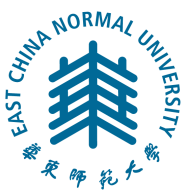
2020 届博士学位论文

分类号：_____ 学校代码：10269

密　级：_____ 学　　号：52151500019



East China Normal University

博 士 学 位 论 文

DOCTORAL DISSERTATION

论文题目：
面向中文短文本的关系抽取算法设计

院　　系:	软件工程学院
专 业 名 称:	软件工程
研 究 方 向:	数据科学与工程
指 导 教 师:	何晓丰 研究员
学位申请人:	汪诚愚

2020 年 05 月

Dissertation for doctor degree in 2020

University Code: 10269
Student ID: 52151500019

EAST CHINA NORMAL UNIVERSITY

Algorithmic Studies on Relation Extraction from Chinese Short Texts

Department:	School of Software Engineering
Major:	Software Engineering
Research direction:	Data Science and Engineering
Supervisor:	Prof. HE Xiaofeng
Candidate:	WANG Chengyu

2020.05

华东师范大学学位论文原创性声明

郑重声明：本人呈交的学位论文《面向中文短文本的关系抽取算法设计》，是在华东师范大学攻读硕士/博士（请勾选）学位期间，在导师的指导下进行的研究工作及取得的研究成果。除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确说明并表示谢意。

作者签名：汪诚愚　　日期：2020 年 5 月 27 日

华东师范大学学位论文著作权使用声明

《面向中文短文本的关系抽取算法设计》系本人在华东师范大学攻读学位期间在导师指导下完成的硕士/博士（请勾选）学位论文，本论文的著作权归本人所有。本人同意华东师范大学根据相关规定保留和使用此学位论文，并向主管部门和学校指定的相关机构送交学位论文的印刷版和电子版；允许学位论文进入华东师范大学图书馆及数据库被查阅、借阅；同意学校将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于（请勾选）

（　）1.经华东师范大学相关部门审查核定的“内部”或“涉密”学位论文*，于　　年　　月　　日解密，解密后适用上述授权。

（√）2.不保密，适用上述授权。

导师签名　　本人签名汪诚愚

2020 年 5 月 27 日

* “涉密”学位论文应是已经华东师范大学学位评定委员会办公室或保密委员会审定过的学位论文（需附获批的《华东师范大学研究生申请学位论文“涉密”审批表》方为有效），未经上述部门审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权）。

汪诚愚 博士学位论文答辩委员会成员名单

姓名	职称	单位	备注
周傲英	教授	华东师范大学	**主席**
黄萱菁	教授	复旦大学	
赵海	教授	上海交通大学	
贺樑	教授	华东师范大学	
兰曼	教授	华东师范大学	

摘 要

海量互联网数据的异构、多源和异质等问题使得高效、精准的知识获取成为巨大的挑战。关系抽取是自然语言处理中的一项基础性任务，从无结构化的文本数据中自动获取结构化的关系型事实，为大规模知识图谱的构建和互联网智能知识服务提供支持。随着深度学习技术的广泛应用，神经关系抽取模型的精度获得了很大提升。然而，现有的主流研究一般关注英语语言的、句子级别的关系抽取。与英语不同，中文表述灵活多变，语法和构词规则相对不固定，大量语义知识蕴含在中文短文本中，通常很难被现有算法有效抽取。

本文主要研究面向中文短文本的关系抽取问题。由于中文短文本独特的语言学特征，其关系抽取任务与传统工作相比具有诸多挑战。短文本的语法结构和语义一般不完整，部分短文本蕴含的语义关系属于常识性知识，关系表述的上下文高度稀疏。与英语相比，中文基础自然语言分析较低的准确度，以及短文本关系抽取标注数据集的缺乏也增大了这一问题的难度。本文分别从基于词嵌入的上下位关系抽取、知识增强的语义关系抽取、以及非上下位关系抽取与语义理解等三个方面进行深入研究，设计了面向中文短文本的关系抽取框架，较好地解决了上述挑战。本文的主要工作和贡献概述如下：

(1) **基于词嵌入的上下位关系抽取：** 分类体系是知识图谱中概念的层次化表示和重要组织形式，由大量上下位关系构成。与英语相比，由于中文语言表述高度灵活，中文上下位关系抽取不能简单采用文本匹配算法来实现。本文结合神经语言模型和中文语言学特性，采用词嵌入作为中文术语的特征表示，建模中文上下位关系在词嵌入空间的表示，即学习中文下位词在词嵌入空间中投影到对应上位词的过程。本文首先提出了半监督式上下位关系扩展模型，即迭代地从互联网数据中发现新的上下位关系元组，解决了中文上下位关系数据集大小有限的问题。为了精确建模中文上下位关系与非上下位关系分类的决策边界，我们进一步提出基于转导学习和模糊正交投影学习的两个上下位关系分类模型。实验效果表明，提出的模型在精度上超过了现有最佳方法，有效实现中文上下位关系抽取。

(2) **知识增强的语义关系抽取：** 上述基于词嵌入的上下位关系抽取模型依赖于特定领域的训练集，对其他类别的数据源和相关任务没有加以良好运用。本文以词嵌入投影模型为基础，探索知识增强的语义关系抽取算法，从多知识源、多语言、多词汇关系三个角度，抽取多种类型的语义关系。首先，由于大规模分类体系中含有大量上下位关系，本文提出分类体系增强的对抗学习框架，利用双重深度对抗学习机制，将互联网中的海量上下位关系知识融入基于特定

训练集的词嵌入投影神经网络中。其次，本文扩展了模糊正交投影模型，分别提出了迁移模糊正交投影模型和其扩展版本迭代迁移模糊正交投影模型，结合了深度迁移学习和双语术语对齐技术，在小样本学习场景下，实现了面向小语种的跨语言上下位关系抽取。最后，由于知识本体中一般包含多种类别的词汇关系，本文提出超球关系嵌入模型，对多种类别的词汇关系分别进行语义建模，学习其超球嵌入表示，使投影模型可以对多种词汇关系进行分类。相应自然语言处理任务的实验效果证明了这三种模型的有效性。

(3) **非上下位关系抽取与语义理解：** 中文短文本中通常具有类别繁多的非上下位关系，前述模型预测的关系类别由人工定义，难以扩展至开放领域，而且缺乏常识性关系检测和深度关系理解的能力。在这一部分研究中，首先提出基于模式的非上下位关系抽取算法，它采用图挖掘技术，从中文短文本中挖掘出表达丰富语义关系的频繁语言模式，无监督地抽取出与这些模式相对应的非上下位关系三元组。由于上述方法只能抽取出频繁模式对应关系，本文进一步提出数据驱动的非上下位关系抽取算法，它采用三阶段的数据驱动架构，实现从中文短文本的切分到关系生成的完整流程，提升关系抽取的覆盖率。最后，我们观察到，基于习语性分析的语义理解技术可以从中文短文本中推导出更多关系，实现深度知识推理。本文据此提出了关系性与组合性表示学习框架，对中文复合名词的习语性程度进行分类，并且探究这一算法对自然语言理解的提升作用。实验结果表明，上述算法在面向中文短文本的关系抽取中，不局限于人工定义关系类别，可以在多个领域准确地抽取出多种非上下位关系。

综上所述，本文从三个方面解决从中文短文本中抽取语义关系的问题，在多个自然语言处理任务相关的公开数据集上进行实验，实验结果证明了提出方法的有效性。本文的研究工作也为实现面向互联网海量中文短文本的关系自动抽取和语义理解系统提供技术基础，在尽可能减少人工干预的情况下，充分挖掘短文本中蕴含的知识，从而对现有大规模中文知识图谱系统进行扩展和补全。

关键词: 关系抽取，中文短文本，上下位关系，词嵌入，语义理解

ABSTRACT

Due to the heterogeneity, multi-source and varied qualities of massive Web data, it is significantly challenging to harvest knowledge from them efficiently and accurately. Relation extraction, a basic task in Natural Language Processing (NLP), aims at obtaining structured relational facts from unstructured textual data automatically, providing technical support for large-scale knowledge graph construction and intelligent Web knowledge services. With the widespread application of deep learning techniques, the accuracy of neural relation extraction models has been greatly improved. However, existing research generally focuses on sentence-level relation extraction for English language. Different from English, Chinese expressions are more flexible, with relatively unfixed grammatical structures and word formation rules. Hence, a large amount of semantic knowledge expressed in short Chinese texts is difficult to be extracted by existing algorithms effectively.

This thesis mainly studies the problem of relation extraction from Chinese short texts. According to the uniqueness of linguistic characteristics of Chinese short texts, the corresponding relation extraction task has many challenges compared with traditional work. The grammatical structures and semantics of short texts are generally incomplete. The semantic relations expressed in some short texts belong to the category of commonsense knowledge. Therefore, the contextual expressions of such relations are highly sparse. Compared with English, the low accuracy of Chinese language analysis, together with the lack of annotated datasets for short-text relation extraction, also increases the difficulty of this problem. We conduct in-depth research from the following three aspects: i) hypernymy extraction based on word embeddings, ii) knowledge-enhanced semantic relation extraction, and iii) non-hypernymy relation extraction and semantic understanding. The framework of relation extraction from short Chinese texts is also illustrated, addressing these challenges well. Major contributions of this thesis are summarized as follows:

(1) **Hypernymy Extraction Based on Word Embeddings.** The taxonomy is a hierarchical representation and an important organization form of concepts in knowledge graphs, consisting of a large number of hypernymy relations. Compared with English, the language expressions in Chinese are highly flexible. Hence, it is infeasible to extract Chinese hypernymy relations by simple text matching algorithms. In this thesis, we integrate neural language models and Chinese linguistic characteristics to address this issue by employing word embeddings as the representations of Chinese terms. The proposed algorithms model the representations of Chinese hypernymy relations, that is, learning the projections of Chinese hyponyms to their corresponding hypernyms in the word embedding space. We first propose a semi-supervised hyper-

nymy extension model, which iteratively discovers new hypernymy relations from Web data, and solves the problem of the limited sizes of Chinese hypernymy datasets. To model the decision boundary of Chinese hypernymy and non-hypernymy relations accurately, two hypernymy classification models are further presented, based on transductive and fuzzy orthogonal projection learning, respectively. Experimental results show that the proposed models outperform state-of-the-arts, achieving the accurate extraction of Chinese hypernymy relations.

(2) **Knowledge-enhanced Semantic Relation Extraction.** The above hypernymy extraction models based on word embedding rely on training sets in specific domains. They do not leverage other types of data sources and related tasks. Based on word embedding projection models, we explore the design of knowledge-enhanced relation extraction algorithms. Briefly, such algorithms harvest semantic relations from three perspectives, namely i) multiple knowledge sources, ii) multiple languages, and iii) multiple types of lexical relations. We first propose the Taxonomy-Enhanced Adversarial Learning framework, which exploits numerous hypernymy relations in large-scale taxonomies. It injects such knowledge into projection models trained over specific datasets by deep coupled adversarial learning. Next, the Transfer Fuzzy Orthogonal Projection Model and the semi-supervised version, the Iterative Transfer Fuzzy Orthogonal Projection Model, are proposed by extending the Fuzzy Orthogonal Projection Model. They combine the techniques of deep transfer learning and bilingual lexicon induction for few-shot cross-lingual hypernymy extraction, especially for lower-resourced languages. Finally, due to the existence of multiple types of lexical relations in ontologies, the learning process of Hyperspherical Relation Embeddings are presented, which learns the representations of different lexical relations in the hyperspherical embedding space. Therefore, the projection models can be extended for multi-way classification of lexical relations. Experimental results on the corresponding NLP tasks prove the effectiveness of these models.

(3) **Non-hypernymy Relation Extraction and Semantic Understanding.** There exist a variety of non-hypernymy relations expressed in Chinese short texts. Previous models can only deal with a finite set of pre-defined relation types, which are difficult to extend to open domains and lack the ability of extracting commonsense relations by deep text understanding. In this part, the Pattern-based Non-hypernymy Relation Extraction model is first proposed. It employs graph mining techniques to acquire frequent textual patterns that express rich semantic relations from Chinese short texts. Relations related to these patterns can be extracts by unsupervised learning. As the algorithm can only deal with frequent patterns, we further present the Data-driven Non-hypernymy Relation Extraction model. It has a three-stage data-driven archi-

tecture, from Chinese short text segmentation to relation generation, improving the coverage of relation extraction. Finally, we observe that idiomaticity analysis based semantic understanding results in the extraction of more relations from Chinese short texts by deep knowledge reasoning. Hence, a Relational and Compositional Representation Learning framework is proposed, which classifies the idiomaticity degrees of Chinese noun compounds and improves the machine's ability of Natural Language Understanding. Experimental results show that the above algorithms can extract relation accurately and are not restricted to manually defined relation types.

In summary, this thesis addresses the problem of relation extraction from Chinese short texts in three aspects. Experiments over public datasets of several NLP related tasks prove the effectiveness of the proposed algorithms. Our research also provides technical foundations for building automatic relation extraction and semantic understanding systems for massive Chinese short texts from the Web. With minimal human intervention, the knowledge in short texts can be fully extracted, beneficial for large-scale Chinese knowledge graph expansion and completion.

Keywords: *Relation Extraction, Chinese Short Texts, Hypernymy Relations, Word Embeddings, Semantic Understanding*

目录

插图

表格

第一章　绪论

随着互联网技术的广泛普及，互联网数据呈现出爆炸式的增长趋势。海量互联网数据的异构、多源和异质等问题，使得高效精准的信息获取成为巨大的挑战。关系抽取是自然语言处理中的一项基础性任务，可以将隐藏在半结构化、无结构化的文本中的关系型事实分进行抽取与精炼，并将其结构化，为大规模语义网络与知识图谱的构建提供技术支持。现有的主流研究一般聚焦基于英语语言的、句子级别的关系抽取。反观中文领域，由于中文表达灵活多变，语法规则相对不固定，大量中文语义知识蕴含在中文短文本中，通常不能被现有算法有效抽取。此外，中文短文本的语法结构和语义通常很不完整，将关系抽取技术应用于这些文本仍然面临诸多挑战。因此，研究如何高精度地从中文短文本中挖掘出丰富的语义关系，对构建和补全大规模中文知识图谱至关重要。本章在回顾相关研究背景的基础上，概述了本文的算法研究工作和主要贡献。

1.1　研究背景

1.1.1　知识图谱

互联网技术的快速发展从根本上改变了人们发布和获取信息的方式，这一现象使得互联网成为人们主要的信息来源。根据第 44 次《中国互联网络发展状况统计报告》[1]，截止至 2019 年 6 月，我国网民规模已经达到 8.54 亿，互联网普及率达到 61.2%，网站数量达到 518 万个。国际数据公司（International Data Corporation）则进一步预测，整个互联网的数据量将从 2018 年的 33ZB 增长到 2025 年的 175ZB [2]。互联网上的海量信息极大地满足了人们的信息需求。然而，互联网数据往往呈现出碎片化特征，数据质量参差不齐，这些海量的碎片化数据导致了“信息过载”的问题，使用户不堪重负；人们对信息获取的需求逐渐从获得足够数量的相关信息转变为快速获取与信息需求直接相符合的知识 [3]。以搜索引擎这一典型互联网应用为例，它不仅需要返回与用户查询最相关的网络信息，还需要拥有智能问答、阅读理解、实体推荐、知识推送等能力，以直接返回用户需要的答案或信息。

为满足上述信息需求，谷歌于 2012 年首先提出“**知识图谱**”(Knowledge Graph) 这一概念，并且将知识推送技术运用于搜索引擎中 [4]，对给定用户搜索推荐答案、

知识卡片和相关实体知识。知识图谱是一种大规模语义网络，以实体或概念作为节点，以边代表实体或概念之间的关系。它有效整合了不同数据源的知识，并且使用图作为灵活、可扩展性高的数据表示模型，将知识的表达结构化。与谷歌知识图谱 (Google Knowledge Graph) 类似，产业界的其他搜索引擎提供者也构建了自己的知识图谱，包括用于微软必应搜索引擎的 Satori Entity Engine [5]、百度的 “百度知心” [6]、用于搜狗知识搜索引擎的 “知立方” [7] 等。图 1.1给出了相应知识图谱在谷歌和必应搜索引擎上的知识推荐和智能问答应用示例。

知识图谱的广泛应用也推动了学术界对相关**自然语言处理**（Natural Language Processing，缩写为 NLP）和知识图谱构建技术的研究。典型的知识图谱系统有德国 Max Planck 研究所研制的基于维基百科的 YAGO 系统 [8]、美国卡内基梅隆大学推出的 NELL 自动关系抽取系统 [9]、微软亚洲研究院提出的 Probase 系统 [10] 等，他们将不同类别的英文数据源中的知识进行抽取和融合。对于中文数据源，著名的知识图谱系统包括复旦大学的中文百科知识图谱 CN-DBPedia [11]、华东理工大学的中文语义知识网络 zhishi.me [12]、清华大学的中英跨语言知识图谱 XLore [13] 等。这些知识图谱的研究工作，也促进了其他相关任务的研究，例如面向知识图谱的实体链接 [14]、基于知识图谱的用户查询理解 [15] 等。此外，上述研究还为诸多人工智能技术提供背景知识的支持。现有基于深度学习的人工智能技术存在难以进行结果解释、常识推断、归因分析等缺点，这些问题使得深度神经网络的研究距离实现通用人工智能（Artificial General Intelligence）的目标尚为遥远。知识图谱显式地提供了知识服务和知识推理功能，为上述问题的解决提供了更大的空间 [16]。

1.1.2 分类体系与上下位关系抽取

概念与概念、概念与实体之间往往存在层次隶属关系。这种关系在语言学中称为 “**上下位关系**”（Hypernymy）。例如在 “棕熊-哺乳动物”、“飞机-交通工具” 中，“棕熊” 和 “飞机” 是**下位词**（Hyponym），“哺乳动物” 和 “交通工具” 分别是其对应的**上位词**（Hypernym）。在知识图谱中，上下位关系的层次化结构通常被称为**分类体系**（Taxonomy），它对知识图谱中的实体和概念进行了严格的类别划分，是知识图谱在语义层次上组织、展现知识的基础。图 1.2展示了一个小型中文分类体系的示例。其中，箭头即表示上下位关系。在传统知识库中，分类体系大多由专家人工制定，数据规模较小，例如英语的词汇语义网络 WordNet [17] 和中文的词汇义元语

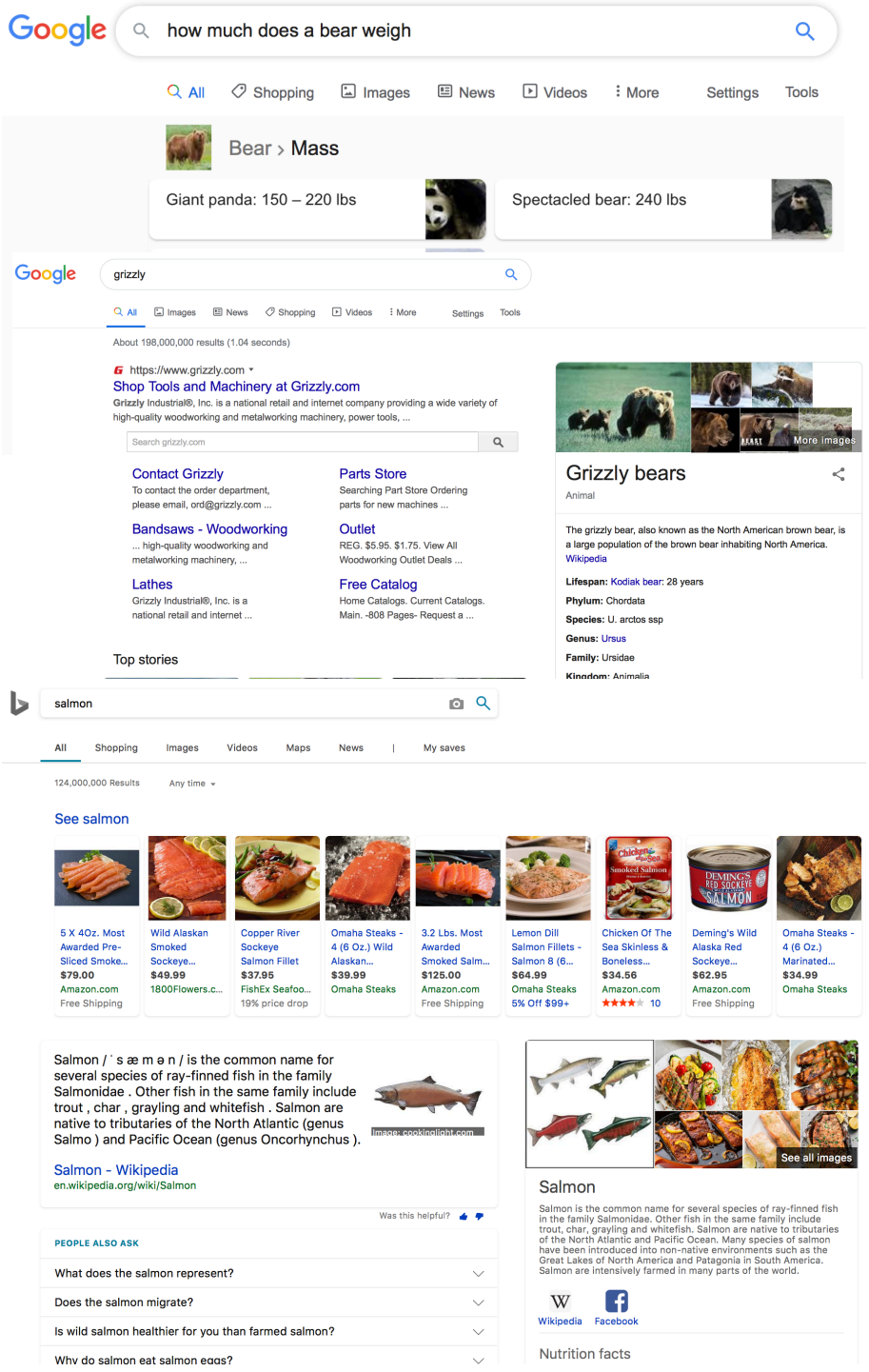


图 1.1: 基于知识图谱的知识推荐和智能问答（示例来自谷歌、必应搜索引擎返回的知识图谱内相关信息）

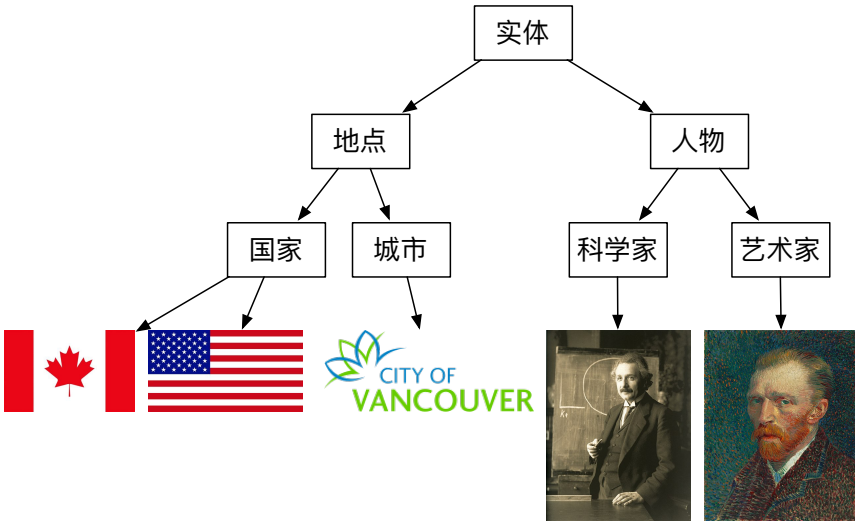


图 1.2: 中文分类体系示例

义网络 HowNet [18] 中的概念层次结构等。在海量互联网环境下，自动构建大规模分类体系成为支持搜索引擎、推荐系统的必要任务。在学术界，经典的大规模分类体系包括英语的 Probase [10]、WikiTaxonomy [19] 和中文对应的 CN-Probase [20]、CN-WikiTaxonomy [21] 等。

构建分类体系的核心任务是上下位关系抽取。上下位关系是一种比较特殊的语义关系，即**词汇关系**（Lexical Relation）。词汇关系描述了概念之间的语义是怎样相互联系的，包括上下位关系、近义词关系（Synonymy）、反义词关系（Antonymy）等。比起通用的语义关系，词汇关系在文本中的表达更加稀疏，这是因为词汇关系一般属于**常识性知识**（Commonsense Knowledge），这些知识的表述往往在文本中被省略。例如，“计算机” 和 “电脑” 是近义词，而在 “计算机” 和 “电脑” 共同出现的上下文中，很少出现描述 “近义词” 这一关系的词或其他相似词汇。在现有的研究中，词汇关系（特别是上下位关系）的获取一般依赖于小部分高精度的语言模式，以及**分布式语义**（Distributional Semantics）的建模 [22]。特别地，上下位关系获取 [23, 24] 将两个概念之间的关系预测过程建模成 “上下位关系-非上下位关系” 二分类问题。由于在中文领域，中文语言模式表述灵活，模式匹配方法的可扩展性差，中文上下位关系的抽取更依赖于基于词嵌入表示的语义关系建模 [25]，采用深度神经网络学习上下位关系的表示。这也是本文上下位关系抽取研究的突破方向。

1.1.3 通用语义关系抽取

关系型知识一般表示为“主语-谓词-宾语”（Subject-Predicate-Object）三元组的形式，是知识图谱的基本构成单元。由于其重要性，关系抽取是 NLP 中最基本的任务之一。根据其输入输出不同，关系抽取可以被进一步分为若干相似的子任务。其中，**关系分类**（Relation Classification）给定两个实体的上下文作为输入，目的是学习一个分类模型，对这两个实体的语义关系进行预测 [26]。这一任务的局限性有两个主要方面，i）只能对固定若干种关系进行预测，且所有关系类别需要预先指定；ii）需要大量人工标注数据作为分类器的输入。**开放关系抽取**（Open Relation Extraction，缩写为 ORE）[27, 28] 克服了第一个局限性，它自动地从未标注的文本中抽取“主语-谓语-宾语”结构，作为候选关系元组。在这一系列算法中，实体之间的关系类别不需要提前指定。为了解决第二个问题，**远程监督**（Distant Supervision）[29] 采用知识图谱中已有的关系为未标注的文本提供关系标注信息，从而大大减少人工标注的工作量。近年来，关系抽取的研究逐渐与深度学习算法相结合，包括对抗学习 [29]、强化学习 [30] 等。

除了基于无结构化文本的通用和词汇关系抽取，有少量研究旨在从更短的文本中获得知识，例如基于名词的开放关系抽取 [31]、基于搜索引擎查询记录的关系抽取 [32] 等。在这些短文本中，维基百科的类别频繁受到学术界的关注，因为它的语言表述形式更加规则，而且概括地描述了实体的信息。例如，在文献 [33] 中，作者设计了基于维基百科类别的语法和构词特征，并提出了相应关系抽取算法。

值得指出的是，上述研究工作大多是基于英语语言的。由于关系抽取算法一般不为语言独立的（即为 Language-dependent），在中文上需要做更多语言相关研究 [34–36]。特别地，由于中文语言表达的特殊性，大量介词和修饰词被省略，中文短文本中蕴含着大量未被抽取出的语义关系。从中文短文本中获取海量的语义关系，能极大地扩展中文知识图谱的知识覆盖率，是本文的研究重点。

1.2 面临的挑战

尽管目前深度学习在 NLP 和知识图谱构建上有广泛的应用，将这些技术应用于中文短文本的关系抽取上仍然具有诸多挑战。我们概述研究挑战如下：

中文基础 NLP 分析的低准确度： 关系抽取的首要工作是对文本进行基础 NLP 解析。例如，在 ORE 和 NELL 系统中 [9, 27]，都必须先利用 NLP 技术解析英文句

子（包括词性标注、语法分析、依存分析等），在此基础上提取特征抽取出关系三元组。由于中文语言表述灵活的特点，中文 NLP 技术还有很大提高的空间。例如，根据文献 [37] 汇总指出，中文语法分析的精度低于 80%，语义角色标注的精度低于 70%，内容分析和文本分析的精度低于 65%。这一现象使得算法导致的错误和噪声从基础 NLP 分析到 NLP 应用逐步传播，影响到后续关系的抽取。所以，依赖于这些技术的中文关系抽取质量相比英文有很大距离，这给中文关系抽取方法带来很大的技术挑战。

常识性关系的上下文稀疏性： 如上文所述，知识图谱中包括多种词汇关系，例如上下位关系、近义词关系等。他们通常表达了人类的常识性知识，因此在文本中很少出现直接描述这些关系的上下文模式。特别地，在英语中，上下位关系可以通过 Hearst 模式（Hearst Patterns）来抽取 [38]，例如从 “cities such as Beijing, Shanghai” 中我们可以用 Hearst 模式 “such as” 抽取出两个上下位关系元组 “Beijing-city” 和 “Shanghai-city”。其余词汇关系（例如近义词关系）对应的语言模式较少而且不固定 [39]。与英语相比，中文中相应的语言模式更稀少更灵活，文献 [25, 40] 指出，中文中甚至不包含覆盖率足够高的类似 Hearst 模式的上下位关系模式。

短文本的语法结构和语义不完整性： 短文本在语法结构和语义上高度不完整，这与与完整句子的特性有显著不同。从语法结构上说，短文本通常不包含 “主语-谓词-宾语” 这一结构。例如，在 “中国首都北京” 这一同位语表达中，“主语-谓词-宾语” 结构的缺乏使得基于语法分析、依存分析等 NLP 分析特征 [41] 无法应用于短文本的关系抽取工作上。从语义结构上说，部分短文本的语义是不完整的。例如，从 “巴黎” 的修饰语 “西欧城市” 中，我们可以推导出一个关系三元组 “巴黎-位于-西欧”。然而，关系谓词 “位于” 只能通过人类的常识进行推理，无法从文本中直接获取。因此，语义的不完整性使得关系获取算法无法从这些短文本中直接抽取出完整的候选关系三元组。这两种不完整性均降低了关系抽取的召回率。

标注数据集的缺乏： 由于现有研究大部分集中于英语语言，中文研究的欠缺也导致中文任务相关的标注数据集比较少。SanWen [42] 和 ACE 2005 中文数据集 [1]是两个比较通用的中文关系抽取标注数据集，然而这两个数据集着眼于中文句子级别的关系抽取，不适用于评测中文短文本的关系获取任务。此外，由于中文短文本在语法语义上均有不完整的问题，语言的歧义性较高，人工标注这些数据需要较多的人力资源和较高的专业知识。这一问题给相关研究工作带来进一步挑战。

[1] http://curtis.ml.cmu.edu/w/courses/index.php/ACE_2005_Dataset

图 1.3: 中文短文本的数据源

设计数据驱动的、只需要少量或不需要训练数据的关系抽取算法对这一方向的研究和应用会有促进和推动作用。

值得进一步讨论的是，上述挑战并非互相独立，而是相互作用和加强。具体来说，语法结构和语义的不完整性使得待抽取关系的上下文更加稀疏。这些问题和中文 NLP 分析的低准确度又增加了文本处理引起的噪声，降低关系抽取的精度。由于中文短文本领域标注数据集的缺乏和特征抽取的严重噪声问题，本文的研究目标很难依靠传统监督学习的方法来完成。因此，非常有必要同时考虑文本挖掘、深度词嵌入模型和中文语言学分析技术，设计数据驱动的关系抽取框架，才能从整体上提升中文短文本知识抽取的精准度和召回率。

1.3 整体研究内容与研究思路

为了丰富现有中文知识图谱中的语义知识，在本文中，我们考虑的输入数据为中文短文本对 (x_i, y_i) 的集合，其中，x_i 表示一个中文实体或概念，y_i 为描述中文实体或概念 x_i 的中文短文本。在实际应用场景中，中文短文本 y_i 的来源有：i）维基类百科全书（例如维基百科、百度百科、互动百科等）中关于实体 x_i 的页面

图 1.4: 本文的研究框架

中的类别标签 [8, 19, 21]；ii）垂直领域网站中关于实体 x_i 的用户生成标签（例如豆瓣电影、Rotten Tomatoes、Netflix 等）；以及 iii）从任意网络语料库中挖掘出的与 x_i 语义相关的概念或关键词 [43, 44] 等。在这种情况下，我们并不限制实体 x_i 与短文本 y_i 共现在同一句中，因为中文表述中省略现象比较多，句子级别的共现这一限制条件容易降低知识获取的召回率 [45]。在图 1.3中，我们给出了几个中文实体-短文本对数据源的示例。基于上述数据源，我们给出了本文的整体研究框架，如图 1.4。

值得说明的是，中文短文本涉及的关系类别繁多、范畴庞杂，很多严格加以定义和区分。除了上述中文实体相关的短文本，中文短文本之间的关系还包括动词之间的关系（例如“上升-升高”）、概念之间的联想关系（例如“月亮-嫦娥”、“玫瑰-浪

漫”）等。由于现有知识图谱主要关注的是实体的信息、实体之间的语义关系、实体与类别之间的关系，以利于支持下游 NLP 应用，在本文中，我们将关注点限于抽取中文实体或概念 x_i 和描述中文实体或概念 x_i 的中文短文本 y_i 之间的关系，其他类别的关系抽取工作有待进一步探索。

由于上下位关系是分类体系和知识图谱中的重要组成部分，也是知识推理和语义理解的关键知识源，在第一部分，我们首先研究从中文短文本对 (x_i, y_i) 中抽取出正确的上下位关系。因为中文语言表述高度灵活，缺少同英语中 Hearst 模式对应的语言模式 [25], 我们不直接采用模式匹配方法，而是利用词嵌入作为中文短文本的特征表示，学习中文下位词的词向量是如何投影其上位词的词向量的。为了克服中文标注数据缺乏的问题，我们首先研究半监督式学习场景下，中文上下位关系数据的自动扩展问题，以少量中文上下位关系元组作为“种子”，从中文短文本对 (x_i, y_i) 中自动获得更多上下位关系。然而，这一研究没有显式地建模上下位关系和非上下位关系的区别，不适合对中文上下位和非上下位关系元组进行监督式的分类。因此，我们进一步研究如何同时学习上下位关系和非上下位关系在词嵌入空间的表示，采用多种复杂数学模型进行投影建模和学习，并且探究如何在关系表示学习基础上实现关系分类。此外，如何结合中文语言学知识，改进投影模型的学习过程，也是这一部分研究的一大关注点。

第一部分的研究主要关注单一数据源中上下位关系和非上下位关系的学习问题，对其他知识源和辅助任务没有涉及，因此应用场景较窄。在第二部分中，我们研究知识增强的语义关系抽取算法，我们分别从多知识源、多语言和多词汇关系三个角度，进一步探究前述研究的拓展空间。相应研究也从解决如下三个问题为中心展开。第一，由于基于网络数据的分类体系中蕴含大量上下位知识 [10, 19]，如何融入分类体系中的上下位关系知识，在特定数据集、特定领域下，提升已有上下位关系预测模型的预测精度。第二，大部分已有上下位关系训练集为英语语言，小语种数据集获得难度很大。如何设计跨语言的上下位关系预测算法，采用深度迁移技术，将源语言的知识迁移到目标语言，使模型能准确地对目标语言（特别是小语种）的上下位和非上下位关系进行分类。第三，除上下位关系外，其他词汇关系（如近义词关系、反义词关系等）也对知识图谱中本体的构建至关重要，对不同类别的词汇关系分别进行语义建模和表示学习，使得模型在识别上下位关系之外，也能对其他词汇关系元组进行准确分类。

除了有限几种类别的词汇关系外，中文短文本中还存在大量其他类别的非上

下位关系，前两部分的研究涉及的关系类别是人工定义的，难以扩展至开放领域，而且对中文短文本的语义缺乏深度理解。在这一部分中，首先，我们探索如何在没有任何人工标注的情况下，自动从海量中文短文本中挖掘出非上下位关系的类别及其对应的关系元组。其次，由于语义关系在数据中的分布一般呈现“长尾效应”[46]，如何继续扩展前述研究，使之能够尽可能多地挖掘出“长尾关系”也是我们关注的一大研究焦点。最后，从中文短文本中抽取知识的最大挑战在于算法缺乏对文本的深度理解。受到中文语言学现象启发，我们研究中文短文本的习语性现象，并且探讨如何利用此现象进行更深层次的语义理解和知识推理。

1.4 研究意义

虽然关系抽取的相关研究比较充分，根据上述分析和介绍可知，从中文短文本抽取出语义关系与经典的关系抽取任务差异性较大。研究这一课题无论在算法层面还是应用层面均有重要的意义。结合我们的研究内容，简要概述这一课题的研究意义如下：

推动关系抽取研究的发展： 正如 Mausam [47] 和 Yao 等人 [48] 指出，深度神经网络的发展促进了关系抽取的发展；然而，关系的表达不仅仅局限于句子级别。仅仅考虑经典句子级别的关系抽取会极大地影响关系抽取的召回率。在本文中，我们考虑基于短文本的关系抽取，突破了传统研究的框架。为了适应中文短文本的语义语法特征，我们提出的模型也与端到端的神经网络有显著区别。这些研究思路对关系抽取的研究起到促进作用。

为短文本的处理提供计算框架： 由于缺乏足够的上下文作为语义支撑，短文本的处理一向被认为是 NLP 的一大难点。特别地，由于短文本语义很难分析而且很多短文本具有暗喻含义，在 NLP 迅猛发展的阶段仍然被认为是“A Pain in the Neck”[49]。在本文的研究中，我们利用词嵌入技术作为表示，建模短文本中词的语义；与之同时，不简单采取编码技术将短文本视为一个整体，而是结合了语言模式（例如 IPM 算法和 RCRL 算法，参见图 1.4，下同）、语言规则（例如 TPM 算法和 DNRE 算法）、频繁模式挖掘（例如 PNRE 算法）和统计量信息（例如 DNRE 算法）等多种技术协同作用，互相促进，提升短文本关系抽取的效果。我们的研究也为 NLP 中短文本的处理提供了一种可能的计算框架，对解决短文本处理的其他难题或任务提供有效的参考。

构建具有更高解释性的深度学习算法： 深度学习算法的一大弱点在于可解释

性较低，特别在安全攸关领域，这一弱点阻碍了其应用范围。在短文本的关系抽取工作中，我们注意到，由于上下文的高度缺乏，很多端到端的深度学习模型很难自动学习到解决相关任务的方法。在本文的研究中，特别是上下位关系抽取方面，我们提出多种投影模型，显示地建模在词嵌入空间如何将下位词的词向量投影到其对应的上位词上（例如 IPM、TPM、FOPM、TFOPM 等模型）。在用于词汇关系分类的 SphereRE 模型中，我们将具有不同词汇关系类别的术语对投射至超球嵌入空间的不同区域。从下文的分析的实验中，不难发现，相比经典基于神经网络的关系抽取模型，这些模型能对需要解决的任务本身建模更加精准，也同时具有更高的可解释性。

提升现有知识图谱的知识覆盖率： 从应用层面，本文研究的直接应用为提升现有知识图谱的知识覆盖率。在经典的知识图谱系统中，知识图谱的数据源大多集中于无结构的网络语料库（例如 NELL [9]、Probase [10] 等）和半结构化的百科词条信息框（例如 YAGO [8]、CN-DBPedia [11]）等。本文提出的算法，从短文本中抽取关系型知识，可以对现有知识图谱系统构成补充，提升其知识覆盖率。在下文的实验中，我们也确认，抽取出的关系元组大多不能被现有知识图谱覆盖[2]，因而对提升知识覆盖率的作用明显。由此也可推之，这些研究对基于知识图谱的下游 NLP 任务也有帮助，例如智能问答、查询理解等。受到篇幅和主题的限制，本文不对这些应用的相关算法进行讨论。

初步探索认知计算对于 NLP 发展的作用： 如上文所述，中文短文本中表述了大量人类的常识性知识，因此关系表达的上下文常常被省略。解决这一问题不仅仅对短文本进行**处理**（Processing），还需要深度**理解**（Understanding）。这一目标离不开对输入文本的**认知计算**（Cognitive Computation），即指模仿人类的大脑，对文本进行计算和理解 [50]。在本文的研究中，我们从认知计算和语言学的角度研究了中文短文本的习语性现象，对一类特殊的中文短文本进行深度语义理解，从而实现知识推理。这一研究也可以作为认知计算在 NLP 上的一种初步应用。

1.5 主要贡献

本文针对中文短文本的关系抽取中三个主要问题，包括基于词嵌入的上下位关系抽取、知识增强的语义关系抽取、以及非上下位关系抽取与语义理解进行了重点研究和探索。本文的主要贡献概述如下：

[2]详见第 4.3.2 节中 PNRE 算法的实验分析。

- 针对基于词嵌入的上下位关系抽取，我们提出了三个模型：**半监督式上下位关系扩展模型**（Iterative Projection Model，IPM）、**基于转导学习的上下位关系分类模型**（Transductive Projection Model，TPM）和**基于模糊正交投影的上下位关系分类模型**（Fuzzy Orthogonal Projection Model，FOPM）。

- 针对知识增强的语义关系抽取，我们提出了：i）**分类体系增强的对抗学习框架**（Taxonomy Enhanced Adversarial Learning，TEAL），用于融合跨知识源的上下位关系知识；ii）**迁移模糊正交投影模型**（Transfer Fuzzy Orthogonal Projection Model，TFOPM），及其扩展算法**迭代迁移模糊正交投影模型**（Iterative Transfer Fuzzy Orthogonal Projection Model，ITFOPM），用于实现跨语言上下位关系抽取；以及 iii）**超球关系嵌入模型**（Hyperspherical Relation Embedding Model，SphereRE），用于区分多种类别的词汇关系。

- 针对非上下位关系抽取与语义理解，我们提出了**基于模式的非上下位关系抽取算法**（Pattern-based Non-hypernymy Relation Extraction，PNRE）和**数据驱动的非上下位关系抽取算法**（Data-driven Non-hypernymy Relation Extraction，DNRE），用于从中文短文本中自动抽取多种非上下位关系。接着，提出了中文复合名词的习语性程度分类问题，并且在**关系性与组合性表示学习**（Relational and Compositional Representation Learning，RCRL）框架下分析中文短文本的习语性，探究相关知识推理和语义理解的应用。

1.6 章节安排

本文余下内容的章节安排与图 1.4中表述相一致。在第二章中，我们关注基于词嵌入的上下位关系抽取问题。首先从模式匹配和分布式学习两个角度综述上下位关系抽取的相关研究，接着依次介绍 IPM、TPM 和 FOPM 三个模型的算法细节，及相应的实验分析。

在第三章中，我们围绕知识增强的语义关系抽取这一主题，研究工作从多知识源、多语言、多词汇关系三个角度展开。在相关工作中，我们扩展了第二章的综述研究，分析了对抗学习和迁移学习是怎样提升多个 NLP 任务的效果（特别是上下位关系抽取任务），并且概述了词汇关系分类的研究现状。在算法研究部分，依次介绍了多知识源、多语言、多词汇关系的三个相关任务，及对应的 TEAL、TFOPM（包括其扩展版本 ITFOPM）和 SphereRE 三种模型。

第四章的主题为非上下位关系抽取与语义理解，即在开放域下抽取任意类别的非上下位关系，并对中文短文本深度理解。在这一章中，我们首先对基于短文本的关系抽取、常识性关系抽取及文本理解等研究进行综述，接着介绍 PNRE 和 DNRE 两种关系挖掘模型，最后围绕中文短文本的习语性问题提出了 RCRL 模型，探索了这一研究是如何提升中文短文本的语义理解能力的。

最后，在第五章中，我们总结了全文的工作，并且针对研究工作的不足、以及 NLP 和深度学习的发展趋势，提出了多个未来研究方向和展望。

第二章　基于词嵌入的上下位关系抽取

根据第一章的概要分析，精准的中文上下位关系抽取是大规模中文知识图谱构建和补全的必要过程，也对中文自然语言理解的研究有重要的推动作用。由于中文语言模式的特殊性和灵活性，中文上下位关系抽取不能简单参照英语语言中基于 Hearst 模式（及其扩展模式）的文本匹配算法来实现 [10, 38]。在本节中，结合深度神经语言模型的最新研究成果和中文语言本身的特点，我们采用词嵌入表示作为中文术语的特征，学习中文上下位关系在词嵌入空间的数学表示，并提出三种模型，有效地抽取出中文上下位关系。这三种模型分别为**半监督式上下位关系扩展模型**、**基于转导学习的上下位关系分类模型**和**基于模糊正交投影的上下位关系分类模型**。接下来，我们首先对本章关注的抽取任务、相关研究工作和背景知识进行简单介绍，其次我们详细描述这三种模型的算法细节，及其相应实验结果。

2.1　引言

回顾第一章的研究目标，在面向中文短文本的上下位关系抽取任务中，输入为中文术语对 (x_i, y_i) 的集合，其中 x_i 为某一中文实体，y_i 是与 x_i 语义相关的概念或术语，典型示例如“（棕熊，哺乳动物）”、“（汽车，交通工具）”等。由于上位词和下位词一般均为名词或名词短语，我们采用基于词性标注的启发式过滤方法，去除不为名词或名词短语的 x_i 和 y_i，以剩下的中文术语或概念对作为本章算法的输入。此外，我们也构建一个大规模中文无标注语料库作为背景语料，并以此为输入训练词嵌入模型。记 $\vec{x}_i$ 和 $\vec{y}_i$ 分别为 x_i 和 y_i 的词嵌入表示，可以通过任何深度神经语言模型进行预训练得到，经典的模型包括 Word2Vec [51]、GloVe [52]、fastText [53] 等。

给定两个中文术语对训练集：正例训练集 $D^P = \{(x_i, y_i)\}$ 和负例数据集 $D^N = \{(x_i, y_i)\}$，及一个未标注数据集 $D^U = \{(x_i, y_i)\}$，本章的任务目标是利用部分或全部上述三个数据集训练模型；在测试阶段，对中文术语对测试集 $D^T = \{(x_i, y_i)\}$ 中的术语对之间是否具有上下位关系进行准确预测。在训练集中，对于每个 $(x_i, y_i) \in D^P$，我们要求 y_i 是 x_i 的上位词；对于每个 $(x_i, y_i) \in D^N$，y_i 是 x_i 的非上位词（x_i 和 y_i 可以有任意其他类型的语义关系，或 x_i 和 y_i 仅在语义上有一定关联度）。

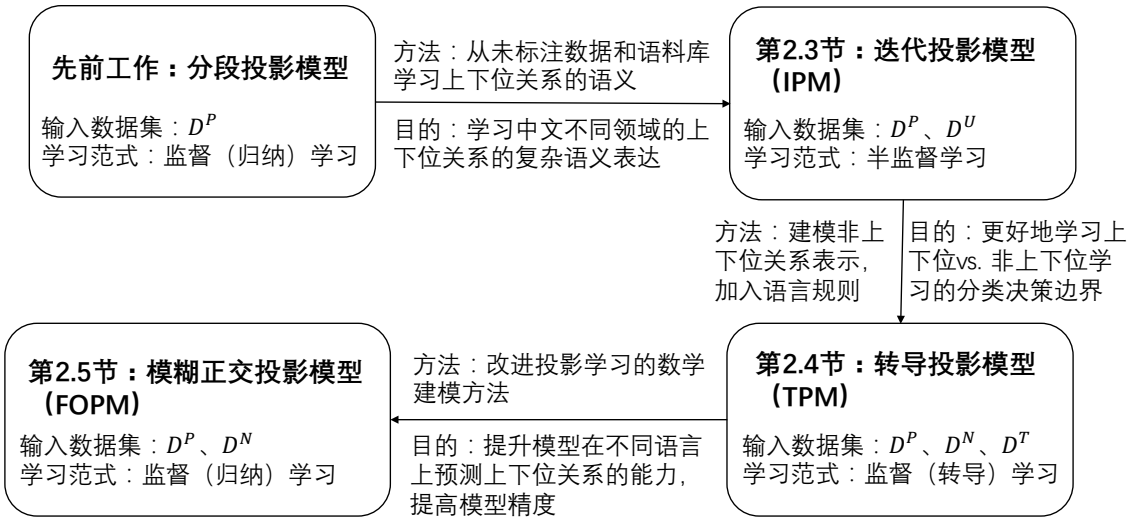


图 2.1: 第二章模型研究思路汇总

本章的研究从 Fu 等人 [25] 提出的分段投影模型出发，这一模型学习多个投影矩阵，将中文下位词的词向量投影到对应上位词的词向量空间中。这一模型的训练需要人工标注的中文上下位关系数据集。由于在中文语境下，人工标注的训练集（特别是上下位关系）的数据量一般较小，很难扩展到海量互联网环境，覆盖不断出现的新兴领域的上下位语义关系，我们首先提出**半监督式上下位关系扩展模型**，它利用正例训练集 D^P 和未标注数据 D^U 进行迭代训练，以发现更多新的上下位关系元组。

然而，这一模型没有建模非上下位关系的语义，因此上下位关系与非上下位关系在词向量空间的分类决策边界并不清晰。此外，中文中有相当数量的高质量语言规则，也不能有效被先前的模型所覆盖。我们进一步提出**基于转导学习的上下位关系分类模型**，在 D^P 和 D^N 上训练模型，并同时考虑测试集 D^T 上的待预测关系元组，以更好地区分中文上下位关系和非上下位关系。

在上述多个模型研究的基础上，我们从数学模型建模的角度改进了投影学习的方法，采用模糊正交矩阵建模从下位词到上位词之间的映射，提出了**基于模糊正交投影的上下位关系分类模型**。这一模型直接在 D^P 和 D^N 完成训练，在中文数据集上保持高预测精度的同时，也可以在多个英语公开评测数据集上超过或相当现有的最佳方法。

由于这三种模型都学习如何在词嵌入空间内，将中文下位词的词向量投影到中文上位词的词向量空间，在下文中，为了简单起见，分别将这三种模型简称为**迭代投影模型**（Iterative Projection Model，缩写为 IPM）、**转导投影模型**（Transductive

表 2.1: 第二章使用的重要符号及其意义

符号	说明
(x_i, y_i)	一个概念/术语对
$\vec{x}_i$	x_i 的词向量
D^P	上下位关系训练集
D^N	非上下位关系训练集
D^U	未标注概念/术语对数据集
D^T	上下位关系测试集
K	IPM 和 FOPM 中簇的数量
T	IPM 和 TPM 中的迭代训练次数
$\mathbf{M}_k^{(t)}$	IPM 中第 k 个簇第 t 个迭代的上下位关系投影矩阵
$\vec{b}_k^{(t)}$	IPM 中第 k 个簇第 t 个迭代的偏置投影向量
$P^{(t)}$	IPM 中第 t 个迭代的上下位关系训练集
$\mathbf{M}^P$	TPM 初始阶段中上下位关系投影矩阵
$\mathbf{M}^N$	TPM 初始阶段中非上下位关系投影矩阵
F_i	TPM 中 (x_i, y_i) 的上下位关系预测得分
$\mathbf{M}_k^P$	FOPM 中第 k 个簇的上下位关系投影矩阵
$\mathbf{M}_k^N$	FOPM 中第 k 个簇的非上下位关系投影矩阵

Projection Model，缩写为 TPM）和**模糊正交投影模型**（Fuzzy Orthogonal Projection Model，缩写为 FOPM）。表 2.1 总结了第二章中使用的重要符号及其意义。读者也可以参阅图 2.1中给出的本章模型的研究思路汇总。

