文献[22，25]中查阅参数和实验设置对性能的影响. ALPM采用Microsoft Concept Graph（Probase系统的公开数据集[1]）中的上下位关系（包括284万个上下位关系元组）和Microsoft Concept Graph中的随机匹配词对作为正负例，训练分类体系增强神经网络，并通过对抗学习改进基础神经网络的效果.

表5汇总了所有投影模型以及所有基线方法在所有这两个英语数据集上的实验效果. 其中Nguyen等人[14]的方法取得了非投影算法的最佳效果，分别为0.94和0.91. 通过分析实验结果，我们可以发现，在这些投影模型中，MWPM、NPM和ALPM都在BLESS数据集上超过了非投影算法的最佳效果，而TPM-EN的实验效果与非投影算法的最佳效果比较接近. 在ENTAILMENT数据集上，MWPM和ALPM也超过了Nguyen等人[14]的方法，而NPM和TPM-EN取得了类似的实验效果. 这一实验结果可以验证上述词嵌入投影模型的有效性. 另外，通过对比经典词嵌入投影模型SPM、PPM等与MWPM、NPM和ALPM等新提出的模型，也可以看出，前文概述的改进措施（如采用负样本和更复杂的投影学习方式）是高度有效的.

在图7中，我们给出了参数K对于PPM、IPM-EN、JTPM和MWMP的实验性能的影响. 从中可以看出，这四个模型对于参数K的变化并不非常敏感. 当K的设置过于小时，模型会退化成SPM，导致预测精准度下降；当K过大时，模型的参数空间显著增大，可能会导致模型过拟合. 所以，预测精度随着K的增大，呈现先轻微上升后轻微下降趋势.



（a）数据集：BLESS



（b）数据集：ENTAILMENT

图7 参数K对于模型性能在两个英语数据集上的影响

表5 词嵌入投影模型在英语通用数据集上的精准度评测

方法	BLESS	ENTAILMENT
非词嵌入投影方法		
Mikolov等人	0.84	0.83
Yu等人	0.90	0.87
Luu等人	0.93	0.91
Nguyen等人	0.94	0.91
词嵌入投影模型		
SPM	0.78	0.76
PPM	0.80	0.78
IPM-EN	0.85	0.84
JTPM	0.86	0.84
TPM-EN	0.90	0.89
MWPM	**0.97**	**0.92**
NPM	**0.96**	0.90
ALPM	**0.97**	**0.92**

4.3 英语领域数据集上的实验结果与分析

我们进一步分析不同的模型在英语领域数据集上的实验效果. 在ANIMAL、PLANT和VEHICLE三个领域数据集中，有部分概念由多个英语单词组成（例如American tree、half track等），而不是词语，因此我们采用Luu等人[9]的做法，将词组中各个词的词向量进行平均，作为这个词组在词嵌入空间的表示. 我们采用与通用数据集上实验相同的实

验设置和调参方法进行实验. 我们同样采用 Leave-One-Out 的评测方法，细节详见文献[9]. 实验结果和基线方法的结果汇总于表 6.

从实验结果中可以看出，MWPM、NPM、ALPM 等模型的实验效果超过了文献中汇报的非词嵌入投影方法中的最高实验效果（即 Luu 等人[9]的方法）. 我们进一步分析各模型在不同类型数据集上的预测精度差异性. MWPM 对通用领域数据集更为有效，这是因为 MWPM 学习的投影矩阵的参数空间比较大，而通用领域一般有较大的训练集. 与之不同，ALPM 的对抗学习技术对较小的数据集（例如 VEHICLE）的效果提升比较明显，因为其分类体系增强神经网络的训练可以使用 Probase 中的额外数据源，起到了数据增强的效果. 综上所述，我们提出的多个词嵌入投影模型的预测精度不仅超越了经典模型（如 SPM、PPM 等），还高于现有的其他种类的上下位关系预测算法.