2.2 相关工作

上下位关系抽取在 NLP 领域研究广泛。参考先前工作的分类 [54]，本节从模式匹配法和分布式学习法两大类出发，概述相关研究，并且对这两种方法的优缺点进行深入讨论。

2.2.1 基于模式匹配的上下位关系抽取

模式匹配法采用固定或相对灵活的语言模式从文本中自动挖掘出上下位关系元组。这一研究方法可以追溯至 Marti A. Hearst 教授于 1992 年提出的英语 Hearst 模式 [38]。典型的 Hearst 模式包括 "..., such as ..." 和 "..., including ..." 等。例如，在英语语句 "mammals, such as dogs and cats" 中，我们可以抽取出两个上下位关系元组 "（dog，mammal）"、"（cat，mammal）"。尽管 Hearst 模式形式简单，完全由人工制定，它们广泛被用于英语大规模分类体系的构建工作中。比较经典的利用 Hearst 模式构建的分类体系包括 Probase [10]、WebIsADB [55] 等。

直接采用 Hearst 模式进行文本匹配会降低上下位关系抽取的精准度和覆盖率，这是由于 Hearst 模式是完全固定的，缺乏灵活性和上下文自适应性。在实际应用中，为了提高精准度，抽取出的上下位关系候选元组可以通过置信度评分的方式进行筛选，获得高置信度的元组。在 Probase 系统中，Wang 等人采用“似然性比率”（Likelihood Ratio）这一统计指标，对每个上位词仅筛选出最有可能正确的下位词，对每个下位词仅筛选出最有可能正确的上位词，将相应的关系元组添加入 Probase 系统中 [10]。Luu 等人 [56] 在 Hearst 模式中加入了基于语法结构的限制条件，以减少单纯依靠字面信息匹配造成的噪声。除了采用单一评分机制，NLP 研究者也采用分类器对抽取出的候选元组进行正确、错误分类，以判断其正确性。Snow 等人首先提出了这一机制，并且采用语法统计信息作为分类器的特征 [57]。Bansal 等人 [58] 引入两个术语本身的特征（例如单词是否首字母大写、字符子串匹配），同时考虑他们在维基百科百科对应实体摘要的统计信息，增强分类器的效果。

比起精准度，更多的工作目的在于提高模式匹配方法的覆盖率，一类常见方法称为“模式泛化（Pattern Generalization）”，即采用更泛化的语言元素来替代 Hearst 模式的某些部分。Navigli 和 Velardi [59] 提出了“Star（*）模式”，即采用通配符（“*”）来取代上下位关系模式中最频繁的词，以增强上下位关系抽取模型的泛化能力。Nakashole 等人 [60] 设计了 PATTY 系统，把模式中部分的词替换为词性标注的标签（例如名词、动词）、这个词在知识本体中所属的类别（例如乐曲、歌手、国家）等。

另一类常见的方法为迭代式抽取法，即用少量人工给定的“种子”（“种子上下位关系元组”或“种子上下位关系模式”）作为输入，从海量语料库中自动挖掘出新的上下位关系元组及相关语言模式，将新发现的高置信知识加入抽取系统中，然后开始下一轮迭代抽取。这一类方法的典型代表是 [61]。给定少量某领域的上下位关系知识，它可以自监督地从网络文本中自动构建出整个领域分类体系。迭代式抽取存在的典型问题是“语义漂移（Semantic Drift）”，即在迭代学习过程中，错误的关系元组和语言模式被加入训练集中，使得抽取出的关系元组的语义与种子元组的语义逐渐不同。为了避免这一问题，Carlson 等人 [62] 提出了基于多视图学习的关系抽取算法，当且仅当从不同类别数据源训练的模型对该术语对的预测结果相同时，才将对应的关系元组抽取出来，加入训练集。在中文语言，也有模式匹配法在上下位关系抽取的实验分析 [25, 63]，其结果表明，这种方法的准确度和召回率很低，对中文灵活的语言模式不能适应。因此，本文并不直接采用语言模式进行

匹配，而是在部分模型中利用某些中文模式或语言知识（详见 IPM 和 TPM 两个模型），提升基于词嵌入模型的投影学习的效果。

除了考虑术语对之间的上下位关系模式，利用术语本身的语言特征（即构词法）进行扩展也可以获得更多上下位关系。例如，“哺乳动物” 的中心词是 “动物”，我们可以推断出 “哺乳动物” 和 “动物” 之间也具有上下位关系。这一启发式规则在 YAGO 系统 [8]、Taxify 系统 [64] 和基于上位词子序列的分类体系构建系统 [65] 中均有所应用。

2.2.2 分布式上下位关系预测

尽管模式匹配法能较为有效地从语料库中抽取出上下位关系，它的最大缺陷在于，当且仅当两个术语在同一句句子共现时，他们之间的上下位关系才有可能被算法抽取出来。具有上下位关系的术语对并非频繁出现在同一句子中，这一现象导致了共现稀疏性问题（Occurrence Sparsity），降低了模式匹配法的召回率 [66]。分布式方法（Distributional Methods）缓解了这一问题，直接利用两个术语的分布式表示作为模型输入的原始特征，来推断这两个术语是否具有上下位关系[1]。根据学习范式和学习任务的不同，分布式方法可大致分为两大类别：**非监督式上下位关系度量**（Unsupervised Hypernymy Measure）和**监督式上下位关系分类器**（Supervised Hypernymy Classifier）。

上下位关系度量： 上下位关系度量是典型的非监督分布式算法，对于任意术语对 (x_i, y_i)，它计算用于衡量 x_i 和 y_i 之间有上下位关系可能性评分。在一个上下位关系元组中，上位词的上下文语义往往比相应下位词的上下文语义更宽泛，这一假设被称为 “分布式包含假设”（Distributional Inclusion Hypothesis）。例如，“动物” 在语料库中的上下文涉及的语义比 “狗” 更宽泛。根据这一假设，学术界提出多个上下位关系度量，例如 WeedsPrec [67]、cosWeeds [68]、ClarkeDE [69]、invCL [68] 等，分别计算如下：

$$WeedsPrec(x_i, y_i) = \frac{\sum_{f \in F(x_i) \cap F(y_i)} \vec{x}_i[f]}{\sum_{f \in F(x_i)} \vec{x}_i[f]}$$

$$cosWeeds(x_i, y_i) = \sqrt{\cos(x_i, y_i) \cdot WeedsPrec(x_i, y_i)}$$

[1]在本文中，分布式的含义为利用词的上下文分布描述词的语义，可翻译为 “Distributional”，含义与 “分布式计算” 中的 “分布式”（Distributed）不同。

$$ClarkeDE(x_i, y_i) = \frac{\sum_{f \in F(x_i) \cap F(y_i)} \min\{\vec{x}_i[f], \vec{y}_i[f]\}}{\sum_{f \in F(x_i)} \vec{x}_i[f]}$$

$$invCL(x_i, y_i) = \sqrt{ClarkeDE(x_i, y_i) \cdot (1 - ClarkeDE(y_i, x_i))}$$

其中，$F(x_i)$ 是 x_i 的上下文特征集合，$\vec{x}_i[f]$ 是 x_i 的特征表示中特征 f 的值。例如在 WeedsPrec 中，如果 y_i 是 x_i 的上位词，根据分布式包含假设，x_i 的上下文极有可能被 y_i 的上下文所包含，因此，$\sum_{f \in F(x_i) \cap F(y_i)} \vec{x}_i[f]$ 与 $\sum_{f \in F(x_i)} \vec{x}_i[f]$ 的值较为接近，WeedsPrec 的评分会比较高。

Santus 等人 [70] 进一步观察到，上位词与它对应的下位词相比由于更加抽象，因而具有更少的信息量，即为 “分布式信息量假设”（Distributional Informativeness Hypothesis）。根据上述假设，他们提出了基于信息熵的上下位关系度量 SLQS：

$$SLQS(x_i, y_i) = 1 - \frac{\text{Medium}_{n=1}^{N}(H(c_n(x_i)))}{\text{Medium}_{n=1}^{N}(H(c_n(y_i)))}$$

其中，$\text{Medium}_{n=1}^{N}(\cdot)$ 表示取中位数的运算符，$c_n(x_i)$ 表示 x_i 在语料库上下文中词频为 Top-N 个词中的第 n 个词，$H(c_n(x_i))$ 表示 $c_n(x_i)$ 在对应语料库中的信息熵。SLQS 通过比较两个术语上下文的信息熵，判断他们之间的关系。此外，Roller 等人 [71] 进一步提出了 “选择性分布式包含假设”（Selective Distributional Inclusion Hypothesis），即上位词的上下文仅在部分维度上包含了下位词的上下文。因为上下位关系度量不是本文的研究重点，下文中我们不再列举这一系列的经典研究工作。读者也可以参阅文献 [24] 对上下位关系度量的综合实验评测。从评测结果可知，没有任何一个经典的上下位关系度量在所有的数据集中比其他度量结果更好，他们各有各的优势和不足。

随着深度神经网络技术在 NLP 的广泛应用，近年来的研究工作致力于将表示学习和上下位关系度量结合起来。其中，一个重要的方向为上下位关系嵌入（Hypernymy Embedding）学习，即为设计上下位关系敏感的词嵌入学习算法和相应度量，使得术语表示更加符合上下位关系的语义。HyperScore 度量 [72] 考虑了上下位关系的层次结构，利用负采样学习算法将词嵌入学习和上下位关系学习的目标统一到同一个模型 HyperVec 中。Chang 等人的工作 [73] 直接结合了分布式包含假设和词嵌入学习，并提出了分布式包含嵌入（Distributional Inclusion Embedding）模型，将相应的词向量作为词的特征。

除了少数情况，上下位关系一般具有传递性（Transitivity）[2]，而这种传递性很难在欧式空间利用词嵌入模型表达出来。为了学习更好的上下位关系表达，可以将术语嵌入在双曲嵌入空间中（Hyperbolic Embedding Space）。Nickel 和 Kiela [75] 提出双曲嵌入空间中的 Lorentz 模型，将分类体系的层次化概念结构嵌入到这一空间中，从几何学角度描述了分类体系中的上下位关系。另一项类似的研究工作由 Ganea 等人 [76] 在同一时期完成。根据前述分布式包含假设，上位词的语义一般 “蕴含” 下位词的语义，他们在双曲嵌入空间中引入双曲蕴含圆锥（Hyperbolic Entailment Cone）的概念，使上位词在双曲蕴含圆锥对应的位置下尽可能 “蕴含” 其下位词。

上下位关系分类器： 上下位关系度量的一个缺点在于，它没有直接利用训练集中的人工标注数据。监督式上下位关系分类器利用两个术语的词向量表示作为原始特征，训练二分类关系分类器，自动学习这些特征是如何帮助预测的。在经典的研究中，常常通过训练神经语言模型 Word2Vec [51]、GloVe [52] 等得到词向量，并且直接利用一个术语对 (x_i, y_i) 的词向量拼接 $\vec{x}_i \oplus \vec{y}_i$ [77]、词向量之差 $\vec{x}_i - \vec{y}_i$ [25, 78] 等组合作为特征，训练 SVM、逻辑斯蒂回归等模型进行关系分类。

由于上下位关系具有反对称性（Anti-symmetry），Roller 等人 [71] 提出了 asym 模型，利用一个术语对的词向量之差，及其平方差作为分类器的特征。对于术语对 (x_i, y_i)，这两组特征 $\vec{f}(x_i, y_i)$ 和 $\vec{g}(x_i, y_i)$ 分别定义为：

$$\vec{f}(x_i, y_i) = \frac{\vec{x}_i}{\|\vec{x}_i\|} - \frac{\vec{y}_i}{\|\vec{y}_i\|}$$

$$\vec{g}(x_i, y_i) = (\frac{\vec{x}_i}{\|\vec{x}_i\|} - \frac{\vec{y}_i}{\|\vec{y}_i\|})^2$$

simDiff 模型是上述模型的一种变体，由 Turney 和 Mohammad 提出 [79]。simDiff 分别计算两个术语与其他词之间的语义相似度，将两个语义相似度向量之差当成分类器的特征。此外，Yu 等人 [80] 观察到术语作为上位词和下位词时的语义有所区别，利用 Probase 中的海量上下位关系元组，对一个术语分别学习其作为上位词时和下位词时的词向量。他们提出最大间隔神经网络模型，学习 Probase 中术语的上位词表示和下位词表示，并进行关系分类。Luu 等人 [81] 利用维基百科作为数据源，提出动态加权神经网络，同时考虑上下位词的上下文信息学习词向量。前述

[2] 文献 [74] 指出，并非所有上下位关系都具有传递性，并提出算法检测分类体系中上下位关系是否具有传递性。在下文的研究中，暂不对这个问题进行详细讨论。

HyperVec 模型 [72] 除了能用于计算非监督式的 HyperScore 度量，也能作为监督式关系分类器的输入。

与上下位关系度量相比，这一类基于分类器的算法在精度上通常更高。这是由于上下位关系的表示不需要显示地定义出来，而是通过分类算法自动学习得到。然而，Levy 等人指出 [82]，这些方法存在“词汇记忆”（Lexical Memorization）问题，即分类器容易学习到两个术语本身的特征，而不是他们之间具有什么语义关系。所以，当训练集和测试集中涉及的术语在语义方面差异较大时，预测精度会明显下降。因此，不同类别的上下位关系预测方法各有优劣，如何结合不同类别方法的优点提升算法的准确性，成为一大研究焦点。

2.2.3 讨论

学术界对于模式匹配法和分布式学习法哪一种更有利于上下位关系预测并无明确定论。例如，很多研究都支持分布式学习法更为有效 [72, 81]。与之相反，Levy 等人 [82] 则认为分布式学习法不能学习上下位关系的语义表示。

近年来，研究者提出可以将这两种方法互相结合、互相吸收，统一到同一算法中。Shwartz 等人的研究最为典型 [54]，该算法采用长短期记忆网络（Long Short Term Network，LSTM）[83] 编码和上下位关系相关的语言模式，并将语言模式的向量表示与相应两个术语的词向量进行拼接，训练混合神经网络。这一工作在分布式上下位关系分类的框架下，利用深度神经网络在模型中加入了语言模式。Roller 等人 [66] 计算某一术语对在海量语料库中与 Hearst 模式 [38] 匹配的统计量，据此设计了基于模式匹配的上下位关系度量，其精度超过了经典的上下位关系度量。Le 等人 [84] 结合了模式匹配法和双曲嵌入学习，使得术语在双曲嵌入空间的表示与基于 Hearst 模式 [38] 的统计量相符合。Held 和 Habash [85] 指出，融合多种简单的上下位关系预测算法也能取得较好结果。

为了解决分布式模型的“词汇记忆”问题 [82]，分布式投影学习方法显示地建模下位词的词向量是如何映射到上下位的词向量空间的。Fu 等人提出了分段线性投影模型 [25]，是这一类方法的开创性研究。这一模型包括多个线性投影模型，将下位词的词向量映射至上位词的词向量，使得映射的误差最小化。在模型的预测阶段，对于某一术语对 (x_i, y_i)，如果 $\vec{x}_i$ 可以通过该模型较为精确地映射至 $\vec{y}_i$，则 x_i 和 y_i 之间很可能有上下位关系；反之，x_i 和 y_i 之间很可能没有上下位关系。由此可见，这一模型在对上下位关系的语义在分布式框架下建模的同时，实现了上下位

关系分类。由于 Fu 等人的模型中需要预先指定投影模型的数量，Yamane 等人 [86] 改进了这一算法，同时学习线性投影模型的数量，以及每个线性投影模型的参数。Biemann 等人 [87] 在线性投影模型上加入了负采样正则化项，这一正则化项建模了非上下位关系的分布式语义表示，从而更容易区分上下位关系与非上下位关系。

本章中提出的三种模型也属于分布式投影学习方法，分别从迭代学习、转导学习和模糊正交投影的角度提升模型学习的效果。此外，这些模型的设计也考虑了上下位关系分类机制和中文语言的特性，从而对上下位关系和非上下位关系（特别是中文语境下）进行精确分类。

2.3 半监督式上下位关系扩展模型

从本节开始，我们详细介绍提出的三种基于词向量的投影模型，用于从中文短文本中自动挖掘上下位关系。回顾第2.1节介绍，这些模型的训练需要人工标注的数据集作为训练数据，由于公开的中文上下位关系训练集（特别是正例数据集 D^P）大小一般很有限，我们首先考虑如何利用半监督迭代学习方法，自动从网络数据源中挖掘出更多的中文上下位关系。

2.3.1 算法模型

这一半监督迭代学习的模型（即 IPM）改进自 Fu 等人基于词向量的的中文概念层次构建工作 [25]。在他们的工作中，他们采用**分段线性投影模型**（Piecewise Linear Projection Model），将下位词的词向量投影至上位词的词向量。投影模型的参数从中文语义词典《同义词词林（扩展板）》[3]学习到。在本工作中，我们进一步研究中文上位词和下位词词向量的关系。我们从上述语义词典和大规模中文分类体系 CN-WikiTaxonomy [21] 采样得到更多数量的上下位关系元组 (x_i, y_i)，并观察上下位词的词向量之差的 l_2 范数（即 $\|\vec{x}_i - \vec{y}_i\|_2$）与他们之间语义关系的联系。根据观察结果，我们发现有如下三种现象。示例如表 2.2所示[4]：

现象 1 对于固定的下位词 x，若 y_1 和 y_2 是不同层次粒度的上位词，则很有可能 $\vec{x} - \vec{y}_1 \not\approx \vec{x} - \vec{y}_2$。例如，“国家” 和 “亚洲国家” 都是 “日本” 的上位词；然而，“亚洲国家” 抽象度较低，“国家” 抽象度较高。

[3] http://www.ltp-cloud.com/download/

[4] 在表 2.2中，$\vec{v}(\cdot)$ 表示对应中文的词向量，词向量的计算过程详见实验设置。

表 2.2: 上下位词的词向量之差与其语义关系的示例

类别	**示例**	$\|\vec{x}_i - \vec{y}_i\|_2$
真正例	$\vec{v}$(日本) $-$ $\vec{v}$(国家) $\approx$ $\vec{v}$(澳大利亚) $-$ $\vec{v}$(国家)	$1.03 \approx 0.99$
现象 1	$\vec{v}$(日本) $-$ $\vec{v}$(国家) $\not\approx$ $\vec{v}$(日本) $-$ $\vec{v}$(亚洲国家)	$1.03 \not\approx 0.71$
现象 2	$\vec{v}$(日本) $-$ $\vec{v}$(国家) $\not\approx$ $\vec{v}$(主权国) $-$ $\vec{v}$(国家)	$1.03 \not\approx 1.32$
现象 3	$\vec{v}$(日本) $-$ $\vec{v}$(国家) $\not\approx$ $\vec{v}$(西瓜) $-$ $\vec{v}$(水果)	$1.03 \not\approx 0.39$

现象 2 若存在两个不同语义类别的上下位关系 "(x_1, instanceOf, y_1)" 和 "(x_2, subClassOf, y_2)", 则很有可能 $\vec{x}_1 - \vec{y}_1 \not\approx \vec{x}_2 - \vec{y}_2$。这是因为 instanceOf 和 subClassOf 虽然都表示广义的上下位关系，但是 instanceOf 表示实体与概念之间的从属关系，而 subClassOf 表示概念之间的类别覆盖关系。他们的语义略有不同，因而他们在词嵌入空间的表示有区别。

现象 3 若两个上下位关系 "(x_1, y_1)" 和 "(x_2, y_2)" 属于的语义话题不同，则很有可能 $\vec{x}_1 - \vec{y}_1 \not\approx \vec{x}_2 - \vec{y}_2$。这一结论与先前的研究 [25] 相吻合，表明上下位可以根据主题划分成更加细粒度的子关系（例如政治、军事、娱乐等），每个子关系在词嵌入空间的语义表达不同。

上述观察印证了建模上下位关系语义的困难性，也反映出先前工作的局限性。例如在文献 [25] 中，作者在中文语义词典数据集上训练模型，将其运用于中文百科数据，用于构建互联网中文分类体系。由于中文语义词典数据大多为通用领域的 subClassOf 关系，中文百科数据中大多为 instanceOf 关系，且包括很多互联网新兴领域术语；这两类数据源涉及到的上下位关系的语义相差较大。与上述技术不同，在 IPM 算法中，我们采用两阶段的训练方法。在初始阶段，我们在较小的人工标注训练集 D^P 上进行分段投影模型的初始训练。接着，从大规模互联网未标注数据集 D^U 中选出高置信度的正例，将其加入 D^P，并更新相应投影模型的参数，开始下一轮迭代。其算法框架如图 2.2所示。

初始训练阶段： 本算法的初始阶段是方法 [25] 的一种变体。我们采用 K-Means 算法对 D^P 中的上下位关系聚成 K 个簇，采用词向量之差 $\vec{x}_i - \vec{y}_i$ 作为特征。设归属于第 k 个簇的上下位关系集合为 C_k（$k = 1, 2, \cdots, K$），其对应的投影矩阵和偏置向量为 $\mathbf{M}_k$ 和 $\vec{b}_k$。在初始阶段，IPM 对于第 k 个簇的优化目标为最小化如下函数的值：

$$J(\mathbf{M}_k, \vec{b}_k; C_k) = \frac{1}{|C_k|} \sum_{(x_i, y_i) \in C_k} \|\mathbf{M}_k \cdot \vec{x}_i + \vec{b}_k - \vec{y}_i\|^2$$

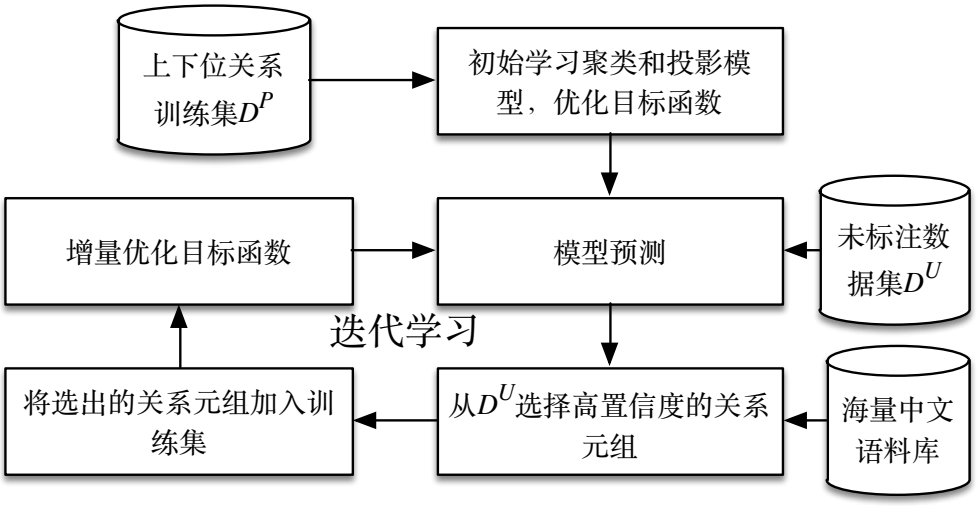


图 2.2: IPM 的算法框架

对于 $k = 1, 2, \cdots, K$，我们采用随机梯度下降算法分别学习 $\mathbf{M}_k$ 和 $\vec{b}_k$ 的值。

迭代训练阶段： 这一阶段的训练集在 T 个迭代周期内动态扩大。记为第 t 个迭代的上下位关系训练集为 $P^{(t)}$（$t = 1, 2, \cdots, T$）。由于在迭代过程中，这 K 个簇的簇中心和属于该簇的数据也在不停变化中。我们设第 t 个迭代的第 k 个簇（关系元组集合）为 $C_k^{(t)}$，簇的中心向量为 $\vec{c}_k^{(t)}$，对应的参数为 $\mathbf{M}_k^{(t)}$ 和 $\vec{b}_k^{(t)}$。这些参数和数据集的初始化设置为：$P^{(1)} = D^P$、$C_k^{(1)} = C_k$、$\vec{c}_k^{(1)} = \frac{1}{|C_k|}\sum_{(x_i,y_i)\in C_k} \vec{x}_i - \vec{y}_i$，$\mathbf{M}_k^{(1)} = \mathbf{M}_k$ 和 $\vec{b}_k^{(1)} = \vec{b}_k$。

对于每个迭代 $t = 1, \cdots, T$，模型参数和数据集按照如下过程更新：

步骤 1 随机从 D^U 采样 $\delta \cdot |D^U|$ 个关系元组并记为 $U^{(t)}$，其中 δ 为采样因子。对于每个元组 $(x_i, y_i) \in U^{(t)}$，根据当前迭代轮次的 K-Means 模型，按照下式分配其所属的簇的 ID，记为 p_i：

$$p_i = \text{argmin}_{k=1,\cdots,K} \|\vec{x}_i - \vec{y}_i - \vec{c}_k^{(t)}\|$$

计算当前元组 $(x_i, y_i) \in U^{(t)}$ 对于选定的第 p_i 个簇的投影差值 $d^{(t)}(x_i, y_i)$：

$$d^{(t)}(x_i, y_i) = \|\mathbf{M}_{p_i}^{(t)} \cdot \vec{x}_i + \vec{b}_{p_i}^{(t)} - \vec{y}_i\|$$

根据计算得到的 $d^{(t)}(x_i, y_i)$ 的值，我们定义预测 (x_i, y_i) 关系标签的函数为：

$$f_M^{(t)}(x_i, y_i) = I(d^{(t)}(x_i, y_i) < \epsilon)$$

其中 $I(\cdot)$ 是指示函数, ϵ 是预定义的阈值。$f_M^{(t)}(x_i, y_i) = 1$ 表示第 t 个迭代的投影模型预测 (x_i, y_i) 有上下位关系（即正例），$f_M^{(t)}(x_i, y_i) = 0$ 则反之。记 $U_*^{(t)}$ 为 $U^{(t)}$ 的子集，它包含 $U^{(t)}$ 中所有模型 $f_M^{(t)}(x_i, y_i)$ 预测为正例的未标注关系元组，即为：$U_*^{(t)} = \{(x_i, y_i) \in U^{(t)} | f_M^{(t)}(x_i, y_i) = 1\}$。

步骤 2 对于 $U_*^{(t)}$ 中每个关系元组 (x_i, y_i)，再用基于模式的关系选择算法（详见下文）预测其标签。若 (x_i, y_i) 被预测为有上下位关系，则定义 $f_P^{(t)}(x_i, y_i) = 1$；否则，$f_P^{(t)}(x_i, y_i) = 0$。定义集合 $U_+^{(t)}$ 为选出的高置信度上下位关系元组集合，即为 $U_+^{(t)} = \{(x_i, y_i) \in U_*^{(t)} | f_P^{(t)}(x_i, y_i) = 1\}$。利用集合 $U_+^{(t)}$ 更新数据集：$D^U = D^U \setminus U_+^{(t)}$，$P^{(t+1)} = P^{(t)} \cup U_+^{(t)}$。

步骤 3 记 $U_k^{(t)}$ 为 $U_+^{(t)}$ 内属于第 k 个簇的关系元组集合。我们采用以下公式更新簇的中心 $\vec{c}_k^{(t)}$，求得下一轮迭代的簇中心 $\vec{c}_k^{(t+1)}$：

$$\vec{c}_k^{(t+1)} = \vec{c}_k^{(t)} + \lambda \cdot \frac{1}{|U_k^{(t)}|} \sum_{(x_i, y_i) \in U_k^{(t)}} (\vec{x}_i - \vec{y}_i - \vec{c}_k^{(t)})$$

其中，$\lambda \in (0, 1)$ 是控制簇中心 "漂移" 的学习率。根据新的簇中心计算结果，对于每个 $P^{(t+1)}$ 中的关系元组 (x_i, y_i)，重新计算他们所属于的簇。这样，我们计算出了第 $t+1$ 个迭代的第 k 个簇 $C_k^{(t+1)}$。

步骤 4 对于每个簇 $C_k^{(t+1)}$，按最小化如下目标函数的方式，更新模型参数 $\mathbf{M}_k^{(t+1)}$ 和 $\vec{b}_k^{(t+1)}$：

$$J(\mathbf{M}_k^{(t+1)}, \vec{b}_k^{(t+1)}; C_k^{(t+1)}) = \frac{1}{|C_k^{(t+1)}|} \sum_{(x_i, y_i) \in C_k^{(t+1)}} \|\mathbf{M}_k^{(t+1)} \cdot \vec{x}_i + \vec{b}_k^{(t+1)} - \vec{y}_i\|^2$$

为了缩短训练时间，我们按如下公式初始化模型参数：$\mathbf{M}_k^{(t+1)} = \mathbf{M}_k^{(t)}$ 和 $\vec{b}_k^{(t+1)} = \vec{b}_k^{(t)}$。这保证了模型参数只需要在每个迭代中增量更新。

模型预测阶段： 模型训练结束后, 对于测试集中的每个关系元组 $(x_i, y_i) \in D^T$, 我们预测 y_i 是 x_i 的上位词，当且仅当如下两个条件满足至少一个：

条件 1 (x_i, y_i) 在 $P^{(T+1)}$ 的**传递闭包**（Transitive Closure）中。

根据上下位关系的传递性，已知 $P^{(T+1)}$ 中的关系元组有很大概率是正确的，则其传递闭包的正确概率也可以保证。

条件 2 $f_M^{(T+1)}(x_i, y_i) = 1$。

根据模型假设，最终模型的训练数据融合了最初的训练集以及从互联网数据源中挖掘的未标注关系元组。我们直接采用最终模型对测试数据预测关系标签，不再使用关系选择算法。

从上述描述可见，这一算法结合了语义和语法的抽取过程，同时支持增量学习。步骤 2 限制了当且仅当投影学习和关系选择预测结果都为正例时，我们才将相应未标注关系元组加入增量更新的训练集。这一思路与 Carlson 等人提出的多视图抽取算法相似 [9]。求解步骤 3 中簇中心更新的递归方程的闭式解，我们可以得到如下结果：

$$\vec{c}_k^{(T+1)} = (1-\lambda)^T \cdot \vec{c}_k^{(1)} + \lambda \cdot \sum_{t=1}^{T} \left(\frac{(1-\lambda)^{T-t}}{|U_k^{(t)}|} \cdot \sum_{(x_i, y_i) \in U_k^{(t)}} (\vec{x}_i - \vec{y}_i - \vec{c}_k^{(t)}) \right)$$

所以，$\vec{c}_k^{(T+1)}$ 对于 $\vec{c}_k^{(1)}$ 的增量可以看成加入簇的关系元组词向量差的加权平均值，权重随着迭代时间按指数方式衰减。随着簇和模型参数的更新，这一模型可以逐渐学习到未标注数据集中上下位关系的表示。

关系选择算法： 本节详细介绍步骤 2 中的关系选择算法。尽管用于中文关系抽取的语言模式不能保证高精准度和覆盖率，我们可以用这一方法在半监督学习迭代过程中 “校验” 被逐步加入训练集的数据的正确性。在先前的研究中，Fu 等人设计了几个中文对应的 Hearst 模式 [63]。我们进一步扩展了这一工作，并把和中文上下位关系有关的语言模式归为三类：“Is-A（是模式）”、“Such-As（例如模式）” 和 “Co-Hyponym（同下位词模式）”。这三种模式的示例详见表 2.3：[5]

我们可以观察到两个现象：

现象 4 如果 x_i 和 y 匹配 “Is-A” 或 “Such-As” 模式，有很大可能 x_i 是 y 的下位词。

定义 $n_1(x_i, y)$ 为 x_i 和 y 在语料库中匹配 “Is-A” 或 “Such-As” 模式的次数。

现象 5 如果 x_i 和 x_j 匹配 “Such-As” 或 “Co-Hyponym” 模式，有很大可能 x_i 和 x_j 之间没有上下位关系。

[5]在实际的中文文本中，“Such-As” 和 “Co-Hyponym” 模式中可能包括多个候选下位词。为简单起见，表中只列出两个，记为 x_i 和 x_j。

表 2.3: 三种中文上下位关系模式的示例

类别	**示例**
Is-A（是模式）	x_i 是一个 y x_i 是一种 y x_i 是 y 之一
Such-As（例如模式）	y，例如 x_i、x_j y，包括 x_i、x_j x_i、x_j 等 y，特别是 x_i、x_j
Co-Hyponym（同下位词模式）	x_i、x_j 等 x_i 和 x_j x_i 以及 x_j

定义 $n_2(x_i, x_j)$ 为 x_i 和 x_j 在语料库中匹配 “Such-As” 或 “Co-Hyponym” 模式的次数，$n_2(x_i)$ 为 x_i 和 x^* 匹配的次数（其中，x^* 是除 x_i 以外的任意下位词）。

本算法中，我们同时利用前述投影模型的预测结果和基于中文上下位关系模式的统计信息来决定 $U_*^{(t)}$ 中哪些关系元组应该被加入 $U_+^{(t)}$。对于 $U_*^{(t)}$ 中的每个关系元组 (x_i, y_i)，分别记 $PS^{(t)}(x_i, y_i)$ 和 $NS^{(t)}(x_i, y_i)$ 为正负向得分。其中，正向得分 $PS^{(t)}(x_i, y_i)$ 的计算基于投影模型预测结果和现象 4 的相关统计量：

$$PS^{(t)}(x_i, y_i) = \alpha \cdot (1 - \frac{d^{(t)}(x_i, y_i)}{\max_{(x,y)\in U_*^{(t)}} d^{(t)}(x, y)}) + (1 - \alpha) \cdot \frac{n_1(x_i, y_i) + \gamma}{\max_{(x,y)\in U_*^{(t)}} n_1(x, y) + \gamma}$$

其中，$\alpha \in (0, 1)$ 是可调的权重，γ 是平滑系数。这两个系数的经验值可以设置为 $\alpha = 0.5$ 和 $\gamma = 1$。基于现象 5，我们定义如下负向得分 $NS^{(t)}(x_i, y_i)$：

$$NS^{(t)}(x_i, y_i) = \log \frac{n_2(x_i, y_i) + \gamma}{(n_2(x_i) + \gamma) \cdot (n_2(y_i) + \gamma)}$$

如果 x_i 和 y_i 负向得分高，x_i 和 y_i 在 “Such-As” 和 “Co-Hyponym” 模式中频繁出现。x_i 和 y_i 有可能为同下位词（Co-hyponym），他们之间有上下位关系的概率低。

为了同时最大化正向得分和最小化负向得分，我们将这一优化问题定义成带负向得分约束的正向得分最大化问题：

$$\begin{aligned} &\max \sum_{(x_i,y_i)\in U_+^{(t)}} PS^{(t)}(x_i, y_i) \\ &\text{s. t.} \sum_{(x_i,y_i)\in U_+^{(t)}} NS^{(t)}(x_i, y_i) < \Delta, U_+^{(t)} \subset U_*^{(t)}, |U_+^{(t)}| = n_+ \end{aligned}$$

Algorithm 1 贪心关系选择算法

初始化 $U_+^{(t)} = \emptyset$
while $|U_+^{(t)}| < n_+$ **do**
  选择正向得分最大的关系元组：$(x_i, y_i) = \text{argmax}_{(x,y)\in U_*^{(t)}} PS^{(t)}(x, y)$
  从 $U_*^{(t)}$ 中移除该元组：$U_*^{(t)} = U_*^{(t)} \setminus \{(x_i, y_i)\}$
  if $NS^{(t)}(x_i, y_i) + \sum_{(x,y)\in U_+^{(t)}} NS^{(t)}(x, y) < \Delta$ **then**
    将关系元组加入 $U_+^{(t)}$：$U_+^{(t)} = U_+^{(t)} \cup \{(x_i, y_i)\}$
  end if
end while
return 关系元组集合 $U_+^{(t)}$

其中，n_+ 为要求获得的 $U_+^{(t)}$ 的集合大小，Δ 是负向得分约束上限。这个问题是**带预算的最大覆盖问题**（Budgeted Maximum Coverage Problem）[88] 的特例，是 NP 难问题。根据文献 [88] 的证明，目标函数是**单调**（Monotonic）且**次模**（Submodular）的。我们设计贪心的关系选择算法优化这一问题，其准确度为 $1 - \frac{1}{e}$。具体过程见算法 1。在不违反约束的前提下，算法从 $U_*^{(t)}$ 中贪心地选择正向得分最大的关系元组，将其加入 $U_+^{(t)}$。从上述过程可见，我们只需要对关系元组进行依次正向得分排序，然后选择符合条件的元组即可。这一算法将一个 NP 难问题的复杂度降为 $O(n \log n)$，其中 $n = |U_*^{(t)}|$。

最后，对于每个关系元组 $(x_i, y_i) \in U_*^{(t)}$，我们设计基于关系选择的预测函数如下：$f_P^{(t)}(x_i, y_i) = I((x_i, y_i) \in U_+^{(t)})$。

2.3.2 实验分析

实验设置： 我们使用四个中文上下位关系数据集和中文语料库作为实验数据，这四个数据集的统计信息见表 2.4。算法的训练集从包括 131.8 万个上下位关系的中文分类体系 CN-WikiTaxonomy[21] 中采样得到，一共含有 7312 个关系元组，我们用它训练初始投影模型。CN-WikiTaxonomy 的数据集已在 GitHub 上开源[6]。

对于未标注数据集，我们随机采样了 10 万个百度百科实体，使用规则过滤了不包括类别标签的实体，最后得到了约 7.8 万个 “实体-类别” 对。我们使用了 Fu 等人 [25] 公开发布的标注数据集评测我们的算法，包括了 1391 个上下位关系和 4294 个非上下位关系。我们随机划分了 1/4 的数据作为验证集，余下 3/4 的数据作为测试集。

[6] https://chywang.github.io/data/apweb2015.zip

表 2.4: 四个中文上下位关系数据集的统计信息

数据集	正例数量	负例数量	未标注关系数量
CN-WikiTaxonomy 数据集（训练集）	7,312	-	-
百度百科未标注数据集	-	-	78,080
验证集	349	1,071	-
测试集	1,042	3,223	-

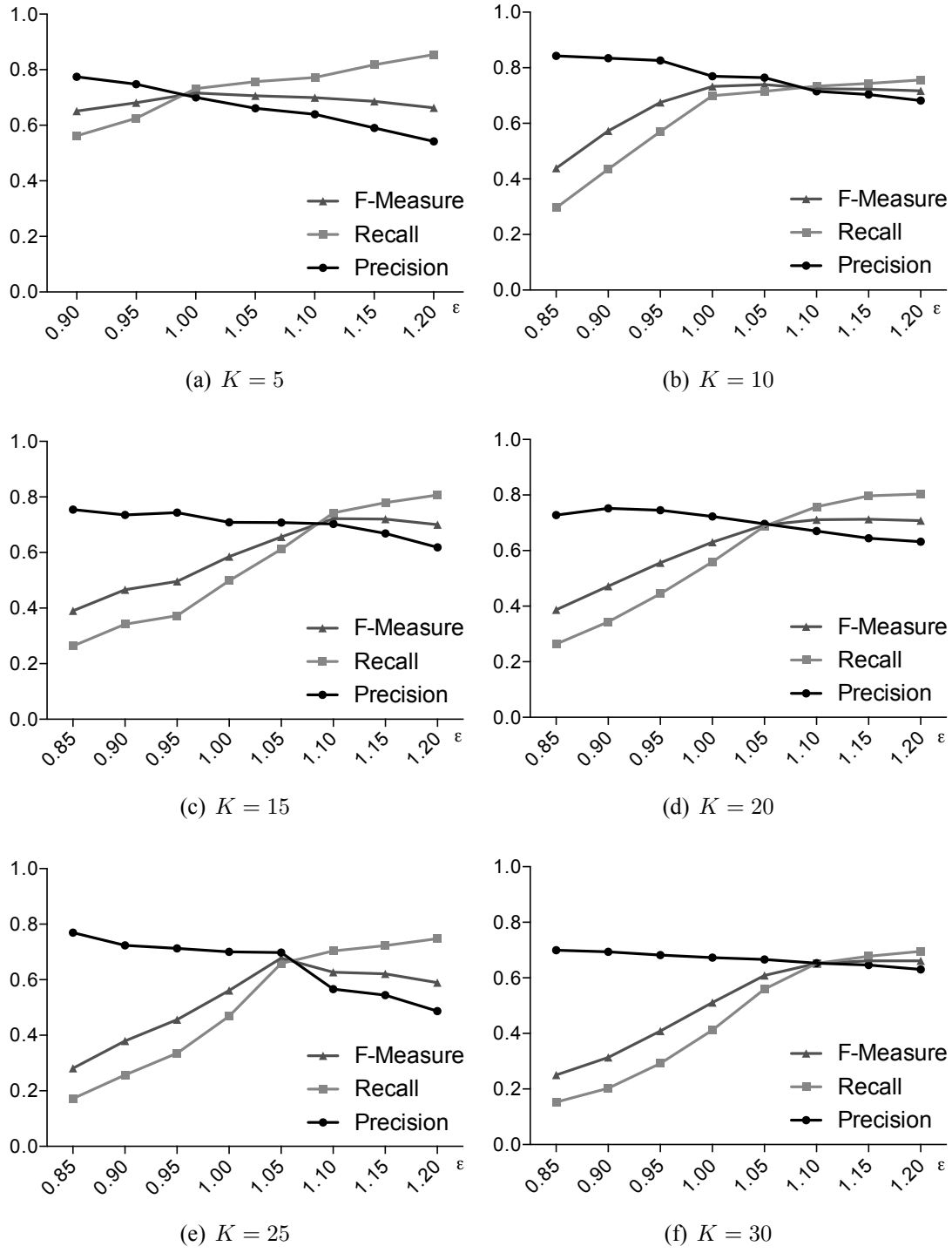
为了学习中文词语的词向量，我们爬取了约 130 万个百度百科的实体信息页面，得到了一个大规模中文网络语料库，包含了约 10.88 亿个词（采用开源工具 Ansj[7]分词后的统计数据）。我们在这些数据上训练了 Skip-Gram 模型 [51]，得到了 580 万个中文词语的 100 维度的词向量。

模型评测： 在模型的初始训练阶段，我们首先调节簇的个数 K 和阈值 ϵ 这两个超参数的值。其中，K 从 5 逐渐增大到 30，ϵ 从 0.85 逐步增大至 1.20，在验证集上完成效果的评测，实验结果详见图 2.3。可以看出，当簇的个数设置为 10，$\epsilon = 1.05$ 时，初始模型的效果最好，最高的 F 值为 73.9%。

在迭代学习阶段，我们设置迭代次数 $T = 20$，并记录每个迭代结束后模型的效果。与初始阶段过程相同，在验证集上调节超参数的效果，最终设置 $\delta = 0.2$、$\lambda = 0.5$，在每个迭代中按照算法 1选择 500 个新的关系元组进入训练集。如图 2.4(a)，在前 10 个迭代中，F 值从 74.9% 逐步上升至 78.5%，这一实验结果说明了在训练集中加入新的上下位关系有助于提升模型的效果。随着迭代的继续，F 值轻微下降，并在 15 个迭代后在 76.7% 左右保持稳定。F 值下降最有可能的原因是尽管我们设计关系选择算法避免“语义漂移” [62]，小部分伪正例仍然会不可避免地被加入训练集。通过人工检查，我们发现部分伪正例包括“（脂肪，健康）”、“（萧亚轩，时尚）”、“（信息，科学）”等，他们表达两个术语之间的语义相关性，而非严格的上下位关系。在图 2.4(b) 中，我们直接随机采样 500 个被先前一轮迭代的模型预测为正例的关系元组加入训练集，不进行关系选择。从实验结果中可以看出，尽管在初始阶段模型的表现有上升趋势，其 F 值迅速明显下降，因为大量伪正例被加入了训练集，“污染”了训练数据集，造成学习到的投影模型参数有偏差。

算法比较： 我们在测试集上评测 IPM 算法的准确性，并与其他基线算法对比，结果汇总在表 2.5中。为了在百度百科的中文语料库上评测我们的算法，我们重新在中文语言上实现了三个基于语料库的上下位关系抽取算法。Hearst [38] 算法最初用于英语语言的抽取，我们用 Fu 等人 [63] 人工翻译成中文的语言模式替换 Hearst

[7]http://nlpchina.github.io/ansj_seg/

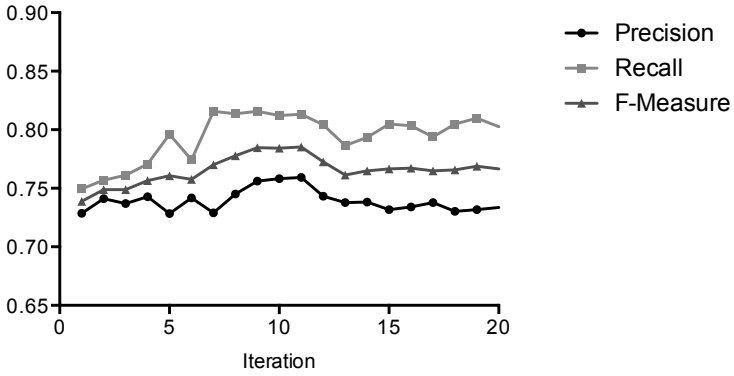


(a) $K = 5$ (b) $K = 10$

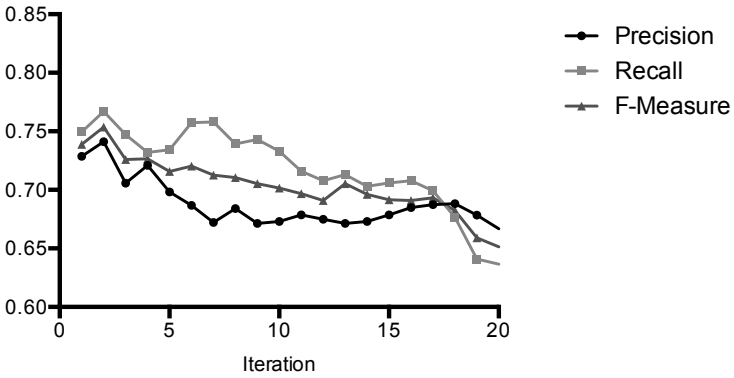
(c) $K = 15$ (d) $K = 20$

(e) $K = 25$ (f) $K = 30$

图 2.3: IPM 中参数簇的个数 K 和阈值 ϵ 的变化对于验证集模型预测效果的影响



(a) 加入关系选择算法



(b) 去除关系选择算法

图 2.4: IPM 在每个迭代的测试结果

设计的语言模式。结果表明，这种人工订制模式的算法对中文语言的覆盖率很低，不适合具有灵活语言表达的情况。Snow [57] 算法采用自动语言模式检测算法，比起 Hearst [38] 算法有更高的灵活性。它把召回率从 19.8% 提升至 28.1%，然而精准度下降至 28.9%，因为中文语言的解析和模式分析仍然不够精准。分布式相似性 invCL 度量 [68] 的 F 值为 58.1%，因为在无结构文本中，实体的上下文噪声仍然比较多，使得上下位关系难以在其他关系中被区分出来。此外，我们直接用基于中文维基百科的分类体系 CN-WikiTaxonomy [21] 在测试集上进行关系匹配，其精准度达到 98.5%，但是由于该系统对于中文概念实体的覆盖率仍然有限，其召回率比较低。与我们的方法对比，最具有竞争力的算法为基于词嵌入的模型 [25]，其 F 值为 73.3%。这一实验结果体现出，学习词嵌入的投影能有效建模中文上下位关系的语义。

我们进一步讨论本节提出的 IPM 算法及其变体在测试集上的表现。在表 2.5中，IPM-Initial 指的是仅采用我们初始阶段训练的模型进行预测，这一模型与 Fu 等

表 2.5: IPM 在测试集上的效果，及其与基线算法的对比

方法	**精准度**	**召回率**	**F 值**
基线方法			
Hearst [38]	0.962	0.198	0.328
Snow 等人 [57]	0.673	0.281	0.396
CN-WikiTaxonomy [21]	**0.985**	0.254	0.404
invCL [68]	0.485	0.581	0.529
Fu 等人 [25]	0.717	0.749	0.733
IPM 及其变体			
IPM-Initial	0.741	0.767	0.753
IPM-Random	0.690	0.757	0.722
IPM-Positive	0.754	0.801	0.776
IPM	0.758	0.814	0.786
IPM&CN-WikiTaxonomy	0.788	**0.847**	**0.816**

人 [25] 的模型类似，提升了 2% 的 F 值。IPM-Random、IPM-Positive 和 IPM 采用迭代学习算法更新投影模型的参数，IPM-Random 不使用关系选择算法，直接将上一轮迭代中预测为正例的关系随机加入训练集，IPM-Positive 使用关系选择算法，但是优化目标只设为最大化正向得分，不考虑负向得分的限制，IPM 是我们算法的完整实现。与 IPM-Initial 相比，IPM-Positive 和 IPM 提升了 2.3% 和 3.3% 的 F 值，显示出迭代学习算法能更好地学习上下位关系的表示。整体而言，我们的算法超过先前最佳算法 [25]，提升了 5.3% 的 F 值。如果将我们的算法和 CN-WikiTaxonomy [21] 相结合（在表中表示为 IPM&CN-WikiTaxonomy），其 F 值达到了 81.6%，这也比 Fu 等人的算法结合中文语义词典的效果更好 [25]。

错误分析： 我们对算法预测的错误原因进行分析。大部分错误案例（大约 72%）源自于很难区分语义相关性和严格的上下位关系。在测试集中，一些中文术语对的语义关系非常紧密，但是不是上下位关系，包括“（中药，药草）”、“（元帅，军事）”等。以前者为例，大部分中药的成分是药草，但是从严格的医学角度来说，并非所有中药材为药草，因此他们之间没有严格的上下位关系。如果没有额外知识，模型很难对这种微妙的语义关系进行准确判断。此外，在迭代学习中，错误的未标注数据仍会不可避免地被加入训练集，也与这一类错误的发生有关。

其余的错误来自于细粒度上位词表示学习的不准确性。以测试集中的下位词“兰科”为例，算法能准确识别“植物”是正确的上位词，但是错误判断“单子叶植物纲”的正确性。最有可能的原因是“单子叶植物纲”出现在中文语料库的词频太低，它在词嵌入空间的表示学习效果相对较差。

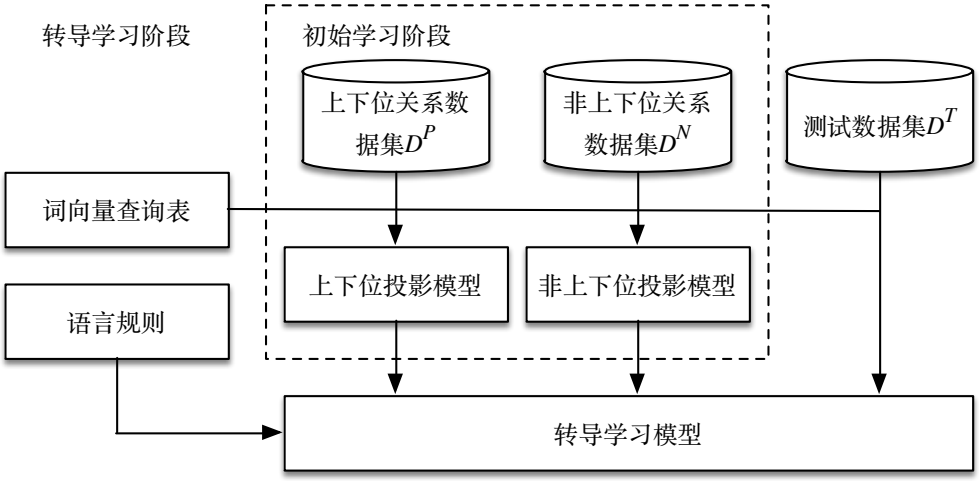


图 2.5: TPM 的算法框架

2.4 基于转导学习的上下位关系分类模型

在上一节中，我们设计了半监督式的 IPM 算法，迭代地从无标注数据中获取更多的知识。然而，这一算法有如下三个局限性：i）这一算法只利用正例训练集进行训练，没有利用非上下位关系，所以没有建模如何在词嵌入空间内区分上下位关系与非上下位关系；ii）中文语言有其特殊性，之前的模型无法加入高精度的语言学规则协同学习；iii）它只假设下位词与上位词之间在词嵌入空间存在线性投影关系，非线性关系没有能够有效建模。本节介绍基于转导学习的上下位关系分类模型 TPM，解决上文提出的三个缺陷。

2.4.1 算法模型

本节介绍的算法为分类模型，给定正负例训练集 D^P 和 D^N，其目标是训练分类算法对测试集 D^T 中的关系元组 (x_i, y_i) 直接进行上下位/非上下位关系分类。如图 2.5所示，本算法有两个阶段：初始阶段和转导学习阶段。在初始阶段，我们在词嵌入空间分别学习上下位关系和非上下位关系的表示；在转导学习阶段，我们同时结合了初始阶段的预测结果、语言规则、非正则投影过程，对测试集中的关系元组进行预测。具体算法详述如下：

初始模型训练： 初始阶段的模型参考了前述 IPM 算法和 Fu 等人 [25] 的工作。对正例训练集中的元组 $(x_i, y_i) \in D^P$，假设存在一个 $|\vec{x}_i| \times |\vec{x}_i|$ 的正向投影矩阵 $\mathbf{M}^P$，使得 $\mathbf{M}^P\vec{x}_i \approx \vec{y}_i$。然而，这一矩阵没有考虑到非上下位关系的语义。因此，对于负例训练集中的元组 $(x_i, y_i) \in D^N$，也学习一个 $|\vec{x}_i| \times |\vec{x}_i|$ 的负向投影矩阵 $\mathbf{M}^N$，使

得 $\mathbf{M}^N\vec{x}_i \approx \vec{y}_i$。这种方法相当于分别学习两个“翻译”模型，将一个实体在词嵌入空间分别“翻译”成它的上位词和非上位词。学习矩阵 $\mathbf{M}^P$ 和 $\mathbf{M}^N$ 相当于最小化如下两个目标函数：

$$J(\mathbf{M}^P) = \frac{1}{2}\sum_{(x_i,y_i)\in D^P} \|\mathbf{M}^P\vec{x}_i - \vec{y}_i\|_2^2 + \frac{\lambda^P}{2}\|\mathbf{M}^P\|_F^2$$

$$J(\mathbf{M}^N) = \frac{1}{2}\sum_{(x_i,y_i)\in D^N} \|\mathbf{M}^N\vec{x}_i - \vec{y}_i\|_2^2 + \frac{\lambda^N}{2}\|\mathbf{M}^N\|_F^2$$

其中，$\lambda^P > 0$ 和 $\lambda^N > 0$ 是 Tikhonov 正则化的超参数。

当这两个矩阵学习完毕后，对于 D^T 中的每对 (x_i, y_i)，分别计算其通过 $\mathbf{M}^P$ 和 $\mathbf{M}^N$ 两个矩阵进行投影产生的残差：

$$d^P(x_i, y_i) = \|\mathbf{M}^P\vec{x}_i - \vec{y}_i\|_2$$

$$d^N(x_i, y_i) = \|\mathbf{M}^N\vec{x}_i - \vec{y}_i\|_2$$

并且计算初始模型的**预测评分** $score(x_i, y_i) \in (-1, 1)$ 如下所示：

$$score(x_i, y_i) = \tanh(d^N(x_i, y_i) - d^P(x_i, y_i))$$

根据 $d^P(x_i, y_i)$ 和 $d^N(x_i, y_i)$ 的性质可知，如果 (x_i, y_i) 的预测评分 $score(x_i, y_i)$ 越高，x_i 和 y_i 之间有上下位关系的概率越大。我们在上式中使用双曲正切函数，而不是 Sigmoid 函数，以避免 Sigmoid 函数的**广泛饱和性**（Widespread Saturation）[89]。

$d^P(x_i, y_i)$ 与 $d^N(x_i, y_i)$ 之间的差值也能用来推断上述模型对预测结果的置信度高低。我们定义 (x_i, y_i) 的**置信度评分** $conf(x_i, y_i) \in (0, 1)$ 如下所示：

$$conf(x_i, y_i) = \frac{|d^P(x_i, y_i) - d^N(x_i, y_i)|}{\max\{d^P(x_i, y_i), d^N(x_i, y_i)\}}$$

如果 (x_i, y_i) 的置信度评分高，则模型正确预测 x_i 和 y_i 之间是否具有上下位关系的可能性更高。在转导学习阶段，置信度评分可以给不同的数据不同的权重。

转导非线性学习： 在本模块中，定义 F_i 是关系对 (x_i, y_i) 的最终预测得分。根据算法初始阶段的结果，如果 $(x_i, y_i) \in D^P$，我们设置 $F_i = 1$；如果 $(x_i, y_i) \in D^N$，$F_i = -1$；如果 $(x_i, y_i) \in D^T$，相应 F_i 的值在 $(-1, 1)$ 范围内随机设置。为了方便

后续模型的推导，我们记 **F** 为 $(|D^P|+|D^N|+|D^T|)\times 1$ 的最终预测得分向量，F_i 为 **F** 中第 i 个元素。转导学习的任务为，对于所有的 $(x_i,y_i)\in D^T$，学习 F_i 的值。这一模型一共包括三个组成部分，分别对应三个假设，详述如下：

假设 1 转导学习模型对于 $(x_i,y_i)\in D^T$ 的最终预测得分 F_i 应该与初始模型的预测结果相近。

定义 **S** 为 $(|D^P|+|D^N|+|D^T|)\times 1$ 的初始预测向量。如果 $(x_i,y_i)\in D^P$，我们设 $S_i=1$；如果 $(x_i,y_i)\in D^N$，$S_i=-1$；如果 $(x_i,y_i)\in D^T$，根据初始模型的预测，有 $S_i=score(x_i,y_i)$。为了在转导学习模型中加入初始模型预测的置信度，定义 **W** 为 $(|D^P|+|D^N|+|D^T|)\times(|D^P|+|D^N|+|D^T|)$ 的对角权重矩阵。矩阵 **W** 的第 i 行第 j 列元素 $W_{i,j}$ 的值计算如下：

$$W_{i,j}=\begin{cases} conf(x_i,y_i) & i=j,(x_i,y_i)\in D^T \\ 1 & i=j,(x_i,y_i)\in D^P\cup D^N \\ 0 & \text{其他} \end{cases}$$

根据 **S** 和 **W** 的计算方式，我们定义目标函数 $\mathscr{O}_s=\|\mathbf{W}(\mathbf{F}-\mathbf{S})\|_2^2$。最小化 $\mathscr{O}_s$ 可以保证对于 D^T 最终预测的结果与初始模型的预测结果接近。权重矩阵 **W** 赋予训练数据最大的权重；初始模型对 $(x_i,y_i)\in D^T$ 的预测置信度越高，则权重越大。

假设 2 转导学习模型对于 $(x_i,y_i)\in D^T$ 的最终预测得分 F_i 应该尽可能不违反专家定制的语言规则。

尽管语言规则的覆盖率有限，他们拥有很高的预测精度。对于中文语言，Li 等人 [21] 研究了中文维基百科概念类别的构词法。在本模型中，我们进一步引入了正向和负向的带权重语言规则。令 C 为所有语言规则的集合。对于每条语言规则 $c_i\in C$，γ_i 是这一正向（或负向）语言规则的真正率（或真负率），可以通过人工标注的数据估计出，或由专家人工制定。在本文中，考虑到中文实体和上位词的构词法，我们设计了一条正向规则（P1）和两条负向规则（N1 和 N2），详见表 2.6。

令 **R** 为 $(|D^P|+|D^N|+|D^T|)\times 1$ 的语言规则向量，R_i 是 **R** 中的第 i 个元素。对于训练数据，若 $(x_i,y_i)\in D^P$，我们设置 $R_i=1$；若 $(x_i,y_i)\in D^N$，$R_i=-1$。这一赋值方法与 **S** 相同，因为我们认为训练数据由人工标注，不受这些规则限制。对于测试数据中每个关系元组 $(x_i,y_i)\in D^T$，如果 (x_i,y_i) 不与任何语言规则 C 相匹

表 2.6: 中文上位词预测的三条语言规则

规则	描述
P1	若 x_i 的核心词与候选上位词 y_i 相匹配，y_i 很可能是 x_i 的上位词。例如，“动物” 是 “哺乳动物” 的上位词。
N1	若 x_i 的核心词与候选上位词 y_i 的非核心词相匹配，y_i 很可能不是 x_i 的上位词。例如，“动物学” 不是 “哺乳动物” 的上位词。
N2	若候选上位词 y_i 的核心词与扩展自 [21] 的中文主题词词典相匹配，y_i 很可能不是 x_i 的上位词。该词典包括了 184 个非概念主题词，例如政治、军事等。

配，我们在学习过程中的每个迭代都锁定 $R_i = F_i$。在这种情况，最终的预测结果不会受到违背任何语言规则的惩罚。对于其他情况，令 $C_{(x_i,y_i)} \subseteq C$ 是 $(x_i, y_i) \in D^T$ 匹配的语言规则集合。如果 $C_{(x_i,y_i)}$ 均为正向语言规则，我们设 R_i 为：

$$R_i = \max\{F_i, \max_{c_j \in C_{(x_i,y_i)}} \gamma_j\}$$

类似地，如果 $C_{(x_i,y_i)}$ 均为负向语言规则，我们有：

$$R_i = -\max\{-F_i, \max_{c_j \in C_{(x_i,y_i)}} \gamma_j\}$$

通过以上设置，易知如果 $(x_i, y_i) \in D^T$ 匹配了正向语言规则，当且仅当 $F_i < \max_{c_j \in C_{(x_i,y_i)}} \gamma_j$，$F_i$ 会受到违反正向语言规则的惩罚；如果 $(x_i, y_i) \in D^T$ 匹配了负向语言规则，当且仅当 $F_i > -\max_{c_j \in C_{(x_i,y_i)}} \gamma_j$，$F_i$ 会受到违反负向语言规则的惩罚[8]。基于上述定义，我们令基于语言规则的目标函数为 $\mathscr{O}_r = \|\mathbf{F} - \mathbf{R}\|_2^2$，可以将任何 “软性” 规则加入模型中，对假正例或者假负例更加具有鲁棒性。

假设 3 如果两个实体 x_i 和 x_j 的语义相似，则对于某概念 y，(x_i, y) 和 (x_j, y) 有相似的上下位或非上下位关系标签。

根据假设 3，如果 y_i 是 x_i 的上位词，y_i 也很可能是与 x_i 语义相似的其他词的上位词。例如，如果我们知道 “美国” 是 “国家”，我们可以推断出 “加拿大”、“澳大利亚” 等也是 “国家”。所以，我们可以定义两对关系元组 $p_i = (x_i, y_i)$ 和 $p_j = (x_j, y_j)$

[8]在本研究中，我们没有考虑一个关系元组同时匹配正向规则和负向规则的情况，因为这种情况意味着这些语言规则相互矛盾，一般不可能发生。我们也可以通过加入启发式规则来解决这一问题。例如，我们可以设置 $R_i = F_i$ 或者 $R_i = F_i + \sum_{c_j \in C_{(x_i,y_i)}} \gamma_j$。

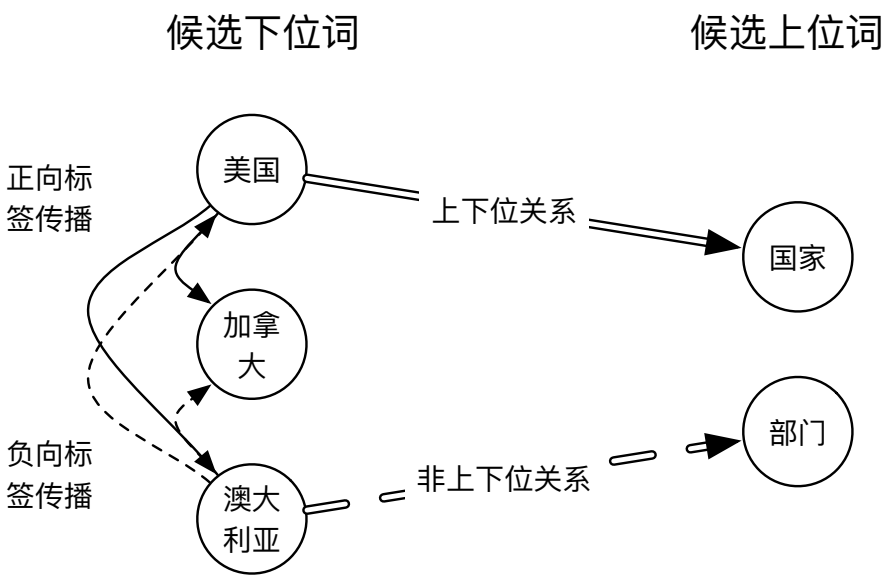


图 2.6: TransLP 框架在中文上下位关系预测的应用

的相似度 $sim(p_i, p_j)$，如下所示：

$$sim(p_i, p_j) = \begin{cases} \cos(\vec{x}_i, \vec{x}_j) & y_i = y_j \\ 0 & \text{其他} \end{cases} \tag{2.1}$$

值得注意的是，本模型只考虑了下位词的相似度，而没有对上位词进行扩展。这是因为上位词的语义一般更加宽泛。再次考虑上文的例子，在我们训练的 Skip-Gram 模型中，与 “国家” 最为接近的词包括 “区域”、“部门” 等，从 “美国” 是 “国家” 无法推断出 “美国” 是 “部门”。

为了将假设 3 数学化，本文采用基于 **TransLP** 的转导学习框架 [90]，这一框架在先前的研究中用于链接预测、文本分类等 [90, 91]。本文应用此框架学习从中文实体到其上位词在词嵌入空间的非线性映射。基于 TransLP，我们可以通过随机游走模型，将标签信息（上下位关系或非上下位关系）根据式 2.1从一个关系元组传播到另一个关系元组。例如，关系元组 “（美国，国家）” 的正向得分 $F_i > 0$ 可以传播至 “（加拿大，国家）”、“（澳大利亚，国家）” 等；“（澳大利亚，部门）” 的负向得分 $F_i < 0$ 可以传播至 “（美国，部门）”、“（加拿大，部门）” 等。示例如图 2.6所示。

令 $\mathbf{F}^*$ 是优化问题 $\min \mathscr{O}_s + \mathscr{O}_r$ 的最优解。受到文献 [90, 91] 启发，我们在 $\mathbf{F}$ 上加上一个多维高斯先验分布 $N(\mathbf{F}^*, \mathbf{P})$，其中，$\mathbf{P}$ 是 $\mathbf{F}$ 的协方差矩阵，$P_{i,j} = sim(p_i, p_j)$。因此，非线性优化目标函数 $\mathscr{O}_n$ 可以定义为：$\mathscr{O}_n = \mathbf{F}^T\mathbf{P}^{-1}\mathbf{F}$。它与高斯随机场先验

的负似然性呈正比关系，使得 **F** 的最终预测结果叠加了式 2.1的平滑效果。

协同优化算法： 将转导非线性学习的三大优化目标合并，整体优化目标函数如下所示：

$$J(\mathbf{F}) = \mathscr{O}_s + \mathscr{O}_r + \frac{\mu_1}{2}\mathscr{O}_n + \frac{\mu_2}{2}\|\mathbf{F}\|_2^2 \tag{2.2}$$

其中，$\|\mathbf{F}\|_2^2$ 在 **F** 上加入了额外的平滑 l_2 正则项。μ_1 和 μ_2 是人工可调的正则化超参数。由于这一优化问题是凸性的，我们可以使用梯度下降求得 **F** 的最优解。$J(\mathbf{F})$ 关于 **F** 的导数为：

$$\frac{\mathrm{d}J(\mathbf{F})}{\mathrm{d}\mathbf{F}} = \mathbf{W}^2(\mathbf{F}-\mathbf{S}) + (\mathbf{F}-\mathbf{R}) + \mu_1\mathbf{P}^{-1}\mathbf{F} + \mu_2\mathbf{F}$$

当 $|D^P|+|D^N|+|D^T|$ 很大的时候,上式的优化耗费大量计算时间。当 $\mathbf{W}^2$、$\mathbf{S}$、$\mathbf{R}$ 和 $\mathbf{P}^{-1}$ 预先计算完毕后，使用梯度下降优化的计算复杂度为 $O(T\cdot(|D^P|+|D^N|+|D^T|)^2)$，$T$ 为迭代次数。

为了加快学习过程，可以采用**分块梯度下降**（Blockwise Gradient Descent）算法加以改进。从式 2.2的定义可知，如果 $y_i \neq y_j$，关于 (x_i, y_i) 和 (x_j, y_j) 的最优化结果 F_i 和 F_j 是互相独立的。因此，原始优化问题可以根据不同候选上位词分解，分别优化。令 H 为 D^T 中的所有候选上位词集合。对于每一个候选上位词 $h \in H$，我们用 D_h 表示 $D^P \cup D^N \cup D^T$ 中所有候选上位词为 h 的关系元组集合。据此，原始的优化问题可以分解为 $|H|$ 个子优化问题，每个问题分别对应一个候选上位词 $h \in H$ 和数据集 D_h。分别定义 $\mathbf{W}_h$、$\mathbf{S}_h$、$\mathbf{R}_h$、$\mathbf{F}_h$ 和 $\mathbf{P}_h$ 为基于 $h \in H$ 的权重矩阵、初始预测得分向量、语言规则得分向量、最终预测得分向量和实体相似性协方差矩阵。原始优化问题可以改写为：$J(\mathbf{F}) = \sum_{h\in H}\tilde{J}(\mathbf{F}_h)$，其中，

$$\tilde{J}(\mathbf{F}_h) = \|\mathbf{W}_h(\mathbf{F}_h - \mathbf{S}_h)\|_2^2 + \|\mathbf{F}_h - \mathbf{R}_h\|_2^2 + \frac{\mu_1}{2}\mathbf{F}_h^T\mathbf{P}_h^{-1}\mathbf{F}_h + \frac{\mu_2}{2}\|\mathbf{F}_h\|_2^2$$

用上标 $^{(t)}$ 表示相应矩阵或向量在第 t 个迭代的值，$\mathbf{F}_h^{(t)}$ 的值可以根据下式迭代更新：

$$\mathbf{F}_h^{(t+1)} = \mathbf{F}_h^{(t)} - \eta\cdot\frac{\mathrm{d}\tilde{J}(\mathbf{F}_h^{(t)})}{\mathrm{d}\mathbf{F}_h^{(t)}}$$

其中，η 是学习率。非线性转导学习算法的总结性描述见算法 2。

这一算法的时间复杂度是 $O(\sum_{h\in D_h} T_h|D_h|^2)$，其中，$T_h$ 是优化关于 D_h 的子问题所需的迭代数。与采用梯度下降算法求解原始优化问题相比，这一算法的复杂

Algorithm 2 非线性转导学习算法

根据初始模型，初始化 $\mathbf{W}_h$ 和 $\mathbf{S}_h$
随机初始化 $\mathbf{F}_h^{(0)}$
根据实体语义相似度，计算 $\mathbf{P}_h^{-1}$
初始化计数器 $t = 1$
for 每条语法规则 $c_i \in C$ **do**
    在训练集上估计 γ_i 的值
end for
while $\|\mathbf{F}_h^{(t)} - \mathbf{F}_h^{(t+1)}\|_2$ 足够小 **do**
    根据 C 和 $\mathbf{F}_h^{(t)}$ 计算 $\mathbf{R}_h^{(t)}$
    计算 $\frac{\mathrm{d}\tilde{J}(\mathbf{F}_h^{(t)})}{\mathrm{d}\mathbf{F}_h^{(t)}} = \mathbf{W}_h^2(\mathbf{F}_h^{(t)} - \mathbf{S}_h) + (\mathbf{F}_h^{(t)} - \mathbf{R}_h^{(t)}) + \mu_1\mathbf{P}_h^{-1}\mathbf{F}_h^{(t)} + \mu_2\mathbf{F}_h^{(t)}$
    为下一次迭代计算 $\mathbf{F}_h^{(t+1)}$ ：$\mathbf{F}_h^{(t+1)} = \mathbf{F}_h^{(t)} - \eta \cdot \frac{\mathrm{d}\tilde{J}(\mathbf{F}_h^{(t)})}{\mathrm{d}\mathbf{F}_h^{(t)}}$
    更新计数器 $t = t + 1$
end while
return 最终预测得分向量 $\mathbf{F}_h^{(t+1)}$

度有明显下降。有两个原因：i) $\sum_{h\in H}|D_h| \leq |D^P| + |D^N| + |D^T|$; ii) 由于 D_h 中的数据量远比原始数据集小，T_h 有很大可能比 T 小。这一优化过程也可以看成一种基于式 2.2分块矩阵优化算法。

最后，对于每个关系元组 $(x_i, y_i) \in D^T$，如果 $F_i > \theta$，我们预测 y_i 是 x_i 的上位词，其中，$\theta \in (-1, 1)$ 是在验证集上可调的阈值。

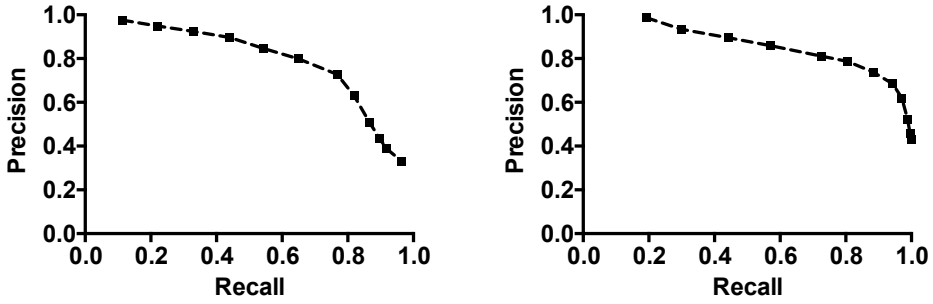
2.4.2 实验分析

本节对提出的转导投影模型进行实验分析，并与基线方法对比。

实验设置： 如上文所述，Fu 等人 [25] 公开的数据集是唯一开源的中文上下位关系数据集，包括 1391 个上下位关系元组和 4294 个非上下位关系元组，以下简称为 FD。为了在更大的数据集上进行评测，我们从百度百科的词条类别进行采样，并且给多个人工标注员标注 “实体-类别” 对之间的的关系类别 (上下位关系或非上下位关系)。如果不同的人工标注员对于同一个关系元组给的关系标签不一致，我们将这个关系元组从数据集中移除。最终，我们得到一个更大的中文上下位关系标注数据集，包括 3870 个上下位关系元组和 3582 个非上下位关系元组，以下简称为 BK。这个数据集已在 Github 上开源[9]。

本实验采用与第2.3.2节相同的中文语料库和词向量训练模型，此处不再赘述。

[9]https://chywang.github.io/data/acl17.zip



(a) 数据集：FD　　(b) 数据集：BK

图 2.7: TPM 中参数 θ 调节的 PR 曲线

在以下的实验中，我们统一使用这两个数据集的随机 60% 部分进行训练，20% 作为验证集，剩下的 20% 作为测试集。如果将这两个数据集随机 5 折划分，我们分别进行 5 次实验，并汇报这 5 次结果的平均值作为最终结果。

参数分析： 我们首先在验证集上调节参数的值。其中，正则化参数的值默认设置为 $\lambda^P = \lambda^N = 10^{-3}$ 和 $\mu_1 = \mu_2 = 10^{-4}$。参数 θ 的大小设置反映了精准度和召回率之间的权衡，如果 θ 的值比较大，比起召回率，模型更重视提升精准度。在图 2.7中，我们给出了 FD 和 BK 两个数据集上的 PR 曲线（Precision-Recall Curve）。从实验结果可以看出，我们的模型在 BK 数据集上的表现比起 FD 数据集上更好。最有可能的原因是 BK 数据集比 FD 更大，而且正负例的比例更加平衡。最终，我们在 FD 数据集上设置 θ 为 0.05，在 BK 数据集上设置为 0.1。

实验结果： 在多个先前的工作 [25, 63] 以及第2.3.2节中，多个基于英语数据集上的模式匹配和分布式算法被改成适用于中文的场景，但是实验结果表明，这些方法对于中文数据集的效果比较差，一般 F 值低于 60%。所以，这些算法一般不作为中文上下位关系预测的强基线算法，本节只考虑基于词嵌入模型的方法。

在本组实验中，我们只考虑两种适用于中文的算法作为强基线算法：Fu 等人 [25] 的算法及我们先前提出的 IPM。对 Fu 等人的算法，我们分别实现采用单投影模型和分段投影模型的版本，记为 Fu-S 和 Fu-P。对于 IPM，为了保证所有算法的公平性，我们使用训练数据训练初始模型，在迭代训练阶段，我们只利用测试数据，并且隐去其人工标注的标签。此外，由于基于词向量的分类模型对于关系分类有较好的效果，而且这一组方法不依赖与语言模式 [92]，因此这些方法可以直接用于中文术语间的关系分类。我们分别采用三种方法结合两个中文术语的词向量作为特征，包括向量连接 $\vec{x}_i \oplus \vec{y}_i$、向量和 $\vec{x}_i + \vec{y}_i$ 和向量差 $\vec{x}_i - \vec{y}_i$，并且使用 SVM 分类

表 2.7: TPM 在两个中文上下位关系数据集上的实验结果

数据集	FD			BK		
方法	**精准度**	**召回率**	**F 值**	**精准度**	**召回率**	**F 值**
Fu-S [25]	0.641	0.560	0.598	0.714	0.648	0.679
Fu-P [25]	0.664	0.593	0.626	0.727	0.675	0.700
$\vec{x}_i \oplus \vec{y}_i$	0.677	**0.752**	0.697	0.803	0.759	0.780
$\vec{x}_i + \vec{y}_i$	0.653	0.607	0.629	0.727	0.656	0.689
$\vec{x}_i - \vec{y}_i$	0.719	0.606	0.657	0.784	0.607	0.684
IPM	0.693	0.645	0.669	0.739	0.698	0.718
TPM-Initial	0.707	0.692	0.699	0.817	0.785	0.800
TPM	**0.728**	0.705	**0.716**	**0.836**	**0.806**	**0.821**

表 2.8: 三条语言规则的真正率或真负率

真正率/真负率	**P1**	**N1**	**N2**
数据集 FD	0.986	0.923	0.941
数据集 BK	0.976	0.968	0.973

器在这三组特征上进行实验。对于我们的算法，我们实现了两种版本：完整 TPM 算法的实现和 TPM 算法的初始学习部分（记为 TPM-Initial）。实验结果见表 2.7。

从实验结果可以看出，Fu-P 的实验效果比 Fu-S 有提升，我们提出的 IPM 算法则进一步提升了 F 值，因为它考虑了上下位关系的复杂表示，同时加入了中文语言模式。对比三种基于词向量的关系分类方法，基于向量和 $\vec{x}_i + \vec{y}_i$ 和向量差 $\vec{x}_i - \vec{y}_i$ 的模型表现类似，而基于向量连接 $\vec{x}_i \oplus \vec{y}_i$ 的方法比前两种方法 F 值更高。与所有基线方法相比，TPM-Initial 的准确度更高，如果加入转导学习的步骤，F 值能在两个数据集上进一步提升 1.7% 和 2.1%。所以，TPM 对预测上文上下位关系高度有效。我们进一步使用成对 t 检验（Paired t-test）比较 TPM 和 IPM 两种模型，结果证明 TPM 比起 IPM 的准确度有显著提升（$p < 0.01$）。

语言规则的有效性： 为了验证本模型采用的语言规则的有效性，我们在训练数据上直接评测每条语言规则的准确性。对于正向规则，我们汇报真正率；对于负向规则，我们汇报真负率。这些值在转导学习阶段作为 γ_i 加入模型，具体数值见表 2.8。由是可知，采用的这三条规则在两个数据集上的准确度都远超过了 90%。值得进一步指出的是，在本节中我们只使用了很少的语言规则，这一算法框架的通用性高，可以加入任何数量、基于任何语言的规则。

错误分析： 我们对转导学习模型的预测错误进行分析。在表 2.9中，我们列举出一些预测正确和错误的例子。在表中，$\surd$ 表示上下位关系，$\times$ 表示非上下位关系。从上述示例看出，模型预测的结果在大部分情况下是正确的。对于少数情况，

表 2.9: TPM 的部分预测结果

候选上位词	预测值	真实值	候选上位词	预测值	真实值
中文实体：乙烯			中文实体：孙燕姿		
化学品	√	√	歌手	√	√
有机化学	×	×	明星	√	√
有机物	√	√	人物	√	√
气体	√	√	金曲奖	√	×
自然科学	×	×	音乐人	√	√
中文实体：显卡			中文实体：核反应堆		
硬件	√	√	建筑学	×	×
电子产品	√	√	核科学	×	×
电脑硬件	√	√	核能	√	×
数码	×	×	自然科学	×	×

表 2.10: IPM 和 TPM 的算法变体在两个英语数据集上的实验结果

数据集	BLESS			Shwartz		
方法	**精准度**	**召回率**	**F 值**	**精准度**	**召回率**	**F 值**
Fu-S [25]	0.653	0.624	0.638	0.656	0.661	0.658
Fu-P [25]	0.681	0.642	0.661	0.623	0.719	0.673
$\vec{x}_i \oplus \vec{y}_i$	0.794	**0.841**	0.817	0.793	**0.809**	**0.801**
$\vec{x}_i + \vec{y}_i$	0.807	0.723	0.763	0.791	0.796	0.794
$\vec{x}_i - \vec{y}_i$	0.780	0.812	0.796	**0.805**	0.775	0.790
IPM-EN	0.762	0.754	0.758	0.751	0.763	0.756
TPM-EN	**0.844**	0.795	**0.819**	0.791	0.775	0.783

与第2.3.2节的分析类似，虽然本模型同时建模了上下位关系和非上下位关系的语义，然而，严格的上下位关系和宽松的语义关联性的区分仍有挑战性。例如，我们的模型错误预测 “核能” 是 “核反应堆” 的上位词。在 FD 和 BK 两个数据集中，这种类型的错误分别占比 80.2% 和 78.6%。

为了减少这一类错误的发生，本模型在转导学习过程中采用了 Li 等人 [21] 提出的主题语义字典，但是其覆盖率仍然很有限。可能的改进方案有：i）进一步扩展现有中文的语义词典；ii）从网络数据构建大规模的语义网络，为模型增加更多语义知识。

在英语数据集上的实验： 本文提出的 IPM 和 TPM 两个模型都加入了中文语言知识，目的是提升中文上下位抽取的效果。然而，如果将这两个模型稍加改变，他们也能对英语数据进行预测，分别记为 IPM-EN 和 TPM-EN。在本组实验中，我们在两个英语数据集上进行实验。对于 IPM-EN，我们用原始的 Hearst 模式 [38] 进行关系选择，不使用 Co-Hyponym 模式；对于 TPM-EN，我们去除了模型中的中文

语言规则优化部分，其转导学习的目标学习函数改为：

$$J(\mathbf{F}) = \mathscr{O}_s + \frac{\mu_1}{2}\mathscr{O}_n + \frac{\mu_2}{2}\|\mathbf{F}\|_2^2$$

我们在两个英语数据集进行实验。第一个是标准的分布式语义评测基准数据集 BLESS [93]。在实验中，我们把所有 HYPER 关系视为正例（包括 1337 个关系元组），随机采样 30% 的 RANDOM 关系作为负例（包括 3754 个关系元组）。由于 BLESS 数据集相对较小，我们也使用 Shwartz 数据集 [54] 进行实验，其中，对于 Shwartz 中随机划分数据，我们同样只使用 30% 的负例。在这一数据集中，我们一共有 14135 个正例和 16956 个负例。预训练的英语词向量通过 Glove 算法 [52] 得到，向量维度为 100 维，英语语料库为英文维基百科。实验结果见表 2.10。

从结果可见, TPM-EN 在 BLESS 数据集上的精度超过了其他所有基线方法, 其 F 值达到 81.9%。它在 Shwartz 数据集上比 $\vec{x}_i \oplus \vec{y}_i$ 稍低。IPM-EN 的实验效果略低于 TPM-EN, 但是也超过了数个基线算法。这体现出, 我们的方法为中文语言提出, 也可以在英语数据集上取得较好的效果。虽然 IPM-EN 和 TPM-EN 不是英语上下位关系抽取的最佳方法，这一设计思路有助于我们进一步研究如何设计更为通用的算法，在中英语上下位关系预测上都能取得领先的效果。

2.5 基于模糊正交投影的上下位关系分类模型

前述 IPM 和 TPM 两个模型分别基于词嵌入模型从两个角度解决问题，前者旨在更为精确地学习上下位关系在词嵌入空间的表示，后者通过分别建模上下位关系和非上下位关系的语义，使得上下位/非上下位关系的分类决策边界更为明晰。在本节中，我们将两者的优点结合起来。此外，我们试图从数据建模的角度做到更加语言独立（Language-Independent），使得我们的算法在中英文语言上都能对上下位关系进行精准抽取。

2.5.1 算法模型

本节介绍的算法同样为分类模型，其输入为训练集 D^P 和 D^N，并在测试集 D^T 中的关系元组 (x_i, y_i) 上进行上下位/非上下位关系分类，不再赘述。它在词嵌入模型的基础上，同时学习上下位关系和非上下位关系的**模糊正交投影**（Fuzzy Orthogonal Projection），分为三个步骤：上下位关系投影学习、非上下位关系投影

学习、关系分类器训练，详述如下：

上下位关系投影学习： 本模型扩展了基于词嵌入投影模型的经典工作 [25, 86]，以及前两节提出的模型，上下位关系投影假设存在一个 $|\vec{x}_i| \times |\vec{x}_i|$ 的正向投影矩阵 $\mathbf{M}^P$，使得对于上下位关系元组 $(x_i, y_i) \in D^P$，有 $\mathbf{M}^P\vec{x}_i \approx \vec{y}_i$。由于这一上下位关系投影也能看成从下位词 "翻译" 到上位词的过程，根据 Xing 等人 [94] 的研究，在设定优化目标上，我们最大化下位词投射到上位词的向量 $\mathbf{M}^P\vec{x}_i$ 及其真实上位词的向量 $\vec{y}_i$ 的余弦相似度。对于正例训练集 D^P，我们的学习目标为：

$$\max \sum_{(x_i,y_i)\in D^P} \cos(\mathbf{M}^P\vec{x}_i, \vec{y}_i)$$

若词向量 $\vec{x}_i$ 和 $\vec{y}_i$ 都是归一化的，为了保证 $\mathbf{M}^P\vec{x}_i$ 也是归一化的，我们将目标函数改写为：

$$\min \sum_{(x_i,y_i)\in D^P} \|\mathbf{M}^P\vec{x}_i - \vec{y}_i\|^2 \text{ s. t. } (\mathbf{M}^P)^T \cdot \mathbf{M}^P = \mathbf{I}$$

其中，$\mathbf{I}$ 是 $|\vec{x}_i| \times |\vec{x}_i|$ 的单位阵。因此，我们需要学习一个正交的投影矩阵 $\mathbf{M}^P$。

然而，这一设定没有考虑到上下位关系的复杂语义表示。根据 [25] 和上文的研究，不同类别的上下位关系有不同的语义表达，因此需要学习不同的投影矩阵。与 IPM 相同，令 K 为簇的数量，不同的簇代表上下位关系的不同语义分量。我们首先在 D^P 的所有关系元组上使用 K-Means 进行聚类，特征为向量之差 $\vec{x}_i - \vec{y}_i$。令第 k 个簇的中心向量为 $\vec{c}_k^P$。定义 $a_{i,k}^P$ 为某关系元组 $(x_i, y_i) \in D^P$ 对于第 k 个簇的权重，计算如下：[10]

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

令 $\mathbf{M}_k^P$ 为 $|\vec{x}_i| \times |\vec{x}_i|$ 的对于第 k 个簇的正向投影矩阵，加入正交性限制，第 k

[10]在算法探索阶段，我们也尝试了很多模糊聚类算法，例如高斯混合模型、模糊 c-Means 等，将每个关系元组 $(x_i, y_i) \in D^P$ 属于每个簇的概率当成对应的权重 $a_{i,k}^P$。然而，由于词向量的维度较高，这些算法的效果并不好。因此，我们采用 K-Means 算法对关系元组聚类并且用启发式算法计算权重 $a_{i,k}^P$。

个簇的优化目标为最小化带权重的投影误差：

$$\begin{aligned} J(\mathbf{M}_k^P) =& \frac{1}{2} \sum_{(x_i,y_i)\in D^P} a_{i,k}^P \|\mathbf{M}_k^P \vec{x}_i - \vec{y}_i\|^2 \\ \text{s. t. } & (\mathbf{M}_k^P)^T \cdot \mathbf{M}_k^P = \mathbf{I}, \sum_{(x_i,y_i)\in D^P} a_{i,k}^P = 1 \end{aligned} \tag{2.3}$$

令 $\mathcal{M}^P = \{\mathbf{M}_1^P, \mathbf{M}_2^P, \cdots, \mathbf{M}_K^P\}$ 为全部 K 个簇对应的投影矩阵的集合。学习上下位关系投影相当于最小化如下目标函数：

$$\begin{aligned} \tilde{J}(\mathcal{M}^P) =& \frac{1}{2} \sum_{k=1}^{K} \sum_{(x_i,y_i)\in D^P} a_{i,k}^P \|\mathbf{M}_k^P \vec{x}_i - \vec{y}_i\|^2 \\ \text{s. t. } & (\mathbf{M}_k^P)^T \cdot \mathbf{M}_k^P = \mathbf{I}, \sum_{(x_i,y_i)\in D^P} a_{i,k}^P = 1, k = 1, \ldots, K \end{aligned} \tag{2.4}$$

在应用数学领域，式 2.3是 Wahba 问题从三维空间到高维空间的扩展，Wahba 问题原来用于两个卫星定位系统中三维坐标的转换 [95]。式 2.4是 K 个扩展 Wahba 问题的独立组合。由于不同 k 值的矩阵 $\mathbf{M}_k^P$（$k = 1, 2, \cdots, K$）的优化结果互相独立，将 K 个式 2.3的最优化结果结合起来即为式 2.3的最优化结果。本节扩展基于奇异值分解（Singular Value Decomposition，SVD）的 Wahba 问题闭式解 [96]，将其扩展至到高维空间，如下所示：

定理 2.5.1. 高维 *Wahba* 问题（式 *2.3*）具有如下闭式解:

1. $\mathbf{B}_k^P = \sum_{(x_i,y_i)\in D^P} a_{i,k}^P \vec{y}_i \cdot \vec{x}_i^T$
2. $\mathbf{U}_k^P \mathbf{S}_k^P (\mathbf{V}_k^P)^T = SVD(\mathbf{B}_k^P)$
3. $\mathbf{R}_k^P = diag(\underbrace{1, \ldots, 1}_{|\vec{x}_i|-1}, det(\mathbf{U}_k^P) \cdot det(\mathbf{V}_k^P))$
4. $\mathbf{M}_k^P = \mathbf{U}_k^P \mathbf{R}_k^P (\mathbf{V}_k^P)^T$

其中，$diag(\cdot)$ 将向量转化为对角矩阵，$det(\cdot)$ 返回矩阵的行列式。

Proof. 为了简单起见，我们省略了式 2.3中的所有变量的上标 P 和下标 k，式 2.3可以简写为：

$$J(\mathbf{M}) = \frac{1}{2} \sum_i a_i \|\mathbf{M}\vec{x}_i - \vec{y}_i\|^2 \text{ s. t. } \mathbf{M}^T\mathbf{M} = \mathbf{I}$$

在上式中，每一个元组 (x_i, y_i) 对于整个数据集有权重 a_i。因为 $\sum_i a_i = 1$，优化目标函数可以重写为：

$$J(\mathbf{M}) = 1 - \sum_i a_i \vec{y}_i^T \mathbf{M} \vec{x}_i = 1 - \text{tr}(\mathbf{M}\mathbf{B}^T) \tag{2.5}$$

其中，$\mathbf{B} = \sum_i a_i \vec{y}_i \vec{x}_i^T$。我们对矩阵 $\mathbf{B}$ 进行 SVD 分解：$\mathbf{U}\mathbf{S}\mathbf{V}^T = \text{SVD}(\mathbf{B})$，$\mathbf{S}$ 是奇异值对角矩阵：$\mathbf{S} = \text{diag}(\lambda_1, \ldots, \lambda_{|\vec{x}_i|})$。

基于 SVD 的结果，定义两个正交矩阵和一个对角矩阵：

$$\mathbf{U}_+ = \mathbf{U}\text{diag}(\underbrace{1, \ldots, 1}_{|\vec{x}_i|-1}, \det(\mathbf{U}))$$

$$\mathbf{V}_+ = \mathbf{V}\text{diag}(\underbrace{1, \ldots, 1}_{|\vec{x}_i|-1}, \det(\mathbf{V}))$$

$$\mathbf{S}^{'} = \text{diag}(\lambda_1, \ldots, \lambda_{|\vec{x}_i|-1}, \lambda_{|\vec{x}_i|}\det(\mathbf{U})\det(\mathbf{V}))$$

由于矩阵 $\mathbf{U}$ 和 $\mathbf{V}$ 是正交的，我们有 $\det(\mathbf{U})\det(\mathbf{V}) = \pm 1$。所以，矩阵 $\mathbf{B}$ 可以重新分解为：

$$\mathbf{B} = \mathbf{U}_+ \mathbf{S}^{'} \mathbf{V}_+^T$$

令 $\mathbf{W} = \mathbf{U}_+^T \mathbf{M} \mathbf{V}_+$，基于迹的循环不变性（ Cyclic Invariance Property ），我们有：

$$\text{tr}(\mathbf{M}\mathbf{B}^T) = \text{tr}(\mathbf{M}\mathbf{V}_+\mathbf{S}^{'}\mathbf{U}_+^T) = \text{tr}(\mathbf{S}^{'}\mathbf{U}_+^T\mathbf{M}\mathbf{V}_+) = \text{tr}(\mathbf{S}^{'}\mathbf{W})$$

所以，我们将式 2.5重写为：$J(\mathbf{M}) = 1 - \text{tr}(\mathbf{S}^{'}\mathbf{W})$。

将正交矩阵 $\mathbf{W}$ 欧拉轴角参数化（ Euler Axis-angle Parameterization ）$\mathbf{W} = \mathcal{R}(\mathbf{e}, \phi)$，我们可以得到下式：

$$\begin{aligned} J(\mathbf{M}) =& 1 - (\sum_{i=1}^{|\vec{x}_i|-1} \lambda_i + \lambda_{|\vec{x}_i|}\det(\mathbf{U})\det(\mathbf{V})) \\ &+ (1 - \cos\phi)[\sum_{i=2}^{|\vec{x}_i|-1} \lambda_i + \lambda_{|\vec{x}_i|}\det(\mathbf{U})\det(\mathbf{V}) + \sum_{i=2}^{|\vec{x}_i|-1} (\lambda_1 - \lambda_i)e_i^2 \\ &+ (\lambda_1 - \lambda_{|\vec{x}_i|}\det(\mathbf{U})\det(\mathbf{V}))e_{|\vec{x}_i|}^2] \end{aligned}$$

其中，$\mathbf{e} = [e_1, e_2, \cdots, e_{|\vec{x}_i|}]$ 是单位向量，ϕ 是旋转角度。当 $\cos\phi = 1$ 时，$J(\mathbf{M})$ 可

以取到最小值，如下所示：

$$\min J(\mathbf{M}) = 1 - (\sum_{i=1}^{|\vec{x}_i|-1} \lambda_i + \lambda_{|\vec{x}_i|}\text{det}(\mathbf{U})\text{det}(\mathbf{V})) = 1 - \text{tr}(\mathbf{S}^{'})$$

此时，我们有 $\mathbf{W} = \mathbf{I}$，因此 $\mathbf{M}$ 的最优解为：

$$\mathbf{M}_{opt} = \mathbf{U}_{+}\mathbf{V}_{+}^{T} = \mathbf{U}\text{diag}(\underbrace{1, \ldots, 1}_{|\vec{x}_i|-1}, \text{det}(\mathbf{U})\text{det}(\mathbf{V}))\mathbf{V}^{T}$$

□

非上下位关系投影学习： 非上下位关系的语义比较复杂。在语义关系分类任务中，术语的非上位词可以为同下位词、近义词、反义词等 [78]。所以，前述用于建模上下位关系的模型（即 K 个高维 Wahba 投影模型）也能用于建模非上下位关系的语义。我们同样采用 K-Means 算法对 D^N 中的关系元组聚成 K 个簇，簇中心为 $\vec{c}_1^N, \cdots, \vec{c}_K^N$。计算关系元组 $(x_i, y_i) \in D^N$ 对于第 k 个簇的权重 $a_{i,k}^N$：

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

令 $\mathcal{M}^N = \{\mathbf{M}_1^N, \mathbf{M}_2^N, \cdots, \mathbf{M}_K^N\}$ 为 K 个投影矩阵的集合。定义非上下位关系投影学习的目标函数如下：

$$\begin{aligned}\tilde{J}(\mathcal{M}^N) =& \frac{1}{2}\sum_{k=1}^{K}\sum_{(x_i,y_i)\in D^N} a_{i,k}^N \|\mathbf{M}_k^N\vec{x}_i - \vec{y}_i\|^2 \\ &\text{s. t. } (\mathbf{M}_k^N)^T \cdot \mathbf{M}_k^N = \mathbf{I}, \sum_{(x_i,y_i)\in D^N} a_{i,k}^N = 1, k = 1, \ldots, K\end{aligned}$$

这个问题同样可以用基于 SVD 的闭式解解决，此处省略具体过程。

关系分类器训练： 当两类关系投影模型训练完毕之后，我们训练基于神经网络的上下位/非上下位关系分类器。对于正负训练集的任一关系元组 $(x_i, y_i) \in D^P \cup D^N$，我们分别使用两类关系投影模型将输入实体的词向量 $\vec{x}_i$，分别投影至其 K 个上位词和非上位词的词向量，之后使用上下位和非上下位关系投影残差作为分类器的特征（即 $\mathcal{F}^P(\vec{x}_i, \vec{y}_i)$ 和 $\mathcal{F}^N(\vec{x}_i, \vec{y}_i)$），计算如下：

$$\mathcal{F}^P(\vec{x}_i, \vec{y}_i) = (\mathbf{M}_1^P\vec{x}_i - \vec{y}_i) \oplus (\mathbf{M}_2^P\vec{x}_i - \vec{y}_i) \oplus \cdots \oplus (\mathbf{M}_K^P\vec{x}_i - \vec{y}_i)$$

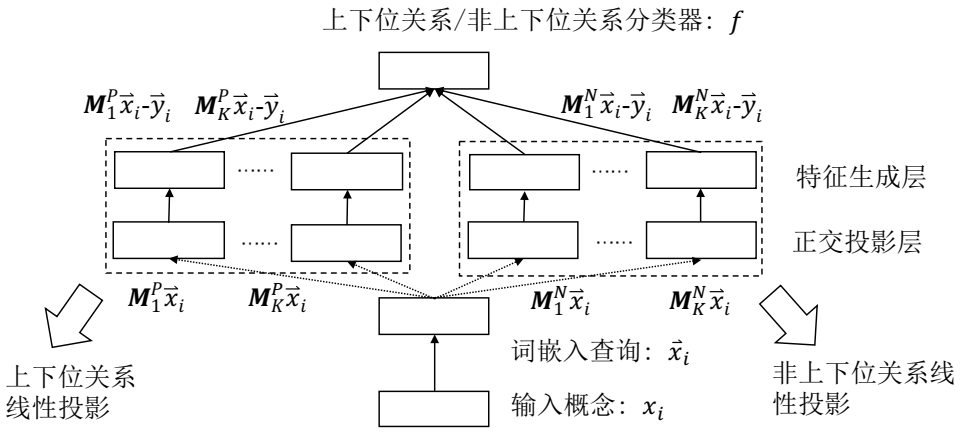


图 2.8: FOPM 采用的神经网络架构

Algorithm 3 FOPM 训练算法

对 D^P 中的关系元组进行 K-Means 聚类，特征采用 $\vec{x}_i - \vec{y}_i$
对 D^N 中的关系元组进行 K-Means 聚类，特征采用 $\vec{x}_i - \vec{y}_i$
for $k = 1$ to 簇的数量 K **do**
    利用基于 SVD 的闭式解，学习投影矩阵 $\mathbf{M}_k^P$ 和 $\mathbf{M}_k^N$
end for
for 每个关系元组 $(x_i, y_i) \in D^P \cup D^N$ **do**
    计算特征 $\mathcal{F}^P(\vec{x}_i, \vec{y}_i)$ 和 $\mathcal{F}^N(\vec{x}_i, \vec{y}_i)$
end for
在数据集 D^P 和 D^N 上训练神经网络关系分类器 f

$$\mathcal{F}^N(\vec{x}_i, \vec{y}_i) = (\mathbf{M}_1^N\vec{x}_i - \vec{y}_i) \oplus (\mathbf{M}_2^N\vec{x}_i - \vec{y}_i) \oplus \cdots \oplus (\mathbf{M}_K^N\vec{x}_i - \vec{y}_i)$$

图 2.8展示了 FOPM 算法中采用的用于关系分类的神经网络架构图。与经典的关系分类神经网络不同，对于一个关系元组 $(x_i, y_i) \in D^P \cup D^N$，它只采用 $\vec{x}_i$ 作为输入，分别计算特征 $\mathcal{F}^P(\vec{x}_i, \vec{y}_i)$ 和 $\mathcal{F}^N(\vec{x}_i, \vec{y}_i)$，然后在此之上训练神经网络分类器 f。在训练神经网络的过程中，我们只调整其对于特征 $\mathcal{F}^P(\vec{x}_i, \vec{y}_i)$ 和 $\mathcal{F}^N(\vec{x}_i, \vec{y}_i)$ 的权重，而投影矩阵本身不进行更新。当模型训练结束后，对于测试集中的关系元组 $(x_i, y_i) \in D^T$，使用预训练得到的投影模型计算上述特征，并用分类器 f 预测 x_i 和 y_i 之间是否具有上下位关系。

接下来简述图 2.8中模型架构设计原理。在分布式语义领域，根据 Levy 等人的研究 [82]，如果直接用 x_i 和 y_i 的词向量 $\vec{x}_i$ 和 $\vec{y}_i$ 作为特征来预测 x_i 和 y_i 之间的关系，可能会使神经网络严重过拟合，造成“词汇记忆”（Lexical Memorization）问题。例如，当分类器在训练数据中得到正例“（狗，动物）”、“（猫，动物）”、“（绵羊，动物）”等进行训练，神经网络会“记住”“动物”的部分词向量特征，将

表 2.11: 三种上下位关系预测算法的特点对比

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	是否利用额外统计信息
IPM	K	否	否	是	是
TPM	2	是	是	是	是
FOPM	$2K$	是	否	否	否