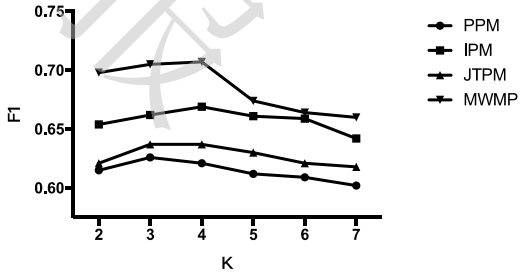
表 6 词嵌入投影模型在英语领域数据集上的精准度评测

方法	ANIMAL	PLANT	VEHICLE
非词嵌入投影方法			
Mikolov 等人	0.80	0.81	0.82
Yu 等人	0.67	0.65	0.70
Luu 等人	0.89	0.92	0.89
Nguyen 等人	0.83	0.91	0.83
词嵌入投影模型			
SPM	0.79	0.76	0.75
PPM	0.82	0.82	0.76
IPM-EN	0.85	0.87	0.78
JTPM	0.86	0.84	0.82
TPM-EN	0.89	0.90	0.84
MWPM	**0.92**	0.92	0.87
NPM	0.89	0.92	**0.92**
ALPM	**0.92**	**0.94**	**0.93**

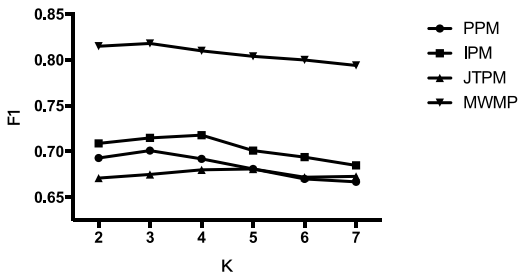
4.4 中文数据集上的实验结果与分析

我们在两个公开的中文数据集 BK 和 FD 上评测了词嵌入投影模型的预测精度. 在这组实验中，我们采用文献[11]的实验设置，将这两个数据集随机分成 5 等份，分别作为测试集进行 5 折交叉验证，最后将实验预测结果平均，用 Precision、Recall 和 F1 作为评测指标. 我们首先利用两个概念词向量的拼接、相减和相加作为特征训练关系预测分类器. 他们也在文献[13，42，43]等中作为强基线方法. 这些方法在表 5 中分别记为 Concat、Offset 和 Addition，我们也采用 Li 等人[24]提出的中文上下位关系预测算法作为基线方法. 对于本文提出的算法，IPM 的参数设置为$t=5$，每次选出 Top-50 个元组作为候选增量数据，加入数据集. 在 TPM 中，我们设置$\mu_1=\mu_2=10^{-4}$，实验细节和参数调整参见文献[11]. 我们同样调整模型 PPM、IPM、JTPM 和 MWPM 中参数K的值，实验结果参见图 8. 因为中文分类体系的规模一般不大，不适合运用于 ALPM 的模型，所以这个模型退化为 NPM，下文不再讨论. 其余模型与中文的语言特性无关，所以我们采取与英语实验相同的设置. 实验结果见表 7.

从实验结果中，我们可以看出 TPM、MWPM 和 NPM 比所有基线算法在两个数据集上都有明显的提升. 特别地，与英语实验结果进行对比，TPM 在中文上的提升最为明显；这是因为 TPM 模型中可以加入中文语言规则. IPM 在上下位关系分类的精度比 TPM、MWMP 等模型低. 与文献[11]中的实验对比可以发现，IPM 更适合在仅有少量正例，有大量未标注数据的 PU 学习（Positive Unlabeled Learning）的情况；而 TPM、MWMP 等模型更适合传统的关系二分类任务. 综合比较 TPM、MWPM 和 NPM，这三个方法在 F1 上相差不大. 其中，TPM 在 FD 数据集表现最好，而 NPM 更适合 BK 数据集. 与多个基线方法（例如 Concat、Offset 和 Addition[13, 42, 43]）和针对中文数据的算法（例如 Li 等人[24]）相比，我们提出的 TPM、MWPM 和 NPM 等模型也有明显的优势.



（a）数据集：FD



（b）数据集：BK

图 8 参数 K 对于模型性能在两个中文数据集上的影响

4.5 对实验结果的综合分析与讨论

综合分析各个词嵌入投影模型在中英文数据集上的实验结果，我们可以得出以下结论：

1. MWPM、NPM 较复杂的投影模型无论在中文还是英语数据集上效果都明显优于强基线算法，说明这些方法对上下位关系预测高度有效.
2. 整体而言，上下位关系预测算法在英语数据集上比中文更精确. 其原因在于中文在语言模式上更加灵活，而由于其语法的灵活性，词嵌入学习的效果也比英语更差. 这一现象也在文献[10, 11, 20, 60]等中也有相关实验验证与充分讨论.
3. 比较不同词嵌入投影模型在相同数据集上的效果，可以发现通过引入图 1 中所示的改进技术，词嵌入投影模型的精准度从最初的 SPM、PPM 到 MWPM、NPM、ALPM 等不断提高
4. 考虑 IPM 和 TPM 两个为中文定制的投影模型. IPM 主要用于半监督学习，在完全监督的上下位关系分类任务中精度不高；而 TPM 加入了中文语言规则，尽管投影模型的参数量比 MWPM 等模型更小，在中文数据集上也能取得较好的效果. 这说明中文语言学规则对中文上下位关系的预测仍有较大的贡献.
5. 通过引入外部知识源，ALPM 的精度比较 NPM 有较大提升，说明额外的知识源对模型学习有帮助. 目前这一工作仅在英语环境下有详细的研究，对于中文语言仍然需要更多探索.