“狗” 当成 “典型上位词”（Prototypical Hypernym）。在预测的时候，神经网络会错误识别 “（玫瑰，动物）” 也是上下位关系元组。

对比我们的方法，如果 x_i 和 y_i 之间有上下位关系，根据两组特征的计算方式，$\mathcal{F}^P(x_i, y_i)$ 的范数会小，$\mathcal{F}^N(x_i, y_i)$ 的范数会大；如果 x_i 和 y_i 之间有非上下位关系，我们也可以类似地得出相反结论。所以，特征 $\mathcal{F}^P(x_i, y_i) \cup \mathcal{F}^N(x_i, y_i)$ 对上下位关系和非上下位关系有很强的区分能力。因此，这一方法有效避免了 “词汇记忆” 问题。FOPM 算法的训练过程见算法 3。

最后，我们汇总了 IPM、TPM 和 FOPM 三种提出的上下位关系预测算法的特点，比较结果见表 2.11。

2.5.2 实验分析

在本节中，我们在两个中文公开数据集上评测 FOPM 算法的有效性，并与基线方法相对比。此外，我们进一步在数据英语上下位关系预测的基准测试集上对提出的模型进行评测，证明了 FOPM 也能有效解决英语上下位关系预测问题。

中英文数据集与实验设置： 为了容易与先前方法对比，本节中实验同样采用与第2.3.2节相同的中文语料库和基于 Skip-Gram 的词向量训练上下位关系预测模型，其维度为 100。在中文实验中，我们同样采用 FD [25] 和 BK 两个数据集作为训练和测试集。与第2.4.2节设置相同，分别将这两个数据集采用同样的方式 5 折划分，进行交叉验证实验。

由于英语单词的构词法对于英语语义学习有很重要的作用，我们采用 fastText 模型 [53] 在英语维基百科的语料库上训练语言模型，其维度参照 Bojanowski 等人 [53] 的默认参数设置，设为 300。我们在两个英语上下位关系评测任务上对 FOPM 及其我们上文提出的模型进行综合评测。第一个任务为通用领域的**监督的上下位关系检测任务**（Supervised Hypernymy Detection），其任务目标为对一个术语对进行上下位/非上下位关系分类。其实验在 NLP 评测中经常使用的基准数据集上进行，分别为 BLESS [93] 和 ENTAILMENT [77]，包括 14547 和 2770 带有标注的通用领

表 2.12: 两个通用领域英语上下位关系检测数据集的统计信息

关系	BLESS	ENTAILMENT
上下位关系	1337	1385
其他（非上下位关系）	13210	1385

表 2.13: 三个特定领域英语上下位关系数据集的统计信息

统计信息	ANIMAL	PLANT	VEHICLE
术语数量	659	520	117
上下位关系数量	4164	2266	283
随机术语对数量	8471	4520	586

域英语词对。这两个数据集的统计信息见表 2.12。我们采用先前一系列英语上下位关系研究的 NLP 论文 [72, 80, 81] 中相同的“留一法（Leve-One-Out）”进行评测。在 BLESS 数据集 [93]，因为其中的语义关系与 WordNet 中 200 个最频繁出现的名词有关，我们随机选择一个名词有关的关系进行测试，在余下的其他关系上训练模型。在 ENTAILMENT 数据集 [77]，每次实验中，我们随机选取一个上下位关系进行测试，模型在其他数据上训练。实验结果的评价指标为平均准确度。

第二个英语语言的任务在于重构特定领域的分类体系。在这组实验中，数据集从三个特定领域的分类体系中采样得到，分别为 ANIMAL（动物）、PLANT（植物）和 VEHICLE（车辆）[97]。在这些数据集中，正例为分类体系中所有正确的上下位关系，负例为随机匹配的、不属于任何上下位关系的术语对，由 Luu 等人 [81] 采样和公开。这三个数据集的统计信息见表 2.13。这个任务的实验评测步骤与 Luu 等人 [81] 的设置相同。每次我们从数据集中留出一个关系元组进行测试，并在余下的数据集下训练模型。实验效果用平均准确度来衡量。因为在这三个数据集中少数术语由多个英语单词构成（例如 American tree、half track），我们将这些词组看成一个整体，其词向量为构成词组的每个词的词向量的算术平均值。其他实验设置均与第一个任务相同，细节不再赘述。

中文数据集上的实验结果： 我们扩展了第2.4.2节中的中文实验，对提出的 IPM、TPM 和 FOPM 三种算法进行综合评测，采用的基线方法同第2.4.2节中相同。IPM 和 TPM 两个模型的参数设置对实验结果的影响已在第2.3.2节和第2.4.2节中进行分析。对于 FOPM 算法，我们在 FD 和 BK 两个数据集上调整了参数 K 的值，并用随机梯度下降训练关系分类神经网络，其 F 值的变化结果与参数 K 的关系如图 2.9所示。从实验结果可见，当 $K = 3$（数据集 BK）或 $K = 4$（数据集 FD）时，实验效果最好。整体而言，预测的精度对参数 K 的变化敏感度不大。此外，由于

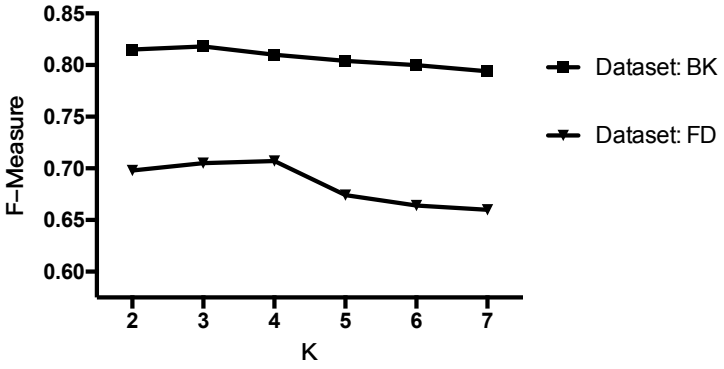


图 2.9: FOPM 中参数 K 的变化对模型在中文数据集预测效果的影响

表 2.14: FOPM 在中文测试集上的效果，及其与基线算法的对比

数据集	FD			BK		
方法	**精准度**	**召回率**	**F 值**	**精准度**	**召回率**	**F 值**
Fu-S [25]	0.641	0.560	0.598	0.714	0.648	0.679
Fu-P [25]	0.664	0.593	0.626	0.727	0.675	0.700
$\vec{x}_i \oplus \vec{y}_i$	0.677	**0.752**	0.697	0.803	0.759	0.780
$\vec{x}_i + \vec{y}_i$	0.653	0.607	0.629	0.727	0.656	0.689
$\vec{x}_i - \vec{y}_i$	0.719	0.606	0.657	0.784	0.607	0.684
IPM	0.693	0.645	0.669	0.739	0.698	0.718
TPM	**0.728**	0.705	**0.716**	**0.836**	**0.806**	**0.821**
FOPM	0.713	0.698	**0.705**	**0.825**	**0.812**	**0.818**

在学习模糊正交投影矩阵时，我们采用了基于 SVD 的闭式解得到全局最优值，这也保证了 FOPM 算法的稳定性。

我们进一步将各个基线方法和本文提出的模型结果汇总在表 2.14中。我们可以看出 TPM 和 FOPM 这两种算法比所有基线算法在 FD 和 BK 数据集上都有明显的提升。进一步观察实验结果，我们发现，IPM 适合在仅有少量上下位关系作为训练集，有大量未标注关系元组的 PU 学习（Positive Unlabeled Learning）的情景。TPM 和 FOPM 更适合区分上下位关系和非上下位关系。对比 TPM 和 FOPM 在两个数据集上的表现，他们在 F 值上相差不大，其中 TPM 比 FOPM 更适合中文数据集，在两个数据集上的 F 值分别比 FOPM 高 1.1% 和 0.3%。这说明了，TPM 中采用的中文语言学规则对中文上下位关系的预测有较大的贡献。值得进一步说明的是，FOPM 在中文数据集的预测精度虽然略低于 TPM，但是它在语言通用性上更好。在进一步的实验中，我们探究 FOPM 在英语上下位关系预测的准确度，并对中英文的相关任务效果进行综合对比和讨论。

英语数据集上的实验结果： 由于上下位关系预测在英语上的研究比较充分，

表 2.15: 不同方法在通用领域监督的上下位关系检测任务中的精确度比较

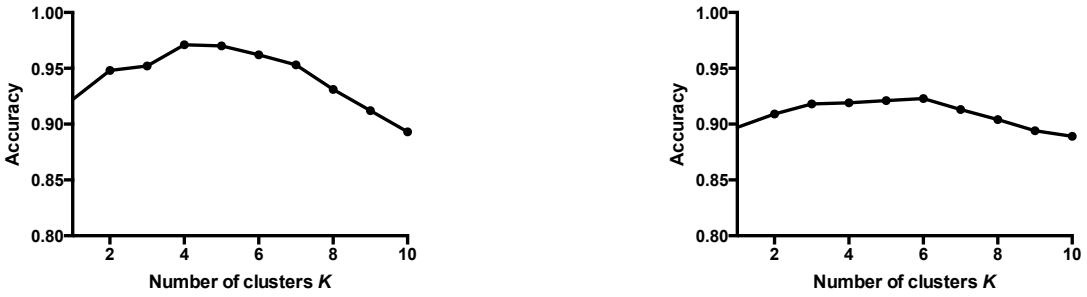
方法	**BLESS**	**ENTAILMENT**
Word2Vec [51]	0.84	0.83
Yu 等人 [80]	0.90	0.87
Luu 等人 [81]	0.93	0.91
HyperVec [72]	0.94	0.91
IPM-EN	0.85	0.84
TPM-EN	0.90	0.89
FOPM-N	**0.95**	0.90
FOPM	**0.97**	**0.92**

相关基线方法比较多。在本文中，我们考虑以下 4 个基线方法：

- Word2Vec [51]：它采用标准的 Skip-Gram 语言模型得到词向量，采用负采样方式训练。
- Yu 等人 [80]：它基于 Probase 分类体系 [10] 中的上下位关系知识，分别学习上位词和下位词的词向量，这一算法利用最大间隔神经网络来优化。
- Luu 等人 [81]：它基于英语维基百科数据学习英语术语的上下位关系词向量，并且提出了一个动态权重神经网络来优化相关词向量。
- HyperVec [72]：它结合 Skip-Gram 模型 [51] 的负采样技术和 WordNet 中的上下位关系知识 [17]，协同学习术语词向量。

根据上述论文的具体实现，算法 [51, 80, 81] 的特征为术语对的上下位关系词向量，及其这两个向量之差。算法 [72] 的特征包括相应两个术语词向量的向量差、余弦相似度和向量差范数。我们在上述分类器上训练基于多项式核的 SVM 分类器进行关系预测。对于本章我们提出的方法，我们在英语数据集上评测了 IPM 算法的英语版本（IPM-EN）、TPM 算法的英语版本（TPM-EN）、FOPM 算法的变体（在模糊正交投影模型中去除了正交化约束，以下记为 FOPM-N）和 FOPM 的完整实现版本。IPM-EN 和 TPM-EN 的实现方法见上文第2.4.2节。

我们汇总了所有本章提出的算法及基线算法在通用领域的实验效果,如表 2.15所示。在 FOPM 算法中, 簇的数量 K 默认设为 4, 我们会在下文中详细讨论参数 K 对实验性能的影响。从实验结果可以看出，FOPM 的精准度超过了所有基线方法。相对简单的模型例如 Word2Vec [51] 等的精度不高，因为上下位关系的语义没有被建模。与最强的基线方法 HyperVec [72] 相比，FOPM 的精度在 BLESS 数据集上提升



(a) 数据集：BLESS　　(b) 数据集：ENTAILMENT

图 2.10: FOPM 中参数 K 的变化对于模型预测效果的影响

表 2.16: 不同方法在特定领域监督的上下位关系检测任务中的精确度比较

方法	**ANIMAL**	**PLANT**	**VEHICLE**
Word2Vec [51]	0.80	0.81	0.82
Yu 等人 [80]	0.67	0.65	0.70
Luu 等人 [81]	0.89	0.92	0.89
HyperVec [72]	0.83	0.91	0.83
IPM-EN	0.85	0.87	0.78
TPM-EN	0.89	0.90	0.84
FOPM-N	**0.90**	**0.92**	0.87
FOPM	**0.92**	**0.94**	**0.90**

了 3%，在 ENTAILMENT 数据集上提升了 1%。与 IPM-EN、TPM-EN 和 FOPM-N 比较，FOPM 的精度也有明显的提升。

因为没有标准方法来确定一个数据集中有多少个上下位关系的分量，参数 K 的设置采用启发式规则，在 BLESS 和 ENTAILMENT 两个数据集中，我们将 K 的值从 1 调整到 10，实验结果见图 2.10。可以发现，如果 K 的值不设为极端大或极端小的值，模型的实验性能对 K 的变化并不敏感。这是因为在我们的模型中，我们引入了 $a_{i,k}^P$ 和 $a_{i,k}^N$，使得投影学习是“模糊的”。当 $K = 1$ 时，FOPM 退化成具有单个正交投影矩阵的模型，与之比较，聚类的方法在两个数据集上分别提升了 5% 和 2% 的精度。

我们进一步在三个特定领域数据集上进行了实验，结果见表 2.16。同样可见，FOPM 在精度上超过了所有基线方法，FOPM-N 在两个领域数据集（PLANT 和 ANIMAL）上超过了最强的基线算法 [81]，在 VEHICLE 数据集上与之相当。在特定领域，使用通用词向量的算法（如 [51, 81] 和我们提出的算法）效果比使用特定上下位关系词向量的算法 [72, 80] 更好。这是因为在 Yu 等人 [80] 和 Nguyen 等人 [72] 的方法中，作者分别利用了 Probase 和 WordNet 中的上下位关系数据学习上

下位关系词向量。这些数据在特定领域的覆盖率往往有限，导致特定领域术语的语义没有在词嵌入空间得到良好的表示。

2.6 小结

在本章中，我们详细介绍了面向中文上下位关系抽取的三种基于词嵌入的模型，分别为 IPM、TPM 和 FOPM。实验结果表明，IPM 对于中文上下位关系扩展、TPM 和 FOPM 对于中文上下位关系分类的效果明显，超过了现有最佳方法。此外，FOPM 也在英语上下位关系预测的多个基线评测任务上取得了与现有最佳方法相近与略高的准确度。我们的研究表明，学习基于词嵌入的上下位关系投影模型，能有效实现高精度中文上下位关系抽取的目标。

第三章　知识增强的语义关系抽取

在第二章中，我们提出了多种基于词嵌入模型的投影学习算法，学习如何将下位词的词向量投影至上位词的词向量，从而准确预测术语间的上下位关系。然而，这些算法中的大部分模块依赖于有限的、人工标注的训练集，对外部知识和其他辅助任务没有加以良好运用。在本节中，我们在第二章研究的基础上，深入地探索知识增强的语义关系抽取算法，以扩展基于词嵌入模型的投影学习算法的应用空间。特别地，我们关注三个问题：i）如何在上下位关系预测模型中，融入海量网络知识源，以提高现有模型的预测精度；ii）如何设计跨语言的上下位关系抽取算法，将源语言的知识自动迁移到目标语言，使得当目标语言训练数据量极小时，模型也能准确地对目标语言中的上下位和非上下位关系进行分类；以及 iii）如何对非上下位关系数据中不同类别的词汇关系分别进行语义建模，使得模型在识别上下位关系之外，也能对其他多种词汇关系进行分类。在下文中，我们详细介绍本章关注的语义关系抽取算法、相关研究以及实验结果。

3.1　引言

回顾第二章的研究主题，对于两个语义相关的术语 x_i 和 y_i，将其对应的词嵌入向量 $\vec{x}_i$ 和 $\vec{y}_i$ 作为算法的输入，判断 x_i 和 y_i 之间是否存在上下位关系。围绕这一研究问题，我们提出了 IPM、TPM 和 FOPM 三种算法。然而，这些算法对人工标注的训练集依赖程度较高，在特定领域、特定小语种情况下很难实现高精度的预测。此外，在第二章中，我们笼统地将所有不属于上下位关系的其他词汇关系称之为“非上下位关系”，没有分别进行建模。因此，模型在学习不同类别的非上下位关系之间的区别、以及不同类别的非上下位关系与上下位关系之间的区别存在困难。

在本章中，我们继续深入研究如何学习 x_i 和 y_i 之间的语义关系。特别地，我们分别从**多知识源**、**多语言**和**多词汇关系**这三个切入点深入探究知识增强的语义关系抽取算法。这三个方面的研究分别概述如下：

多知识源： 上下位关系分类基本输入为正负例关系数据集 D^P 和 D^N，分别包含上下位和非上下位关系术语对 (x_i, y_i)。通过第二章的实验分析，我们可以发现，当训练集的数据量较小时（特别是中文训练集 FD [25] 和英语特定领域训练集

表 3.1: 自动构建的大规模分类体系关系数量

分类体系	上下位关系数量
WikiTaxonomy（英语）[19]	105418
YAGO（英语）[8]	8277227
WiBi（英语）[100]	2736022
Probase（英语）[10]	16285393
Probase+（英语）[101]	21332357
CN-WikiTaxonomy（中文）[21]	1317956
CN-Probase（中文）[20]	32925306

ANIMAL、PLANT 和 VEHICLE [97] 等），预测精度随之下降。在我们提出的 IPM 算法中，尽管可以自动从网络数据中挖掘出更多上下位关系，为了保证迭代学习的准确性，新挖掘出的关系往往与训练数据语义高度相似；因此，IPM 算法精度的提升上限受到这一机制本身的制约。

另一方面，随着分类体系构建技术的快速发展，研究者从海量网络数据中自动构建诸多大规模分类体系（统计信息见表 3.1），这些分类体系数据量大、精度较高而且覆盖领域广，可以作为高价值的网络知识源。然而，分类体系中具有的知识在数据量、数据质量和领域分布与待预测数据有差异，很难直接进行知识迁移。深度对抗学习技术在 NLP 诸多任务中应用广泛 [29, 98, 99]。我们可以分别在训练集 D^P 和 D^N，和基于分类体系正负例关系数据集 T^P 和 T^N 上分别学习词嵌入投影神经网络，利用对抗学习自动将这两种数据源所需的知识相融合，从而在 D^P 和 D^N 数据量较少时，提升投影学习的效果。这一框架即为**分类体系增强的对抗学习框架**（Taxonomy Enhanced Adversarial Learning，缩写为 TEAL）。

多语言： TEAL 框架的缺点在于，当且仅当训练集和使用的分类体系为同一语言时，不同知识源的投影学习模型才可以通过对抗学习实现知识迁移。当某目标语言的训练集数据量较小，而且比较难以获得与该训练集语言相同、领域相似的分类体系时，TEAL 框架适用性会显著降低。随着神经机器翻译技术的深入发展，不同语言之间可以实现无平行语料库情况下的术语翻译 [102]，即自动对齐两种语言的词向量空间。这一技术给上下位关系预测算法在不同语种之间的迁移提供了空间。

在上述条件下，我们研究了跨语言上下位关系预测问题。给定四个训练集：源语言上下位、非上下位关系训练集 D_S^P 和 D_S^N，目标语言上下位、非上下位关系训练集 D_T^P 和 D_T^N（$|D_S^P| \gg |D_T^P|$，$|D_S^N| \gg |D_T^N|$），该任务的目标是实现**小样本学习**（Few-shot Learning）场景下目标语言的上下位/非上下位关系的准确分类。利用双

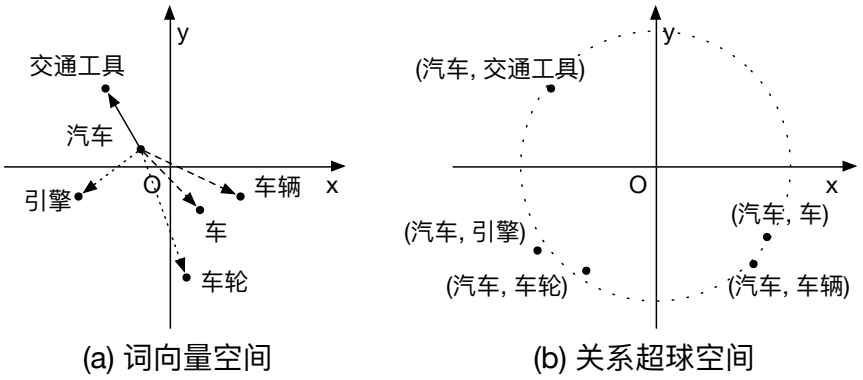


图 3.1: 超球学习关系嵌入学习示例

语术语对齐模型 [102], 我们扩展了先前的 FOPM 算法, 提出了**迁移模糊正交投影模型**（Transfer Fuzzy Orthogonal Projection Model，缩写为 TFOPM）。此外，为了进一步学习目标语言中特有主题的上下位关系, 我们将 TFOPM 与 IPM 中的迭代学习思想相结合，提出了**迭代迁移模糊正交投影模型**（Iterative Transfer Fuzzy Orthogonal Projection Model，缩写为 ITFOPM）。

多词汇关系： 在前述所有提出的模型中，非上下位关系的语义没有得到细粒度建模。例如, 在 TPM 中, 我们采用单一投影矩阵学习一个术语到它的非上下位词在词嵌入空间的投影；在 FOPM 中，我们采用模糊正交矩阵作为非上下位关系的投影模型。然而，这种建模方式没有对非上下位关系中的多种词汇关系加以区分, 例如同义词关系、反义词关系等。**词汇关系分类**（Lexical Relation Classification）任务 [103–106] 能有效解决这一问题，它的训练数据为术语对集合 $D = \{(x_i, y_i)\}$, 每个术语对 (x_i, y_i) 之间有唯一的词汇术语类别 $r_i \in \mathcal{R}$, 其中, $\mathcal{R}$ 为预定义的词汇关系类别集合。词汇关系分类模型不仅能对未知关系的术语对进行多路关系分类，而且也是知识图谱中**本体构建**（Ontology Construction）的重要支持任务 [107, 108]。

由于具有不同词汇关系的术语在语义上都高度相关, 在原始词向量空间，这些术语之间的词汇关系很难进行建模和区分。受到计算机视觉的相关研究启发 [109, 110]，我们对每个术语对 (x_i, y_i)，都学习唯一的**超球关系嵌入**（Hyperspherical Relation Embedding, 缩写为 SphereRE）向量 $\vec{r}_i$。这一向量处于在超球嵌入空间中，使得具有相同词汇关系类别的术语对具有相似的 SphereRE 向量。这些术语的词汇关系在原始词嵌入空间和超球嵌入空间的表示对比示例见图 3.1 。从上述示例可见, 在超球嵌入空间，术语对的词汇关系更容易被分类模型学习并准确分类。

综上所述，我们分别概述了知识增强的语义关系抽取在三个角度的研究，扩

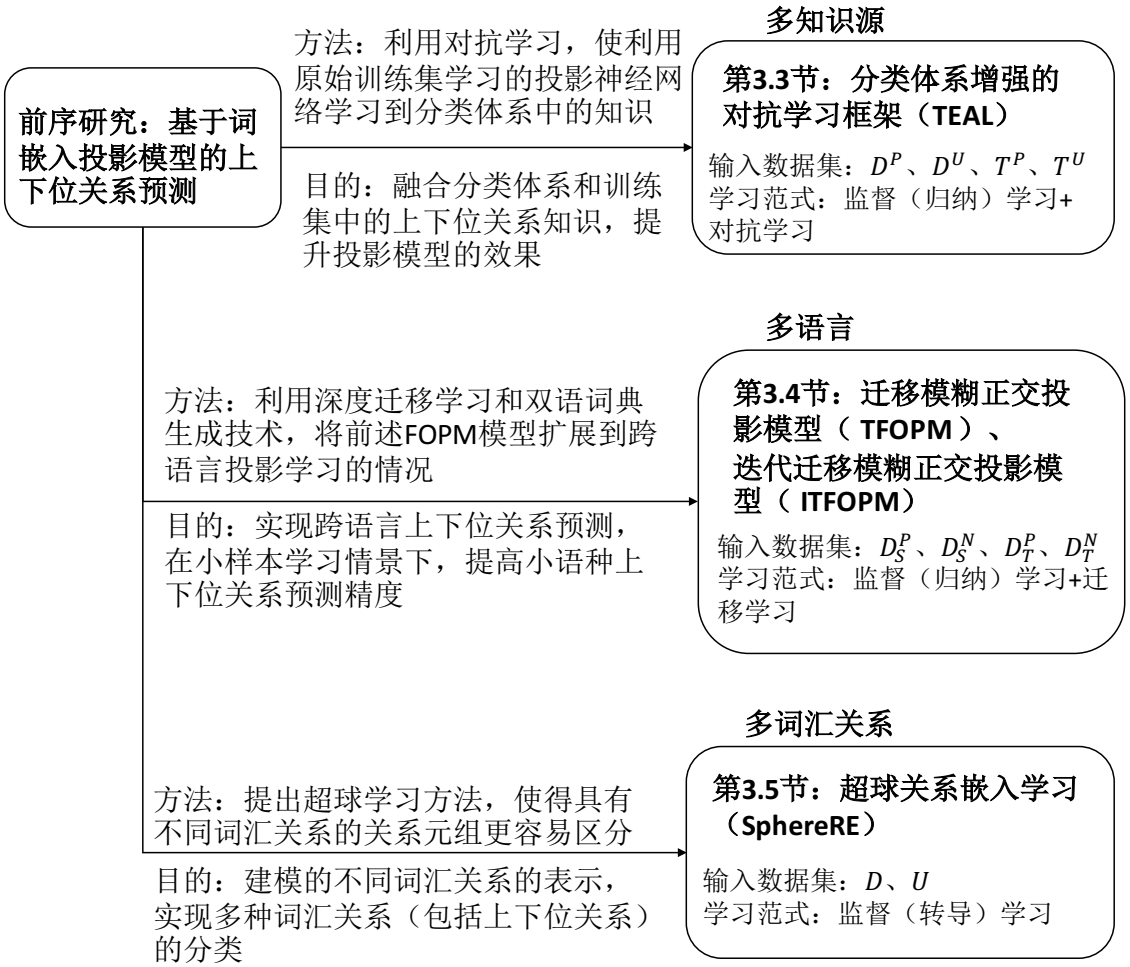


图 3.2: 第三章模型研究思路汇总

展了第二章的研究工作。本章提出的 TEAL、TFOPM（包括其扩展模型 ITFOPM）和 SphereRE 模型的研究思路汇总于图 3.2中。表 3.2总结了第三章中使用的重要符号及其意义。

3.2 相关工作

由于在第二章中，我们已经详述了上下位关系抽取（特别是基于模式匹配和分布式学习方法）的相关研究进展，在本节中，我们对上下位关系抽取的基础方法不再赘述。在本节中，我们从多知识源、多语言和多词汇关系三个知识增强的角度，概述了相关工作。从多知识源角度，我们概述了对抗学习在关系抽取和其他 NLP 任务中的应用；从多语言角度，我们介绍了跨语言迁移学习的相关算法；此外，我们也对多词汇关系分类的相关研究进行总结和评述。

表 3.2: 第三章使用的重要符号及其意义

符号	说明
(x_i, y_i)	一个概念/术语对
$\vec{x_i}$	x_i 的词向量
D^P	TEAL 中上下位关系训练集
D^N	TEAL 中非上下位关系训练集
T^P	TEAL 中基于分类体系的上下位关系数据集
T^N	TEAL 中基于分类体系的非上下位关系数据集
$H(\vec{x_i}; \vec{\theta})$	TEAL 中 x_i 通过以 $\vec{\theta}$ 为参数集的神经网络生成的词向量（$\vec{\theta}$ 可以为不同形式、不同数量的参数，此处从简表示）
D_S^P	TFOPM 中源语言上下位关系训练集
D_S^N	TFOPM 中源语言非上下位关系训练集
D_T^P	TFOPM 中目标语言上下位关系训练集
D_T^N	TFOPM 中目标语言非上下位关系训练集
U_T	TFOPM 中目标语言术语对测试集
K	TFOPM 中簇的数量
$\mathbf{S}$	TFOPM 中跨语言词向量空间映射矩阵
$\mathbf{M}_k^P$	TFOPM 中第 k 个簇的目标语言上下位关系投影矩阵
$\mathbf{M}_k^N$	TFOPM 中第 k 个簇的目标语言非上下位关系投影矩阵
D	SphereRE 中词汇关系分类训练集
U	SphereRE 中词汇关系分类测试集
$\mathcal{R}$	SphereRE 中词汇关系类别集合
r_i	SphereRE 中 (x_i, y_i) 的词汇关系类别
$\mathbf{M}_m$	SphereRE 中词汇关系 $r_m \in \mathcal{R}$ 对应的投影矩阵
$p_{i,m}$	SphereRE 中模型预测 (x_i, y_i) 有词汇关系 $r_m \in \mathcal{R}$ 的概率
$\vec{r_i}$	(x_i, y_i) 的 SphereRE 向量

3.2.1 对抗学习在 NLP 的应用

对抗学习（Adversarial Learning）是深度学习的重要研究方向之一。它的核心技术在于学习对抗神经网络，误导已有的模型做出错误的预测。对抗学习的一个典型的应用是**对抗生成网络**（Generative Adversarial Network，缩写为 GAN）[111]，它在学习图像概率生成模型的基础上，同时训练对抗分类器，用于判断给定的图像为真实图像或由模型生成。概率生成模型同时需要尽可能拟合真实图像的概率分布，以及欺骗对抗分类器，使得生成的图像尽可能真实。对抗学习在多个机器学习领域都有广泛的应用，例如人脸生成 [112]、眼动追踪 [113]、视频生成 [114]。

在 NLP 领域，由于文本是高度离散化的，经典的对抗生成模型很难直接产生自然语言。SeqGAN [115] 是文本生成的经典模型，它采用强化学习中的随机策略框架作为生成器，并且利用 GAN 判别器的输出结果作为强化学习算法中的奖励。de Masson d'Autume 等人 [116] 提出，不需要基于极大似然的预训练，就可以从零

开始训练生成自然语言的 GAN。

除了直接生成自然语言，对抗学习在 NLP 的研究重点是如何利用对抗学习算法，生成 NLP 中多个任务相关的对抗样本，以测试深度学习模型的鲁棒性。其中，Alzantot 等人 [117] 提出了基于深度神经网络的黑盒优化模型，用于生成与训练数据在语法和语义上足够相似的对抗样本，以误导情感分类和文本蕴含模型。Zhang 等人 [118] 提出了基于 Metropolis-Hastings 采样的攻击模型，用于改进自然语言生成的梯度训练方法，从而使模型生成更为流利的自然语言文本。另一种类似的对抗文本生成算法由 Ren 等人 [119] 提出，这一算法基于同义词替换的策略，使被误导的分类器准确率下降极大化的同时，控制被替换的词的数量。

对抗学习在 NLP 中最常用的技术为对抗训练（Adversarial Training），即不以生成自然语言的任何单元作为任务的直接目标，而是使得多个 NLP 模型互相对抗，使得所需的模型在对抗训练中不断增强。在关系抽取任务中，对抗训练技术应用广泛。Wu 等人 [120] 首次将对抗分类器的损失作为关系分类中机器学习模型正则项，提升基于 CNN 和 RNN 的关系分类器的效果。由于利用远程监督的关系抽取算法容易受到标注噪声的影响，这种用于关系抽取的训练数据会包含伪正例，Wang 等人 [29] 利用类似 GAN 的架构，同时学习关系抽取器和鉴别器，使得关系抽取器尽可能排除伪正例，从真正例的数据集上学习参数。Shi 等人 [121] 考虑到在不同题材文本中抽取关系的困难性，设计了基于对抗学习的题材分割神经网络解决这一问题。这一网络包括题材独立和题材共享的特征抽取器，并且对抗关系分类器，使得在源题材学习的分类器可以直接用于目标题材上。

在下文中，我们也举例论述其他 NLP 任务中采用的对抗学习技术：

- **命名实体识别** [122]：分别在高资源和低资源语言训练命名实体识别神经网络，采用对抗学习融合两种语言的特征，实现低资源语言的命名实体识别。

- **中文分词** [99]：由于中文分词的标注数据集的数据量相对较小，且分词的标准很难统一，在训练用于中文分词的序列标注模型的基础上，同时学习中文分词标准鉴别器，最终使得中文分词模型融合了多个中文分词标准的知识。

- **文本摘要** [123]：采用 “生成器-重构器-鉴别器” 三重架构，生成器获得原始文档作为输入，产生较短的文本摘要；重构器试图从生成的文本摘要重构原始输入文档，保证摘要包含原始输入的必要信息；鉴别器采用对抗学习机制，使生成器输入的文本摘要更加接近人工撰写的摘要。

- **篇章关系分析** [124]：篇章关系分析的难点在于篇章之间一般缺乏关系连接词（例如“但是”、“因此”）。这一模型分别训练两个篇章关系分类神经网络，在原始数据和关系连接词补全的情况下进行训练，此外，模型优化鉴别器的损失作为关系分类模型的正则项，使得基于原始数据的篇章关系分类网络逐步学习到另一个网络的知识。

- **知识图谱的表示学习** [125]：知识图谱一般只包含正例（即正确的关系元组），对比较难以判别正确性的负例进行采样比较困难。KBGAN 的鉴别器使用已有的表示学习模型产生的正例作为负例，协助其他知识图谱表示学习的模型进行训练，提升这些模型的表示学习能力。

此外，对抗学习还能运用于其他 NLP 任务中，例如词性标注 [126]、对话系统 [127] 等，不再详细列举。我们提出的 TEAL 框架也属于这一类别。与通常 NLP 任务中使用的 GAN 架构不同，由于我们的模型在多任务框架下同时学习上下位关系和非上下位关系的投影，TEAL 中采用双重对抗学习机制，学习两个对抗分类器用于上下位知识的融合。

3.2.2 跨语言迁移学习在 NLP 的应用

对很多语言和 NLP 任务而言，获取足够多的训练数据难度比较大，这个问题使精确训练目标语言所需的 NLP 模型比较困难。**跨语言迁移学习**（Cross-lingual Transfer Learning，缩写 CTL）是解决这一问题的一种方法，指从其他语言中获取相关的资源，帮助目标语言模型的训练。按照 CTL 学习范式的不同，一般可以分为**基于资源的迁移**（Resource-based Transfer）和**基于模型的迁移**（Model-based Transfer）两类方法 [128]。

基于资源的迁移将源语言的训练数据通过跨语言投射映射到目标语言。在有平行数据资源的情况下，源语言训练数据的标签可以直接进行投射，并用于目标语言模型的训练。Wang 和 Manning [129] 提出投射标签的期望分布而非标签本身，降低了 CTL 过程中引入的噪声。双语词典在基于资源的 CTL 中通常扮演重要角色。Prettenhofer 和 Stein [130] 利用两种语言在词级别的对齐，进行跨语言文本分类。Xu 等人 [91] 指出，部分语言对的双语词典规模较小不适合直接使用，他们计算两种语言术语之间模糊概率的对齐关系，实现双语词典有限情况下的文本分类。部分研究直接借鉴了机器翻译技术 [131–133]，将其结果应用于多个 NLP 任务中。

本文提出的 TFOPM 和 ITFOPM 两种算法也属于这一类别，通过双语字典生成学习跨语言词汇对齐，将源语言上下位关系训练集投射到多个小语种上。

与基于资源的方法相比，基于模型的迁移在 NLP 发展的早期应用并不广泛，因为不同语言区别较大。即使涉及的 NLP 任务相同，模型在源语言和目标语言的特征空间也往往不同，很难实现迁移。早期基于模型的迁移常使用去词汇化（Delexicalizaion）技术，即不依赖于具体词汇，实现特征空间的迁移。具体的应用场景包括依存句法分析 [134]、词性标注 [135] 等。随着深度学习技术的广泛应用，基于模型的迁移可以通过学习跨语言表示实现。这一类神经网络的架构分别学习源语言和目标语言各自的特征表示，通过添加共享层学习跨语言特征表示，实现模型参数的迁移。典型的跨语言研究工作包括共指消解 [136]、命名实体识别 [137]、实体对齐 [138] 等。

上节概述的对抗训练方法（例如 [29, 99, 122, 123] 等）也可以被归类成基于模型的深度迁移方法，因为它将源领域训练的模型隐示地迁移到目标领域，不需要建模迁移的方式。用于跨语言的对抗迁移学习的研究工作有 Cao 等人提出的中文命名实体识别模型 [139]、Li 等人提出的命名实体边界检测 [140] 等。特别地，对于跨语言关系抽取工作，也有不少相关的算法，例如 [141, 142] 等。

与 CTL 有关的另一个任务为多语言表示学习，即将多种语言的词汇单元映射到同一向量空间中，使下游任务的模型可以在多个语言之间任意迁移。Chen 和 Cardie [143] 提出非监督式的多语言词嵌入模型，考虑了多种语言之间的相似关系。Wada 等人 [144] 提出多语言神经语言模型（Multilingual Neural Languge Model），在同一空间学习多种语言的句子表示，并且实现多语言的部分参数共享。随着深度学习在 NLP 的进一步发展，深度 CTL 的应用空间将会进一步增大。

3.2.3 词汇关系分类

词汇关系分类可以看成上下位/非上下位关系分类任务从二分类到多分类的扩展。与前述任务相似，词汇关系分类的方法也可以分为模式匹配法和分布式学习法两个主要类别。

对于上下位关系，Hearst 模式 [38] 是模式匹配法中最经典的英语语言模式。随着深度学习的发展，基于 LSTM 的神经网络可以用于学习语言模式的嵌入表示，典型的模型包括 Shwartz 等人提出的混合神经网络 [54] 等。LexNET [104] 将这一二分类模型扩展至多分类，采用与 Shwartz 等人 [54] 相似的神经网络结构用于词汇关

系分类。与上下位关系不同，其他词汇关系（例如同义词关系、反义词关系等）在语料库中相对固定的表达更为稀少，这导致模式匹配法在词汇关系分类任务中获得的召回率较低。为了缓解这一问题，Washio 和 Kato [105] 提出了“增广依存路径”（Augmented Dependency Path）这一概念，通过丰富术语对在语料库中挖掘到的依存路径的语义表达，提高了语言模式对于相应术语对的覆盖率。Nguyen 等人 [145] 提出了 AntSynNET 模型，从句子的语法解析树中提取词汇语法模式，作为神经网络的原始输入，对同义词、反义词两种关系分类。

用于词汇关系分类的分布式学习方法也与上下位关系分类的该类方法相似。基本的方法可以采用术语对 (x_i, y_i) 的词向量拼接 $\vec{x_i} \oplus \vec{y_i}$ [23, 77]、词向量之差 $\vec{x_i} - \vec{y_i}$ [71, 78, 146] 等组合作为特征，训练用于多分类的模型。这一类方法同样容易受到“词汇记忆”问题 [82] 的影响，使这些模型的泛化能力大幅下降。为了更好地学习各种词汇关系的语义，研究者提出了多种结构更为复杂的神经网络模型。Glavas 和 Vulic [106] 假设在预测不同类别的词汇关系时，同一个词往往需要学习不同的词向量来表征相应词汇关系的语义，并且提出了特化张量模型（Specialization Tensor Model）解决这一问题。Attia 等人 [147] 将多类别词汇关系的学习任务建模成多任务学习问题，采用多任务卷积神经网络进行关系分类。Bouraoui 等人 [148] 系统性地研究了两个词的词向量是怎样与他们之间的语义关系有关，并从概率角度改进了基于 Word2Vec 的经典关系预测模型。

除了直接采用预训练的词嵌入模型作为模型的原始输入，部分研究旨在改进词的表示学习方法。在上下位关系方面，已经有多个方法 [72, 80, 81] 学习上下位关系嵌入，以更好地从词向量角度表达上下位关系的层次性质，在本节中不再详细介绍。对于其他词汇关系，Nguyen 等人 [149] 观察到同义词和反义词关系在词向量空间通常难以区分，因此提出词汇关系对比敏感的词向量学习算法，改进了基于 Word2Vec 中负采样学习算法,着重考虑了同义词和反义词的区分性问题。Hashimoto 等人 [150] 提出面向语义关系分类这一特定任务的词向量学习算法，并在这一模型的基础上给出了多组描述名词之间语义关系的特征表示。Mrksic 等人 [151] 提出了单一语言和跨语言的 Attract-Repel 模型，在词向量学习模型中显示地加入了限制条件，这些条件从现有的词汇资源中抽取出来，表达了这些词的单语言语义相似性、跨语言语义相似性和反义关系。

另一类方法直接学习怎样在向量空间学习词对的关系表示，使这种向量表示与该词对的语义关系相关联。Washio 和 Kato [39] 提出了神经潜在关系分析（Neural

Latent Relational Analysis）模型，其生成的关系向量表达了两个词在语料库匹配的词汇语法模式以及各自的语义。Chen 等人 [152] 改进了 Fu 等人的投影模型 [25]，设计了基于上下位关系的词关系自编码器，使两个词在测试阶段不在语料库中同一个句子共现时，模型也可以进行自动关系发现新的上下位关系元组。Camacho-Collados 等人 [153] 等人提出了关系词嵌入（Relational Word Embedding）模型，这一模型基于词在语料库的共现统计信息，当外部语言资源或者词汇关系训练集不存在时，这一模型仍然能学习关系向量表示。

上述研究工作分别从不同的角度解决词汇关系分类的问题，然而部分方法不可避免地具有一定局限性。例如，文献 [72, 80, 81, 151, 152] 等只关注一种或少数几种词汇关系的分类或检测。另一些神经网络模型（例如 [39, 150]）则需要在整个语料库重新训练模型，算法复杂度较高。与之相比，我们提出的基于超球学习的 SphereRE 模型基于现有神经语言模型，并可以对任意种类的词汇关系分类。值得注意的是，超球学习相关研究工作大部分关注计算机视觉领域 [109, 110]，在 NLP 应用比较有限。例如，Masumura 等人 [154] 提出超球查询似然性模型，用于基于概率的查询检索。Lv 等人 [155] 将超球建模方法应用于知识图谱的表示学习，在超球模型上更好地区分知识图谱中概念与实体的表示。我们提出的 SphereRE 算法首次将超球学习应用于词汇关系分类，并且提出高效的随机采样学习算法，克服了先前研究的部分缺点。

3.3 基于对抗学习的跨知识源上下位关系融合

从本节开始，我们详细介绍提出的三种知识增强的上下位关系抽取模型。在本节中，我们首先描述**分类体系增强的对抗学习框架**（TEAL）的技术细节，这一框架用于将大规模网络分类体系中的关系型知识与上下位关系抽取中的训练集相融合，提升模型的效果。

3.3.1 算法模型

在 TEAL 模型中，我们利用基于深度神经网络的对抗学习方法达到融合大规模分类体系和训练集的目的。为了容易理解模型构造，在下文中，我们首先概述不考虑对抗学习的情况下，如何利用神经网络学习上下位关系和非上下位关系在词向量空间的非线性投影，实现上下位、非上下位关系的分类，其次详述 TEAL 模型的技术细节。

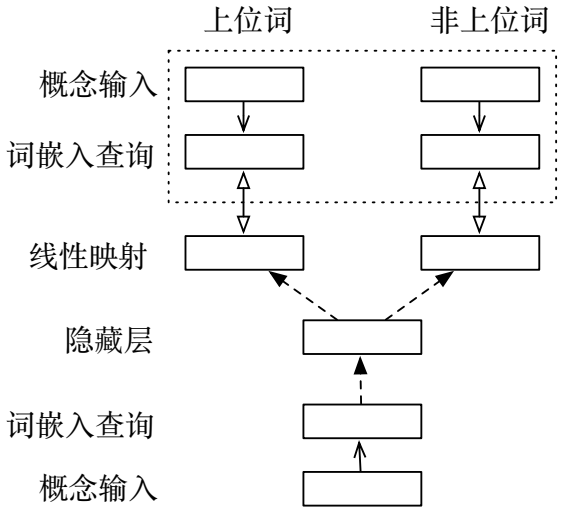


图 3.3: TEAL 的基础神经网络架构

监督式神经网络架构： 根据 Levy 等人 [82] 和我们先前的研究，直接利用词向量作为特征训练分类器会造成 “词汇记忆” 的问题。因此，如果利用神经网络，我们同样不能直接将分类错误的损失当成神经网络的损失函数。在本工作中，我们将神经网络替代了 FOPM 算法中的投影矩阵，其架构如图 3.3 所示：

该神经网络利用正负例的训练集 $D^P = \{x_i, y_i\}$ 和 $D^N = \{x_i, y_i\}$ 进行训练。与经典前馈神经网络分类器不同，它的输入为候选下位词 x_i 词向量 $\vec{x}_i$，有两种类型的输出：x_i 的上位词的词向量和非上位词的词向量。神经网络的学习目标为分别学习 $\vec{x}_i$ 到上位词的词向量和非上位词的词向量的映射，使投影误差最小化。据此，我们定义基础神经网络的损失函数 $\mathcal{L}_D$，如下所示：[1]

$$\mathcal{L}_D = \mathbb{E}_{(x_i,y_i)\sim D^P}\|H(\vec{x}_i;\vec{\theta}_D^P) - \vec{y}_i\|^2 + \mathbb{E}_{(x_i,y_i)\sim D^N}\|H(\vec{x}_i;\vec{\theta}_D^N) - \vec{y}_i\|^2$$

其中，$\vec{\theta}_D^P$ 和 $\vec{\theta}_D^N$ 分别为神经网络关于上下位映射和非上位词映射的参数（我们省略了各层具体参数表达，将其记为 $\vec{\theta}_D^P$ 和 $\vec{\theta}_D^N$），$H(\vec{x}_i;\vec{\theta}_D^P)$ 和 $H(\vec{x}_i;\vec{\theta}_D^N)$ 分别为模型预测 x_i 的上位词的词向量和非上位词的词向量[2]。$\mathbb{E}_{(x_i,y_i)\sim D^P}(\cdot)$ 和 $\mathbb{E}_{(x_i,y_i)\sim D^N}(\cdot)$ 分别为数据集 D^P 和 D^N 上，对应投影词向量误差平方的期望值。

由于这一神经网络有两个预测子任务，我们也可以将其看成多任务学习的神经网络。我们利用参数共享技术 [156] 改进神经网络的损失函数。对于某术语 x_i，

[1]在本节中，为了公式表达简洁，我们省略所有神经网络损失函数的正则化项。

[2]因为一个词 x_i 可能会有多个上位词和非上位词，所以从严格意义上，$H(\vec{x}_i;\vec{\theta}_D^P)$ 和 $H(\vec{x}_i;\vec{\theta}_D^N)$ 不代表具体某个词的词向量，而可以认为是 x_i 的多个上位词和非上位词的词向量 “中心”。

Algorithm 4 TEAL 中基础神经网络训练算法

随机初始化 $\vec{\theta}_D^P$、$\vec{\theta}_D^N$ 和 $\vec{\theta}_D^*$
for 每个神经网络的 Epoch **do**
  for 每一次神经网络的迭代 **do**
    从 D^P 中随机采样一个 Mini-Batch 数据集 $\tilde{D}^P$
    固定 $\vec{\theta}_D^N$，通过最小化 $\sum_{(x_i,y_i)\in\tilde{D}^P}\|H(\vec{x}_i;\vec{\theta}_D^P,\vec{\theta}_D^*)-\vec{y}_i\|^2$ 更新 $\vec{\theta}_D^P$ 和 $\vec{\theta}_D^*$
    从 D^N 中随机采样一个 Mini-Batch 数据集 $\tilde{D}^N$
    固定 $\vec{\theta}_D^P$，通过最小化 $\sum_{(x_i,y_i)\in\tilde{D}^N}\|H(\vec{x}_i;\vec{\theta}_D^N,\vec{\theta}_D^*)-\vec{y}_i\|^2$ 更新 $\vec{\theta}_D^N$ 和 $\vec{\theta}_D^*$
  end for
end for

表 3.3: TEAL 中关系分类采用的特征

词向量	**词向量的范数**
$H(\vec{x}_i;\vec{\theta}_D^P,\vec{\theta}_D^*)-\vec{y}_i$	$\|H(\vec{x}_i;\vec{\theta}_D^P,\vec{\theta}_D^*)-\vec{y}_i\|_1$
	$\|H(\vec{x}_i;\vec{\theta}_D^P,\vec{\theta}_D^*)-\vec{y}_i\|_2$
$H(\vec{x}_i;\vec{\theta}_D^N,\vec{\theta}_D^*)-\vec{y}_i$	$\|H(\vec{x}_i;\vec{\theta}_D^N,\vec{\theta}_D^*)-\vec{y}_i\|_1$
	$\|H(\vec{x}_i;\vec{\theta}_D^N,\vec{\theta}_D^*)-\vec{y}_i\|_2$

我们首先学习其深度共享表示，然后将其映射到上位词和非上位词的词向量上。我们将 $\mathcal{L}_D$ 改写为下式：

$$\mathcal{L}_D=\mathbb{E}_{(x_i,y_i)\sim D^P}\|H(\vec{x}_i;\vec{\theta}_D^P,\vec{\theta}_D^*)-\vec{y}_i\|^2+\mathbb{E}_{(x_i,y_i)\sim D^N}\|H(\vec{x}_i;\vec{\theta}_D^N,\vec{\theta}_D^*)-\vec{y}_i\|^2$$

其中，$\vec{\theta}_D^*$ 为神经网络共享层参数，学习到的上位词和非上位词的词向量可以据此改写为 $H(\vec{x}_i;\vec{\theta}_D^P,\vec{\theta}_D^*)$ 和 $H(\vec{x}_i;\vec{\theta}_D^N,\vec{\theta}_D^*)$。

在这一神经网络中，我们在中间层使用非线性激活函数，在输出层使用线性函数，保证模型输出为连续实数值的词向量。其训练算法参见算法 4。当神经网络训练完毕后，与 FOPM 类似，我们训练 SVM 分类器进行上下位、非上下位关系分类。SVM 分类器使用的特征汇总在表 3.3中，这些特征考虑了神经网络非线性投影的残差，减轻了分布式语义学习中的 “词汇记忆” 问题 [82]。

TEAL 神经网络架构： TEAL 同时在已有分类体系中采样生成的数据集和原有训练集上分别学习神经网络投影模型，并且学习对抗分类器，判断数据来源于分类体系或训练集。这两个神经网络在对抗分类器的作用下相互对抗，从而达到知识迁移、知识融合的效果。我们先给出其神经网络架构示意图（如图 3.4），在下文中详细介绍其细节。

具体而言，TEAL 模型包括了两个子神经网络：基础神经网络和分类体系增强

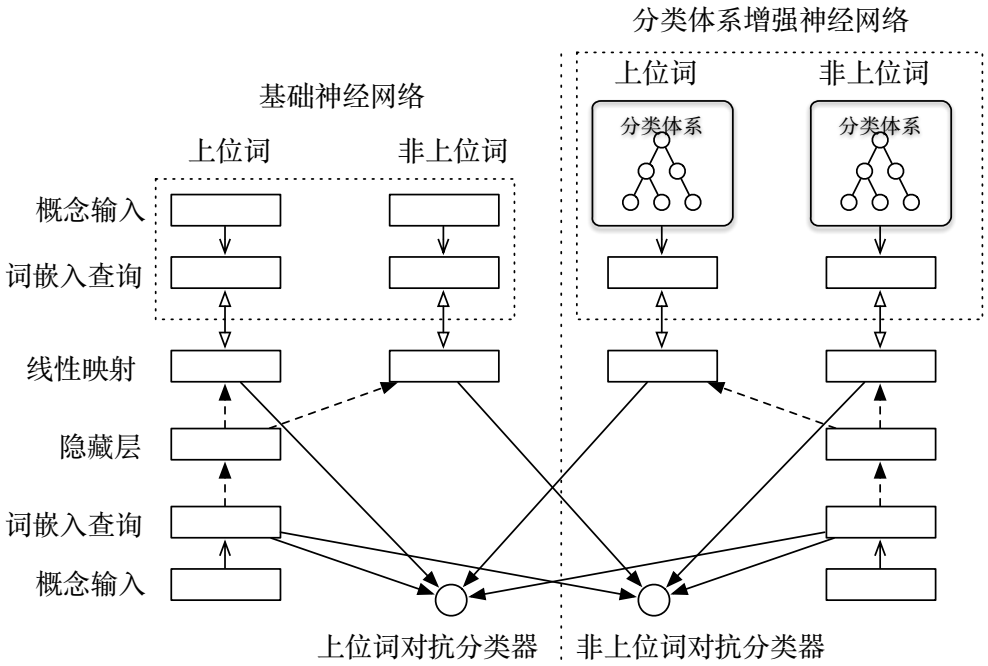


图 3.4: TEAL 的完整神经网络架构

的神经网络。基础神经网络架构在上文已经进行介绍。分类体系增强的神经网络的输入从大规模分类体系中进行采样，具体方法如下：

- 分类体系正例数据集 $T^P = \{(x_i, y_i)\}$：从分类体系中的直接上下位关系 (x_i, y_i)（Direct Hypernymy，即 x_i 和 y_i 在分类体系中有直接边相连）中进行随机采样。我们不考虑上下位关系的传递性，因为自动生成的网络分类体系中传递性不一定成立 [74]。

- 分类体系负例数据集 $T^N = \{(x_i, y_i)\}$：可以使用 T^P 中的反向上下位关系，即 $T^N = \{(y_i, x_i)|(x_i, y_i) \in T^P\}$，也可加入随机匹配的术语对和同下位词关系等。对于一个术语对 (x_i, y_i)，如果存在某术语 z_i 使得 x_i 和 z_i、y_i 和 z_i 之间都有上下位关系，则 (x_i, y_i) 是同下位词关系。

该神经网络的架构与基础神经网络相似。令 $\vec{\theta}_T^P$、$\vec{\theta}_T^N$ 和 $\vec{\theta}_T^*$ 分别为分类体系增强的神经网络关于上下位关系映射、非上下位关系映射和共享层的参数。它的损失函数 $\mathcal{L}_T$ 定义为：

$$\mathcal{L}_T = \mathbb{E}_{(x_i,y_i)\sim T^P} \|H(\vec{x}_i; \vec{\theta}_T^P, \vec{\theta}_T^*) - \vec{y}_i\|^2 + \mathbb{E}_{(x_i,y_i)\sim T^N} \|H(\vec{x}_i; \vec{\theta}_T^N, \vec{\theta}_T^*) - \vec{y}_i\|^2$$

为了使 TEAL 起到对抗学习的效果，我们另外加入了两个对抗分类器。第一个

为上位词对抗分类器，它的输入为两个正例数据集 D^P 和 T^P，目标是给定一个术语 x_i 及其投影的结果，区分其投影后的词向量来自于基础神经网络（即训练集中的知识）或分类体系增强的神经网络（即分类体系中的知识）。它最小化分类器做出错误区分的对数概率，损失函数 $\mathcal{L}_P$ 定义如下：

$$\mathcal{L}_P = \mathbb{E}_{(x_i,y_i)\sim D^P} \log(1 - \delta(H(\vec{x}_i; \vec{\theta}_D^P, \vec{\theta}_D^*), \vec{x}_i)) + \mathbb{E}_{(x_i,y_i)\sim T^P} \log \delta(H(\vec{x}_i; \vec{\theta}_T^P, \vec{\theta}_T^*), \vec{x}_i)$$

其中，$\delta(\vec{y}_i, \vec{x}_i) = \frac{1}{1+e^{-\vec{x}_i \oplus \vec{y}_i}}$ 表示逻辑斯蒂回归分类器，使用 $\vec{x}_i$ 和 $\vec{y}_i$ 的向量拼接作为特征。TEAL 模型的一部分也可以看成是条件对抗生成网络 [157] 的变体，它在输入 $\vec{x}_i$ 的条件下对上位词的词向量进行区分。

类似地，另一个分类器为非上位词对抗分类器，它的输入为两个负例数据集 D^N 和 T^N，同样对生成的非上位词词向量的来源进行区分，损失函数 $\mathcal{L}_N$ 定义为：

$$\mathcal{L}_N = \mathbb{E}_{(x_i,y_i)\sim D^N} \log(1 - \delta(H(\vec{x}_i; \vec{\theta}_D^N, \vec{\theta}_D^*), \vec{x}_i)) + \mathbb{E}_{(x_i,y_i)\sim T^N} \log \delta(H(\vec{x}_i; \vec{\theta}_T^N, \vec{\theta}_T^*), \vec{x}_i)$$

在 TEAL 中，分类体系增强的神经网络在 T^P 和 T^N 上进行训练，基础神经网络在学习给定训练集 D^P 和 D^N 中的知识的同时，也要模拟分类体系增强的神经网络的行为，以误导两个对抗分类器，使他们做出错误的预测。在训练过程中，我们限定基础神经网络只需要学习 T^P 中与 D^P 在语义上足够类似的上下位关系。这是因为不同领域的上下位关系通常有不同的语义表达（参见文献 [25] 和我们提出的 IPM 算法），学习与 D^P 语义无关的上下位关系知识可能会降低基础神经网络的预测准确性。

为了解决这一问题，我们使用语义过滤的方法，从 T^P 中只选择一部分上下位关系元组进行训练，将其记为 $\tilde{T}^P$。我们对 D^P 中关系元组的下位词词向量 $\vec{x}_i$ 进行 K-Means 聚类。当且仅当某一关系元组 $(x_i, y_i) \in T^P$ 的下位词词向量 $\vec{x}_i$ 与 D^P 中的某一簇中心向量 $\vec{c}$ 足够相似（记词向量余弦相似度阈值为 γ），才将其加入 $\tilde{T}^P$。具体流程见算法 5。

我们也对 T^N 采用相同的算法进行过滤，并令领域相关的非上下位关系数据集为 $\tilde{T}^N$。综上，我们给出加入对抗分类器后基础神经网络的完整损失函数 $\mathcal{L}_D^*$：

$$\begin{aligned}\mathcal{L}_D^* = \mathcal{L}_D &+ \lambda_1 \mathbb{E}_{(x_i,y_i)\sim \tilde{T}^P} \log(1 - \delta(H(\vec{x}_i; \vec{\theta}_D^P, \vec{\theta}_D^*), \vec{x}_i)) \\ &+ \lambda_2 \mathbb{E}_{(x_i,y_i)\sim \tilde{T}^N} \log(1 - \delta(H(\vec{x}_i; \vec{\theta}_D^N, \vec{\theta}_D^*), \vec{x}_i))\end{aligned}$$

Algorithm 5 TEAL 的语义过滤算法

初始化 $\tilde{T}^P = \emptyset$
对 $\{\vec{x}_i|(x_i, y_i) \in D^P\}$ 使用 K-Means 算法
for 每个关系元组 $(x_i, y_i) \in T^P$ **do**
  for 每个簇中心 $\vec{c}$ **do**
    if $\cos(\vec{x}_i, \vec{c}) > \gamma$ **then**
      将 (x_i, y_i) 加入 $\tilde{T}^P$
      break
    end if
  end for
end for

Algorithm 6 TEAL 模型训练算法

通过最小化 $\mathcal{L}_T$，初始化参数 $\vec{\theta}_T^P$、$\vec{\theta}_T^N$ 和 $\vec{\theta}_T^*$ 的值
通过最小化 $\mathcal{L}_D$，初始化参数 $\vec{\theta}_D^P$、$\vec{\theta}_D^N$ 和 $\vec{\theta}_D^*$ 的值
while 算法不收敛 **do**
  通过最小化 $\mathcal{L}_P$，训练上下位关系对抗分类器
  通过最小化 $\mathcal{L}_N$，训练非上下位关系对抗分类器
  通过最小化 $\mathcal{L}_T$，训练分类体系增强的神经网络
  通过最小化 $\mathcal{L}_D^*$，训练基础神经网络
end while

其中，λ_1 和 λ_2 为平衡对抗分类器与神经网络损失的超参数。

TEAL 模型的训练算法详见算法 6。这一过程迭代地对四个模型进行训练，直到损失函数 $\mathcal{L}_D^*$ 收敛。当这一神经网络训练完毕之后，我们同样利用基础神经网络计算表 3.3中所列出特征的值，使用 D^P 和 D^N 中的数据训练 SVM 关系分类器。此时所用的特征已经融入了分类体系中上下位关系知识。

3.3.2 实验分析

在本节中，我们在中英文两种语言的分类体系和标注数据集上评测 TEAL 的实验效果，并与基线方法作对比。

数据集和实验设置： 我们使用的英语分类体系来自 Microsoft Concept Graph 知识图谱系统[3]，该系统包含的概念分类体系由 Probase [10] 所构建。在官方开源数据集中，一共包括 33377320 个上下位关系元组，格式为 “（上位词，下位词，计数）” 三元组，“计数” 表示该上下位关系在网络语料库中被抽取出来的次数，部分示例见表 3.4。由于抽取出的上下位关系元组可能包含噪声和错误，我们采用启发

[3]https://concept.research.microsoft.com

表 3.4: Microsoft Concept Graph 数据示例

上位词	下位词	计数
symptom	fatigue	6206
material	aluminum	5518
fruit	strawberry	4824
fish	salmon	3733
organ	kidney	3011
issue	security	2646
medication	aspirin	1998
shape	square	1645
site	youtube	1578
game	chess	1343

式规则进行过滤。我们删除出现次数少于 5 次和包括多词表达的关系元组，最后得到 2844951 个高精度的上下位关系元组，作为分类体系训练数据。我们也将这些术语随机组合形成术语对，如果他们不包含在 Microsoft Concept Graph 的上下位关系集合中，则视为负例。由于 TEAL 需要在大规模分类体系上完成训练，为了保证词向量的高覆盖性，我们采用在 Wikipedia 和 Gigaword 语料库上训练的英语 GloVe 模型 [52] 作为神经语言模型，词向量的维度为 100。在英语的实验中，我们采用上文第2.5.2节采用的数据集进行实验评测，其中包括通用领域的两个数据集 BLESS [93] 和 ENTAILMENT [77]，以及特定领域的三个数据集 ANIMAL、PLANT 和 VEHICLE [97]。具体的实验步骤和相关统计信息同第2.5.2节。

与英语相比，中文语言可用的资源相对较少。在本节的实验中，我们采用 CN-WikiTaxonomy[21] 作为中文分类体系，使用第 2.3.2节中采样得到的分类体系训练集和上一节同样方法得到负例，来学习分类体系增强的神经网络模型。中文神经语言模型也与第 2.3.2节中所用相同，即为维度为 100 的 Skip-Gram 模型 [51]。在评测中，我们利用 FD [25] 和 BK 两个人工标注数据集进行评测，统计信息和评测方法见第2.4.2节。

英语数据集上的实验结果： 在英语实验中，由于基线方法和实验评测标准已在第2.5.2节中介绍，本节中我们只列出现有工作和本文先前工作的最佳结果，以平均准确度为指标，与 TEAL 算法的实验效果相对比。在实验中，我们实现了 TEAL 框架下的两种算法：i）TEAL-S，指的是完全监督式神经网络模型，即只用基础神经网络进行投影学习，不考虑分类体系增强效果；和 ii）TEAL-AS，指的是基于分类体系增强的对抗学习模型。在神经网络中，我们使用有 100 个单元的隐藏层，使用双曲正切函数作为激活函数，Adam [158] 作为神经网络优化算法，运行 500 个

表 3.5: TEAL 在英语上下位关系检测任务中的精确度

方法	BLESS	ENTAILMENT	ANIMAL	PLANT	VEHICLE
现有工作中的强基线算法					
Luu 等人 [81]	0.93	0.91	0.89	0.92	0.89
HyperVec [72]	0.94	0.91	0.83	0.91	0.83
本文先前工作的最佳结果					
FOPM	**0.97**	**0.92**	**0.92**	**0.94**	0.90
TEAL 框架的实验结果					
TEAL-S	0.95	0.87	0.89	0.93	0.91
TEAL-AS	0.96	0.91	**0.92**	**0.94**	**0.93**

Epoch，每批数据量为 64。我们同样调整了 TEAL-AS 中超参数的值，其中正则化超参数设为 $\lambda_1 = \lambda_2 = 0.01$。在基于 SVM 的关系分类模型中，我们利用多项式核函数作为 SVM 核函数。在语义过滤算法中，我们设置阈值 $\gamma = 0.8$，并默认设簇的数量为 10。算法在五个数据集上的实验结果见表 3.5。

从实验结果分析，Luu 等人 [81] 和 HyperVec [72] 的算法是现有研究中的强基线算法，分别在通用和特定领域的数据集上取得较好的表现。与之相比，我们在第二章中提出的 FOPM 算法在通用和领域数据集上均有提升。本节提出的 TEAL 框架在四个数据集（除了 ENTAILMENT）上都超过了基线算法，并与 FOPM 算法的性能类似。在现有的研究中，对抗学习主要被应用于通过反向传播训练的神经网络上，由于 FORM 算法的特征基于 SVD 求解，很难直接运用对抗学习。因此，如何将分类体系中的海量知识应用于 FOPM 算法中，进一步提升模型的精度，是一个有趣的研究方向。

通过对比分析，我们也发现 TEAL-AS 比 TEAL-S 在所有实验上均有优势，在准确度上提升了 1% 到 4% 不等。与 FOPM 比较，TEAL-AS 在三个领域数据集上优势较大，因为这三个特定领域的数据集比较小，很难训练精确的投影模型，例如 FOPM；TEAL-AS 则可以利用 Microsoft Concept Graph 中的额外知识，辅助模型的预测。

中文数据集上的实验结果： 在中文实验中，我们采取与英语实验相同的步骤，在 FD [25] 和 BK 两个数据集上进行实验，实验结果参见表 3.6。从实验结果可以观察到，TEAL-S 和 TPM 的实验效果相似，在数据集 FD 上 TPM 的 F 值较高，TEAL-S 在数据集 BK 上提升了 0.6% 的 F 值。在 TEAL 框架中加入了中文分类体系 CN-WikiTaxonomy[21] 中的知识后，F 值在 TEAL-S 的基础上分别提升了 3.9% 和 0.2%。由上述结果可见，尽管中文语境下公开可用的数据集较少，当使用了 TEAL

表 3.6: TEAL 在中文上下位关系分类任务中的实验结果

数据集	FD			BK		
方法	**精准度**	**召回率**	**F 值**	**精准度**	**召回率**	**F 值**
现有工作中的强基线算法						
$\vec{x}_i \oplus \vec{y}_i$	0.677	**0.752**	0.697	0.803	0.759	0.780
本文先前工作的最佳结果						
TPM	**0.728**	0.705	0.716	**0.836**	0.806	0.821
TEAL 框架的实验结果						
TEAL-S	0.695	0.684	0.689	0.788	0.869	0.827
TEAL-AS	0.721	0.736	**0.728**	0.791	0.870	**0.829**

表 3.7: Microsoft Concept Graph 中新发现的上下位关系准确率检测结果

上位词	**# 正确/# 总数**	**准确率**	**上位词**	**# 正确/# 总数**	**准确率**
material	78/102	0.76	goods	20/20	1.00
person	17/19	0.89	sector	18/20	0.90
group	37/43	0.86	component	76/80	0.95
technology	12/14	0.86	individual	24/24	1.00
provision	15/15	1.00	location	8/9	0.89
合计	302/346	0.87			

框架的对抗学习技术，模型在中文上下位/非上下位关系的预测精度得到了提升。

对 Microsoft Concept Graph 的扩展： 在下文中，我们对 TEAL 算法进行轻微改动，利用 Microsoft Concept Graph 中原有的上下位关系知识，对它进行补全。我们概述分类体系补全的算法。首先，我们利用 T^P 和 T^N 训练 TEAL 中的分类体系增强的神经网络（即最小化目标函数 $\mathcal{L}_T$），不考虑基础神经网络的存在。因为基于 SVM 的分类器很难对一个术语对进行基于置信度的上下位关系评分，我们将神经网络和 TPM 算法结合起来，对于任意术语对 (x_i, y_i)，定义如下评分 $score(x_i, y_i)$：

$$score(x_i, y_i) = \tanh(\|H(\vec{x}_i; \vec{\theta}_T^N, \vec{\theta}_T^*) - \vec{y}_i\|_2 - \|H(\vec{x}_i; \vec{\theta}_T^P, \vec{\theta}_T^*) - \vec{y}_i\|_2)$$

其中，$score(x_i, y_i)$ 越高表示 x_i 和 y_i 之间有上下位关系的概率越大。令 $\alpha_1 > 0$ 是相似度阈值，$\alpha_2 > 0$ 是高置信度评分阈值。对于 Microsoft Concept Graph 中每个上下位关系元组 $(x_i, y_i) \in T^P$，我们通过神经语言模型得到 x_i 在词嵌入空间的邻居集合 $N(x_i) = \{x' \,|\, \cos(\vec{x'}, \vec{x_i}) > \alpha_1\}$。如果 $(x_i, y_i) \in T^P$，$x' \in N(x_i)$ 且 $score(x', y_i) > \alpha_2$，我们预测 x' 和 y_i 之间也有上下位关系。举例来说，若在 Microsoft Concept Graph 中我们知道狗是一种动物，我们也可以推断猫是一种动物，并将其加入系统中。为了保证新加入上下位关系的高精度，我们设 $\alpha_1 = \alpha_2 = 0.8$。

表 3.8: Microsoft Concept Graph 中新发现的上下位关系与其评分示例，预测错误的元组粗体显示

上位词	上位词	评分	上位词	上位词	评分
petrol	provision	0.908	wildfires	threat	0.845
handicrafts	business	0.872	**steroids**	**alternative**	**0.813**
pantsuit	product	0.870	psychiatrist	profession	0.808
bacteria	**measure**	**0.864**	tarragon	food	0.808

在表 3.7中, 我们展示了 Microsoft Concept Graph 中十个上位词的上下位关系扩展结果, “# 正确” 和 “# 总数” 分别表示抽取出的正确及总共的上下位关系数量。对于每个上位词，我们将所有新扩展的上下位关系交给人工标注员进行正确性标注。从实验结果可见，TEAL 预测现有分类体系的新关系整体准确率较高，平均准确率达到了 87%。我们也给出了 8 个具体的预测示例，及模型预测评分值，见表 3.8。从结果中我们可以发现一部分错误来自于模型自身的预测错误，例如 “bacteria”（细菌）和 “measure”（度量）之间无上下位关系。第二类错误主要来源于上位词和下位词的语义不完整性，例如 “steroids”（类固醇）可以作为很多疾病的 “alternative”（即替代治疗方法），但是这两个词本身之间没有严格的上下位关系。在未来的研究工作中，我们的方法也可以与其他数据驱动的方法（例如 [101]）结合，提升现有分类体系的覆盖率。

3.4 基于迁移学习的跨语言上下位关系抽取

前述 TEAL 框架主要用于单语言、跨知识源的上下位关系抽取, 它的一个缺点是异种知识源的关系元组必须来自于同一语言。由于英语上下位关系训练数据量较大, 其他语种（特别是小语种）的训练数量量较小, 很难从中训练到准确率足够高的模型，本节扩展了单语言环境下的 FOPM 算法，介绍**迁移模糊正交投影模型**（TFOPM）和**迭代迁移模糊正交投影模型**（ITFOPM）两种算法，用于跨语言上下位关系抽取。

3.4.1 算法模型

我们首先回顾跨语言上下位关系抽取的任务目标，其次分别介绍 TFOPM 和 ITFOPM 的算法细节。

跨语言上下位关系抽取： 记 D_S^P、D_S^N、D_T^P 和 D_T^N 分别为来自源语言和目标语言的、上下位关系和非上下位关系训练集，其中下标 S 代表源语言，下标 T 代

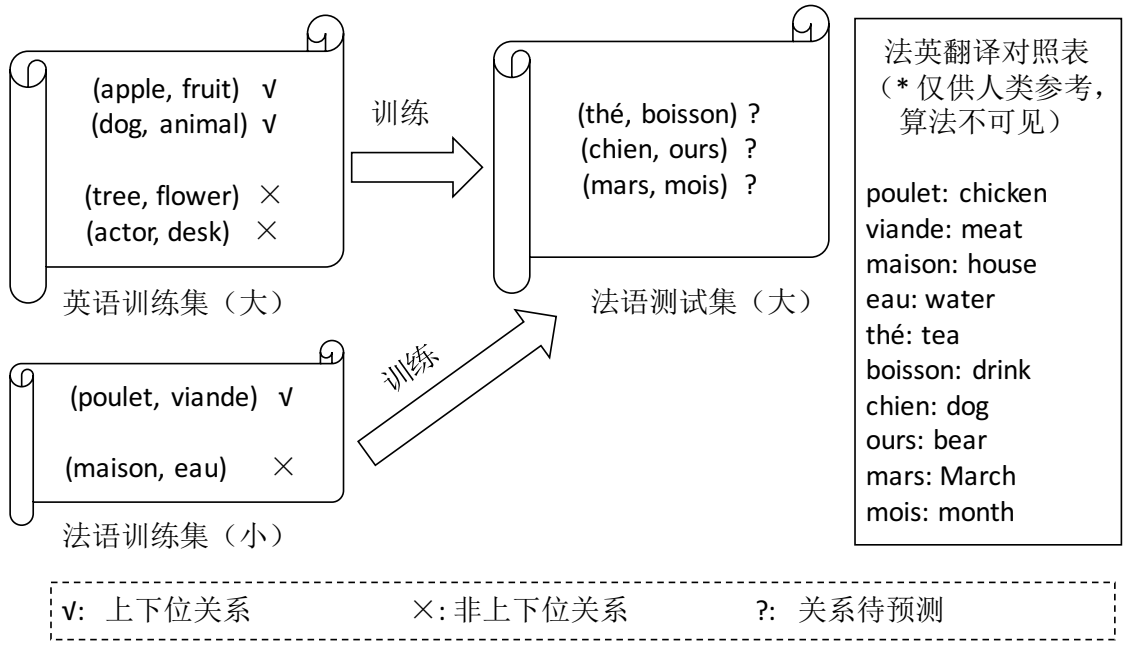


图 3.5: 跨语言上下位关系抽取任务示意

表目标语言。由于在本研究中，我们的目标是小语种的上下位关系预测，我们通常有如下限制条件：

$$|D_S^P| \gg |D_T^P|$$

$$|D_S^N| \gg |D_T^N|$$

我们同时令 $U_T = \{(x_i, y_i)\}$ 为目标语言术语对测试集，一般地，我们也考虑如下限制条件：

$$|U_T| \gg |D_T^P|$$

$$|U_T| \gg |D_T^N|$$

跨语言上下位关系抽取的目标是利用四个训练集 D_S^P、D_S^N、D_T^P 和 D_T^N，训练区分目标语言的上下位和非上下位关系的分类器 f。在测试阶段，我们利用分类器 f 直接在目标语言测试集 U_T 进行关系分类。以英语-法语对为例，图 3.5给出了这一任务的简单示意图。

近年来，神经机器翻译技术发展迅速（例如 [159]）；然而基于句子的机器翻译技术不能有效地解决跨语言上下位关系抽取问题。这是因为为了保证高精度，神经机器翻译模型的训练需要大量高质量的双语平行语料，在我们的小语种任务中，很难甚至不可能获得需要的数据 [160]。神经机器翻译主要着眼于翻译完整的句子，

而在我们的任务中，我们关注的是不包括完整上下文的术语对。

在我们的工作中，我们采用 Lample 等人**双语术语对齐**（Bilingual Lexicon Induction）的工作 [102] 解决这个问题。在训练 TFOPM 和 ITFOPM 之前，我们采用这一算法，在源语言和目标语言的未标注语料库上训练双语词向量对齐模型。令 $\mathbf{S}$ 为 $|\vec{x}_i| \times |\vec{x}_i|$ 的投影矩阵，将源语言空间中的词 x_i 的词向量 $\vec{x}_i$ 映射至它目标语言下的同义词 x_i^* 的词向量 $\vec{x}_i^*$。

上下位关系投影学习： 与单语环境下的 FOPM 类似，我们利用 K-Means 算法，将 D_S^P 和 D_T^P 中的术语对混合聚成 K 个簇。如果 $(x_i, y_i) \in D_T^P$，我们使用与 FOPM 中相同的特征 $\vec{x}_i - \vec{y}_i$；如果 $(x_i, y_i) \in D_P^S$，我们使用的特征为 $\mathbf{S}\vec{x}_i - \mathbf{S}\vec{y}_i$，从而在映射后的目标语言空间完成聚类。令第 k 个簇的中心向量为 $\vec{c}_k^{\mathcal{P}}$。由此可见，所有训练数据都在目标语言的词嵌入空间中进行计算。

我们同样令 $\mathbf{M}_k^P$ 为第 k 个簇的上下位关系投影矩阵，$a_{i,k}^P$ 为上下位关系元组 $(x_i, y_i) \in D_S^P \cup D_T^P$ 的权重，计算为：

$$a_{i,k}^P = \begin{cases} \frac{\cos(\mathbf{S}\vec{x}_i - \mathbf{S}\vec{y}_i, \vec{c}_k^P)}{\sum_{(x,y)\in D_S^P} \cos(\mathbf{S}\vec{x} - \mathbf{S}\vec{y}, \vec{c}_k^P)}, & (x_i, y_i) \in D_S^P \\ \frac{\cos(\vec{x}_i - \vec{y}_i, \vec{c}_k^P)}{\sum_{(x,y)\in D_T^P} \cos(\vec{x} - \vec{y}, \vec{c}_k^P)}, & (x_i, y_i) \in D_T^P \end{cases}$$

第 k 个簇对应的目标函数定义如下：

$$\begin{aligned} J(\mathbf{M}_k^P) =& \frac{\beta}{2} \sum_{(x_i,y_i)\in D_S^P} a_{i,k}^P \gamma_{i,k}^P \|\mathbf{M}_k^P \cdot \mathbf{S}\vec{x}_i - \mathbf{S}\vec{y}_i\|^2 \\ &+ \frac{1-\beta}{2} \sum_{(x_i,y_i)\in D_T^P} a_{i,k}^P \|\mathbf{M}_k^P \vec{x}_i - \vec{y}_i\|^2 \\ \text{s. t. } &(\mathbf{M}_k^{P^T}) \cdot \mathbf{M}_k^P = \mathbf{I}, \sum_{(x_i,y_i)\in D_S^P} a_{i,k}^P \gamma_{i,k}^P = 1, \sum_{(x_i,y_i)\in D_T^P} a_{i,k}^P = 1 \end{aligned} \tag{3.1}$$

其中，$\gamma_{i,k}^P$ 是源语言中上下位关系元组的权重因子，用于衡量这一元组对于目标语言的重要性。$\beta \in (0, 1)$ 是预定义的平衡参数，用于给源语言和目标语言投影损失不同的权重。

为了计算权重 $\gamma_{i,k}^P$ 的值，我们首先定义未归一化权重 $\tilde{\gamma}_i^P$：

$$\tilde{\gamma}_i^P = \cos(\mathbf{S}\vec{x}_i - \mathbf{S}\vec{y}_i, \frac{1}{|D_T^P|} \sum_{(x_j,y_j)\in D_T^P} \vec{x}_j - \vec{y}_j)$$

它计算某一源语言上下位关系元组 $(x_i, y_i) \in D_S^P$ 与所有目标语言的上下位关系元组在目标语言词嵌入空间的语义相似性。因为我们的优化目标必须满足如下限制条件：$\sum_{(x_i,y_i)\in D_S^P} a_{i,k}^P \gamma_{i,k}^P = 1$，$\gamma_{i,k}^P$ 采取如下方式归一化：

$$\gamma_{i,k}^P = \frac{\tilde{\gamma}_i^P}{\sum_{(x_j,y_j)\in D_S^P} a_{j,k}^P \gamma_{i,k}^P}$$

当所有权重计算完毕后，我们通过最小化 $J(\mathbf{M}_k^P)$ 来计算正交投影矩阵 $\mathbf{M}_k^P$。与 FOPM 类似，可以利用基于 SVD 的方法计算 $\mathbf{M}_k^P$ 的最优值，式 3.1可以通过下述定理求解：

定理 3.4.1. *式 3.1为高维 Wahba 问题的一种变体，具有如下闭式解:*

1. $\mathbf{B}_k^P = \beta \sum_{(x_i,y_i)\in D_S^P} a_{i,k}^P \gamma_{i,k}^P \cdot (\mathbf{S}\vec{y}_i) \cdot (\mathbf{S}\vec{x}_i)^T + (1-\beta) \sum_{(x_i,y_i)\in D_T^P} a_{i,k}^P \vec{y}_i \cdot \vec{x}_i^T$
2. $\mathbf{U}_k^P \mathbf{S}_k^P (\mathbf{V}_k^P)^T = SVD(\mathbf{B}_k^P)$
3. $\mathbf{R}_k^P = diag(\underbrace{1, \ldots, 1}_{|\vec{x}_i|-1}, det(\mathbf{U}_k^P) \cdot det(\mathbf{V}_k^P))$
4. $\mathbf{M}_k^P = \mathbf{U}_k^P \mathbf{R}_k^P (\mathbf{V}_k^P)^T$

Proof. 为了简单起见，我们省略了式 3.1中的所有变量的上标 P 和下标 k，式 3.1可以简写为：

$$\begin{aligned} J(\mathbf{M}) =& \frac{\beta}{2} \sum_{(x_i,y_i)\in D_S} a_i \gamma_i \|\mathbf{M}\mathbf{S}\vec{x}_i - \mathbf{S}\vec{y}_i\|^2 + \frac{1-\beta}{2} \sum_{(x_i,y_i)\in D_T^P} a_i \|\mathbf{M}\vec{x}_i - \vec{y}_i\|^2 \\ \text{s. t. } & \mathbf{M}^T\mathbf{M} = \mathbf{I}, \sum_{i=1}^{|D_S|} a_i \gamma_i = 1, \sum_{i=1}^{|D_T|} a_i = 1 \end{aligned}$$

根据该优化问题的约束条件，我们有：

$$\beta \sum_{(x_i,y_i)\in D_S} a_i \gamma_i + (1-\beta) \sum_{(x_i,y_i)\in D_T} a_i = 1$$

所以，每个源语言的元组 $(x_i, y_i) \in D_S$ 有权重 $\beta a_i \gamma_i$；每个目标语言的元组 $(x_i, y_i) \in D_T$ 有权重 $(1-\beta)a_i$。我们定义矩阵 $\mathbf{B}$ 如下所示：

$$\mathbf{B} = \beta \sum_{(x_i,y_i)\in D_S} a_i \gamma_i \mathbf{S}\vec{y}_i (\mathbf{S}\vec{x}_i)^T + (1-\beta) \sum_{(x_i,y_i)\in D_T} a_i \vec{y}_i \vec{x}_i^T$$

利用矩阵 **B**，目标函数 $J(\mathbf{M})$ 可以改写为如下形式：

$$J(\mathbf{M}) = 1 - \text{tr}(\mathbf{M}\mathbf{B}^T)$$

由此，我们将式 3.1的问题转化为高维 Wahba 问题。证明的剩余部分与原始高维 Wahba 问题闭式解的证明相同，此处省略。 □

将 K 个簇对应的目标函数进行加和，我们得到了跨语言情况下上下位关系投影的目标函数 $\tilde{J}(\mathcal{M}^P)$：

$$\begin{aligned}
\tilde{J}(\mathcal{M}^P) =& \frac{\beta}{2} \sum_{(x_i,y_i)\in D_S^P} \sum_{k=1}^{K} a_{i,k}^P \gamma_{i,k}^P \|\mathbf{M}_k^P \cdot \mathbf{S}\vec{x}_i - \mathbf{S}\vec{y}_i\|^2 \\
&+ \frac{1-\beta}{2} \sum_{(x_i,y_i)\in D_T^P} \sum_{k=1}^{K} a_{i,k}^P \|\mathbf{M}_k^P \vec{x}_i - \vec{y}_i\|^2 \\
&\text{s. t. } (\mathbf{M}_k^{PT}) \cdot \mathbf{M}_k^P = \mathbf{I}, \sum_{(x_i,y_i)\in D_S^P} a_{i,k}^P \gamma_{i,k}^P = 1, \sum_{(x_i,y_i)\in D_T^P} a_{i,k}^P = 1 \\
&k = 1, \ldots, K
\end{aligned}$$

其中，$\mathcal{M}^P$ 是 K 个正交投影矩阵的集合。我们将上述投影模型称为**迁移模糊正交投影**（TFOPM）。

非上下位关系投影学习： 相似地，为了学习非上下位关系的投影矩阵，我们将 D_S^N 和 D_T^N 中的非上下位关系元组聚成 K 个簇，分别用 $\mathbf{S}\vec{x}_i - \mathbf{S}\vec{y}_i$ 和 $\vec{x}_i - \vec{y}_i$ 作为特征。设 $\mathcal{M}^N$ 是 K 个非上下位关系的正交投影矩阵集合。非上下位关系投影学习的目标函数 $\tilde{J}(\mathcal{M}^N)$ 定义为：

$$\begin{aligned}
\tilde{J}(\mathcal{M}^N) =& \frac{\beta}{2} \sum_{(x_i,y_i)\in D_S^N} \sum_{k=1}^{K} a_{i,k}^N \gamma_{i,k}^N \|\mathbf{M}_k^N \cdot \mathbf{S}\vec{x}_i - \mathbf{S}\vec{y}_i\|^2 \\
&+ \frac{1-\beta}{2} \sum_{(x_i,y_i)\in D_T^N} \sum_{k=1}^{K} a_{i,k}^N \|\mathbf{M}_k^N \vec{x}_i - \vec{y}_i\|^2 \\
&\text{s. t. } (\mathbf{M}_k^{NT}) \cdot \mathbf{M}_k^N = \mathbf{I}, \sum_{(x_i,y_i)\in D_S^N} a_{i,k}^N \gamma_{i,k}^N = 1, \sum_{(x_i,y_i)\in D_T^N} a_{i,k}^N = 1 \\
&k = 1, \ldots, K
\end{aligned}$$

其中，$\mathbf{M}_k^N$、$a_{i,k}^N$ 和 $\gamma_{i,k}^N$ 分别为非上下位关系投影学习的投影矩阵和对应权重。他们

Algorithm 7 TFOPM 训练算法

1: 对 D_S^P 和 D_T^P 中的关系元组进行 K-Means 聚类，特征分别采用 $\mathbf{S}\vec{x}_i - \mathbf{S}\vec{y}_i$ 和 $\vec{x}_i - \vec{y}_i$
2: 对 D_S^N 和 D_T^N 中的关系元组进行 K-Means 聚类，特征分别采用 $\mathbf{S}\vec{x}_i - \mathbf{S}\vec{y}_i$ 和 $\vec{x}_i - \vec{y}_i$
3: **for** $k = 1$ to 簇的数量 K **do**
4: 　利用基于 SVD 的闭式解，学习投影矩阵 $\mathbf{M}_k^P$ 和 $\mathbf{M}_k^N$
5: **end for**
6: **for** 每个关系元组 $(x_i, y_i) \in D_S^P \cup D_S^N \cup D_T^P \cup D_T^N$ **do**
7: 　计算特征 $\mathcal{F}^P(\vec{x}_i, \vec{y}_i)$ 和 $\mathcal{F}^N(\vec{x}_i, \vec{y}_i)$
8: **end for**
9: 在数据集 D_S^P、D_S^N、D_T^P 和 D_T^N 上训练神经网络关系分类器 f

的计算方式与 $\mathbf{M}_k^P$、$a_{i,k}^P$ 和 $\gamma_{i,k}^P$ 相同，不再赘述。

关系分类器训练： 当 $\mathcal{M}^P$ 和 $\mathcal{M}^N$ 这 $2K$ 个矩阵的值计算得到后，我们训练跨语言的关系分类器 f。对于源语言和目标正负训练集的任一关系元组 $(x_i, y_i) \in D_S^P \cup D_S^N \cup D_T^P \cup D_T^N$，与 FOPM 相同，我们计算得上下位和非上下位关系投影残差（即 $\mathcal{F}^P(\vec{x}_i, \vec{y}_i)$ 和 $\mathcal{F}^N(\vec{x}_i, \vec{y}_i)$）作为特征，定义如下：

$$\mathcal{F}^P(\vec{x}_i, \vec{y}_i) = \begin{cases} (\mathbf{M}_1^P\mathbf{S}\vec{x}_i - \mathbf{S}\vec{y}_i) \oplus \cdots \oplus (\mathbf{M}_K^P\mathbf{S}\vec{x}_i - \mathbf{S}\vec{y}_i), & (x_i, y_i) \in D_S^P \cup D_S^N \\ (\mathbf{M}_1^P\vec{x}_i - \vec{y}_i) \oplus \cdots \oplus (\mathbf{M}_K^P\vec{x}_i - \vec{y}_i), & (x_i, y_i) \in D_T^P \cup D_T^N \end{cases}$$

$$\mathcal{F}^N(\vec{x}_i, \vec{y}_i) = \begin{cases} (\mathbf{M}_1^N\mathbf{S}\vec{x}_i - \mathbf{S}\vec{y}_i) \oplus \cdots \oplus (\mathbf{M}_K^N\mathbf{S}\vec{x}_i - \mathbf{S}\vec{y}_i), & (x_i, y_i) \in D_S^P \cup D_S^N \\ (\mathbf{M}_1^N\vec{x}_i - \vec{y}_i) \oplus \cdots \oplus (\mathbf{M}_K^N\vec{x}_i - \vec{y}_i), & (x_i, y_i) \in D_T^P \cup D_T^N \end{cases}$$

我们在特征 $\mathcal{F}^P(\vec{x}_i, \vec{y}_i) \oplus \mathcal{F}^N(\vec{x}_i, \vec{y}_i)$ 上训练神经网络分类器 f，其架构同 FOPM。TFOPM 的算法流程见算法 7。

迭代迁移模糊正交投影模型： 由于目标语言的训练集大小非常有限，TFOPM 只能从目标语言的训练数据中学到很少的语义知识。因为不同的语言有各自文化特有的词，他们的语义往往很难进行跨语言迁移。在这种情况下，跨语言上下位关系预测的精度会下降。

为了进一步提升跨语言上下位关系预测的精度，我们扩展 TFOPM，提出**迭代迁移模糊正交投影模型**（ITFOPM）。如上文所述，$U_T = \{(x_i, y_i)\}$ 为目标语言关系待预测的术语对。与 IPM 类似，ITFOPM 算法迭代地训练 $\mathcal{M}^P$ 和 $\mathcal{M}^N$，然后从 U_T 选出高置信度的术语对，并加入训练集进行下一迭代的训练。特别地，对于每个术

Algorithm 8 ITFOPM 训练算法

利用算法 7，在数据集 D_S^P、D_S^N、D_T^P 和 D_T^N 上训练 TFOPM
while 算法不收敛 **do**
  for 每一个目标语言术语对 $(x_i, y_i) \in U_T$ **do**
    if $conf(x_i, y_i) > \tau$ **then**
      if 分类器 f 预测 (x_i, y_i) 为上下位关系 **then**
        更新 $D_T^P = D_T^P \cup \{(x_i, y_i)\}$
      else
        更新 $D_T^N = D_T^N \cup \{(x_i, y_i)\}$
      end if
      更新 $U_T = U_T \setminus \{(x_i, y_i)\}$
    end if
  end for
  利用算法 7，在数据集 D_S^P、D_S^N、D_T^P 和 D_T^N 上更新 TFOPM
end while

语对 $(x_i, y_i) \in U_T$，我们计算其预测置信度 $conf(x_i, y_i)$ 如下：

$$conf(x_i, y_i) = \frac{|\|\mathcal{F}^P(\vec{x}_i, \vec{y}_i)\|_2 - \|\mathcal{F}^N(\vec{x}_i, \vec{y}_i)\|_2|}{\max\{\|\mathcal{F}^P(\vec{x}_i, \vec{y}_i)\|_2, \|\mathcal{F}^N(\vec{x}_i, \vec{y}_i)\|_2\}}$$

在本方法中, 我们采用基于置信度的方法, 而非依靠神经网络预测, 这是因为现代神经网络一般不能生成校准的概率分布（Calibrated Probabilistic Distribution）[161]。给定阈值 τ，如果 $conf(x_i, y_i) > \tau$，我们将这一元组加入训练集（D_T^P 或 D_T^N，取决于预测的关系标签）。ITFOPM 迭代地在不断增大的训练集上完成训练，直到其在验证集上的预测精度不再提高为止。ITFOPM 算法流程如算法 8所示。

3.4.2 实验分析

在本节中，我们在多个小语种上展开实验，综合评测 TFOPM 和 ITFOPM 这两个跨语言模型的准确性。

数据集与实验设置： 我们选择英语作为源语言，因为英语在全世界的使用最为广泛，而且有很多上下位关系预测的训练集可供使用。我们将五个人工标注的英语数据集合并起来，包括 BLESS [93]、Shwartz [54]、Kotlerman [162]、Turney [79] 和 ENTAILMENT [77]。在合并的数据集中，我们去除了重复项和多词表达式，一共得到 85234 英语术语对，包括 17394 个上下位关系元组和 67930 非上下位关系元组（混合了各种语义关系）。对于目标语言，我们利用 Open Multilingual Wordnet

表 3.9: 7 个目标语言上下位与非上下位关系数据集的统计信息

关系↓语言 →	fr	zh	ja	it	th	fi	el
# 上下位关系	4,035	2,962	1,448	3,034	1,156	7,157	2,612
# 非上下位关系	8,947	6,382	3,203	6,081	1,977	9,433	1,454

计划 [163] 生成所需的训练集和测试集[4]。目标语言从 Open Multilingual Wordnet 支持的语言中选出，一共包括 7 种，分别为法语（fr）、中文（zh）、日语（ja）、意大利语（it）、泰语（th）、芬兰语（fi）和希腊语（el）。我们采用的 Wordnet 版本分别为：Wordnet Libre du Français（法语）、Chinese Open Wordnet（中文）、Japanese Wordnet（日语）、ItalWordnet（意大利语）、Thai Wordnet（泰语）、FinnWordnet（芬兰语）和 Greek Wordnet（希腊语）。目标语言的上下位关系元组从对应的 Wordnet 中的概念层次类别中随机采样得到，非上下位关系元组合并了对应的 Wordnet 中的其他多种关系（包括整体-部分关系、同义词关系等）。生成的 7 个目标语言数据集的统计信息见表 3.9。所有非英语数据集已在 GitHub 上开源[5]。

我们采用多语言维基百科语料库训练所有涉及到的 8 种语言的 fastText 词嵌入模型 [53]，并且使用 Lample 等人 [102] 开源的算法学习跨语言的映射矩阵，参数设为原始论文的默认设置。在本组实验中，词向量的维度统一设为 300。

评测方法： 我们在两个任务上评测两个跨语言模型 TFOPM 和 ITFOPM 的效果。第一个任务为**跨语言上下位关系方向分类**（Cross-lingual Hypernymy Direction Classification），它的目的在于预测目标语言的上下位关系元组中，哪一个词为上位词。在这组实验中，我们使用数据集中的上下位关系作为正例，反向上下位关系（Reverse-hypernymy）作为负例来评测我们的模型。第二个任务为**跨语言上下位关系检测**（Cross-lingual Hypernymy Detection），它的目的是对目标语言中的上下位关系和非上下位关系进行分类。在实验中，我们使用所有英语数据作为源语言训练数据集，在目标语言上进行 5 折交叉验证。我们使用 5 折的平均准确率作为评测指标，比较所有算法的有效性。

因为基于模式匹配和一部分分布式的上下位关系预测方法是和语言本身的特性高度相关的，他们不适合用于面向任何目标语言的跨语言上下位关系的预测。我们采用 Shwartz 等人的评测过程 [54]，并且采用以下分布式方法作为基线方法：

- Santus 等人 [70]：它是基于信息熵的上下位关系度量 SLQS。

[4] http://compling.hss.ntu.edu.sg/omw/

[5] https://chywang.github.io/data/www2019.zip

- Kiela 等人 [164]：它是基于分布式通用性的关系度量，可以建模上下位关系的层次性质。
- Weeds 等人 [78]：它是监督式的上下位关系分类模型。
- Shwartz 等人 [54]：它采用混合神经网络用于关系分类。由于部分目标语言（例如中文、泰语等）缺乏高质量上下位关系模式，在这些语言中基于模式的子神经网络被省略。
- TFOPM-N 和 ITFOPM-N：他们分别是 TFOPM 和 ITFOPM 的变体，在原始投影矩阵学习的基础上去除了正交性约束。

除了提出的 TFOPM 和 ITFOPM 两个模型及其变体 TFOPM-N 和 ITFOPM-N，其他方法 [54, 70, 78, 164] 并不能直接利用源语言数据训练模型。为了使实验对比更加公平，我们同样使用 Lample 等人 [102] 的模型将源语言数据的词向量映射到目标语言空间，与目标语言训练数据协同学习模型参数。

实验结果： 我们首先利用交叉验证，调整提出 TFOPM 和 ITFOPM 的参数值。其默认值设为 $K = 4$ 和 $\beta = 0.5$。在表 3.10中，我们汇总了所有方法在两个任务和 7 个目标语言的实验结果。从实验结果中，我们可以得出三个结论：i）对于跨语言上下位关系方向分类和关系检测两个任务，TFOPM 和 ITFOPM 两个模型在 7 种目标语言上都超过了所有基线方法。从整体结果而言，我们提出了模型在跨语言上下位关系方向分类任务上提升了 2% 到 9% 的准确度，不同语言的提升程度略有不同。我们的模型在跨语言上下位关系检测任务上的提升程度与前述任务相似。ii）通过利用无监督式双语字典生成技术和正交性投影约束条件，源语言的上下位关系知识可以逐步地迁移到目标语言上，提升小语种的预测精度。iii）模型在跨语言上下位关系方向分类上的预测精度比跨语言上下位关系检测高，说明了后一项任务的难度更大。

我们进一步分析 ITFOPM 是怎样提升模型学习能力的。首先，我们固定阈值 $\tau = 0.7$，在两个任务上都执行迭代训练 8 次，并且记录每个迭代的模型效果。实验结果汇总在图 3.6中。由此可知，在开始的几个迭代，准确度是不断上升的；之后，准确度逐渐变得稳定，这是因为没有足够多的新的目标语言数据可以被加入到训练集中。在我们的实验中，算法运行 5 个迭代效果较好。总体上，采用迭代学习的技术，准确度可以提升至少 2%。

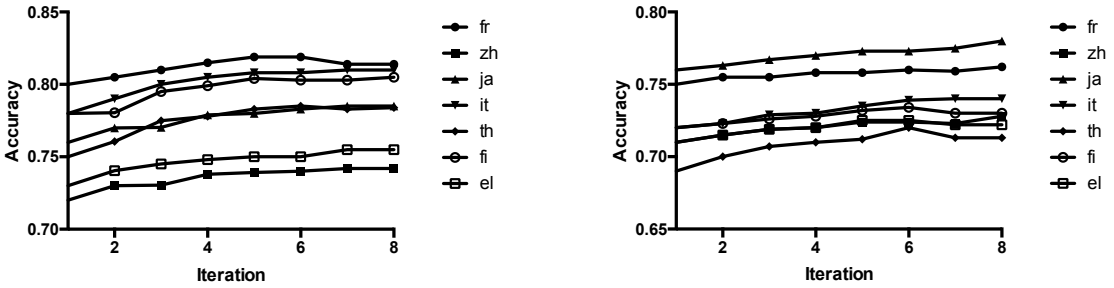
表 3.10: 不同方法在两个跨语言上下位关系预测任务中的精确度比较

方法	fr	zh	ja	it	th	fi	el
任务：跨语言上下位关系方向分类							
Santus 等人 [70]	0.65	0.65	0.68	0.61	0.63	0.70	0.62
Weeds 等人 [78]	0.76	0.71	0.77	0.76	0.72	0.77	0.70
Kiela 等人 [164]	0.67	0.65	0.71	0.68	0.65	0.70	0.62
Shwartz 等人 [54]	0.79	0.67	0.71	0.72	0.66	0.75	0.66
TFOPM-N	0.78	0.71	0.75	0.76	0.73	0.76	0.71
TFOPM	0.80	0.72	0.76	0.78	0.75	0.78	0.73
ITFOPM-N	**0.82**	0.72	0.76	0.78	0.75	**0.81**	0.72
ITFOPM	0.81	**0.74**	**0.78**	**0.81**	**0.78**	**0.81**	**0.75**
任务：跨语言上下位关系检测							
Santus 等人 [70]	0.67	0.63	0.67	0.62	0.64	0.62	0.64
Weeds 等人 [78]	0.74	0.66	0.68	0.71	0.62	0.68	0.69
Kiela 等人 [164]	0.70	0.61	0.65	0.68	0.57	0.61	0.67
Shwartz 等人 [54]	0.72	0.66	0.69	0.64	0.66	0.69	0.70
TFOPM-N	0.72	0.67	0.70	0.70	0.68	0.71	0.70
TFOPM	0.75	0.71	0.76	0.72	0.69	0.72	0.71
ITFOPM-N	0.72	**0.74**	0.77	**0.74**	0.67	0.71	0.72
ITFOPM	**0.76**	0.73	**0.78**	**0.74**	**0.72**	**0.73**	**0.73**

接下来，我们调节 ITFOPM 中阈值 τ 的值，运行 5 个迭代，并汇总当 τ 取不同的值是的模型效果，详见图 3.7。实现效果说明，τ 的调整在不同语言数据集上的表现是类似的。τ 的选择反映了半监督学习过程中，加入训练集的关系元组数量与关系元组的准确性之间的权衡。当 τ 较小时，算法倾向于将更多的未标注关系元组加入训练集，这会导致预测误差被引入训练集；反之，虽然关系元组的准确性提高了，但是迭代学习的客观效果被削弱。实验结果说明当 τ 设为 0.7 左右比较合适。

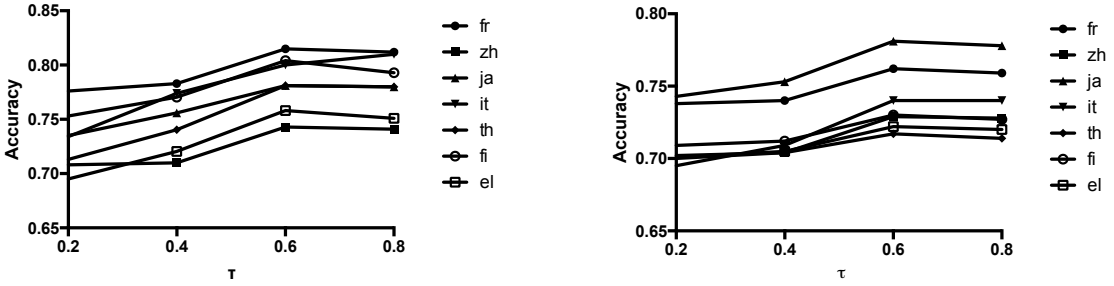
3.5 基于超球学习的词汇关系分类

前述章节介绍了在不同场景下，上下位关系和非上下位关系是如何在词嵌入空间区分的，这些方法的主要部分是建模上下位关系的语义。然而，非上下位关系包括同义词关系、反义词关系、整体部分关系、语义相关关系等，这些关系在之前的任务中往往被看出一个整体，没有被算法区分开来。对这些不同类别的词汇关系的准确分类，不仅使上下位关系和其他关系的分类更精准，还能促进知识图谱的**本体构建**工作。在本节中，我们提出基于**超球学习**（Hyperspherical Learning）的**词汇关系分类**算法，同时对上下位关系、以及多种非上下位关系进行多分类学习。