表 7 词嵌入投影模型在中文数据集上的综合评测（%）

方法	数据集：FD			数据集：BK		
	Precision	Recall	F1	Precision	Recall	F1
非词嵌入投影方法						
Concat	67.7	**75.2**	69.7	80.3	75.9	78.0
Offset	71.9	60.6	65.7	78.4	60.7	68.4
Addition	65.3	60.7	62.9	72.7	65.6	68.9
Li 等人	54.3	38.4	45.0	61.2	47.5	53.5
词嵌入投影模型						
SPM	64.1	56.0	59.8	71.4	64.8	67.9
PPM	66.4	59.3	62.6	72.7	67.5	70.0
IPM	**69.3**	64.5	66.9	73.9	69.8	71.8
JTPM	65.2	62.3	63.7	70.9	65.4	68.0
TPM	**72.8**	70.5	**71.6**	**83.6**	**80.6**	**82.1**
MWPM	**71.3**	69.8	**70.5**	**82.5**	**81.2**	**81.8**
NPM	**71.8**	68.8	**70.3**	**84.1**	**81.3**	**82.7**

5 结论与未来研究展望

在本文中，我们概述了一系列基于词嵌入的投影模型，用于上下位关系预测. 这些基于词嵌入的投影模型包括经典的模型 SPM、PPM、IPM、JTPM，乃至新提出的模型 TPM、MWPM、NPM、ALPM. 这些模型学习上下位关系在词嵌入空间的投影，利用投影结果进行上下位关系关系分类. 在实验中，我们在多个中文和英语的基准数据集上对这些模型在统一的实验环境下利用相同词向量进行评测，实验结果证实了基于词嵌入的投影模型对上下位关系预测是高度有效的. 此外，由于中文语言的复杂性，中文上下位关系预测往往不够精确，本文也在实验结果的基础上做了更多的讨论.

然而，值得指出的是，这些方法仍然有一定的局限性. 尽管 MWPM、NPM、ALPM 等模型在多个英语数据集中取得了 State-of-the-art 的效果，在中文和其他语言上的表现还有很大提升的空间. 此外，我们的方法仍然需要相当数量的人工标注数据集作为训练数据，以训练投影模型. 与文献[32，61]等的结论一致，在处理特定领域或者长尾概念的上

下位关系时，由于相关概念词频出现较低，这些词的词向量学习质量会降低，这也会进一步影响投影模型的学习. 目前，在 NLP 的研究中，文献[61，62]都旨在解决特定领域的、长尾关系抽取问题，这些研究也会未来上下位关系预测的研究起到了指导作用. 此外，本文的研究工作仅限于两个概念间的上下位关系判别，没有涉及整个分类体系或者概念层次网络的构建. 目前仅有少量研究工作将这两个任务统一到一个端到端的模型中[63]，这个问题今后也值得进一步研究.

参 考 文 献

[1] Wu Wen-Tao, Li Hong-Song, Wang Hai-Xun, et al. Probase: a probabilistic taxonomy for text understanding//Proceedings of the ACM SIGMOD International Conference on Management of Data 2012. Scottsdale, USA, 2012: 481-49

[2] Suchanek F, Kasneci G, Weikum G. YAGO: a core of semantic knowledge//Proceedings of the 16th International Conference on World Wide Web. Banff, Canada, 2007: 697-706

[3] Vulic I, Gerz D, Kiela D, et al. HyperLex: a large-scale evaluation of graded lexical entailment. Computational Linguistics, 2017, 43(4):781-835.