(a) 任务：跨语言上下位关系方向分类　　(b) 任务：跨语言上下位关系检测

图 3.6: ITFOPM 在两个跨语言上下位关系预测任务上的迭代训练效果



(a) 任务：跨语言上下位关系方向分类　　(b) 任务：跨语言上下位关系检测

图 3.7: ITFOPM 在两个跨语言上下位关系预测任务上的参数 τ 的变化对效果的影响

3.5.1 算法模型

我们首先简要回顾词汇关系分类的目标，及算法的整体优化过程。在此之后，我们详细介绍提出的**超球关系嵌入**（SphereRE）模型的技术细节。

学习任务与目标： 令 $\mathcal{R}$ 为所有预定义的词汇关系类别的集合，例如上下位关系、同义词关系、反义词关系等。$D = (x_i, y_i)$ 为词汇关系分类任务的训练集，每个术语对 $(x_i, y_i) \in D$ 都对应唯一的词汇关系类别 $r_i \in \mathcal{R}$[6]。词汇关系分类的目标是训练分类器 f，对于未知关系的术语对 (x_i, y_i) 的词汇关系类别进行预测。在本节中，我们记词汇关系分类的测试集为 $U = \{(x_i, y_i)\}$。

因为词汇关系大多为常识性知识，关系的表达经常在文本中被省略，因此我们的研究工作主要从分布式语义建模的角度出发。对于任意关系元组 $(x_i, y_i) \in D$，我们通过预训练神经网络得到他们的向量表示，分别为 $\vec{x}_i$ 和 $\vec{y}_i$。在所有词汇关系类别 $\mathcal{R}$ 中，我们对每种词汇关系 $r_m \in \mathcal{R}$ 都训练一个投影模型 $f_m(\cdot)$，将关系主语

[6]在部分词汇关系分类的评测数据集中，一部分为随机组合的术语对，无特定词汇关系，通常被标注为 “Random”。在本文中，我们将其视为一种特别的词汇关系类别。

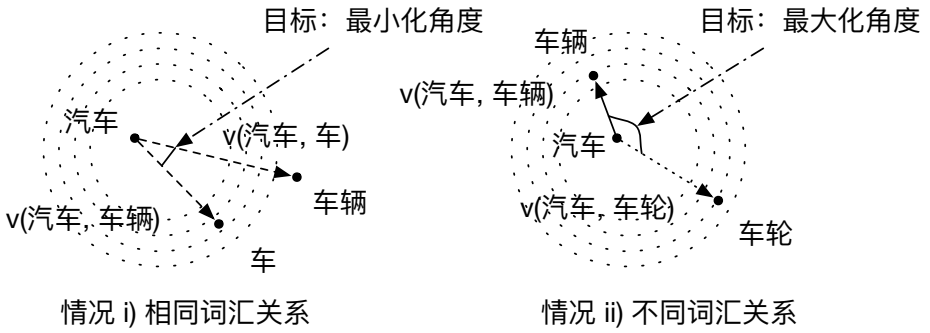


图 3.8: 超球学习的几何学解释

x_i 的词向量 $\vec{x}_i$ 投影至对应关系宾语 y_i 的词向量 $\vec{y}_i$。同第二章，我们用 $I(\cdot)$ 表示指示函数，投影学习的目标 J_f 为最小化投影误差，其函数定义如下：

$$J_f = \sum_{(x_i,y_i)\in D} \sum_{r_m \in \mathcal{R}} I(r_i = r_m) \| f_m(\vec{x}_i) - \vec{y}_i \|^2$$

根据上述模型，我们可以把 x_i 和 y_i 之间词汇关系在原始词向量空间中用向量差模型 [71, 78, 146] 表示。若术语对 $(x_i, y_i) \in D$ 对应的词汇关系为 r_i，这一元组在我们的模型中可以表示为 $f_i(\vec{x}_i) - \vec{x}_i$。

然而，这一模型没有考虑到不同类别的词汇关系应在向量空间模型中的表示问题，我们进一步推导超球学习的目标函数。首先，我们定义一个对称函数 $g(\cdot, \cdot)$，用来衡量两个词汇关系元组在超球空间内表示的距离。对于训练集和测试集中任意两个词汇关系元组 (x_i, y_i) 和 $(x_j, y_j) \in D \cup U$，若其词汇关系的类别分别为 r_i 和 r_j，我们最小化如下函数：

$$\delta(r_i, r_j) g(f_i(\vec{x}_i) - \vec{x}_i, f_j(\vec{x}_j) - \vec{x}_j)$$

其中，$\delta(r_i, r_j)$ 为符号函数。如果 (x_i, y_i) 和 (x_j, y_j) 的词汇关系类别相同（即 $r_i = r_j$），我们有 $\delta(r_i, r_j) = 1$；否则，$\delta(r_i, r_j) = -1$。由此可知，当 $\delta(r_i, r_j) g(f_i(\vec{x}_i) - \vec{x}_i, f_j(\vec{x}_j) - \vec{x}_j)$ 最小化时，有不同词汇关系类别的术语对在超球空间内的距离将最大化，有相同词汇关系类别的术语对在超球空间内的距离将最小化。图 3.8给出了本节使用的超球学习方法的几何学解释。

根据上述目标，在超球嵌入空间下的词汇关系表示学习目标函数 J_g 定义如下

所示：

$$J_g = \sum_{(x_i,y_i)\in D\cup U,(x_j,y_j)\in D\cup U} \delta(r_i,r_j)g(f_i(\vec{x}_i)-\vec{x}_i, f_j(\vec{x}_j)-\vec{x}_j)$$

令 Φ 为模型中所有参数的集合。SphereRE 的整体目标函数定义如下：

$$J(\Phi) = J_f + \xi_1 J_g + \xi_2\|\Phi\|^2$$

其中，ξ_1 和 ξ_2 是可调的平衡超参数。

最小化 $J(\Phi)$ 的优化问题是计算困难的（Computationally Intractable）。主要原因有：i）在最小化 $J(\Phi)$ 前，所有关系元组 $(x_i,y_i)\in U$ 的词汇关系类别 r_i 必须是已知的，而这些词汇关系类别本身就是模型预测的目标；ii）从 J_g 的定义中，并不能直接推出词汇关系元组在超球嵌入空间的表示；iii）最小化 $J(\Phi)$ 需要在多项式时间范围内对 D 和 U 的所有数据进行遍历，其时间复杂度高。

在下文中，我们提出**关系敏感的语义投影模型**（Relation-aware Semantic Projection）作为函数 $f_m(\cdot)$。利用上述模型，我们对每个关系元组 $(x_i,y_i)\in U$，预测其词汇关系类别的分布。其次，我们详细介绍词汇关系的表示学习算法和词汇关系分类算法。

关系敏感的语义投影模型： 对测试集中的每个元组 $(x_i,y_i)\in U$，我们从概率角度预测其词汇关系类别的分布。扩展 Yamane 等人 [86] 和我们先前的工作，对于每种词汇关系类别 $r_m\in\mathcal{R}$，我们分别学习一个投影矩阵 $\mathbf{M}_m$ 作为函数 $f_m(\vec{x}_i)$，将关系主语的词向量 $\vec{x}_i$ 映射到其宾语词向量 $\vec{y}_i$。在投影矩阵 $\mathbf{M}_m$ 上加上一个 Tikhonov 正则项后，对于某词汇关系类别 $r_m\in\mathcal{R}$ 的投影学习目标函数 J_m 可以重写为：

$$J_m = \sum_{(x_i,y_i)\in D} I(r_i=r_m)\|\mathbf{M}_m\vec{x}_i-\vec{y}_i\|^2 + \mu\|\mathbf{M}_m\|_F^2$$

其中，μ 为 Tikhonov 正则化超参数。

所以，我们有 $J_f=\sum_{r_m\in\mathcal{R}}J_m$。函数 J_m 的最小化可以通过梯度下降来实现，或直接计算其闭式解。$\mathbf{M}_m$ 的最优解 $\mathbf{M}_m^*$ 如下式所示：

$$\mathbf{M}_m^* = \text{argmin}_{\mathbf{M}_m} J_m = (\mathbf{X}_m^T\mathbf{X}_m+\mu\mathbf{I})^{-1}\mathbf{X}_m^T\mathbf{Y}_m$$

其中，$\mathbf{X}_m$ 和 $\mathbf{Y}_m$ 为两个 $n_m\times|\vec{x}_i|$ 数据矩阵，n_m 是 D 中具有词汇关系 $r_m\in\mathcal{R}$ 的关系元组的数量。$\mathbf{X}_m$ 和 $\mathbf{Y}_m$ 的第 i 行为对应具有词汇关系 r_m 的 $(x_i,y_i)\in D$ 的两

个术语 x_i 和 y_i 的词向量。$\mathbf{I}$ 是 $|\vec{x}_i| \times |\vec{x}_i|$ 的单位矩阵。

当我们对每种词汇类别 $r_m \in \mathcal{R}$ 对应的投影矩阵 $\mathbf{M}_m$ 都学习完毕后，我们在数据集 D 上训练一个关系分类器，分类器的特征为 $|\mathcal{R}| \times \vec{x}_i$ 维，表示为 $\mathcal{F}(x_i, y_i)$：[7]

$$\mathcal{F}(x_i, y_i) = (\mathbf{M}_1\vec{x}_i - \vec{y}_i) \oplus \cdots \oplus (\mathbf{M}_{|\mathcal{R}|}\vec{x}_i - \vec{y}_i)$$

其中，$\mathbf{M}_1, \cdots, \mathbf{M}_{|\mathcal{R}|}$ 是对应词汇关系 $r_1, \cdots, r_{|\mathcal{R}|}$ 的 $|\mathcal{R}|$ 个投影矩阵。根据 J_m 的定义，如果 (x_i, y_i) 具有词汇关系类别 r_m，向量 $\mathbf{M}_m\vec{x}_i - \vec{y}_i$ 的范数会较小；其他向量 $\mathbf{M}_n\vec{x}_i - \vec{y}_i (1 \leq n \leq |\mathcal{R}|, n \neq m)$ 的范数会较大。因此，该特征向量可以区分不同的词汇关系类别。

对于每个元组 $(x_i, y_i) \in U$，分类器输出一个 $|\mathcal{R}|$ 维的概率分布向量，表示该元组有相应词汇关系类别的概率。在本工作中，我们令 $p_{i,m}$ 为 $(x_i, y_i) \in U$ 具有词汇关系类别 $r_m \in \mathcal{R}$ 的概率。

超球关系嵌入学习： 当所有关系元组 $(x_i, y_i) \in U$ 的概率 $p_{i,m}$ 计算完毕后，我们关注目标函数 J_g。在本步骤中，我们旨在给每个训练集和测试集中的关系元组 $(x_i, y_i) \in D \cup U$ 学习一个 d_r 维的向量 $\vec{r}_i$，表达了这个关系元组的词汇关系。在下文中，我们将其称为 "SphereRE 向量"。

为了降低计算复杂度，我们参考图嵌入表示的研究 [165, 166]，将目标函数 J_g 和对称函数 $g(\cdot, \cdot)$ 用 Skip-Gram 模型 [51] 重新表示。在超球嵌入空间中，设 $Nb(x_i, y_i)$ 是关系元组 (x_i, y_i) 的邻居节点的集合。在这一空间中，(x_i, y_i) 的词汇关系类别与其邻居节点 $(x_j, y_j) \in Nb(x_i, y_i)$ 的词汇关系类别相似。为了保证词汇关系类别相同的关系元组有相似的 SphereRE 向量，我们将优化 J_g 的问题改为给定一个关系元组 (x_i, y_i) 的 SphereRE 向量 $\vec{r}_i$，模型能成功预测其邻居节点的概率最大化。基于负对数概率最小化的原则，我们定义一个新的目标函数 J_g'，用于替换原有的函数 J_g：

$$J_g' = -\sum_{(x_i,y_i)\in D\cup U} \sum_{(x_j,y_j)\in Nb(x_i,y_i)} \log \Pr((x_j, y_j)|\vec{r}_i) \tag{3.2}$$

上述模型的关键问题是正确定义邻域 $Nb(x_i, y_i)$，其原则为保持距离函数 $g(f_i(\vec{x}_i) - \vec{x}_i, f_j(\vec{x}_j) - \vec{x}_j)$ 的几何性质。在本算法中，我们引入两个关系元组 (x_i, y_i) 和 $(x_j, y_j) \in D \cup U$ 之间的权重 $w_{i,j} \in [0, 1]$，用于衡量两个关系元组的 SphereRE 向量在超球空

[7] 在本算法实现中，我们采用多分类的逻辑斯蒂回归模型作为分类器，而不采用深度神经网络。因为它能产生校准的概率分布，体现出模型的预测置信度，而深度神经网络一般不具有此性质，参见 [161]。

表 3.11: SphereRE 中 $w_{i,j}$ 在不同情况下的取值

情况	$w_{i,j}$ **的取值**
$(x_i, y_i) \in D, (x_j, y_j) \in D,$ $r_i = r_j$	1
$(x_i, y_i) \in D, (x_j, y_j) \in D,$ $r_i \neq r_j$	0
$(x_i, y_i) \in D, (x_j, y_j) \in U,$ $r_i = r_m$	$\frac{1}{2}p_{j,m}(\cos(\mathbf{M}_m\vec{x}_i - \vec{x}_i, \mathbf{M}_m\vec{x}_j - \vec{x}_j) + 1)$
$(x_i, y_i) \in U, (x_j, y_j) \in D,$ $r_j = r_m$	$\frac{1}{2}p_{i,m}(\cos(\mathbf{M}_m\vec{x}_i - \vec{x}_i, \mathbf{M}_m\vec{x}_j - \vec{x}_j) + 1)$
$(x_i, y_i) \in U, (x_j, y_j) \in U$	$\frac{1}{2}\sum_{r_m \in R} p_{i,m}p_{j,m} \cdot (\cos(\mathbf{M}_m\vec{x}_i - \vec{x}_i, \mathbf{M}_m\vec{x}_j - \vec{x}_j) + 1)$

间的距离。如果 $(x_i, y_i) \in D$ 且 $(x_j, y_j) \in D$，因为这两个关系元组都在训练集内，他们的类别标签已知，我们采用最简单的方式定义 $w_{i,j}$：

$$w_{i,j} = I(r_i = r_j)$$

我们继续讨论其他更复杂的情况。如果 $(x_i, y_i) \in D$ 具有关系类别 r_m，$(x_j, y_j) \in U$ 的关系类别未知，但是先前训练的分类器预测其有关系类别 r_m 的概率为 $p_{j,m}$，则 (x_i, y_i) 和 (x_j, y_j) 之间的权重 $w_{i,j}$ 定义为带权重的余弦相似度的一种变体，取值范围被归一化到 $(0, 1)$，数学表达如下所示：

$$w_{i,j} = \frac{1}{2}p_{j,m}(\cos(\mathbf{M}_m\vec{x}_i - \vec{x}_i, \mathbf{M}_m\vec{x}_j - \vec{x}_j) + 1)$$

如果 $(x_i, y_i) \in U$，$(x_j, y_j) \in D$，情况与之类似。如果 $(x_i, y_i) \in U$，$(x_j, y_j) \in U$，因为两个关系元组的词汇关系类别都未知，我们通过对所有可能的词汇关系类别，根据相应概率对余弦相似度进行加和。权重 $w_{i,j}$ 按下式计算：

$$w_{i,j} = \frac{1}{2}\sum_{r_m \in \mathcal{R}} p_{i,m}p_{j,m} \cdot (\cos(\mathbf{M}_m\vec{x}_i - \vec{x}_i, \mathbf{M}_m\vec{x}_j - \vec{x}_j) + 1)$$

读者也可参阅表 3.11中对 $w_{i,j}$ 在不同情况下的取值方式汇总。

根据 $w_{i,j}$ 的取值，我们提出了一个基于蒙特卡洛的采样与学习算法以学习所有关系元组的 SphereRE 向量，算法过程见算法 9。在算法的初始阶段，所有关系元组 $(x_i, y_i) \in D \cup U$ 的 SphereRE 向量 $\vec{r}_i$ 都随机初始化。之后，算法开始迭代采样过程，它随机选择一个元组 (x_i, y_i) 作为起始点，以如下概率采样得到下一个 (x_j, y_j)：

Algorithm 9 SphereRE 学习算法

1: **for** 每个关系元组 $(x_i, y_i) \in D \cup U$ **do**
2: 随机初始化 SphereRE 向量 $\vec{r}_i$
3: **end for**
4: **for** $i = 1$ 至最大迭代次数 **do**
5: 根据式 3.3采样得到序列：$\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_{|\mathcal{S}|}, y_{|\mathcal{S}|})\}$
6: 通过最小化 $-\sum_{(x_i,y_i)\in\mathcal{S}} \sum_{j=i-l(j\neq i)}^{i+l} \log \Pr((x_j, y_j)|\vec{r}_i)$，更新 SphereRE 向量 $\vec{r}_i$ 的值
7: **end for**

$$\Pr((x_j, y_j)|(x_i, y_i)) = \frac{w_{i,j}}{\sum_{(x'_j,y'_j)\in D_{mini}} w_{i,j'}} \tag{3.3}$$

其中 D_{mini} 是从 $D \cup U$ 中随机抽取到的小批次样本集。在每个迭代中，算法只需要遍历 $|D_{mini}|$ 个关系元组，而不是原来的 $|D| + |U|$ 个。算法不断采样，得到下述词汇关系元组的序列 $\mathcal{S}$：

$$\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_{|\mathcal{S}|}, y_{|\mathcal{S}|})\}$$

令 l 为窗口的大小。我们利用 Skip-Gram 模型 [51] 来近似式 3.2中的目标函数 J'_g：$-\sum_{(x_i,y_i)\in\mathcal{S}} \sum_{j=i-l(j\neq i)}^{i+l} \log \Pr((x_j, y_j)|\vec{r}_i)$。SphereRE 向量 $\vec{r}_i$ 的值在迭代过程中不断更新至收敛。我们可以看出，计算出的 $\vec{r}_i$ 是超球嵌入空间内词汇关系元组的低维表示。

在实践中, 我们发现采样过程有一个缺陷。因为所有 $(x_i, y_i) \in U$ 的预测结果都是带概率的, 算法在序列 $\mathcal{S}$ 中更倾向于采样 D 中的元组, 而非 U 中的元组。U 的低采样率使得对应元组的表示学习效果降低。我们利用分层抽样的方法增大 U 中关系元组的采样率。对于每个 $(x_i, y_i) \in U$，我们增大其关系预测概率：$p_{i,m} \leftarrow p_{i,m}\gamma$, 其中 $\gamma > 1$ 为概率提升因子。值得注意的是，尽管我们不直接优化目标函数 J_g，或显示地构建超球嵌入空间，算法 9学习到的 SphereRE 向量反应出不同词汇关系的差异性。

词汇关系分类： 最后,我们训练词汇关系分类器。对训练集中每个元组 $(x_i, y_i) \in D$，我们抽取如下 $|\mathcal{R}| \times \vec{x}_i + d_r$ 维度的特征集 $\mathcal{F}^*(x_i, y_i)$：

$$\mathcal{F}^*(x_i, y_i) = \mathcal{F}(x_i, y_i) \oplus \vec{r}_i$$

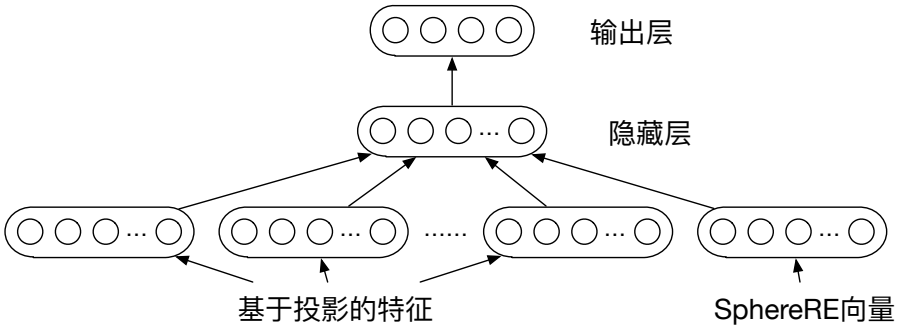


图 3.9: SphereRE 的神经网络架构

Algorithm 10 SphereRE 的词汇关系分类算法

for 每种词汇关系类别 $r_m \in \mathcal{R}$ **do**
    计算投影矩阵的最优解 $\mathbf{M}_m^*$
end for
在训练集 D 上使用特征 $\mathcal{F}(x_i, y_i)$ 训练分类器
for 每个关系元组 $(x_i, y_i) \in U$ **do**
    利用分类器预测词汇关系分类分布 $p_{i,m}$
end for
对所有关系元组 $(x_i, y_i) \in D \cup U$，利用算法 9学习 SphereRE 向量 $\vec{r}_i$
在训练集 D 上 D 使用特征 $\mathcal{F}^*(x_i, y_i)$ 训练神经网络
for 每个关系元组 $(x_i, y_i) \in U$ **do**
    利用神经网络预测词汇关系类别 r_i
end for

其中, $\mathcal{F}(x_i, y_i)$ 是 $|\mathcal{R}| \times |\vec{x}_i|$ 维的基于投影的特征, $\vec{r}_i$ 是 d_r 维的超球空间内的关系特征。分类器参照文献 [104], 采用全连接的前馈神经网络。输入层有 $|\mathcal{R}| \times |\vec{x}_i| + d_r$ 个节点, 仅包括一个隐藏层, 输出层的维度为 $|\mathcal{R}|$, 采用 Softmax 函数作为预测函数, 其架构如图 3.9。这一神经网络采用随机梯度下降训练，用来对每个元组 $(x_i, y_i) \in U$ 的实际词汇类别做最终预测。SphereRE 的整体流程见算法 10。

3.5.2 实验分析

在本节中，我们在多个基准数据集上评测 SphereRE 算法的准确性，并与基线方法进行充分比较。

数据集与实验设置： 在实验中，我们使用与第3.4.2节相同的 fastText 模型 [53] 得到词向量，维度为 300。为了评测算法的有效性，我们在四个公开的多词汇关系分类数据集上进行评测，这四个数据集分别为 K&H+N [103]、BLESS [93]、ROOT09 [167] 和 EVALution [168]。我们也在 CogALex-V 任务的第 2 个子任务 [169] 上评测了 SphereRE 的准确度。这 5 个数据集的统计信息汇总在表 3.12中。

表 3.12: 五个词汇关系分类数据集的统计信息

关系类别	K&H+N	BLESS	ROOT09	EVALution	CogALex
反义词关系	-	-	-	1,600	601
属性关系	-	2,731	-	1,297	-
同下位词关系	25,796	3,565	3,200	-	-
事件关系	-	3,824	-	-	-
整体关系	-	-	-	544	-
上下位关系	4,292	1,337	3,190	1,880	637
部分关系	1,043	2,943	-	654	387
随机关系	26,378	12,146	6,372	-	5,287
实质关系	-	-	-	317	-
同义词关系	-	-	-	1,086	402
合计	57,509	26,546	12,762	7,378	7,314

我们采用文献 [104] 中同样的方式将四个公开数据集分成训练集、验证集和测试集。CogALex 数据集的划分与标准 CogALex-V 评测任务 [169] 中的划分相同。SphereRE 算法的默认参数设置如下：$\mu = 0.001$、$d_r = 300$、$|D_{mini}| = 20$、$|\mathcal{S}| = 100$、$\gamma = 2$ 和 $l = 3$。在迭代算法 9中，我们默认设置为 500 个迭代。我们也在实验中进一步调整参数设置和神经网络架构，并且详细比较了实验结果。需要指出的是，尽管在目标函数 $J(\Phi)$ 中，我们引入了正则化超参数 ξ_1 和 ξ_2，但是在算法实现中，我们采用基于蒙特卡洛的采样方法学习 SphereRE 向量 $\vec{r}_i$，不需要使用 ξ_1 和 ξ_2 直接优化 $J(\Phi)$。

四个公开数据集的实验结果： 我们在四个公开数据集上评测 SphereRE 算法的准确度，并与现有方法比较，考虑如下基线算法：

- $\vec{x}_i \oplus \vec{y}_i$、$\vec{x}_i - \vec{y}_i$ [77, 78]：他们是经典的分布式关系分类算法，采用无隐藏层的神经网络作为分类模型。

- NPB [54]：它采用基于依存路径的 LSTM 神经网络作为分类器，由 [104] 实现。

- LexNET [104]：它在 Shwartz 等人 [54] 工作的基础上，结合了术语的向量表示和依存路径的 LSTM 表示作为神经网络的特征进行分类。

- $(\vec{x}_i \oplus \vec{y}_i)_h$、$(\vec{x}_i - \vec{y}_i)_h$, LexNET$_h$：他们是 $\vec{x}_i \oplus \vec{y}_i$、$\vec{x}_i - \vec{y}_i$ 和 LexNET 的变体，在神经网络输入、输出层之间加入了一个隐藏层。

- NPB+Aug、LexNET+Aug [105]：他们是 NPB 和 LexNET 的变体，采用增强

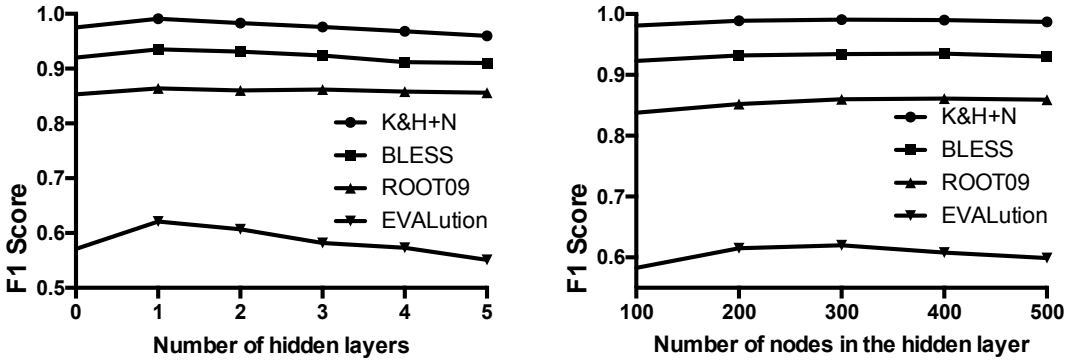
表 3.13: 词汇关系分类算法在四个公开数据集上的比较

方法	K&H+N			BLESS		
	准确度	**召回率**	**F 值**	**准确度**	**召回率**	**F 值**
$\vec{x}_i \oplus \vec{y}_i$	0.909	0.906	0.904	0.811	0.812	0.811
$(\vec{x}_i \oplus \vec{y}_i)_h$	0.983	0.984	0.983	0.891	0.889	0.889
$\vec{x}_i - \vec{y}_i$	0.888	0.886	0.885	0.801	0.803	0.802
$(\vec{x}_i - \vec{y}_i)_h$	0.941	0.942	0.941	0.861	0.859	0.860
NPB	0.713	0.604	0.55	0.759	0.756	0.755
LexNET	0.985	0.986	0.985	0.894	0.893	0.893
LexNET$_h$	0.984	0.985	0.984	0.895	0.892	0.893
NPB+Aug	-	-	0.897	-	-	0.842
LexNET+Aug	-	-	0.970	-	-	0.927
SphereRE	**0.990**	**0.989**	**0.990**	**0.938**	**0.938**	**0.938**
方法	ROOT09			EVALution		
	准确度	**召回率**	**F 值**	**准确度**	**召回率**	**F 值**
$\vec{x}_i \oplus \vec{y}_i$	0.636	0.675	0.646	0.531	0.544	0.525
$(\vec{x}_i \oplus \vec{y}_i)_h$	0.712	0.721	0.716	0.57	0.573	0.571
$\vec{x}_i - \vec{y}_i$	0.627	0.655	0.638	0.521	0.531	0.528
$(\vec{x}_i - \vec{y}_i)_h$	0.683	0.692	0.686	0.536	0.54	0.539
NPB	0.788	0.789	0.788	0.53	0.537	0.503
LexNET	0.813	0.814	0.813	0.601	0.607	0.6
LexNET$_h$	0.812	0.816	0.814	0.589	0.587	0.583
NPB+Aug	-	-	0.778	-	-	0.489
LexNET+Aug	-	-	0.806	-	-	0.545
SphereRE	**0.860**	**0.862**	**0.861**	**0.62**	**0.621**	**0.62**

的依存路径作为 LSTM 的输入，提高语言模式的覆盖率。

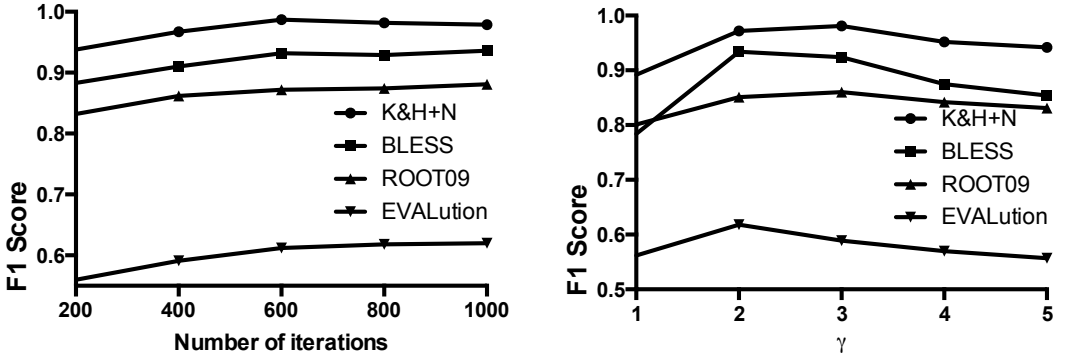
SphereRE 和基线方法的实验结果见表 3.13。我们分别计算每个数据集中每种词汇关系类别的预测精准度、召回率和 F 值，汇报其加权平均值。从实验结果可见，经典的分布式分类算法比带有语言模式表示的神经网络（例如 LexNET）效果更低，这体现出直接使用词向量作为分类器的特征不能学习到不同词汇关系的真正语义。使用平均 F 值作为评测指标，我们提出的 SphereRE 方法在四个公开数据集都超过了所有基线算法。当词汇关系的种类相对较多时（例如 EVALution），SphereRE 的提升效果相对其他数据集（例如 BLESS、ROOT09）不够明显。最有可能的原因是在关系敏感的语义投影模型步骤中，预测的错误可能会传播到下一步骤中。

我们对词汇关系分类的神经网络架构进行了调整，分别将不同设置下在验证集测试的效果变化趋势展示在图 3.10中。这体现中，加入更多数量的隐藏层并不能提升词汇分类模型的分类效果。在部分数据集中（例如 EVALution），模型的预测



(a) 调整隐藏层的层数　　(b) 调整隐藏层的节点数量

图 3.10: SphereRE 中神经网络架构分析



(a) 调整迭代的数量　　(b) 调整参数 γ

图 3.11: SphereRE 中蒙特卡洛算法的参数分析

精度会下降，体现出模型逐渐过拟合的趋势。我们固定只采用一层隐藏层，调整了隐藏层节点的数量（即神经网络的宽度），实验效果表明，这一设置对模型的效果影响不明显。

我们继续研究基于蒙特卡洛的采样算法在不同的设置下是怎么影响 SphereRE 向量的质量的。我们在实验中调整迭代的数量和参数 γ 的值，实验结果见表 3.11。当采样算法不断迭代，关系向量的质量也逐步提升。当模型执行足够数量的迭代后（> 500），模型的精度保持稳定。参数 γ 的取值需要注意平衡：当 γ 过小时，测试数据的采样率会过低，影响模型在测试阶段的表现；当 γ 过大时，关系敏感的语义投影模型会将过多的错误引入 SphereRE 训练算法中。此外，我们也研究加入 SphereRE 向量对词汇关系分类有多大的贡献。我们在实验中移除了 SphereRE 向量使用余下的特征训练词汇关系分类神经网络，对比实验结果见表 3.14。由结果可见，加入 SphereRE 向量在四个数据集上 F 值都有提升。

表 3.14: SphereRE 模型中的特征分析

特征	K&H+N	BLESS	ROOT09	EVALution
不加入 SphereRE 向量	0.968	0.918	0.82	0.581
加入 SphereRE 向量	0.990	0.938	0.861	0.62
提高	**+2.2%**	**+2.0%**	**+4.1%**	**+3.9%**

表 3.15: 词汇关系分类算法在 CogALex-V 任务上的比较

方法	**同义词关系**	**反义词关系**	**上下位关系**	**部分关系**	**总计**
GHHH [147]	0.204	0.448	0.491	0.497	0.423
LexNET [104]	**0.297**	0.425	0.526	0.493	0.445
STM [106]	0.221	**0.504**	0.498	0.504	0.453
SphereRE	0.286	0.479	**0.538**	**0.539**	**0.471**

CogALex-V 任务的实验结果： 我们在 CogALex-V 任务 [169] 上评测 SphereRE 算法的效果。在这一任务的子任务 2 中，要求将 4260 个术语对分成 5 种词汇关系：同义词关系、反义词关系、上下位关系、部分关系和随机关系。训练集包括 3054 个带关系标注的术语对。比起先前 4 个公开数据集，这个任务数据集的评测更具有挑战性，因为 i）任务中将随机关系视为噪声，其预测结果不加入最终评测结果；ii）训练集比较小和 iii）它的训练集和测试集按照词汇划分，模型不会因为 “词汇记忆” 现象 [82] 而结果虚高。

我们在表 3.15中列出 CogALex-V 数据集涉及的每种词汇关系预测的 F 值。在这个任务提交的系统中，GHHH [147] 和 LexNET [104] 是取得最高平均 F 值的两个系统。在 CogALex-V 上进行评测的最近的工作是 STM [106]。SphereRE 的平均 F 值为 47.1%，效果超过了先前的工作。我们进一步发现，分布式上下位关系的预测受到 “词汇记忆” 问题的影响比较严重，SphereRE 在上下位关系的预测 F 值为 53.8%，也明显超过了其他方法。

SphereRE 向量分析： 我们进一步评测 SphereRE 向量的质量。我们的实验任务为 Top-k 相似关系元组检索，即给定任何一个关系元组对应的 SphereRE 向量，在 SphereRE 向量空间中检索到 Top-k 个最相似的向量（相似度用两个向量的余弦相似度来衡量），评测这 Top-k 个向量对应的关系元组是否和输入元组有相同的词汇关系类别。我们采用 Top-k 平均准确度（Average Precision@k，缩写为 AP@k）作为评测指标。在一个词汇关系数据集中，AP@k 越高，则 SphereRE 向量的质量越好。在表 3.16中，我们对 5 个数据集的训练和测试集都进行评测，并且列出其 AP@k（$k = 1, 5, 10$）。

表 3.16: 5 个词汇关系分类数据集的 Top-k 相似关系元组检索结果

数据集	**AP@1**	**AP@5**	**AP@10**	**AP@1**	**AP@5**	**AP@10**
	训练集			测试集		
K&H+N	0.972	0.954	0.951	0.862	0.844	0.839
BLESS	0.962	0.950	0.948	0.868	0.830	0.825
ROOT09	0.987	0.993	0.989	0.814	0.789	0.828
EVALution	0.988	0.987	0.982	0.653	0.650	0.697
CogALex	0.953	0.904	0.918	0.631	0.628	0.649

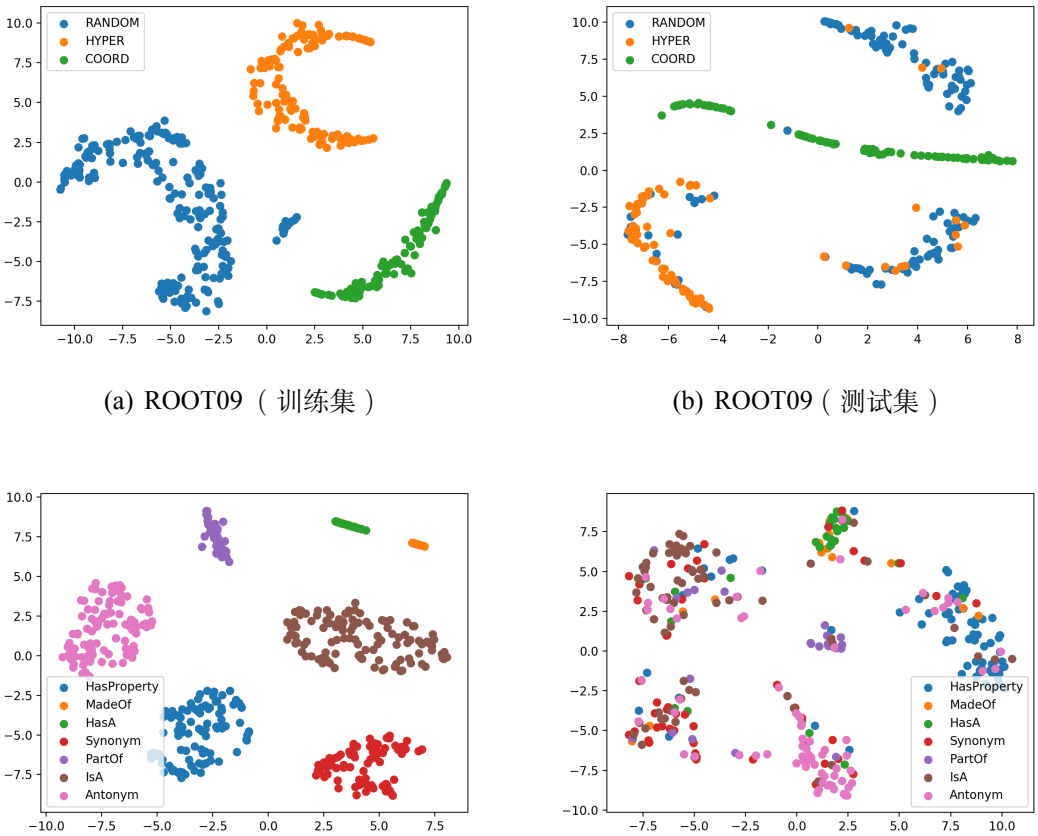
表 3.17: SphereRE 算法的错误案例

术语对	**预测关系类别**	**真实关系类别**
(heart, courage)	随机关系	同义词关系
(wing, animal)	随机关系	部分关系
(mint, pennyroyal)	随机关系	上下位关系
(handlebar, bike)	同下位词关系	部分关系
(grenade, object)	属性关系	上下位关系

在实验中，SphereRE 在 5 个训练集上都有接近完美的学习效果，其中，AP@1 超过 95%, AP@5 和 AP@10 超过 90%。这是因为在关系表示学习阶段, 这些元组的类别标签对算法是完全可见的。由此可见 SphereRE 对于带标注的数据集，能较好地表示不同术语对的词汇关系类别。对测试集而言, SphereRE 的 AP@k 在 K&H+N、BLESS 和 ROOT09 三个数据集上有轻微下降。EVALution 和 CogALex 中的词汇关系类别多、数据杂，因此这两个数据集的 AP@k 相对较低。

为了对 SphereRE 向量有更加直观的理解，在图 3.12中展示了 SphereRE 向量在 t-SNE 算法 [170] 下的可视化结果。对训练集，我们可以观察到不同词汇关系类别的词向量有在二维平面有明显的分隔；对于测试集，不同词汇关系类别的词向量有轻微杂乱的现象，说明了 SphereRE 算法在模型预测方面有部分错误。

错误分析：我们随机采样 300 个预测错误的案例进行人工分析,示例见表 3.17。大部分错误案例的原因可以归结于数据集中的随机关系, 例如在 K&H+N、BLESS、ROOT09 和 CogALex 中，有比较大的比例是随机关系。这些术语对之间没有明确的语义关系, 模型很难进行建模, 所以分类器很可能将有其他词汇关系的元组预测为随机关系。此外，训练集不同类别数据的不平衡性也使得模型的训练造成困难, 例如在数据集 EVALution 中的部分关系、数据集 CogALex 中的同义词关系数量很小，相应关系元组的表示学习效果也会较差。



(a) ROOT09（训练集）　(b) ROOT09（测试集）

(c) EVALution（训练集）　(d) EVALution（测试集）

图 3.12: SphereRE 向量利用 t-SNE 算法的可视化结果

3.6 小结

在本章中，我们分别从多知识源、多语言、多词汇关系三个角度扩展了前述基于词嵌入的投影模型。其中，TEAL 采用深度对抗学习技术，将大规模分类体系中的上下位关系知识融入基于训练集的词嵌入投影神经网络中；TFOPM（及其扩展算法 ITFOPM）基于深度迁移学习和双语术语对齐技术，实现了面向小语种的、小样本学习场景下跨语言上下位关系预测；SphereRE 进一步考虑了非上下位关系中存在多种词汇关系的情况，提出了超球关系嵌入模型，使具有不同词汇关系类别的术语对更容易被区分。实验效果证明了上述模型的有效性。值得注意的是，本文第二章、第三章的研究只局限于固定类别的语义关系预测，如何从中文短文本中自动挖掘出更多类别的语义关系，需要进一步研究。

第四章　非上下位关系抽取与语义理解

在第三章中，我们分别从多知识源、多语言和多词汇关系三个角度，研究了不同情境下基于词嵌入投影模型的语义关系预测算法。其中，TEAL、TFOPM 和 ITFOPM 等三种算法特别关注上下位关系的表示学习和预测问题；SphereRE 将上下位关系预测的研究扩展到多种类别的词汇关系上，实现多词汇关系分类。然而，中文短文本中存在多种类别的非上下位关系，这些关系类别难以被人工穷举，高效的数据标注和完全监督式关系抽取模型训练的挑战性极高。前述算法的预测关系标签空间是固定的，难以扩展至开放领域，即不能实现在没有明确定义待抽取关系类别的情况下，自动抽取相应非上下位关系元组。本章在第二章和第三章研究的基础上，进一步在开放域下探究面向中文短文本的非上下位关系抽取与语义理解的问题。在下文中，我们介绍本章关注的关系抽取与语义理解任务，对相关研究工作进行汇总和讨论，详细描述我们提出的三种算法，并对这些算法的实验结果进行详细分析。

4.1　引言

如前三章所述，语义关系的自动抽取对知识图谱构建有重要作用。在 NLP 相关研究中，与关系抽取相关的任务包括监督式的关系分类 [35, 171, 172]、基于知识库的远程监督式关系抽取 [29, 173, 174]、以及无需指定关系类别的开放关系抽取 [27, 28, 175, 176] 等。这些方法都旨在从完整的句子中抽取关系三元组，然而这些模型很难从较短的文本中抽取语义关系。这是因为，短文本一般不包含句子结构的必要元素，语法结构不完整，表达语义关系的上下文也高度稀疏 [47, 177]。此外，短文本中常常表达人类的常识性知识，现有机器学习算法对常识性知识的获取和处理挑战较大 [178, 179]。

为了从中文短文本中自动抽取关系，我们在第二章和第三章提出基于词嵌入的投影学习方法，然而这些模型处理的关系类别是有限的。在实际应用场景中，中文短文本中表述的语义关系数目繁多，且标注大量数据费时费力。因此，设计无需大量人工干预的、领域自适应的关系挖掘框架，成为面向中文短文本的关系抽取这一研究课题中亟待解决的问题。在本章中，我们继续考虑中文短文本对 $\{(x_i, y_i)\}$

作为算法的输入，其中 x_i 为某中文实体，y_i 为描述 x_i 的中文短文本。图 4.1给出了中文实体 “欧洲联盟”，及其对应描述性短文本（例如 “诺贝尔和平奖获得组织”、“1993 年建立”）作为示例，详述关系抽取与语义理解的研究思路。

首先，在没有任何人工标注信息的情况下，无法直接在中文短文本数据源上训练关系抽取模型。我们观察到，在中文短文本中，语言模式的存在与语义关系有密切的关联性。根据这一发现，我们提出了**基于模式的非上下位关系抽取算法**（Pattern-based Non-hypernymy Relation Extraction，缩写为 PNRE）。它首先在所有描述中文实体 x_i 的短文本 y_i 中，挖掘频繁出现的、最有可能描述某特定语义关系的语言模式 p（例如图 4.1的 “[E] 获得组织”、“[E] 建立”，其中 “[E]” 为实体标签）。对于每一种语言模式 p, PNRE 采用图挖掘算法, 检测出最可能正确的关系元组, 作为种子关系元组 R_p^*。利用种子关系元组，算法可以自动抽取出与 “种子” 足够相似的关系元组，其正确概率较高，因此无需人工训练集的标注工作。值得注意的是，由于在短文本的关系抽取任务中，保证高精度比较困难，PNRE 只采用固定的语言模式，不利用迭代抽取机制修改语言模式，避免了 “语义漂移” 的问题 [62]。

由于 PNRE 不使用任何模式扩展的机制, 它只能从固定的、频繁出现的语言模式抽取对应的关系。因为语义关系在语料库中的类别分布一般具有 “长尾效应” [46], PNRE 只能处理频率位于 “头部” 的关系类别，“长尾关系” 容易被 PNRE 忽略。我们在开放关系抽取 [175] 的框架下，进一步提出了**数据驱动的非上下位关系抽取算法**（Data-driven Non-hypernymy Relation Extraction，缩写为 DNRE），以获取更多数量、更多类别的非上下位关系。它包括三个主要模块：**修饰词敏感的词组切分**（Modifier-sensitive Phrase Segmenter, 缩写为 MPS）、**候选关系元组生成**（Candidate Relation Generator，缩写为 CRG）和**缺失关系谓词检测**（Missing Relation Predicate Detector，缩写为 MRPD）。这一算法抽取中文短文本的修饰词，作为中文实体的潜在关系知识源，并且利用海量中文语料库作为背景知识，利用**语义解释**（Semantic Interpretation）技术 [180] 挖掘关系类别。例如，在图 4.1中，DNRE 从 “联合国大会观察员” 中抽取出关系谓词 “参与” 和关系宾语 “联合国大会”，其中 “参与” 没有出现在短语中，而从语料库中挖掘得到。由此可见，与 PNRE 相比，DNRE 提升了对中文短文本的语义理解能力。

前述 PNRE 和 DNRE 两个算法都关注于中文实体 x_i 和及其描述性短文本 y_i 之间的关系，这两种算法都缺乏对短文本 y_i 本身的语义进行深度理解。我们观察到，y_i 中如果存在复合语言结构，可以通过对 y_i 进行语义解释，间接推断出更多 x_i 和

y_i 之间的深层次的关系。例如，在图 4.1中，如果我们预测出 y_i“国家联盟” 的语义可以通过以下两个关系元组进行解释：

（国家联盟，属于，联盟），（国家联盟，包含，国家）

据此，针对中文实体 x_i“欧洲联盟”，我们可以进行如下知识推理：

（欧洲联盟，属于，联盟），（欧洲联盟，包含，国家）

若 y_i 为中文复合名词, 在计算语言学领域, 我们可以将 y_i 语义解释方式建模为**习语性**（Idiomaticity）程度预测问题。在这一部分研究中，我们提出了用于习语性程度预测的**关系性与组合性表示学习框架**（Relational and Compositional Representation Learning，缩写为 RCRL），对 y_i 分别学习其**关系性表示** $\vec{r}_i$ 与**组合性表示** $\vec{c}_i$。我们进一步研究 RCRL 是如何有利于中文短文本的理解、以及语义关系的扩展。

综上所述, 我们分别概述了面向中文短文本的、非上下位关系抽取与语义理解在三个角度的研究，扩展了第二章和第三章的研究工作。表 4.1总结了第四章中使用的重要符号及其意义。

4.2 相关工作

在本节中，我们汇总了面向短文本的知识获取相关研究工作。特别地，由于本文研究中文语境下的问题，我们重点讨论面向中文短文本的多种类别知识获取的研究挑战。

4.2.1 基于短文本的关系抽取

与基于句子和文档的关系抽取不同，基于短文本的关系抽取研究并不充分，这是因为短文本的关系表达上下文比较稀疏, 经典关系抽取的算法很难应用于这一任务。在英语短文本中，很多关系型知识是通过名词词组表达的，基于名词词组的关系抽取是开放关系抽取（ORE）的最新研究方向之一 [47]。RENOUN [31] 是知名的基于名词词组的 ORE 系统, 它自动扩展英语语言模式, 从英语名词词组中抽取属性关系。例如，从 “Princeton economist Paul Krugman was awarded the Nobel prize in 2008” 中, RENOUN 可以抽取出关系三元组 “（Princeton, economist, Paul Krugman）”，这一关系由名词（“economist”）而非动词（“awarded”）表达。RELNOUN [177] 扩展了 RENOUN 系统，抽取了复合名词中的关系型知识。在基于名词词组的 ORE

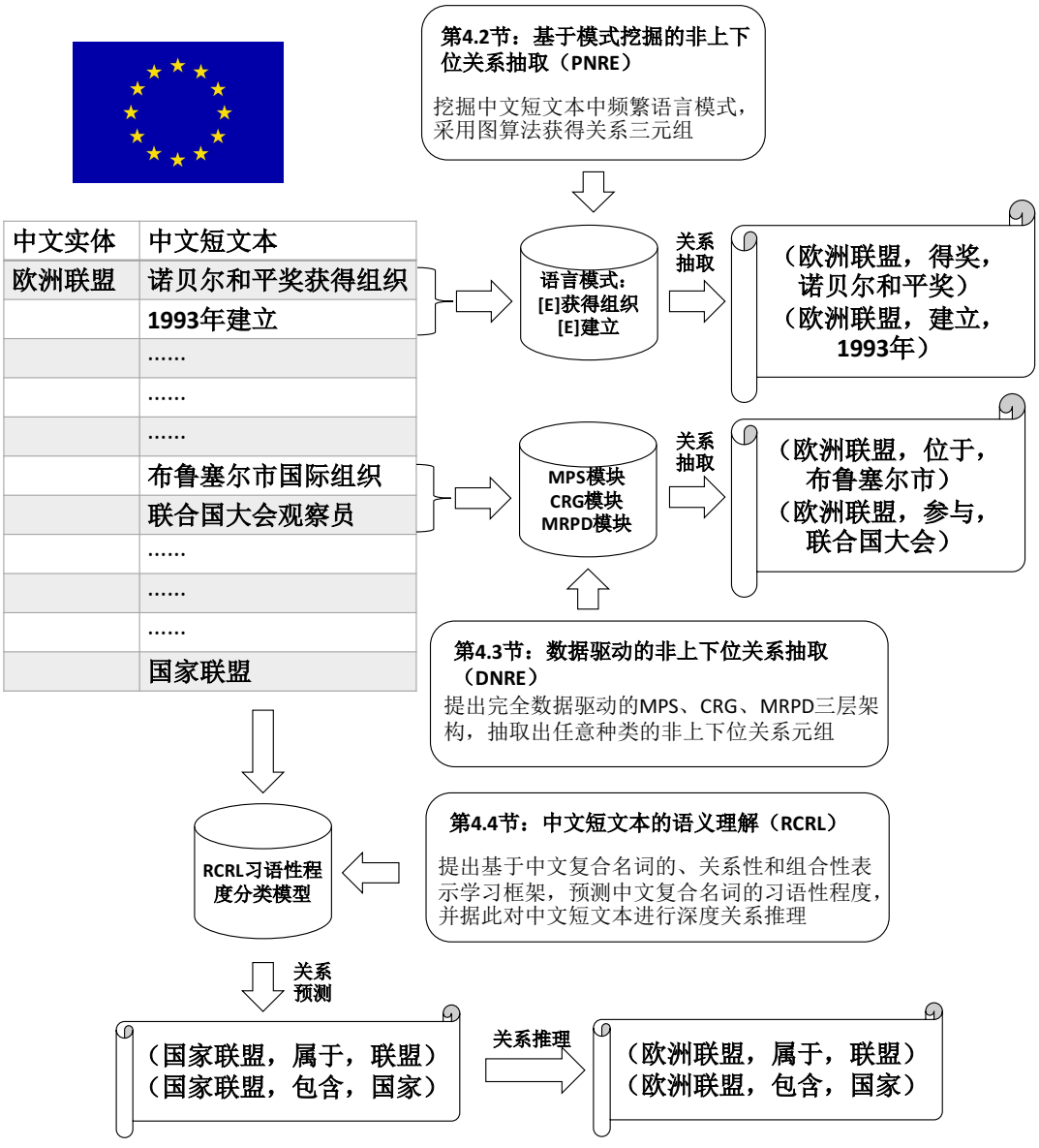


图 4.1: 第四章模型研究思路汇总及其示例

表 4.1: 第四章使用的重要符号及其意义

符号	说明
(x_i, y_i)	中文短文本对，其中 y_i 为描述中文实体 x_i 的短文本
$\vec{x}_i$	x_i 的词向量
p	PNRE 中的语言模式
y_i^p	PNRE 中从 y_i 中利用模式 p 匹配得到实体
R_p	PNRE 中关于模式 p 的候选关系元组集
$G_p(C_p, L_p, W_p)$	PNRE 中关于模式 p 的模式图
C_p^*	PNRE 中 G_p 的最大边权重团
R_p^*	PNRE 中关于模式 p 的种子关系元组集
R_p'	PNRE 中关于模式 p 的抽取出的关系元组集
$ws(y_i)$	DNRE 中 y_i 的中文分词结果
$ps(y_i)$	DNRE 中 y_i 的修饰词敏感切分结果
$(x_i, q_i^{(j)})$	DNRE 中某实体-修饰词对
$r(x_i, q_i^{(j)})$	DNRE 中从 $(x_i, q_i^{(j)})$ 抽取出的候选完整关系元组
$\tilde{r}(x_i, q_i^{(j)})$	DNRE 中从 $(x_i, q_i^{(j)})$ 抽取出的候选部分关系元组
n	DNRE 中的 N-Gram 因子
$G_n(y_i)$	DNRE 中关于 $ws(y_i)$ 的 N-Gram 分割图
$R(x_i, y_i)$	DNRE 中从 (x_i, y_i) 抽取出的候选关系元组集合
v^*	DNRE 中 $\tilde{r}(x_i, q_i^{(j)})$ 最有可能的关系谓词
$H(\mathcal{R}, \mathcal{V})$	DNRE 中基于谓词的超图网络
L	RCRL 中习语性程度分类训练集
U	RCRL 中习语性程度分类测试集
$x_i = N_1N_2$	中文复合名词，其中两个组成名词分别为 N_1 和 N_2
f_i	中文复合名词 x_i 的习语性程度
$\vec{r}_i$	RCRL 中中文复合名词的关系性表示
$\vec{c}_i$	RCRL 中中文复合名词的组合性表示
$\mathcal{F}_i$	RCRL 中中文复合名词 x_i 的关系性特征
$\mathbf{M}_r$	RCRL 中关系性特征的线性投影矩阵

系统中，维基百科的关系类别是比较重要的知识源，因为其数据质量高、语义关系丰富。在 YAGO 系统 [8] 中，Suchanek 等人将关系类别的语言模式建模为 “Pre-modifier + Head word + Post-modifier”（例如 “French people of Italian descent”），并以此为模板构建基于英语维基百科实体的知识本体。Nastase 和 Strube [33] 从词汇、语义、语法等多角度提出混合式的模式匹配法，从维基百科的关系类别中获得更多关系元组。Pasca [181] 研究怎样将维基百科的关系类别分解成属性-值对（Attribute Value Pairs），并且利用英语词汇模式解决这一问题。

与上述任务的另一相似任务为**名词词组解释**（Noun Phrase Interpretation），即自动生成描述性文本，概括名词短语中的关系，这是因为在名词词组中，语义关系的表达往往是隐含的，需要关系抽取系统额外生成。例如，在词组 “olive oil” 中，

存在 “made-from” 关系，因此名词词组解释算法可以针对 “olive oil” 生成关系三元组 “（olive oil，made-from，olive）”。这一任务通常被建模为监督学习的分类任务，即将名词词组分类到预定义的关系标签中，每个标签对应一种动词性关系谓词，解释这一词组 [182–185]。然而，仅仅采用有限数量的动词很难覆盖名词词组的语义。为了生成更细粒度的名词词组解释文本，Cruys 等人 [186] 采用多个动词和介词短语表达名词词组的语义。Grycner 和 Weikum [187] 设计了 POLY 系统，从多语言语料库中自动挖掘名词词组的跨语言关系表达。在 SemEval-2013 Task 4 [188] 这一评测任务中，参与者可以采用任意方法从无结构化文本中挖掘名词词组的解释方法。在基于名词词组的 ORE 研究中，Xavier 和 de Lima [189] 结合 ORE 和名词词组解释两个 NLP 任务，自动从语料库中挖掘关系型的名词词组，并且利用名词词组解释技术生成相应关系三元组，以扩展现有知识图谱。

与英语的研究相比较，中文的相关研究处于初步阶段。由于中文表达比英语更加灵活，相应任务的解决难度更大。对于 ORE 任务，Qiu 和 Zhang [34]、以及 Jia 等人 [190] 都设计了句子级别的 ORE 系统，从中文句子的依存句法分析结果，抽取中文知识三元组。Wei 和 Yuan [191] 构建了中文词组的解释模板，用于中文名词词组的自动解释，然而这一工作涉及的模板覆盖率较低，使得在中文开放领域中很难得到广泛应用。由此可见，在 NLP 领域仍然缺少针对中文短文本的关系抽取算法和系统的研究，本文的研究工作可以在一定程度上弥补这一缺陷。

4.2.2 常识性关系抽取

常识性关系抽取与本章的研究工作也密切相关，这是因为中文短文本中表达的语义关系包含大量常识性关系。常识性知识的抽取与通用关系抽取有较大差别，因为常识性关系一般很少在文本中显示地进行表达，计算机对常识的获取、处理和理解都存在困难。在人工智能技术发展的早期，常识性知识一般通过专家人工编纂，或者采用网络众包的形式获得。在 CYC 项目中 [192]，相关项目专家将超过百万条常识性知识编成机器可读的逻辑表达形式，利用上述知识库进行常识逻辑推理。ConceptNet [179] 起源于 MIT 发起的网络众包 Open Mind Common Sense 计划 [1]，以关系三元组的形式，汇集了不同知识源中概念之间的常识性关系，与 CYC 系统中的常识性知识相比，其可读性、可理解性更高。

常识性知识的自动获取主要依赖于模式匹配法。WebChild 系统 [178] 采取迭代

[1] https://www.media.mit.edu/projects/open-mind-common-sense/

式学习方法，自动从网络语料库中抽取多种类型的常识性知识，包括 “has-shape”、“has-taste”、“evokes-emotion” 等。Narisawa 等人 [193] 关注数值型常识性知识的获取和推理，例如判断身高为 204cm 的男性成年人的身高是高还是矮。在这一算法中，Narisawa 等人从网络文本中抽取数值表达式及其对应的上下文，同时利用分布式词向量和上下文模式对数值的大小进行分类。另一类研究特别着重于空间的常识性知识。Collell 等人 [194] 挖掘隐示的空间关系表达（例如 “glass on table”、“man riding horse”）及其相应文本模板，用于基于文本的空间位置常识推理，例如判断实体在空间中的相对位置。Xu 等人 [195] 提出了句子级别的神经网络关系分类模型，用于判断实体之间是否具有 “location-near”（“位置靠近”）的空间常识性关系。随着深度语言模型的迅猛发展，它在常识性关系的表示学习上也有相应应用。Bosselut 等人 [196] 提出了基于 ConceptNet [179] 的 Transformer 模型，采用多重注意力机制学习 ConceptNet 图谱中概念和关系的表达，用于常识性知识图谱的表示学习和补全。

4.2.3 名词短语的习语性分析

由于名词短语在自然语言中分布广泛，且语义表达丰富多变，名词短语的习语性分析成为理解名词短语含义的关键。**习语分类**（Idiom Token Classification）是习语性分析的重要任务之一，目的是判断名词短在特定语境下表达的是字面含义还是习语性含义。习语分类的研究始于 Hashimoto 和 Kawahara 的研究 [197]，他们提出一系列基于习语的特征，采用 SVM 分类器判断日语词组为字面含义或习语性含义。Peng 等人 [198] 指出，由于习语的含义与通常语言不同，习语的主题分布与其他词汇有明显区别，并且提出基于主题模型和情绪表达的习语分类算法。随着词嵌入技术的快速发展，词嵌入模型在习语分类任务上应用越加广泛。例如，Salton 等人 [199] 使用包含目标短语的句子作为神经网络模型的输入，利用目标短语及其上下位的词嵌入表示学习分类模型。Gharbieh 等人 [200] 在实验研究中指出，使用词嵌入作为特征，无论是监督式还是非监督式习语检测的模型精度都比使用经典特征的模型有明显提升。King 和 Cook [201] 的模型基本架构与前述研究相似，他们加入了词汇和语法的语言学知识，提升了基于词嵌入模型的监督式模型效果。Liu 和 Hwa [202] 同样考虑语言学知识在习语检测任务中的作用，提出了一种习语使用含义度量，用于衡量特定习语在文本中为字面含义的度量。上述算法都对名词短语的习语性分析有相当的贡献，然而这些方法的目标都是英语语言，对中文缺乏特定

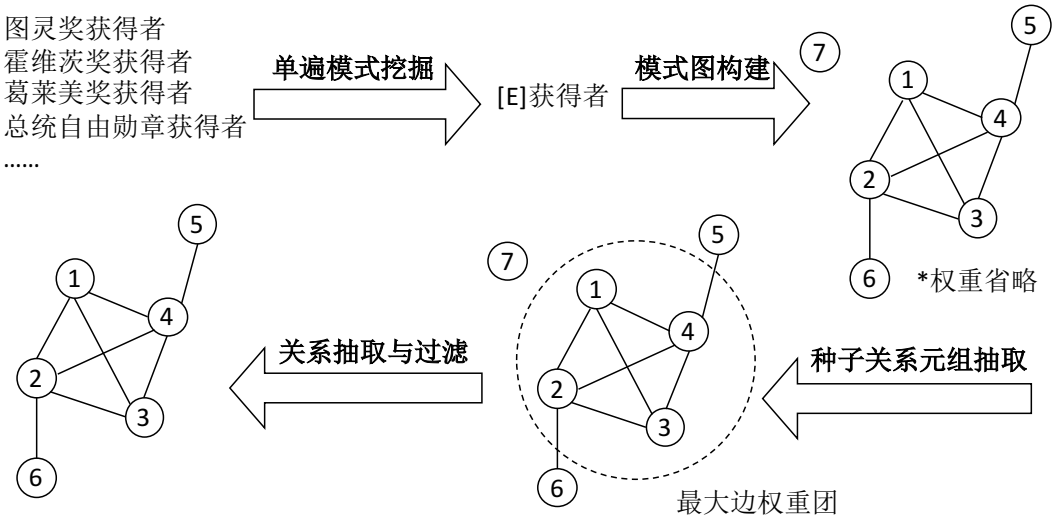
语言特性的研究。

另一个与习语性分析紧密相关的任务为**复合名词**（Noun Compound）的**组合性分析**（Compositionality Analysis），即给定某复合名词（例如“orange juice”、“cloud nine”），判断在这一复合名词中的两个名词语义可分的程度。这一任务的早期研究关注于向量空间中复合名词中构成名词的表示，并且设计一系列**组合性度量**（Compositionality Measures），定量表示其语义可分性，研究工作包括 [203, 204] 等。与习语分类任务相似，词嵌入在复合名词的组合性分析中有广泛应用。例如，Salehi 等人 [205] 结合了词嵌入模型和先前 Reddy 等人 [203] 提出的组合性度量，用于检测语义不可分的英语复合名词。Yazdani 等人 [206] 和 Cordeiro 等人 [207] 分别提出了一系列基于深度学习的分布式语义模型，用于学习多词表达式的不可分性。将复合名词的组合词嵌入与复合名词中的两个词各自的词嵌入相对比，可以计算多个组合性度量的值。

与上述研究相比，中文复合名词的组合性分析在 NLP 领域研究不够充分，因为与之相应的语言学知识很难建模，并通过机器学习模型学习。Qi 等人 [208] 将中文语义字典 HowNet 中的中文义原信息加入组合性预测模型，在已知中文短语所有可能对应义原的情况下，学习中文短语的语义向量表示。然而，这一方法只适应于能被中文义原覆盖的中文短语，对任意中文语言的适应性较低。在我们提出的 RCRL 模型中，我们同时考虑语言模式和词嵌入信息，不依赖于义原字典可以对中文复合名词进行习语性程度分类，对中文自然语言理解的推进更加适合，也有利于词语关系的扩展与推理。

4.3 基于模式挖掘的非上下位关系抽取

从本节开始，我们分别详述三种面向中文短文本的非上下位关系抽取与语义理解算法。首先，由于非上下位关系的语义关系种类繁多，在学术界缺乏相关的人工标注集用于关系抽取模型的训练，而且这一类别的数据集人工标注费时费力。在本节中，我们提出面向中文短文本的**基于模式挖掘的非上下位关系抽取算法**（PNRE）。它首先挖掘中文短文本中频繁出现的、描述语义关系的语言模式，然后采用图算法无监督地挖掘出相应语言模式对应的非上下位关系三元组。



关系映射

关系谓词：
获奖

ID	x_i	y_i^p
1	东尼·霍尔	图灵奖
2	古特·布洛伯尔	霍维茨奖
3	约翰·科克	美国国家科学奖
4	姚期智	图灵奖
5	托尼·布莱尔	总统自由勋章
6	爱德华·路易斯	诺贝尔生理学或医学奖
7	回声音乐奖	奖项

图 4.2: PNRE 的关系抽取流程（以“获奖”关系为例）

4.3.1 算法模型

PNRE 算法包括五个主要步骤：**单遍模式挖掘**、**模式图构建**、**种子关系元组抽取**、**关系抽取与过滤**和**关系映射**。在图 4.2中，我们以“获奖”这一种语义关系为例，简要描述了 PNRE 的算法流程。在下文中，我们详细介绍 PNRE 算法的各个步骤的技术细节。

单遍模式挖掘： 由于中文短文本中包含的关系类别和用于抽取该关系的语言模式都未知，这一模块自动从短文本中挖掘出频繁出现的、具有很大概率描述某种特定语义关系的模式。考虑某中文短文本对 (x_i, y_i)，x_i 为某中文实体或概念，y_i 为描述 x_i 的中文短文本，其中可能包含 x_i 的非上下位关系。典型的示例如“（蒂姆·伯纳斯·李，图灵奖获得者）”、“（马云，1964 年出生）”等。在 PNRE 算法中，我们记某特定语言模式为 p，包括中文词语序列和中文实体占位符“[E]”。例如，从中文短文本“图灵奖获得者”中可以挖掘出语言模式“[E] 获得者”，“[E]”可以代表

表 4.2: 中文术语对及其相应语言模式匹配示例

中文实体 x_i	**关系类别术语** y_i	**语言模式** p
黑客帝国 2: 重装上阵	人工智能题材作品	[E] 题材作品 $\Rightarrow y_i^p$ =“人工智能”
奥尔良	卢瓦雷省市镇	[E] 省市镇 $\Rightarrow y_i^p$ =“卢瓦雷”
教父 (电影)	奥斯卡最佳男主角获奖电影	[E] 获奖电影 $\Rightarrow y_i^p$ =“奥斯卡最佳男主角”
维珍航空	1984 年成立的航空公司	[E] 成立的航空公司 $\Rightarrow y_i^p$ =“1984 年”

任意类型的中文实体或概念。

对于某中文短文本对 (x_i, y_i)，如果 y_i 可以匹配语言模式 p，我们记 y_i^p 为 y_i 匹配模式 p 中实体占位符 “[E]” 的实体。例如，如果 x_i 为 “蒂姆·伯纳斯·李”，y_i 为 “图灵奖获得者”，p 为 “[E] 获得者”，我们可以得到 y_i 中模式 p 对应的关系宾语 y_i^p 为 “图灵奖”。据此，令 $R_p = \{(x_i, y_i^p)\}$ 为所有通过模式 p 匹配生成的候选关系元组集合。我们可以看出，R_p 包括了语言模式 p 对应的所有可能的语义关系元组，我们可以从 R_p 中进一步挖掘和筛选，以生成高精度的非上下位关系元组。表 4.2给出了中文维基百科中的中文实体-关系类别术语对，及相应语言模式匹配示例。在表中，x_i 为中文维基百科的实体，y_i 为 x_i 对应维基百科页面中的某一个关系类别术语。

本模块的关键步骤是获得所有可能表示语义关系的语言模式。我们定义长度 $length(p)$ 为语言模式 p 中包括的字数（不包括中文实体占位符 “[E]”）。模式 p 对应的**支持度** $supp(p)$ 可以按照下式计算：

$$supp(p) = |R_p| \cdot \ln(1 + length(p))$$

其中，$\ln(1 + length(p))$ 增大了更长的语言模式的支持度，因为长语言模式在表达上更加具体，更可能描述某种特定的语义关系。

在本模块的实现中，我们使用基于 CRF 的中文命名实体标注模型 [209] 和包括所有中文维基百科实体的字典，识别出所有关系类别术语 y_i 中的实体，并且挖掘出相应的语言模式。它只需要对所有中文短文本对 (x_i, y_i) 进行单遍扫描，计算出所有语言模式的支持度。因为低支持度的语言模式一般并不频繁出现，多数为噪声，我们只使用 Top-k 支持度的语言模式 p，以及对应的候选关系元组集合 R_p 作为下一步算法的输入。

模式图构建： 从上一步抽取出的候选关系元组 R_p 并不一定正确。考虑图 4.2中的示例，中文维基百科中的实体 “回声音乐奖” 有对应的关系类别术语 “奖项获得者”。利用语言模式 “[E] 获得者”，我们可以抽取出候选关系元组 “（回声音乐奖, 奖项）”，这两个术语之间显然没有 “获奖” 语义关系。为了在没有人工干预的情况下实现高精度的关系抽取，在本模块中，对于每个具有 Top-k 支持度的语言模式 p，我们从 R_p 中选出一个子集 R_p^*，作为种子关系元组集合。这些种子关系元组在大概率上是正确的，为下一步的关系抽取提供知识基础。

我们提出了无监督的、基于图挖掘的算法，用于从 R_p 中筛选出种子元组 R_p^*。令 $G_p = (C_p, L_p, W_p)$ 为关于 p 的无向的、边带权重的**模式图**（Pattern Graph）。C_p、L_p 和 W_p 分别表示 G_p 中的节点集合、边集合和边权重集合。节点集合 C_p 中的每一个元素分别对应模式 p 匹配出的实体 y_i^p，可以表示为：$C_p = \{y_i^p | (x_i, y_i^p) \in R_p\}$。边的权重 W_p 描述了 C_p 中实体之间的语义相似度。在本研究中，描述中文术语非上下位关系的部分实体 y_i^p 比较长（例如 “奥斯卡最佳男主角”），很难直接采用现有的词嵌入向量计算语义相似度。我们同样使用中文维基百科中的实体-关系类别结构作为知识源。对于 C_p 中的某实体 y_i^p，令 $Cat(y_i^p)$ 为维基百科中 y_i^p 对应的类别集合。给定 C_p 中的任意两个实体 y_i^p 和 y_j^p，我们按照下式计算其语义相似度：

$$sim(y_i^p, y_j^p) = \frac{\sum_{c \in Cat(y_i^p)} \sum_{c' \in Cat(y_j^p)} \cos(\vec{c}_h, \vec{c}_h')}{|Cat(y_i^p)| \cdot |Cat(y_j^p)|}$$

其中，我们用 $\vec{c}_h$ 和 $\vec{c}_h'$ 分别表示 $c \in Cat(y_i^p)$ 和 $c' \in Cat(y_j^p)$ 的**核心词**（Head Word）的词向量。在这一步中，我们只利用核心词的词向量进行计算，同样因为这些类别一般长度较长，在语料库中词频过低，很难用词嵌入模型精确计算其语义。

给定相似度阈值 τ，当且仅当 $sim(y_i^p, y_j^p) > \tau$，我们在图 G_p 上加入边 $(y_i^p, y_j^p) \in L_p$，以及权重 $w(y_i^p, y_j^p) = sim(y_i^p, y_j^p)$。所以，在图 G_p 中，如果 C_p 中的实体具有相似的语义，对应的节点之间有边连接。以先前的语言模式 “[E] 获得者” 为例，对应实体 “图灵奖”、“霍维茨奖” 和 “诺贝尔生理学或医学奖” 都比较相似，他们之间应该互相有边连接。

种子关系元组抽取： 从 R_p 中选出子集 R_p^* 的问题可以建模成**最大边权重团问题**（Maximum Edge Weight Clique Problem，缩写为 MEWCP）[210]。NEWCP 的目标是从一个具有边权重的无向图中检测出一个团，使得这个团中节点之间的边的权重之和在所有团中是最大的。在本算法中，我们从 R_p 中检测出最大边权重团

Algorithm 11 MEWCP 的近似求解算法

初始化图 $G_p^* = (C_p^*, L_p^*)$，其中 $C_p^* = \emptyset$，$L_p^* = \emptyset$
while $L_p \neq \emptyset$ **do**
  从 L_p 中采样 (y_i^p, y_j^p)，选出 (y_i^p, y_j^p) 的概率正比于 $w(y_i^p, y_j^p)$
  $C_p = C_p \setminus \{y_i^p, y_j^p\}$
  $C_p^* = C_p^* \cup \{y_i^p, y_j^p\}$
  $L_p = L_p \setminus \{(y_i^p, y_j^p)\}$
  $L_p^* = L_p^* \cup \{(y_i^p, y_j^p)\}$
  for 每条边 $(\tilde{y}_i^p, \tilde{y}_j^p) \in L_p$ **do**
    if $\tilde{y}_i^p \notin C_p^*$ 且 $\tilde{y}_j^p \notin C_p^*$ **then**
      $C_p = C_p \setminus \{\tilde{y}_i^p, \tilde{y}_j^p\}$
      $L_p = L_p \setminus \{(\tilde{y}_i^p, \tilde{y}_j^p)\}$
    end if
  end for
end while
return 最大边权重团 C_p^*

C_p^*，根据 C_p^* 的构成生成种子关系元组集合 R_p^*。该问题的目标函数定义如下：

$$\max \sum_{(y_i^p, y_j^p) \in L_p^{'}} w(y_i^p, y_j^p)$$
$$\text{s.t. } L_p^{'} \subseteq L_p, \forall y_i^p, y_j^p \in C_p^* (y_i^p \neq y_j^p), (y_i^p, y_j^p) \in L_p^{'}$$

其中，$L_p^{'}$ 是边的集合，这些边的每个节点都在需要检测出的最大边权重团 C_p^* 中。

在最优化的研究中，很多算法旨在获得 MEWCP 的精确解，例如非限制性二次规划方法 [210]。然而，由于 MEWCP 是 NP 难问题，这些算法的时间复杂度较大，实际应用空间有限。PNRE 算法采用基于蒙特卡洛的近似算法解决这一问题，其算法过程详见算法 11。在算法的初始阶段，我们用一个空的图 G_p^* 来存放最大边权重团。在每个迭代中，算法从 G_p 中随机采样一条边 (y_i^p, y_j^p)，其概率正比于权重 $w(y_i^p, y_j^p)$。当某一特定边 (y_i^p, y_j^p) 被选出后，算法将其加入 G_p^*，并从原图 G_p 中移除 (y_i^p, y_j^p) 以及任何其他与 C_p^* 中节点都不连接的边。这一过程迭代进行，直到 G_p 中没有边可以加入 G_p^*。所以，G_p^* 中的节点组成了所需的最大边权重团 C_p^*。根据 C_p^* 中的节点元素，我们可以从 R_p 中对应选出种子关系元组 R_p^*，如下所示：

$$R_p^* = \{(x_i, y_i^p) | y_i^p \in C_p^*, (x_i, y_i^p) \in R_p\} \tag{4.1}$$

因为上述算法是随机近似算法，其平均算法复杂度与输入的图结构密切相关；

它在最差情况下的时间复杂度为 $O(|L_p|^2)$。我们运行上述算法多次，获得多个输出结果，然后选出具有最大边权重的团生成最终的结果 R_p^*。因此，这一 NP 难问题可以在多项式的时间复杂度内解决。此外，尽管这一算法并不保证完全的正确性，我们的实验结果表明，即使生成的团并非是最大边权重团，得到的种子关系元组仍然大概率是正确的。所以，提出的这一算法能准确、高效地获取种子关系元组。

关系抽取与过滤： 当某语言模式 p 对应的种子关系元组 R_p^* 抽取出后，我们也可以用 R_p^* 的质量衡量模式 p 的**置信度**。如果模式 p 清晰地描述了某个语义关系，则无论是 R_p^* 的大小还是 C_p^* 中节点间边权重之和都会比较大。我们按这两个特征定义语言模式 p 的未归一化置信度 $conf^*(p)$，如下所示：

$$conf^*(p) = \frac{\ln(1+|R_p^*|)}{|R_p^*| \cdot (|R_p^*|-1)} \sum_{y_i^p, y_j^p \in C_p^*, y_i^p \neq y_j^p} sim(y_i^p, y_j^p)$$

其中，$\frac{\sum_{y_i^p, y_j^p \in C_p^*, y_i^p \neq y_j^p} sim(y_i^p, y_j^p)}{|R_p^*| \cdot (|R_p^*|-1)}$ 是团中实体相似度的平均值，$\ln(1+|R_p^*|)$ 使具有更大的团的语言模式的置信度更高。为了使置信度的数值范围归一化到 $[0,1]$ 内，我们计算归一化置信度 $conf(p)$，如下所示：

$$conf(p) = \frac{conf^*(p)}{\max_{p' \in P} conf^*(p')}$$

其中，P 是所有支持度 Top-k 的语言模式的集合。根据上述公式，我们可以过滤掉具有低置信度的语言模式。

对于剩余的语言模式，给定每个候选关系元组 $(x_i, y_i^p) \in R_p$，如果 $(x_i, y_i^p) \in R_p^*$ 或 (x_i, y_i^p) 与 R_p^* 中的种子关系元组足够相似，我们将其加入最终抽取出的关系元组集合 R_p' 中。例如，在图 4.2中，我们的算法抽取出“（托尼·布莱尔，总统自由勋章）”，而舍弃“（回声音乐奖，奖项）”。这是由于“总统自由勋章”与团中的实体（其他奖项名称）在语义上更加相似，“奖项”这一抽象的概念与团中的实体并不相似。记 γ 为控制关系抽取中精准度与召回率之间的权衡参数，γ 的取值越大，算法越重视精准度，忽略召回率的重要性。如果满足下述条件，我们将候选关系元组 $(x_i, y_i^p) \in R_p \setminus R_p^*$ 加入 R_p'：

$$\frac{\sum_{y_j^p \in C_p^*} sim(y_i^p, y_j^p)}{|C_p^*|} > \frac{\gamma \cdot \sum_{y_j^p, y_k^p \in C_p^*, y_j^p \neq y_k^p} sim(y_j^p, y_k^p)}{|R_p^*| \cdot (|R_p^*|-1)} \tag{4.2}$$

Algorithm 12 PNRE 中的关系抽取算法

从 R_p 构建模式图 $G_p = (C_p, L_p, W_p)$
利用算法 11检测最大边权重团 C_p^*
根据式 4.1生成种子关系元组集合 R_p^*
初始化 $R_p^{'} = R_p^*$
for 每个候选关系元组 $(x_i, y_i^p) \in R_p \setminus R_p^*$ **do**
  if (x_i, y_i^p) 满足式 4.2定义的条件 **then**
    更新 $R_p^{'} = R_p^{'} \cup \{(x_i, y_i^p)\}$
  end if
end for
return 抽取出的关系元组集合 $R_p^{'}$

由此可见，PNRE 中的关系抽取过程检测出最有可能正确的元组当成 “种子”，然后抽取出与种子足够相似的关系元组作为结果输出。因为在面向短文本的关系抽取中，高精度是很难保证的，我们不采用迭代抽取算法来避免 “语义漂移” 问题 [62]。这一关系抽取算法的过程汇总见算法 12。

关系映射： 关系元组集合 R_p、$R_p^{'}$ 和 R_p^* 都只包含关系主语和关系宾语，不包含关系谓词。PNRE 的最后一个步骤为根据语言模式 p 的语义，生成 $(x_i, y_i^p) \in R_p^{'}$ 的关系谓词。一共可以分为如下三种映射情况：

- **直接动词映射**：如果语言模式的核心词为动词，我们可以直接将其作为关系谓词。例如，模式 “[E] 出生” 的核心词 “出生” 为动词，关系三元组 “（蒂姆·伯纳斯·李，出生，1955 年）” 可以直接从上述模式和抽取出的二元组 “（蒂姆·伯纳斯·李，1955 年）” 进行映射得到。

- **直接非动词映射**：如果语言模式中不包含动词, 其语义关系由非动词表达, 我们可以人工定义关系谓词，或利用语义字典生成其语义关系对应的动词。之后，可以采用与直接动词映射相同的方法生成关系三元组。例如，模式 “[E] 获奖者” 中，“获得者” 为名词，表达了 “得奖” 的语义关系。我们可以人工订制这一关系生成的逻辑表达式，从 “（蒂姆·伯纳斯·李，图灵奖）” 中推导出关系三元组 “（蒂姆·伯纳斯·李，得奖，图灵奖）”。

- **间接映射**：与 YAGO [8] 类似，少数语言模式不表达直接的语义关系，而间接暗示了其他语义关系，这种情况下应当人工定制映射关系。例如，模式 “[E] 军事” 暗示了相关实体与 “军事” 有关，我们定义新的关系谓词 “话题”，并生成对应实体与 “军事” 之间的 “话题” 关系。

综上所述，在 PNRE 算法的所有流程中，我们只需要在部分直接非动词映射和间接映射的情况下进行人工定制关系谓词和关系映射的逻辑表达式，不需要任何训练数据标注的过程。因此，这一方法只需要极小的人工干预工作，就能从中文短文本中抽取出多种非上下位关系。

4.3.2 实验分析

在本节中，我们以中文维基百科的实体-关系类别结构作为数据源，从多个方面对 PNRE 算法的效果进行综合评测，并与其他基于维基百科的短文本关系抽取算法进行对比。

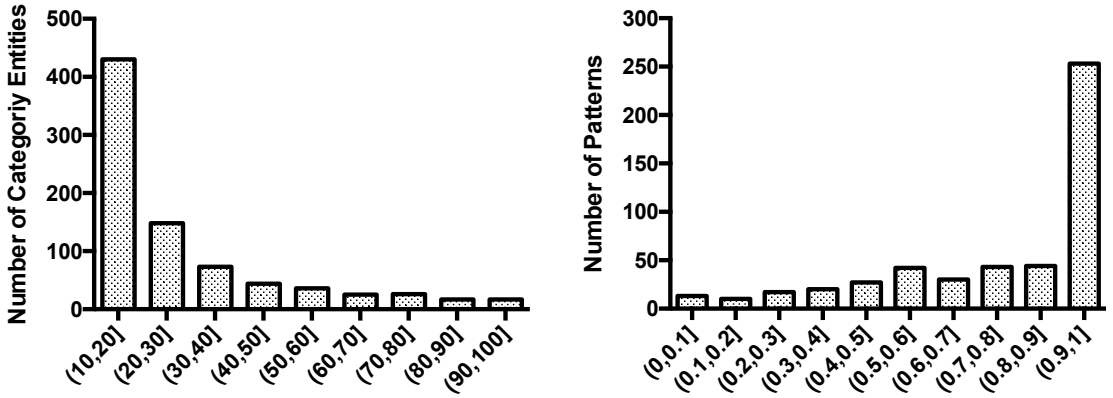
数据源与实验设置： 由于中文维基百科中具有大量中文实体-关系类别对，我们将其作为评测 PNRE 算法的数据源。在本组实验中，我们下载了 2017 年 1 月 20 日版本的中文维基百科全部数据 [2]，使用启发式规则过滤了部分不描述中文实体的维基百科页面数据（包括消歧义、重定向、模板、列表页面），最后获得了约 60 万个中文实体和 240 万个中文实体-关系类别对，作为 PNRE 算法的抽取数据源。在算法实现中，我们采用 FudanNLP 开源工具 [209] 进行基本的中文 NLP 分析，包括中文分词、词性标注等。与第 2.3.2节采用的词嵌入模型相同，我们使用中文词级别的 Skip-Gram 模型 [51] 生成中文词向量，词向量的维度设为 100。从中文维基百科中抽取出的关系数据集已在 GitHub 上开源 [3]。

实验步骤与算法分析： 我们首先运行 PNRE 的单遍模式挖掘算法，从关系类别中抽取频繁语言模式。从输出结果可以发现，对应候选关系元组的数量小于 20 的语言模式一般为噪声，缺乏明显的语义。在图 4.3(a)中，我们给出了语言模式对应候选关系元组数量的分布。因为数值大小过于极端，在 $(10, 100]$ 范围之外的统计数量已在图中省略。根据上述结果，我们选择支持度位于 Top-500 的语言模式及对应的候选关系元组作为下一步骤的输入。这些语言模式对应候选关系元组的数量都在 20 以上。在表 4.3中，我们也给出了高支持度和低支持度的语言模式示例。从中可以发现，高支持度的模式通常有非常明晰的语义，对应特定的语义关系；与之相反，低支持度的模式语义比较模糊。

对于上一步骤中选定的语言模式，我们默认设置 $\tau = 0.7$，构建模式图，并且运行 MEWCP 的近似求解算法三次，选择边权重之和最大的团作为结果，以保证

[2] http://download.wikipedia.com/zhwiki/20170120/

[3] https://chywang.github.io/data/tkde.zip



(a) 语言模式对应候选关系元组的数量分布　　(b) 高支持度语言模式的置信度分布

图 4.3: 语言模式的支持度和置信度分布

表 4.3: 具有较高或较低的支持度和置信度的语言模式示例

类别	语言模式	支持度或置信度得分
高支持度	[E] 校友	2316
	[E] 出生	1253
低支持度	[E] 地区	14
	国际 [E]	12
高置信度	[E] 州城市	0.99
	[E] 副校长	0.98
低置信度	[E] 地理	0.10
	[E] 事故	0.06

选出的种子关系元组的高准确性。图 4.3(b)展示了这些高支持度语言模式的置信度分布，可见绝大部分语言模式的置信度都比较高。在表 4.3中，我们同样列出了高置信度和低置信度的语言模式示例。我们选出了置信度位于 Top-250 的语言模式作为关系抽取和过滤步骤的输入，这些语言模式的置信度都大于 0.9。为了获得 γ 的较优取值，我们进行了初步实验，在不同 γ 取值下随机采样了 200 个关系元组，估计其准确度。我们发现，当 γ 的取值比较小时（例如 0.2），关系抽取的准确度已经可以超过 90%。在最后一个步骤中，26 个关系谓词可以通过直接动词映射生成，例如“建立”、“出生”、“废除”等。我们对于剩下的 16 种关系类别人工定制了关系谓词和映射规则，相关示例参见表 4.4。

准确度和覆盖度测试： 为了评测抽取出的关系的质量，在本组实验中，我们进行了两组测试：准确度测试和覆盖度测试。为了评测抽取出的关系的准确度，我们参考 YAGO 系统 [8] 的评测方法，对每种关系类别，随机采样 200 个关系元组，人工检查其准确性。覆盖度测试用于衡量抽取出的关系是否存在于已有的中文知

表 4.4: 人工定制的关系映射规则示例

语言模式	人工定制的关系谓词
[E] 校友	毕业
[E] 队教练	执教
[E] 省市镇	位于
[E] 获得者	获奖

表 4.5: PNRE 抽取出的 8 种关系的关系元组数量、准确度和覆盖度统计

关系类别	关系元组数量	准确度	覆盖度
毕业	44,118	98.0%	22.9%
位于	29,460	97.2%	8.5%
建立	20,154	95.0%	31.5%
出生	11,671	98.3%	41.4%
成员	8,445	96.0%	4.2%
启用	8,956	98.2%	21.6%
逝世	5,597	100.0%	18.4%
得奖	3,262	90.0%	27.3%

识图谱中。如果不存在，则表明 PNRE 可以抽取出不能被现有方法覆盖的新关系，对现有的中文知识图谱有补全的作用。在实验中，我们使用 CN-DBpedia V2.0 [11] 作为基准中文知识图谱，截止至 2017 年 2 月，它一共包含了约 900 万个中文实体的 4100 万个语义关系。我们使用 CN-DBpedia API [4]获取 CN-DBpedia 中相关中文实体的关系。对于每一种语义关系 r，我们计算其覆盖度 $cov(r)$ 如下：

$$cov(r) = \frac{\text{\#能被 CN-DBpedia 覆盖的、PNRE 抽取出的正确的语义关系}}{\text{\#PNRE 抽取出的正确的语义关系}}$$

因为相同含义的语义关系在不同的知识图谱中表达可能不同，为了使评测更加公平，我们人工评测 PNRE 抽取出的语义关系是否被 CN-DBpedia 覆盖。在表 4.5中，我们给出了 8 种非上下位关系的关系元组数量、准确度和覆盖度。对于这 8 种关系的每一种，PNRE 都抽取出超过 3000 个关系元组。

从实验结果可以看出，这 8 种语义关系的准确度都超过了 90%，对于某些特定的语义关系，准确度超过了 98%，甚至达到 100%。这可以表明，PNRE 可以准确可靠地从中文短文本中自动挖掘出非上下位关系。与准确度的结果不同，不同类别的语义关系覆盖度结果差别很大。某些关系（例如“出生”、“建立”）在 CN-DBpedia 中的覆盖率相对较高，其他语义关系对应的关系元组在 CN-DBpedia 知识图谱中几

[4]http://knowledgeworks.cn:20313/cndbpedia/api/entityAVP

表 4.6: CN-WikiRe 使用的三类语言模式

模式类别	语言模式	示例
成员模式	[E] 成员/总统	中国科学院成员
动词-名词词组模式	[E]+ 动词 +(的)+ 名词词组	1990 年建立的组织
动词模式	[E]+ 动词	1980 年出生

乎不出现。整体而言，这 8 种语义关系的平均覆盖率为 21.1%。尽管 CN-DBpedia 等中文知识图谱的规模相对较大，它包含的知识仍然缺乏。此外，CN-DBpedia 中的语义关系大部分从网络百科的半结构化的信息框中直接抽取，其形式一般为属性-值对 [13, 211]。与之不同，PNRE 直接从无结构化的短文本中进行抽取，可以作为前述方法的补充。

与其他方法的比较： 由于 PNRE 从中文维基百科关系类别中进行关系抽取，在 NLP 已有研究中没有标准的评估和比较方法。回顾相关工作的讨论，由于 PNRE 只抽取频繁出现的语言模式涉及的语义关系，传统和开放关系抽取的算法由于应用情景不同，很难作为基线算法。此外，中英文语言的极大差异性使得现有基于英语维基百科的关系抽取算法不能与 PNRE 直接进行比较。在 YAGO [8] 中，作者采用固定的正则表达式匹配法从维基百科关系类别抽取出 9 种非上下位关系，准确度在 90% 至 98% 之间。我们的方法可以视为在中文语言上的扩展，在中文语境下抽取出的更多种类的关系，而且准确度与 YAGO [8] 相似。

我们进一步将 PNRE 与 [33] 相比较，这一研究旨在从英语维基百科的关系类别中抽取关系，其方法依赖于英语介词短语的语言模式。在中文中，这些表达通常是隐示的，因此 [33] 不能直接在中文维基百科数据上应用。按照这一方法的思想，我们实现了用于中文维基百科的变体（记为 CN-WikiRe）。CN-WikiRe 使用的语言模式见表 4.6。在实验中，CN-WikiRe 抽取了 631 种 165048 个非上下位关系元组。尽管 CN-WikiRe 检测出的关系类别数量远大于 PNRE，然而，只有 14% 为真正的关系谓词，其余均为错误或噪声。这是因为中文短文本的分词和词性标注仍然准确度有限，且被 CN-WikiRe 抽取出的很多动词并不能作为关系谓词（例如“传导”、“缩小”等）。在排除这些非关系谓词相关的三元组后，我们随机采样了 500 个关系元组，人工标注其是否准确，其准确度为 58.6%。由此可见，CN-WikiRe 的实验效果远低于 PNRE，体现出面向英语的这一关系抽取算法不能直接运用于中文语境下。

我们也实现了 PNRE 的两种变体：PNRE-Conf 和 PNRE-Filter。其中，PNRE-Conf 不使用基于图挖掘的算法对候选关系元组进行选择，PNRE-Filter 不进行关系过滤。与 CN-WikiRe 相同，我们计算了 PNRE-Conf 和 PNRE-Filter 抽取出的关系

表 4.7: PNRE 与其变体的准确度比较

方法	**预估准确度**
PNRE-Conf	74.4%
PNRE-Filter	94.2%
PNRE	**97.4%**

表 4.8: DNRE 的输入输出及其示例

类别	**符号**	**示例**
输入	x_i	比利时
	y_i	19 世纪建立的西欧国家
MPS	$ws(y_i)$	$w_i^{(1)}$ =19，$w_i^{(2)}$ = 世纪，$w_i^{(3)}$ = 建立，$w_i^{(4)}$ = 的， $w_i^{(5)}$ = 西欧，$w_i^{(6)}$ = 国家
	$ps(y_i)$	$q_i^{(1)}$ =19 世纪建立的（修饰词），$q_i^{(2)}$ = 西欧（修饰词）， $q_i^{(3)}$ = 国家（核心词）
CRG	$R(x_i, y_i)$	$r(x_i, q_i^{(1)})$ =（比利时，建立，19 世纪） $\tilde{r}(x_i, q_i^{(2)})$ =（比利时，?，西欧）
MRPD	$\tilde{r}_{v^*}(x_i, q_i^{(j)})$	$\tilde{r}_{v^*}(x_i, q_i^{(2)})$ = 位于
输出		（比利时，建立，19 世纪） （比利时，位于，西欧）

的准确度，其结果汇总在表 4.7中。从实验结果可见，我们提出的 PNRE 比这两种变体提升了 23.0% 和 3.2% 的准确度。因此，本节中介绍的 PNRE 算法对抽取频繁模式对应的语义关系是高度有效的。

4.4 数据驱动的非上下位关系抽取

前述 PNRE 的算法可以在无监督的情况下自动从中文短文本中挖掘中多种类型的语义关系。然而，这种方法的一个缺点是它只挖掘了频繁语言模式对应的关系，由于语义关系在文本中的表达一般具有 “长尾效应” [46]，长尾关系容易被 PNRE 忽略，很难被抽取出来。本节在开放关系抽取的框架下，进一步提出了**数据驱动的非上下位关系抽取算法**（DNRE），从短文本解析和理解的角度，抽取更多数量、更多类别的非上下位关系。

4.4.1 算法模型

在本节中，我们首先给出 DNRE 算法的整体流程，以及其中涉及的关键符号和概念。之后，我们详述 DNRE 算法的具体算法细节。

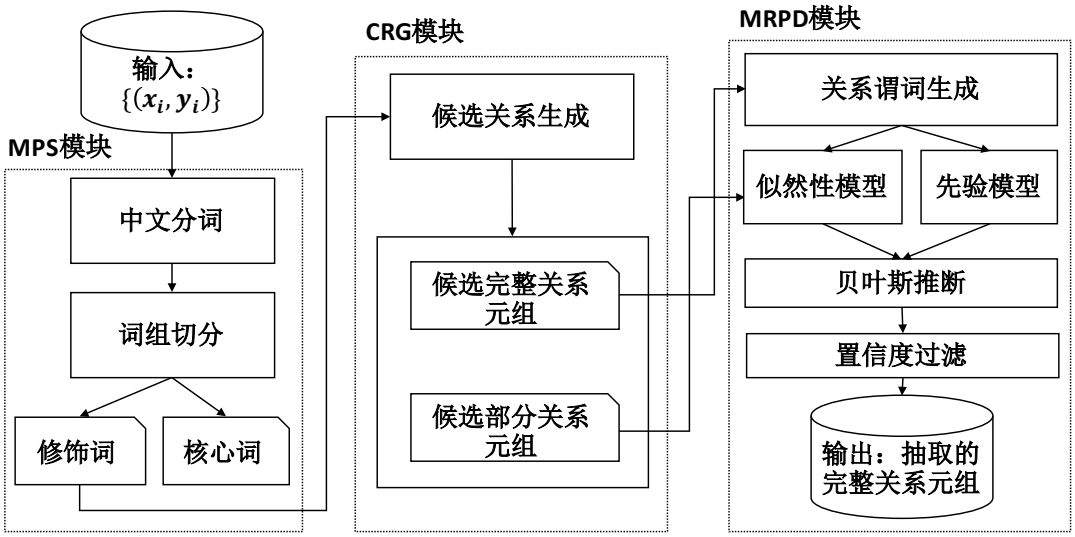


图 4.4: DNRE 系统的整体框架

算法整体流程： DNRE 的算法输入数据与 PNRE 完全相同，它主要包括三个模块：**修饰词敏感的词组切分**（MPS）、**候选关系元组生成**（CRG）和**缺失关系谓词检测**（MRPD）。DNRE 系统的整体框架如图 4.4。为了便于阅读和理解，表 4.8给出了 DNRE 的输入输出涉及的各个重要符号，以及对应的示例。

因为中文短文本的语法结构比较复杂，根据 Pasca 在研究中提出的假设 [212], MPS 的目标将描述中文实体 x_i 的短文本 y_i 切分成多个修饰词和一个核心词[5]。对于每个短文本 y_i，我们首先进行中文分词，可以获得如下结果：

$$ws(y_i) = \{w_i^{(1)}, w_i^{(2)}, \cdots, w_i^{(|ws(y_i)|)}\}$$

其中，$w_i^{(j)} \in ws(y_i)$ 是短文本 y_i 分词之后的第 j 个词。当分词结束后，MPS 生成 y_i 的修饰词敏感的切分结果：

$$ps(y_i) = \{q_i^{(1)}, q_i^{(2)}, \cdots, q_i^{(|ps(y_i)|)}\}$$

其中，$q_i^{(j)} \in ps(y_i)$ 是 y_i 中的一个修饰词或者核心词，包含了 y_i 中的一个或多个词。根据 Pasca 的研究 [212], 我们认为 $q_i^{(|ps(y_i)|)}$ 是 y_i 的核心词，$q_i^{(j)}$（$1 \leq j \leq |ps(y_i)|-1$）的修饰词。

当 MPS 运行完成后，我们生成关于 (x_i, y_i) 的实体-修饰词对 $\{(x_i, q_i^{(j)})\}$（$1 \leq j \leq |ps(y_i)| - 1$），并且利用 CRG 模块生成候选关系元组。令 $R(x_i, y_i)$ 是所有从

[5]在本研究中，我们研究的短文本 y_i 一般为名词词组。

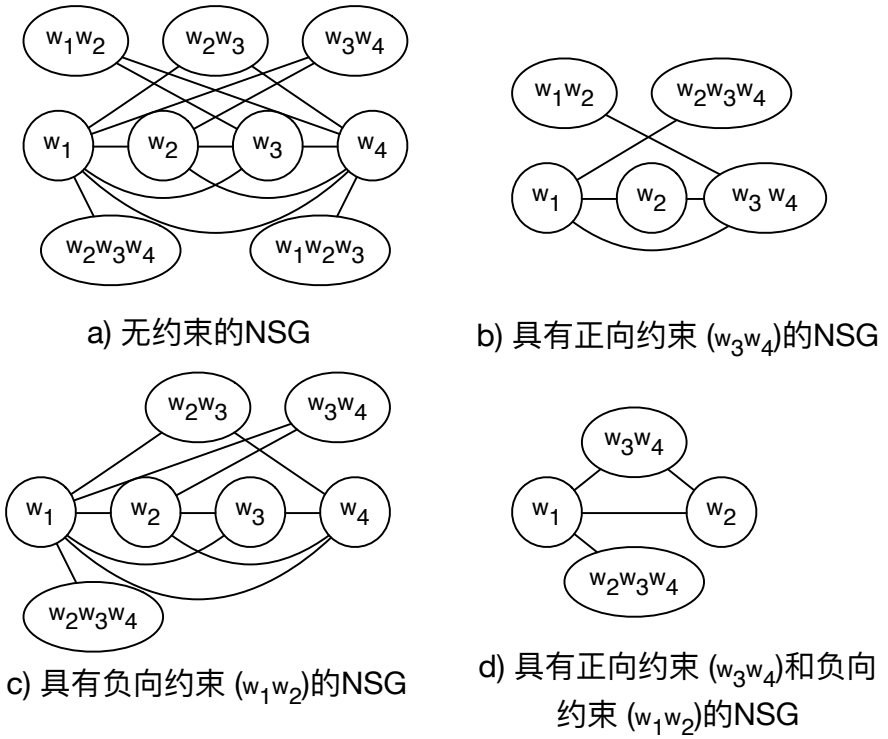


图 4.5: NSG 结构示意图，边的权重省略（在本例中，我们有 $ws(y_i) = \{w_1, w_2, w_3, w_4\}$ 和 $n = 3$）

(x_i, y_i) 中抽取出的候选关系元组集合。对于每个实体-修饰词对 $(x_i, q_i^{(j)})$，如果完整的关系谓词和宾语可以从 $q_i^{(j)}$ 中抽取出来，我们记该候选完整关系元组为 $r(x_i, q_i^{(j)})$，并且将其加入 $R(x_i, y_i)$；如果无法检测到正确的关系谓词，我们将主语 x_i 和抽取出的关系宾语作为候选部分关系元组 $\tilde{r}(x_i, q_i^{(j)})$，同样加入 $R(x_i, y_i)$。

由于 CRG 模块生成的部分候选关系元组中的谓词有缺失，MRPD 采用基于贝叶斯推断的方法补全缺失的关系谓词。令 $\mathcal{V}$ 为所有可能的关系谓词集合。我们分别学习先验模型 $\Pr(v)$ 和似然性模型 $\Pr(\tilde{r}(x_i, q_i^{(j)})|v)$，根据上述模型，推断候选部分关系元组 $\tilde{r}(x_i, q_i^{(j)})$ 的最有可能的关系谓词 $\tilde{r}_{v^*}(x_i, q_i^{(j)})$。最后，对于所有的候选完整关系元组为 $r(x_i, q_i^{(j)})$ 和已发现谓词的候选部分关系元组 $\tilde{r}(x_i, q_i^{(j)})$，我们分别计算其置信度 $conf(r(x_i, q_i^{(j)}))$ 或 $conf(\tilde{r}(x_i, q_i^{(j)}))$，并过滤低置信度关系元组，保证抽取的准确性。