[4] Zhang Yu-Chen, Ahmed A, Josifovski V, et al. Taxonomy discovery for personalized recommendation//Proceedings of the 7th ACM International Conference on Web Search and Data Mining. New York, USA, 2014: 243-252

[5] Wang Zhong-Yuan, Zhao Ke-Jun, Wang Hai-Xun, et al. Query understanding through knowledge-based conceptualization//Proceedings of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 264-3270

[6] Wang Cheng-Yu, He Xiao-Feng, Zhou Ao-Ying. A short survey on taxonomy learning from text corpora: issues, resources and recent advances//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 1190-1203

[7] Hearst M. Automatic acquisition of hyponyms from large text corpora//Proceedings of the 14th International Conference on Computational Linguistics. Nantes, France, 1992: 539-545

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

[9] Luu A, Kim J, Ng S. Taxonomy construction using syntactic contextual evidence//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 810-819

[10] Fu Rui-Ji, Guo Jiang, Qin Bing, et al. Learning semantic hierarchies via word embeddings//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014: 1199-1209

[11] Wang Cheng-Yu, He Xiao-Feng. Chinese hypernym-hyponym extraction from user generated categories//Proceedings of the 26th International Conference on Computational Linguistics. Osaka, Japan, 2016: 1350-1361

[12] Weeds J, Weir D, McCarthy D. Characterising measures of lexical distributional similarity//Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland, 2004

[13] Santus E, Lenci A, Lu Qin, et al. Chasing hypernyms in vector spaces with entropy//Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden, 2014: 38-42

[14] Nguyen K, Köper M, Walde S, et al. Hierarchical embeddings for hypernymy detection and directionality/Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 233-243

[15] Roller S, Erk K, Boleda G. Inclusive yet selective: supervised distributional hypernymy detection//Proceedings of the 25th International Conference on Computational Linguistics. Dublin, Ireland, 2014: 1025-1036

[16] Turney P, Mohammad S. Experiments with three approaches to recognizing lexical entailment. Natural Language Engineering, 2015, 21(3): 437-476

[17] Yu Zheng, Wang Hai-Xun, Lin Xue-Min, et al. Learning term embeddings for hypernymy identification//Proceedings of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 1390-1397

[18] Levy O, Remus S, Biemann C, et al. Do supervised distributional methods really learn lexical inference relations? //Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, USA, 2015: 970-976

[19] Carlson A, Betteridge J, Wang R, et al. Coupled semi-supervised learning for information extraction//Proceedings of the Third International Conference on Web Search and Web Data Mining. New York, USA, 2010: 101-110

[20] Wang Cheng-Yu, Yan Jun-Chi, Zhou Ao-Ying, et al. Transductive non-linear learning for Chinese hypernym prediction//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 1394-1404

[21] Liu Han-Xiao, Yang Yi-Ming. Bipartite edge prediction via transductive learning over product graphs//Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015: 1880-1888

[22] Wang Cheng-Yu, Yan Fan, He Xiao-Feng, et al. A family of fuzzy orthogonal projection models for monolingual and cross-lingual hypernymy prediction//Proceedings of the 2019 World Wide Web Conference. San Francisco, USA, 2019

[23] Chen Jin-Dong, Wang Ao, Chen Jiang-Jie, et al. CN-Probase: A data-driven approach for large-scale Chinese taxonomy construction. arXiv: 1902.10326, 2019

[24] Li Jin-Yang, Wang Cheng-Yu, He Xiao-Feng, et al. User generated content oriented Chinese taxonomy construction//Proceedings of the 17th Asia-Pacific Web Conference. Guangzhou, China, 2015: 623-634

[24] Wang Cheng-Yu, He Xiao-Feng, Zhou Ao-Ying. Improving hypernymy prediction via taxonomy enhanced adversarial learning//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019

[25] Turney P, Pantel P. From frequency to meaning: vector space models of semantics. Journal of Artificial Intelligence Research, 2010, 37: 141-188

[26] Seitner J, Bizer C, Eckert K, et al. A large database of hypernymy relations extracted from the web//Proceedings of the Tenth International Conference on Language Resources. Portoroz, Slovenia, 2016

[27] Bansal M, Burkett D, de Melo G, et al. Structured learning for taxonomy induction with belief propagation//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014: 1041-1051

[28] Navigli R, Velardi P. Learning word-class lattices for definition and hypernym extraction//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden,2010: 1318-1327

[29] Mo Yuan-Yuan, Guo Jian-Yi, Yu Zheng-Tao, et al. Hyponymy extraction of domain ontology concept based on CCRF. Computer Engineering, 2014, 40(06): 138-141 (in Chinese)
(莫媛媛, 郭剑毅, 余正涛等. 基于CCRF的领域本体概念上下位关系抽取. 计算机工程, 2014, 40(06): 138-141)

[30] Chen Jin-Dong, Xiao Yang-Hua. Hypernymy relation extraction based on semantics. Computer Applications and Software, 2019, 36(2): 216-221 (in Chinese)
(陈金栋,肖仰华.一种基于语义的上下位关系抽取方法.计算机应用与软件, 2019, 36(2): 216-221)

[31] Kozareva Z, Hovy E. A semi-supervised method to learn and construct taxonomies using the web//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. MIT Stata Center, USA, 2010: 1110-1118

[32] Alfarone D, Davis J. Unsupervised learning of an is-a taxonomy from a limited domain-specific corpus//Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 1434-1441

[33] Gupta A, Lebret R, Harkous H, et al. Taxonomy induction using hypernym subsequences//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore, 2017: 1329-1338

[34] Wang Cheng-Yu, Fan Yan, He Xiao-Feng, et al. Learning fine-grained relations from Chinese user generated categories//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 2577-2587

[35] Cheng Yun-Ru. Research on domain entity hyponymy extraction automatically [Master Thesis]. Kunming University of Science and Technology, Kunming, 2016 (in Chinese)
(程韵如. 领域实体上下位关系自动获取研究[硕士学位论文].昆明理工大学, 昆明, 2016)

[36] Wang Chang-You, Yang Zeng-Chun. An acquisition method of domain-specific terminological hyponym based on structure feature of sentence. Journal of Chongqing University of Posts and Telecommunication (Natural Science Edition), 2014, 26(03): 385-389 (in Chinese)
(王长有,杨增春. 一种基于句子结构特征的领域术语上下位关系获取方法. 重庆邮电大学学报(自然科学版), 2014, 26(03): 385-389)

[37] Santus E, Shwartz V, Schlechtweg D. Hypernyms under siege: linguistically-motivated artillery for hypernymy detection//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain, 2017: 65-75

[38] Chang Haw-Shiuan, Wang Zi-Yun, Vilnis L, et al. Distributional inclusion vector embedding for unsupervised hypernymy detection//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, USA, 2018: 485-495

[39] Nickel M, Kiela D. Learning continuous hierarchies in the lorentz model of hyperbolic geometry//Proceedings of the 35th International Conference on Machine Learning. Stockholm, 2018: 3776-3785

[40] Ganea OE, Bécigneul G, Hofmann T. Hyperbolic entailment cones for learning hierarchical embeddings//Proceedings of the 35th International Conference on Machine Learning. Stockholm, 2018: 1632-1641

[41] Liu Shen. Chinese entity relation discovery for Big Cilin [Master Thesis]. Harbin Institute of Technology, Harbin, 2016 (in Chinese)
(刘燊. 面向《大词林》的中文实体关系挖掘[硕士学位论文]. 哈尔滨工业大学, 哈尔滨, 2016)

[42] Weeds J, Clarke D, Reffin J, et al. Learning to distinguish hypernyms and co-hyponyms//Proceedings of the 25th International Conference on Computational Linguistics. Dublin, Ireland, 2014: 2249-2259

[43] Shwartz V, Goldberg Y, Dagan I. Improving hypernymy detection with an integrated path-based and distributional method//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, German, 2016:2389-2398

[44] Ma Xiao-Jun, Guo Jian-Yi, Xian Yan-Tuan, et al. Entity hyponymy acquisition and organization combining word embedding and bootstrapping in special domain. Computer Science, 2018, 45(1): 67-72 (in Chinese)
(马晓军,郭剑毅,线岩团等.结合词向量和Bootstrapping的领域实体上下位关系获取与组织.计算机科学, 2018, 45(1): 67-72)

[45] Sun Jia-Wei, Li Zheng-Hua, Chen Wen-Liang, et al. Hypernym relation classification based on word pattern. Acta Scientiarum Naturalium Universitatis Pekinensis, 2019, 55(1): 1-7 (in Chinese)
(孙佳伟,李正华,陈文亮等.基于词模式嵌入的词语上下位关系分类.

北京大学学报(自然科学版), 2019, 55(1): 1-7)

[46] Roller S, Kiela D, Nickel M. Hearst patterns revisited: automatic hypernym detection from large text corpora//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, 2018: 358-363

[47] Le M, Roller S, Papaxanthos L, et al. Inferring concept hierarchies from text corpora via hyperbolic embeddings. arXiv:1902.00913, 2019

[48] Yamane J, Takatani T, Yamada H, et al. Distributional hypernym generation by jointly learning clusters and projections//Proceedings of the 26th International Conference on Computational Linguistics. Osaka, Japan, 2016: 1871-1879

[49] Biemann C, Ustalov D, Panchenko A, et al. Negative sampling improves hypernymy extraction based on projection learning//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain, 2017: 543-550

[50] Khuller S, Moss A, Naor J. The budgeted maximum coverage problem. Information Processing Letters, 1999, 70(1): 39–45

[51] Xing Chao, Wang Dong, Liu Chao, et al. Normalized word embedding and orthogonal transform for bilingual word translation//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado, USA, 2015: 1006-1011

[52] Markley F, Crassidis J. Fundamentals of spacecraft attitude determination and control. Vol. 33. 2014. New York: Springer-Verlag New York, 2014

[53] Markley F. Attitude determination using vector observations and the singular value decomposition. Journal of the Astronautical Sciences, 1988, 36(3): 245–258

[54] Mirza M, Osindero S. Conditional generative adversarial nets. arXiv :1411.1784, 2014

[55] Mikolov T, Chen Kai, Corrado G, et al. Efficient estimation of word representations in vector space//Proceedings of the 1st International Conference on Learning Representations. Scottsdale, USA, 2013

[56] Bojanowski P, Grave E, Joulin, et al. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146

[57] Baroni M, Lenci A. How we blesseed distributional semantic evaluation// Proceedings of the Gems 2011 Workshop on Geometrical MODELS of Natural Language Semantics. Edinburgh, UK, 2011: 1–10.

[68] Baroni M, Bernardi R, Do NQ, et al. Entailment above the word level in distributional semantics//Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France, 2012: 23-32

[59] Velardi P, Faralli S, Navigli R. OntoLearn reloaded: a graph-based algorithm for taxonomy induction. Computational Linguistics, 2013, 39(3): 665–707

[60] Li Hai-Guang Li, Wu Xin-Dong, Li Zhao, Wu Gong-Qing. A relation extraction method of Chinese named entities based on location and semantic features. Applied Intelligence, 2013, 38(1): 1-15

[61] Fan Yan, Wang Cheng-Yu, He Xiao-Feng. Exploratory neural relation classification for domain knowledge acquisition//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, USA, 2018: 2265-2276

[62] Zhang Qing, Wang Hou-Feng. Noise-clustered distant supervision for relation extraction: a nonparametric bayesian perspective//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 1808-1813

[63] Mao Yu-Ning, Ren Xiang, Shen Jia-Ming, et al. end-to-end reinforcement learning for automatic taxonomy induction//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, 2018: 2462-2472

WANG Cheng-Yu, born in 1991, Ph. D. student. His research interests include relation extraction, knowledge graphs and natural language processing.

HE Xiao-Feng, born in 1969, Ph. D., professor. His research interests include data mining, machine learning and information retrieval.

XUE Qing-Gong, born in 1974, Ph. D., professor. His research interests include databases and distributed database management systems.

ZHOU Ao-Ying, born in 1965, Ph. D., professor. His main research interests include big data processing and analysis, data management for data intensive computing, etc.

Background

As a type of important linguistic resources in the fields of NLP, hypernymy relations establish the “is-a” relations among concepts. The accurate automatic harvesting of hypernymy relations helps to construct large-scale semantic networks, ontologies, and knowledge graphs, and also beneficial to a variety of downstream NLP tasks, such as natural language inference, recommendation, etc.

In the NLP research community, pattern-based methods and distributional methods are two main types of learning paradigms for hypernymy prediction. Pattern-based approaches rely more on language patterns, which may have low coverage and are more language-dependent. Distributional methods are more precise, but likely to suffer from “lexical memorization”.

In this work, we introduce the development process and the latest breakthrough of word embedding projection models. Word embedding projection models are distributional models that map the embeddings of concepts to those of their hypernyms, in order to predict hypernymy relations accurately and to avoid the “lexical memorization” problem at the same time. We give a unified mathematical framework of these models and discuss how these models are developed. Additionally, we evaluate all these models under a unified framework. Experimental results over English and Chinese datasets illustrate the effectiveness of word embedding projection models for hypernymy prediction. We also discuss future research directions on domain-specific and long-tail hypernymy prediction.

Currently, the research interests of our research group mainly focus on artificial intelligence and NLP, including knowledge graphs, personalized recommendation, etc. The work described in this paper would be beneficial to NLP researchers who are working on relation extraction, distributional semantics and knowledge graphs.

This work is supported by the National Key Research and Development Program of China (No. 2016YFB1000904) and the Natural Science Foundation of China (No. 61572194).