修饰词敏感的词组切分： 我们介绍基于图挖掘的 MPS 算法细节。在这一步骤中，由于需要应用在开放领域下，MPS 完全由输入数据驱动，不需要任何人工标注训练数据。

MPS 的第一个步骤是中文分词，将短文本 y_i 切分若干个词语，即为 $ws(y_i) =$

$\{w_i^{(1)}, w_i^{(2)}, \cdots, w_i^{(|ws(y_i)|)}\}$。在 DNRE 中，我们将中文分词和修饰词敏感的词组切分作为两个单独的任务处理，降低了算法的复杂性。对于每个短文本 y_i 的分词结果 $ws(y_i)$，我们构建 **N-Gram 分割图**（N-gram Segmentation Graph，缩写为 NSG），表征 $ws(y_i)$ 所有可能的词组切分结果。令正整数 n 为 N-Gram 因子，用于 NSG 的构建过程，下文详细论述。

定义关于 $ws(y_i)$ 的 N-Gram 分割图 $G_n(y_i)$ 是为无向的、边带权重的图。其中，$M_n(y_i)$ 是图 $G_n(y_i)$ 的节点集合，每个节点 $m \in M_n(y_i)$ 对应从 $ws(y_i)$ 中生成的词序列，从 Uni-Gram、Bi-Gram 一直到 n-Gram。例如在图 4.5 中，给定 $ws(y_i) = \{w_1, w_2, w_3, w_4\}$ 和 $n = 3$ 作为输入，我们生成如下节点集合：$M_n(y_i) = \{w_1, w_2, w_3, w_4, w_1w_2, w_2w_3, w_3w_4, w_1w_2w_3, w_2w_3w_4\}$,包括 Uni-Gram、Bi-Gram 和 Tri-Gram。$E_n(y_i)$ 是图 $G_n(y_i)$ 的边集合，对于任意两个节点 $m_i, m_j \in M_n(y_i)$，当且仅当 $m_i \cap m_j = \emptyset$[6]，我们在这两个节点之间加入边，即 $(m_i, m_j) \in E_n(y_i)$。这是因为切分后的短文本中各个元素两两不重叠。$W_n(y_i)$ 是 $|E_n(y_i)|$ 维的边权重向量，给图中的每条边 $(m_i, m_j) \in E_n(y_i)$ 加上数值范围在 $[0, 1]$ 之间的权重 $\alpha_{i,j}$。

从上述图构建过程中可知，图 $G_n(y_i)$ 中的每一个**极大团**（Maximal Clique）分别对应短文本 y_i 的一种切分方式。例如，在图 4.5 中，$\{w_1, w_2w_3, w_4\}$ 可以构成一个极大团，因此，y_i 的一种切分方式 $ps(y_i)$ 为 $m_1 = w_1, m_2 = w_2w_3, m_3 = w_4$。为了准确起见，我们提出以下定理：

定理 4.4.1. 图 $G_n(y_i)$ 中的一个极大团对应短文本 y_i 的一种切分方式。

Proof. 根据 MPS 的任务定义，如果 y_i 的切分方式 $ps(y_i) = \{q_i^{(1)}, q_i^{(2)}, \cdots, q_i^{(|ps(y_i)|)}\}$ 是合法的，它必须构成 $ws(y_i)$ 的一种**划分**（Partition）。正确的划分必须满足两个条件：i）$\forall q_i^{(j)}, q_k^{(j)} \in ps(y_i)$，必须满足 $q_i^{(j)} \cap q_i^{(k)} = \emptyset$；以及 ii）$\bigcup_{q_i^{(j)} \in ps(y_i)} = ws(y_i)$。

考虑图 $G_n(y_i)$ 中的某个极大团 $M^{'}$。根据极大团的性质，对于 $\forall m_j, m_k \in M^{'}$，我们有：$m_j \cap m_k = \emptyset$。将团 $M^{'}$ 中的每个节点 $m_j \in M^{'}$ 分别映射到 $ps(y_i)$ 中的对应元素 $q_i^{(j)}$，易得 $\forall q_i^{(j)}, q_k^{(j)} \in ps(y_i)$，结论 $q_i^{(j)} \cap q_i^{(k)} = \emptyset$ 成立。

接下来，我们采用反证法证明条件 ii）也可以满足。假设存在极大团 $M^{'}$ 不满足条件 $\bigcup_{q_i^{(j)} \in ps(y_i)} = ws(y_i)$，则在 $M^{'}$ 外必然存在某节点 m_i^*，使得 $m_i^* \notin \bigcup_{m_i \in M^{'}}$。所以，$M^{'}$ 不是极大团，因为将 m_i^* 加入 $M^{'}$ 能构成一个更大的新团。因此，原假设不成立，结论成立。 □

[6]在没有歧义的情况下，我们同时使用 m_i 这一符号表示图 $G_n(y_i)$ 中的某个节点，以及这个节点对应的 y_i 中的 N-Gram。例如，在本句中，$m_i \cap m_j = \emptyset$ 指 m_i 与 m_j 对应的 N-Gram 不重叠。

我们同时考虑统计和分布式知识，采用混合方法计算 $W_n(y_i)$ 中的权重。如果 m_i 和 m_j 是 $ws(y_i)$ 中两个连续的 N-Gram（例如 $m_i = w_1$ 和 $m_j = w_2w_3$），统计评分 $w_s(i,j)$ 定义为**归一化点互信息**（Normalized Pointwise Mutual Information）的变体，即为

$$w_s(i,j) = \frac{1}{2} - \frac{\text{PMI}(i;j)}{2h(i,j)} = -\frac{\log \Pr(m_i)\Pr(m_j)}{2\log \Pr(m_i,m_j)}$$

其中，PMI$(i;j)$ 和 $h(i,j)$ 是 N-Gramm_i 和 m_j 在语料库中的点互信息和自信息。概率 Pr(m_i)、Pr(m_j) 和 Pr(m_i,m_j) 可以通过任意语言模型估计出。$w_s(i,j)$ 的定义可以保证统计评分在 $[0,1]$ 范围内。分布式评分 $w_d(i,j)$ 利用计算语言学中的组合性分析 [207] 进行定义，如下所示：

$$w_d(i,j) = \frac{1}{2}(1 - \cos(\vec{m}_{i\oplus j}, \vec{m}_{i+j}))$$

其中，$\vec{m}_{i\oplus j}$ 是 m_i 和 m_j 的组合词嵌入（即将 m_i 和 m_j 看成一个整体的词嵌入）。$\vec{m}_{i+j}$ 是 m_i 和 m_j 的各自词嵌入归一化之和，即为：

$$\vec{m}_{i+j} = \frac{\vec{m}_i}{\|\vec{m}_i\|} + \frac{\vec{m}_j}{\|\vec{m}_j\|}$$

如果 m_i 和 m_j 高度不可分，m_i 和 m_j 各自的上下文应该与 m_im_j 组合的上下文有显著区别。所以，$\vec{m}_{i\oplus j}$ 和 $\vec{m}_{i+j}$ 明显不相似。综上，在图 $G_n(y_i)$ 中，权重 $\alpha_{i,j}$ 的定义为这两个评分的线性组合：

$$\alpha_{i,j} = \gamma w_s(i,j) + (1-\gamma)w_d(i,j)$$

其中，$\gamma \in (0,1)$ 是预定义的超参数。

现在我们对 MPS 的算法复杂度作初步分析。为了简单起见，我们不妨令 $\zeta = |ws(y_i)|$。易知在 MPS 中，至少需要对 y_i 进行 $\lceil \frac{\zeta}{n} \rceil$ 次切分。因此，y_i 的所有可能的切分数量 Δ 为：

$$\Delta = \sum_{i=\lceil \frac{\zeta}{n} \rceil}^{\zeta-1} \binom{\zeta-1}{i} = 2^{\zeta-1} - \sum_{i=0}^{\lceil \frac{\zeta}{n} \rceil - 1} \binom{\zeta-1}{i}$$

其中，$\binom{\zeta}{i} = \frac{\zeta!}{i!(\zeta-i)!}$。在最坏情况下，蛮力搜索找到最优切分的算法复杂度为 $O(2^\zeta)$ 级别。在实际应用中，尽管 n 和 ζ 都是很小的整数，使用蛮力搜索最佳切分的计算代价仍然比较高。在下文中，我们给出高效解决这一问题的方法。

我们注意到，如果在构建 NSG 的过程中引入语言规则，不但能够显著降低 NSG 的节点数，而且可以提升 MPS 的精确度。在 DNRE 中，我们考虑两种约束条件：**正向约束**（Positive Constraint）和**负向约束**（Negative Constraint）。其中，正向约束定义在中文分词结果 $ws(y_i)$ 中连续的两个词 $w_i^{(j)}$ 和 $w_i^{(j+1)}$ 上，限制了 $w_i^{(j)}$ 和 $w_i^{(j+1)}$ 必须在 MPS 中切分到同一个单元中。在图 4.5(b) 中，我们给出了加入一个正向约束 (w_3, w_4) 后 NSG 的结构，其中 w_3 和 w_4 被看出一个整体。

类似地，负向约束同样定义在中文分词结果 $ws(y_i)$ 中连续的两个词 $w_i^{(j)}$ 和 $w_i^{(j+1)}$ 上，限制了 $w_i^{(j)}$ 和 $w_i^{(j+1)}$ 不能切分到同一个单元中。在图 4.5(c) 中，我们在原来的 NSG 上加入了一个负向约束 (w_1, w_2)。在这种情况下，所有包含 w_1w_2 的节点都被移除。将正向约束 (w_3, w_4) 和负向约束 (w_1, w_2) 同时置于原始的 NSG 上，我们可以得到图 4.5(d) 的结构。

我们对上述策略进行定量分析。对于原始的 NSG，如果不加入任何约束，其节点数量 $|M_n(y_i)| = \zeta + (\zeta - 1) + \cdots + (\zeta - n + 1) = n\zeta - \frac{1}{2}n^2$。令 Φ 为 $ws(y_i)$ 中匹配正向约束的词集合。使用正向约束后，去除了 $|\Phi|$ 个 Uni-Gram，NSG 的节点数为 $|M_n(y_i)| - |\Phi|$。使用负向约束的情况比较复杂。对于作用于 $w_i^{(j)}$ 和 $w_i^{(j+1)}$ 的负向约束 $\mathcal{N}$，如果 $j = 1$ 或 $j + 1 = \zeta$（此为最坏情况），图上的节点可以减少 $n - 1$ 个。在最好情况下，我们有 $j + 1 \geq n$ 且 $j > \zeta - n$，减少的节点数量为 $\sum_{j=1}^{n-1} j = \frac{1}{2}n(n-1)$。令 ψ 为某 $ws(y_i)$ 中起作用的负向约束数量，NSG 的节点数量在以下范围内：

$$[n\zeta - \frac{1}{2}n^2 - |\Phi| - \frac{\psi}{2}(n-1), n\zeta - \frac{1}{2}n^2 - |\Phi| - \frac{\psi}{2}n(n-1)]$$

如图 4.5所示，仅加入两条约束，NSG 的节点数量从 9 减少至 4。与之对应，NSG 中边的数量从 14 减少至 4。

因为本研究工作特别针对中文语言。我们设计了三种正向约束条件和一种负向约束条件，见表 4.9。例如，“的” 是中文重要助词，通常表示了修饰词的结尾，所以应该在作为助词的 “的” 后进行切分。值得一提的是，DNRE 算法的这一模块灵活方便，可以加入任意数量、面向任意语种的语言规则。

当加入正负向约束的 NSG 构建完毕，我们旨在选择最合适的极大团作为最终的切分结果。由于在图 $G_n(y_i)$ 中，边的权重 $\alpha_{i,j}$ 刻画了两个 N-Gram 的可分割性，选择最佳团的问题也可以建模成**最大边权重团问题**（MEWCP）。在第 4.3.1节中，我们给出了高效的 MEWCP 算法，这一算法对于密集图比较合适。然而，加入约束的 NSG 的图结构特征与第 4.3.1节的模式图具有区别，NSG 的边是高度稀疏的。直接

表 4.9: DNRE 中采用的正向和负向约束

作用于 $w_i^{(j)}$ 和 $w_i^{(j+1)}$ 的正向约束
约束 1：POS($w_i^{(j)}$)=VERB 且 POS($w_i^{(j+1)}$)=PREP
约束 2：POS($w_i^{(j)}$)=CONJ 或 POS($w_i^{(j)}$)=CONJ
约束 3：$w_i^{(j+1)}$=“的”
作用于 $w_i^{(j)}$ 和 $w_i^{(j+1)}$ 的负向约束
约束 1：$w_i^{(j)}$=“的”

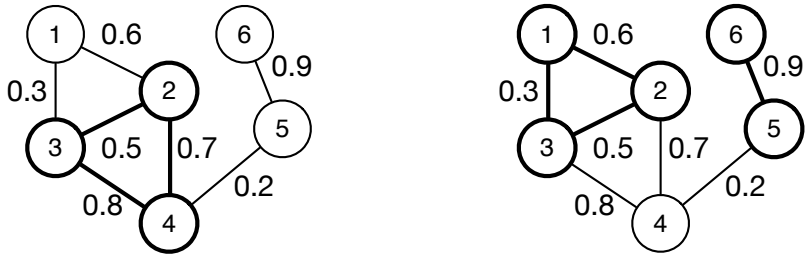


图 4.6: 正确的最大边权重团及 MEWCP 算法生成的反例

使用第 4.3.1节的算法可能会检测到一个图中的多个团，而不是一个最大边权重团，其简单示例如图 4.6 。

在 DNRE 中，我们采用改进的 MEWCP 近似算法检测最大边权重团，同样基于蒙特卡洛算法，其流程见算法 13。在初始阶段，算法选出一条边 $(m_i, m_j) \in E_n(y_i)$，选出 (m_i, m_j) 的概率正比于 $\alpha_{i,j}$。令 $G'(M', E')$ 为仅含有该边的初始图。令 $N(M')$ 为原图 $E_n(y_i)$ 中 M' 的所有邻居节点集合。对于 $N(M')$ 为中的每个节点 m_i，算法检查将 m_i 加入 M' 是否仍然构成团，将可以构成团的节点集合记为候选节点结合 $Can(M') \subseteq N(M')$。在此之后，算法迭代地从 $Can(M')$ 中采样节点 m_i，其概率正比于 $\sum_{m_j \in M'} \alpha_{i,j}$。它将 m_i 和对应的边加入 $G'(M', E')$，并且更新 $N(M')$ 和 $Can(M')$。当没有更多的边可选时，M' 即为所选的最大边权重团。

如果使用哈希图作为图的存储结构，该算法在最差情况下的时间复杂度为 $O(|M_n(y_i)|^2|E_n(y_i)|)$，比第 4.3.1节的算法有略微提高。然而，因为加入约束的 NSG 的节点数量一般少于 10 个，这一算法仍然能高效地解决切分的问题。我们同样运行该算法若干次，并且记选出的团集合为 $\mathcal{C} = \{M'\}$。我们根据下式选择最佳的团生成 MPS 的最终结果：

$$M^* = \text{argmax}_{M' \in \mathcal{C}} \frac{\sum_{(m_i, m_j) \in M'} \alpha_{i,j}}{\log(1 + \beta |M'|)} \tag{4.3}$$

Algorithm 13 MEWCP 的改进算法

从 $E_n(y_i)$ 中采样边 (m_i, m_j)，选出 (m_i, m_j) 的概率正比于 $\alpha_{i,j}$
初始化图 $G^{'}(M^{'}, E^{'})$，其中 $M^{'} = \{m_i, m_j\}$，$E^{'} = \{(m_i, m_j)\}$
构建邻居节点集合 $N(M^{'})$
生成候选节点集合 $Can(M^{'}) \subseteq N(M^{'})$
while $Can(M^{'}) \neq \emptyset$ **do**
    从 $Can(M^{'})$ 采样节点 m_i，选出 m_i 的概率正比于 $\sum_{m_j \in M^{'}} \alpha_{i,j}$
    将 m_i 和对应的边加入 $G^{'}$
    更新 $N(M^{'})$ 和 $Can(M^{'})$
end while
return 最大边权重团 $M^{'}$

Algorithm 14 MPS 的整体算法流程

对短文本 y_i 进行中文分词，生成结果 $ws(y_i)$
根据 $ws(y_i)$ 和正负向约束，构建 NSG$G_n(y_i)$
初始化团集合 $\mathcal{C} = \emptyset$
for 每个迭代 i **do**
    利用算法 13生成最大权重团 $M^{'}$
    更新 $\mathcal{C} = \mathcal{C} \cup \{M^{'}\}$
end for
根据式 4.3从 $\mathcal{C}$ 中选择最合适的团 M^{*}
根据 M^{*} 生成 MPS 结果 $ps(y_i)$
return MPS 结果 $ps(y_i)$

其中，$\beta > 0$ 是缩放因子，使得算法更倾向于选择更小的团作为最终结果。这是因为小的团对短文本进行更少次数的切分，避免将短文本分成过多语义不完整的词语集合。综上，我们将 MPS 的整体流程汇总在算法 14中。

候选关系元组生成： 当 MPS 执行完毕后，对于每个 (x_i, y_i)，我们可以获得实体-修饰词对 $\{(x_i, q_i^{(j)})\}$（$1 \leq j \leq |ps(y_i)| - 1$）。如果某修饰词 $q_i^{(j)}$ 中不包含任何实体，很有可能该修饰词不表示任何语义关系；反之，我们有可能抽取出从 $q_i^{(j)}$ 中抽取出关系谓词和宾语，作为 x_i 的候选关系三元组。我们记 $R(x_i, y_i)$ 为所有从 (x_i, y_i) 中抽取出的候选关系元组集合。如果可以抽取出完整的关系谓词和宾语，我们记该**候选完整关系元组**（Candidate Full Relation）为 $r(x_i, q_i^{(j)})$，并且将其加入 $R(x_i, y_i)$；在某些情况下，正确的关系谓词无法检测到，我们将主语 x_i 和抽取出的关系宾语作为**候选部分关系元组**（Candidate Partial Relation）$\tilde{r}(x_i, q_i^{(j)})$，同样加入 $R(x_i, y_i)$。根据修饰词 $q_i^{(j)}$ 的语法结构，CRG 中一共包含三种类型的操作，详见下文。此外，我们也给出了这三种类型的操作对应的示例，如图 4.7。

- **情况 i**：如果 $q_i^{(j)}$ 是动词从句，我们可以对其进行依存语法解析 [209]，从依

情况	中文短文本	抽取的关系谓词	抽取的关系宾语
i)	DEP 1947年 建立 的 行政区划	建立	1947年
ii)	DEP 罹患 肺结核 逝世者	罹患	肺结核
	DEP 福克斯电视台 播放 的 电视连续剧	播放	福克斯电视台
iii)	意大利 作曲家	?	意大利 (Italy)

图 4.7: CRG 中三种类型的操作示例，$\xrightarrow{DEP}$ 表示依存语法树中的依赖关系

存语法树中直接抽取相应关系谓词和宾语，生成候选关系元组 $r(x_i, q_i^{(j)})$。

- **情况 ii**：因为 MPS 是完成非监督的算法，在少数情况下，关系谓词和宾语没有被 MPS 切分到同一个簇中。在这种情况下，我们进一步搜索 $q_i^{(j-1)}$ 和 $q_i^{(j+1)}$，如果根据依存语法解析的结果，检测到相应关系谓词和宾语，我们同样生成候选关系元组 $r(x_i, q_i^{(j)})$。

- **情况 iii**：如果没有检测到关系谓词，我们抽取候选部分关系元组 $\tilde{r}(x_i, q_i^{(j)})$。

缺失关系谓词检测： 在 MRPD 模块中，我们基于贝叶斯推断的方法补全候选部分关系元组中缺失的关系谓词。令 $\mathcal{V}$ 为所有可能的关系谓词集合。因为中文关系抽取准确度较低 [34]，在 DNRE 系统中，我们规定只从两种知识源中获取关系谓词 $\mathcal{V}$：i）所有候选完整关系元组中的关系谓词；以及 ii）人类常识定制的关系谓词。受到 Zhang 等人关系生成研究 [213] 的启发，我们将关系谓词预测的问题建模成如下生成模型：

$$\Pr(v, \tilde{r}(x_i, q_i^{(j)})) = \Pr(v) \Pr(\tilde{r}(x_i, q_i^{(j)})|v)$$

其中，$\Pr(v)$ 是关系谓词 v 的**先验模型**（Prior Model），$\Pr(\tilde{r}(x_i, q_i^{(j)})|v)$ 为**似然性模型**（Likelihood Model）。根据贝叶斯公式，我么可以推断候选部分关系元组 $\tilde{r}(x_i, q_i^{(j)})$

表 4.10: 时间和空间常识性知识示例，地点和事件表达粗体显示

常识性知识类别	**实体**	**中文短文本**
时间	复旦大学	**上海**高等院校
	故宫博物院	**北京**博物馆
空间	诺曼底战役	**1944 年**欧洲战场战役
	安史之乱	**8 世纪**中国战争

中最有可能的关系谓词 $\tilde{r}_{v^*}(x_i, q_i^{(j)})$，如下所示：

$$\begin{aligned}\tilde{r}_{v^*}(x_i, q_i^{(j)}) &= \text{argmax}_{v' \in \mathcal{V}} \Pr(v'|\tilde{r}(x_i, q_i^{(j)})) = \text{argmax}_{v' \in \mathcal{V}} \frac{\Pr(v') \Pr(\tilde{r}(x_i, q_i^{(j)})|v')}{\Pr(\tilde{r}(x_i, q_i^{(j)}))} \\ &= \text{argmax}_{v' \in \mathcal{V}} \Pr(v') \Pr(\tilde{r}(x_i, q_i^{(j)})|v')\end{aligned}$$

在下文中，我们分别介绍先验模型 $\Pr(v)$ 和似然性模型 $\Pr(\tilde{r}(x_i, q_i^{(j)})|v)$ 的学习过程。

先验模型 $\Pr(v)$ 融合了 CRG 的抽取结果和人类的常识性知识。$\Pr(v)$ 的第一部分可以通过极大似然估计学习，即 $\Pr(v)^{MLE} = \frac{N_v}{N}$，其中 N 和 N_v 分别为 CRG 抽取出的候选完整关系元组数量及关系谓词为 v 的候选完整关系元组数量。根据我们的分析和先前的研究 [178, 193, 194]，大部分关系谓词缺失源于常识性知识。特别地，在中文短文本中，时间和空间常识性知识是最常见的两种关系谓词缺失情况，分别指“发生于”和“位于”两种关系，其示例见表 4.10。

令 N_s、N_t 和 N_p 分别为 CRG 生成的候选关系元组中地点、时间及其他关系宾语的数量。基于常识性知识的先验模型 $\Pr(v)^{CS}$ 定义如下：

$$\Pr(v)^{CS} = \begin{cases} \frac{N_s}{N_p}, & \text{空间常识性知识} \\ \frac{N_t}{N_p}, & \text{时间常识性知识} \\ \frac{1}{|\mathcal{V}|-2}(1 - \frac{N_s+N_t}{N_p}), & \text{其他} \end{cases}$$

结合 $\Pr(v)^{MLE}$ 和 $\Pr(v)^{CS}$，先验模型 $\Pr(v)$ 定义为

$$\Pr(v) = \lambda_1 \Pr(v)^{MLE} + \lambda_2 \Pr(v)^{CS} + (1 - \lambda_1 - \lambda_2)\frac{1}{|\mathcal{V}|} \tag{4.4}$$

其中，λ_1 和 λ_2 为平衡超参数，具有限制条件 $0 < \lambda_1 < 1$、$0 < \lambda_2 < 1$ 和 $\lambda_1 + \lambda_2 < 1$。$(1 - \lambda_1 - \lambda_2)\frac{1}{|\mathcal{V}|}$ 对关系谓词分布 $\Pr(v)$ 加上 Jelinek-Mercer 平滑效果 [214]。

似然性模型 $\Pr(\tilde{r}(x_i, q_i^{(j)})|v)$ 通过**超图随机游走过程**（Hypergraph-based Random Walk Process）进行估计。我们首先定义这一游走过程中的两个重要评分。**谓词**

契合评分（Predicate Coherence Score）$w_p(v, \tilde{r}(x_i, q_i^{(j)}))$ 衡量某关系谓词 $v \in \mathcal{V}$ 是否可以描述候选部分关系元组 $\tilde{r}(x_i, q_i^{(j)})$ 中主语和宾语之间的语义关系。记 $\tilde{r}(x_i, q_i^{(j)})$ 中的宾语分别为 $o(x_i, q_i^{(j)})$。我们在海量网络语料库中用 Apache Lucene[7]构建句子级别倒排索引，并且检索得 x_i 和 $o(x_i, q_i^{(j)})$ 共现的句子集合。根据 Banko 等人 [175] 的研究，如果在 x_i 和 $o(x_i, q_i^{(j)})$ 之间的依存语法链中存在动词 v，我们将其视为 x_i 和 $o(x_i, q_i^{(j)})$ 的候选关系谓词。令 $V(x_i, q_i^{(j)})$ 为候选关系谓词集合，$c(v)$ 为动词 v 的计数。关系谓词 v 与候选部分关系元组 $\tilde{r}(x_i, q_i^{(j)})$ 的谓词契合评分 $w_p(v, \tilde{r}(x_i, q_i^{(j)}))$ 定义为：

$$w_p(v, \tilde{r}(x_i, q_i^{(j)})) = \frac{1}{Z} \sum_{v' \in V(x_i, q_i^{(j)})} c(v') \cos(\vec{v}, \vec{v}')$$

其中 $Z = \sum_{v' \in V(x_i, q_i^{(j)})} c(v')$ 为归一化常数。

关系相似评分（Relation Similarity Score）衡量两个候选关系元组 $r(x_i, q_i^{(j)})$ 和 $r(x_k, q_k^{(l)})$ 之间是否具有相同的关系谓词[8]。它计算关系主语和宾语在词嵌入空间的余弦相似度的平均值：

$$w_r(r(x_i, q_i^{(j)}), r(x_k, q_k^{(l)})) = \frac{1}{2}(\cos(\vec{x}_i, \vec{x}_k) + \cos(\vec{o}(q_i^{(j)}), \vec{o}(q_k^{(l)})) \tag{4.5}$$

其中，$\vec{o}(q_i^{(j)})$ 和 $\vec{o}(q_k^{(l)})$ 分别指 $r(x_i, q_i^{(j)})$ 和 $r(x_k, q_k^{(l)})$ 的关系宾语词向量。

根据这两个评分，我们正式定义超图随机游走的具体过程。令 $H(\mathcal{R}, \mathcal{V})$ 为**基于谓词的超图网络**（Predicate-based Hypergraph Network，缩写为 PHN）。其中，$\mathcal{R}$ 是超图中的节点集合，每个节点分别对应所有 CRG 模块生成的候选部分关系元组和候选完整关系元组；$\mathcal{V}$ 是**超边**（Hyper-edge）的集合，每条超边分别对应某关系谓词 $v \in \mathcal{V}$。在 PHN$H(\mathcal{R}, \mathcal{V})$ 中，每条超边 $v \in \mathcal{V}$ 中的节点对应候选部分关系元组和候选完整关系元组的集合。如果候选完整关系元组 $r(x_i, q_i^{(j)})$ 的谓词为 v，$r(x_i, q_i^{(j)})$ 在对应的超边 v 中；如果候选部分关系元组 $\tilde{r}(x_i, q_i^{(j)})$ 与谓词 v 的评分 $w_p(v, \tilde{r}(x_i, q_i^{(j)})) > \tau_1$（$\tau_1 \in (0, 1)$ 是预定义阈值），$\tilde{r}(x_i, q_i^{(j)})$ 在对应的超边 v 中。图 4.8给出了 PHN 的简单示意图。我们进一步考察 PHN 的邻域。在 PHN$H(\mathcal{R}, \mathcal{V})$ 中，节点 $r(x_i, q_i^{(j)})$ 的邻域 $Nb(r(x_i, q_i^{(j)}))$ 为节点集合，在这个集合中的每个元素 $r(x_k, q_k^{(l)}) \in Nb(r(x_i, q_i^{(j)}))$ 都与 $r(x_i, q_i^{(j)})$ 在同一超边中。在图 4.8中，节点 1 和 4

[7] http://lucene.apache.org/

[8] 在关系相似评分的计算中，我们不区分候选部分关系元组和候选完整关系元组，统一表示为 $r(x_i, q_i^{(j)})$ 和 $r(x_k, q_k^{(l)})$，下同。

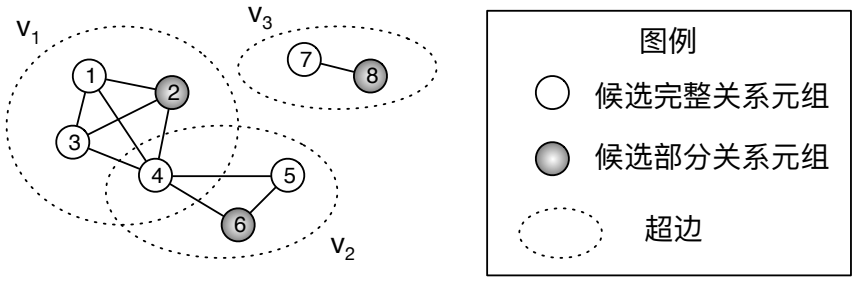


图 4.8: PHN 的简单示意

Algorithm 15 缺失关系谓词检测算法

```
生成关系谓词集合 V
for 每个关系谓词 v ∈ V do
    根据式 4.4计算先验概率 Pr(v)
    生成关系谓词为 v 的候选完整元组节点集合 R_v
end for
构建 PHN H(R, V)
for 每个关系谓词 v ∈ V do
    根据式 4.5运行随机游走算法
    for 每个候选部分元组 r̃(x_i, q_i^(j)) do
        根据式 4.6计算概率 Pr(r̃(x_i, q_i^(j))|v)
        预测 r̃(x_i, q_i^(j)) 的关系谓词 r̃_{v*}(x_i, q_i^(j))
    end for
end for
```

的邻域分别为 {2,3,4} 和 {1,2,3,5,6}。

基于超图的随机游走过程如下。令 $\mathcal{R}_v$ 为超图 $H(\mathcal{R},\mathcal{V})$ 中超边 $v \in \mathcal{V}$ 对应的所有候选完整元组节点集合。对每个超边 $v \in \mathcal{V}$ 中的每个候选完整元组对应节点 $r(x_i, q_i^{(j)}) \in \mathcal{R}_v$，一个**随机游走者**（Random Walker）从该节点出发，随机跳转一个邻居节点 $r(x_k, q_k^{(l)}) \in Nb(r(x_i, q_i^{(j)}))$，概率正比于 $w_r(r(x_i, q_i^{(j)}), r(x_k, q_k^{(l)}))$，并且迭代足够多次数。最后，每个候选部分元组节点 $\tilde{r}(x_i, q_i^{(j)})$ 可以对应分数 $s_v(\tilde{r}(x_i, q_i^{(j)}))$，表示所有随机游走者的访问次数。根据上述过程，概率 $\Pr(\tilde{r}(x_i, q_i^{(j)})|v)$ 可以按照下式近似计算：

$$\Pr(\tilde{r}(x_i, q_i^{(j)})|v) = \frac{s_v(\tilde{r}(x_i, q_i^{(j)}))}{\sum_{\tilde{r}(x_k, q_k^{(l)}) \in \mathcal{R}} s_v(\tilde{r}(x_k, q_k^{(l)}))} \tag{4.6}$$

缺失关系谓词检测算法的步骤汇总在算法 15中。

DNRE 的最后一个步骤为置信度过滤。我们观察到大部分抽取的错误源于无意义的“关系谓词”,这一错误发生的原因也与经典的开放关系抽取研究相似 [28, 175]。令 $\tilde{c}(v)$ 为所有关系谓词为 v 的候选完整和部分元组数量。若候选完整关系元组

$r(x_i, q_i^{(j)})$ 的关系谓词为 v，其置信度即为 $conf(r(x_i, q_i^{(j)})) = \tilde{c}(v)$。对于候选部分关系元组 $\tilde{r}(x_i, q_i^{(j)})$，我们加入缺失关系谓词检测的正确性作为额外的置信度因子。记 $\text{secmax}_{v\in\mathcal{V}} \Pr(v) \Pr(\tilde{r}(x_i, q_i^{(j)})|v)$ 为所有概率 $\Pr(v) \Pr(\tilde{r}(x_i, q_i^{(j)})|v)$（$v \in \mathcal{V}$）中的第二大值。最有可能的关系谓词为 $\tilde{r}_{v^*}(x_i, q_i^{(j)})$ 的候选部分关系元组 $\tilde{r}(x_i, q_i^{(j)})$ 置信度为：

$$conf(\tilde{r}(x_i, q_i^{(j)})) = \frac{\tilde{c}(\tilde{r}_{v^*}(x_i, q_i^{(j)})) \cdot \max_{v\in\mathcal{V}} \Pr(v) \Pr(\tilde{r}(x_i, q_i^{(j)})|v)}{\max_{v\in\mathcal{V}} \Pr(v) \Pr(\tilde{r}(x_i, q_i^{(j)})|v) + \text{secmax}_{v\in\mathcal{V}} \Pr(v) \Pr(\tilde{r}(x_i, q_i^{(j)})|v)}$$

我们使用预定义的阈值 τ_2 过滤低置信度的候选完整关系元组 $r(x_i, q_i^{(j)})$（即 $conf(r(x_i, q_i^{(j)})) < \tau_2$）和候选部分关系元组 $\tilde{r}(x_i, q_i^{(j)})$（即 $conf(\tilde{r}(x_i, q_i^{(j)})) < \tau_2$）。

4.4.2 实验分析

在本节中，我们对 DNRE 的实验性能进行详细评测。特别地，由于 DNRE 完全由数据驱动，可以抽取任意类别的关系，我们在 ORE 的框架下评测 DNRE 算法的实验效果，并与其他基线方法进行对比。

数据源与实验设置： 由于 DNRE 与 PNRE 处理的 NLP 任务相同，我们采用与评测 PNRE 相同的中文维基百科数据作为输入数据源，用 FudanNLP 工具 [209] 进行基本的中文 NLP 分析。我们同样使用 Skip-Gram 算法 [51] 获得所需的词向量。在 DNRE 的实现中，超参数的默认设置为：$n = 3$、$\gamma = 0.3$、$\beta = 5$、$\lambda_1 = 0.6$、$\lambda_2 = 0.3$、$\tau_1 = 0.7$ 以及 $\tau_2 = 20$。在 MPS 模块中，我们运行 MEWCP 算法三次来生成团集合 $\mathcal{C}$；在 MRPD 模块的随机算法中，我们在超图的每个出发点同时发出 10 个随机游走者，每个随机游走者在超图中走 500 步。我们同时研究超参数的调整是怎样影响 DNRE 的实验效果。为了研究 DNRE 的时间效率问题，我们用 JAVA 语言实现了 DNRE 系统，并且在单机上运行，该单机的 CPU 主频为 2.9GHz，内存为 16GB。

基线算法： 因为 DNRE 可以抽取任意开放领域的关系，我们使用多个 ORE 算法、基于维基百科的关系抽取算法及 PNRE 作为基线算法。

- **经典句子级别 ORE 算法**：我们采用经典的中文 ORE 系统 ZORE [34] 作为强基线算法。为了适应中文短文本的输入，我们将每个短文本对 (x_i, y_i) 作为输入在前述中文语料库中获得 x_i 与 y_i 共现的 Top-5 句句子作为 ZORE 的输入。

ZORE 的实现和参数设置与作者公开的源码[9]相同。

- **基于神经网络的句子级别 ORE 算法**：采用 Cui 等人 [176] 提出的编码解码器网络（Encoder-Decoder Network）作为基于神经网络的句子级别 ORE 模型，其输入与 ZORE [34] 相同。编码器和解码器都采用三层的双向 LSTM 架构。

- **词组级别 ORE 算法**：因为我们实验中的短文本主要是名词性词组，我们使用基于名词词组的 ORE 系统 RELNOUN [177] 作为基线算法。在本研究工作中，我们将 RELNOUN [177] 中的英语语言模型翻译成中文，实现了 CN-RELNOUN 来抽取关系。

- **基于维基百科的算法**：我们采用两种基于维基百科类别系统的关系抽取算法作为基线算法：其一为 Nastase 和 Strube [33] 提出算法的中文版本，即 CN-WikiRe（见第 4.3.2节）；其二为我们先前提出的 PNRE 算法。

在少数情况下，部分中文维基百科的类别为动词词组，例如 “1946 年出生” 。为了使 DNRE 可以公平地与其他基线算法对比，我们规定，如果输入的词组为动词词组，我们将所有切分后的得到短语均视为修饰词（而非一个或多个修饰词和一个核心词），进行基于 CRG 和 MPRD 的关系抽取。

评测度量： 由于 DNRE 系统从开放领域自动抽取多种类别的关系，很难确定所有可抽取关系的 Ground Truth 来计算这一系统的召回率和 F 值。在 ORE 的研究中 [28]，通常采用 Yield 分数来评测 ORE 系统的有效性，这一分数为抽取出关系数量与他们的准确度之积。在本研究中，我们采用三个评测度量：抽取出的关系数量（简记为 # 关系）、准确度和 Yield 分数。

整体实验结果： 为了评测 DNRE 的实验效果，我们在中文维基百科数据集上选定了四个领域：通用、政治、娱乐、军事，对于每个领域任选 300 个中文实体，分别评测不同方法从他们对应的实体-类别对中抽取出的关系的数量、准确度和 Yield 分数。这三个特定领域及每个领域对应的中文实体示例参见表 4.12。在表 4.11中，我们汇总了这四个领域的实验结果。由此可以，句子级别的 ORE 系统 [34, 176] 没有产生令人满意的输出，这是因为句子级别的 ORE 系统往往从句子中的 “主语-谓语-宾语” 结构抽取出关系；在我们的数据中，很多中文短文本在语料库中并没有类似的表达，因此无法抽取出这些短文本中蕴含的关系。CN-RELNOUN [177] 抽

[9]https://sourceforge.net/projects/zore/

表 4.11: 四个领域的中文短文本关系抽取实验效果对比

方法	# 关系	准确度	Yield	# 关系	准确度	Yield
领域	**通用**			**政治**		
CN-WikiRe [33]	87	41.7%	41	84	57.1%	48
CN-RELNOUN [177]	31	93.5%	29	35	88.6%	31
ZORE [34]	28	75.0%	21	34	76.4%	26
Cui 等人 [176]	52	51.9%	27	51	43.1%	22
PNRE	193	**94.3%**	182	193	**95.9%**	185
DNRE	**289**	92.7%	**268**	**314**	93.9%	**295**
提升	**+49.7%**		**+47.3%**	**+62.7%**		**+59.5%**
领域	**娱乐**			**军事**		
CN-WikiRe [33]	102	39.2%	40	76	53.9%	41
CN-RELNOUN [177]	42	88.1%	37	34	82.3%	28
ZORE [34]	21	76.2%	16	32	81.2%	26
Cui 等人 [176]	54	48.1%	26	44	56.8%	25
PNRE	204	**95.1%**	194	188	**96.3%**	181
DNRE	**324**	92.3%	**299**	**274**	94.2%	**258**
提升	**+58.8%**		**+54.1%**	**+45.7%**		**+42.5%**

表 4.12: 中文维基百科的三个特定实验领域及其中文实体示例

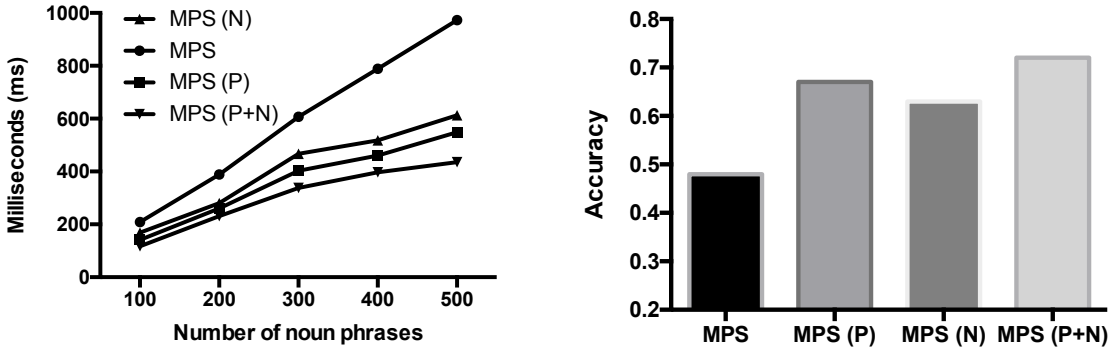
领域	中文实体示例
政治	唐纳德·特朗普、罗纳德·里根、英国议会
娱乐	王菲、肖申克的救赎、奥斯卡金像奖
军事	航空母舰、氢弹、中途岛海战

取出的关系精度较高，然而其整体的 Yield 分数较低。在英语中，由于介词的频繁使用，例如 “[...] is [...] of [...]” 和 “[...] is [...] from [...]” 的语言模式对 RELNOUN 中的关系抽取高度有效；而在中文中，类似的语言模式的召回率较低，所以抽取出的关系数量少。与之相对比，我们的算法 DNRE 不依赖于固定的语言模式，而是采用数据驱动的分治抽取方法，可以更有效地解决这一问题。与我们先前的 PNRE 算法相比，在通用领域，DNRE 抽取出的关系数量是 PNRE 的 149.7%，其精度与 PNRE 类似，为 92.7%。在三个特定领域，DNRE 的实验结果与通用领域相似，超过了所有基线算法。

因为先前的实验只在部分维基百科数据上进行评测，我们进一步估计整个中文维基数据集的实验效果。在所有抽取出的关系中，我们随机抽取出 500 个元组，通过人工标注估计其准确度。根据抽取出的关系数量和准确度的估计值，我们也可以估计各个算法的 Yield 分数，实验结果见表 4.13。整体而言，DNRE 抽取了 55.4 万个关系元组，准确度为 95.4%。它比 PNRE 多抽取出 55.2% 的关系，其 Yield 分

表 4.13: 中文维基百科的整体中文短文本关系抽取实验效果对比

方法	# 关系	准确度（估计值）	Yield（估计值）
CN-WikiRe [33]	165K	58.6%	96.7K
CN-RELNOUN [177]	65K	92.8%	60.3K
ZORE [34]	42K	82.3%	34.6K
Cui 等人 [176]	89K	51.2%	45.6K
PNRE	357K	**97.4%**	347.7K
DNRE	**554K**	95.4%	**528.5K**
提升	**+55.2%**		**+52.0%**



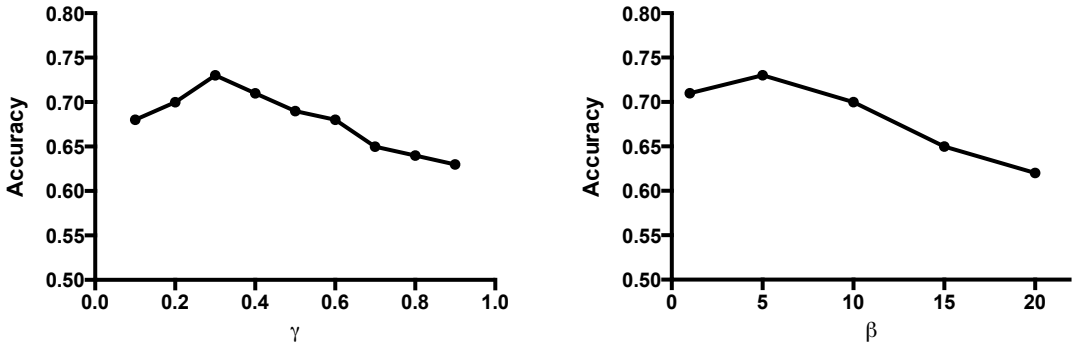
(a) 算法效率实验　　(b) 算法正确性实验

图 4.9: MPS 算法的效率和正确性评测结果

数提升了 52.0%，在没有任何语言模式的定制情况下，没有牺牲过多的准确度。

DNRE 的算法细节分析： 我们调整 DNRE 中超参数的值，并且分析 DNRE 中不同模块的实验效果。

MPS 模块. 在 MPS 模块中，我们发现超过 90% 的修辞词包括少于 4 个中文词语（经过中文分词后的结果）。所以，我们可以设置 N-Gram 因子为 $n = 3$。我们也可以将 N-Gram 因子的值设为更大，然而，这一设置会使算法所需的运行时间和计算资源更大，而且不利于语言模型的学习。其次，我们评测 MEWCP 算法的效率和正确性。我们将超参数 γ 和 β 设为默认值，在四种实验设置条件上进行两组实验：“MPS”（不加任何正负向约束的 MEWCP 算法）、“MPS (P)”（仅加正向约束的 MEWCP 算法）、“MPS (N)”（仅加负向约束的 MEWCP 算法）和 “MPS (P+N)”（同时加入正负向限制的 MEWCP 算法）。在算法效率实验中，我们随机采样 100 到 500 个中文维基百科类别名词词组，分别运行四种设置下的 MEWCP 算法，并且记录了算法运行时间。在算法正确性实验中，我们人工标注通用领域数据集中的 MPS 算法切分中文短文本的准确性。这两组实验结果参见图 4.9中。从实验结果



(a) 调整参数 γ　　(b) 调整参数 β

图 4.10: MPS 模块中 γ 和 β 的参数分析

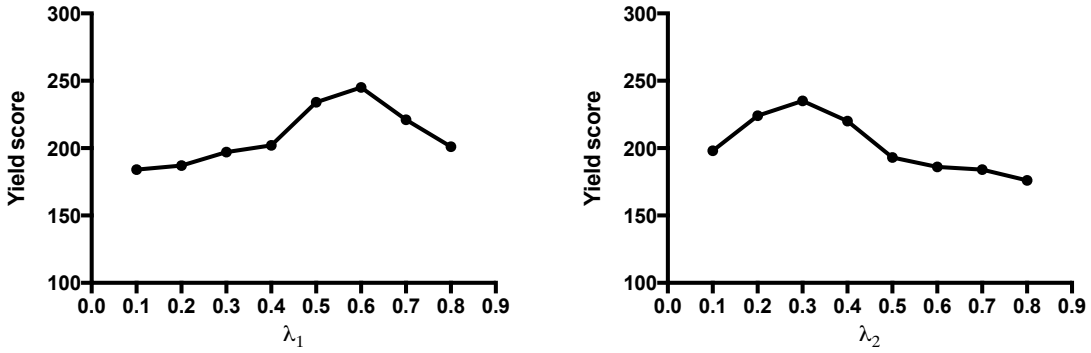
表 4.14: CRG 中候选完整关系元组占所有候选关系元组的比例

设置	**通用**	**政治**	**娱乐**	**军事**
不采用 CMRG	12.2%	11.0%	15.4%	13.3%
采用 CMRG	**16.8%**	**14.6%**	**17.8%**	**15.8%**
提升	**+4.6%**	**+3.6%**	**+2.4%**	**+2.5%**

可见，同时加入正负向约束时，我们提出的算法效率最高，减少了原来运行时间的 50% 左右。我们也可以发现，当加入基于中文语言规则的正负向约束时，MPS 的准确度也加以提升。此外，我们调整了两个参数 γ 和 β 的值，结果见图 4.10。实验结果表明，基于分布式模型的分数 $w_d(i,j)$ 比统计评分 $w_s(i,j)$ 重要性更高。当 $\beta = 5$ 时，我们取得了最佳的实验效果。

CRG 模块. 在 CRG 模块中，我们采用了**跨修饰词关系生成**（Cross-Modifier Relation Generation，缩写 CMRG）策略提升候选完整关系元组在所有生成候选关系元组的比例。为了评测这一策略的有效性，我们分别统计采用 CMRG 和不采用 CMRG 情况下，候选完整关系元组的比例，其实验结果见表 4.14。从实验结果可见，这一策略在不同数据集中均获得实验效果的提升，候选完整关系元组的提升比例从 2.4% 到 4.6% 不等。

MRPD 模块. 在 MRPD 模块中，我们调整参数 λ_1 和 λ_2 的值，并且汇报当 λ_1 和 λ_2 在不同取值在通用领域数据集的 Yield 分数。实验结果参见图 4.11。在每组实验中，我们将每个参数固定为 0.1，并且调整另一个参数的值。我们可见参数 λ_1 和 λ_2 的调整反映出三种类别的先验概率分布的相对重要性。在基于超图的随机游走过程中，我们设置 $\tau_1 = 0.7$，因为我们观察到当 $\tau_1 \geq 0.7$ 时，关系主语-宾语对往往具有相同的关系谓词。我们进一步调整 τ_2 的值，进行基于置信度的过滤。在



(a) 调整参数 λ_1　　(b) 调整参数 λ_2

图 4.11: MRPD 中参数 λ_1 和 λ_2 的分析

表 4.15: 较高及较低置信度的关系谓词示例

关系谓词	**置信度** $\tilde{c}(v)$	**关系谓词**	**置信度** $\tilde{c}(v)$
位于	124K	警告	19
发生	53K	民变	16
毕业	44K	冷藏	14
建立	23K	集会	8

表 4.15中，我们给出了具有较高及较低置信度 $\tilde{c}(v)$ 的关系谓词示例。我们可见具有低置信度的动词通常不是正确的关系谓词，他们来源于 POS 和文本解析的错误。我们在系统中最终设置 $\tau_2 = 20$，因为当 $\tau_2 < 20$ 时，生成的关系谓词正确率较低。

错误分析： 我们对 DNRE 算法的错误案例进行分析，其示例见表 4.16。其中第一类常见的错误类别为**不完整宾语抽取**（Incomplete Object Extraction，缩写为 IOE）。在这种情况下，从短文本中抽取出的关系宾语在语义上不完整。例如在表 4.16中，“首尔特别市” 是完整的实体名称，然而 MPS 错误地将 “首尔特别市” 切分成两个部分，使得抽取出的关系元组变为 “（梁耀燮，出身，特别市）”。这一错误也在很多其他 ORE 系统中出现 [28, 175, 215]。在这些系统中，一般采用加入语法约束的方法使得抽取出的结果尽可能完整，这一个问题在中文短文本的语境下很难解决。另一种常见的错误为**错误谓词检测**（Error Predicate Detection，缩写为 EPD）。例如在 DNRE 中，有很大概率预测中文实体与另一个地点实体之间的关系为 “位于”，当 NER 算法错误地将其他实体标注为地点时，关系谓词可能会错误地预测为 “位于”。在基于超图的随机游走算法中，关系谓词也有可能被算法错误预测。除了关系抽取的错误，一个更重要的问题是关系抽取的缺失问题。ORE 的最新研究 [216] 指出，ORE 的抽取缺失问题的严重程度甚至很难被评测，完全解决这

表 4.16: 两种主要类别的抽取错误及其示例

错误类别	示例
IOE	抽取的关系：(王家骥，职务，校长)
	更正的关系：(王家骥，职务，国立台东大学校长)
	抽取的关系：(梁耀燮，出身，特别市)
	更正的关系：(梁耀燮，出身，首尔特别市)
EPD	抽取的关系：(第 65 届戛纳电影节，担任，戛纳)
	更正的关系：(第 65 届戛纳电影节，位于，戛纳)
	抽取的关系：(台北 101，生于，台湾)
	更正的关系：(台北 101，位于，台湾)

一问题更加困难。因此，如何从中文短文本中抽取更多的语义关系，值得进一步加以研究。

4.5 中文短文本的语义理解

前述 PNRE 和 DNRE 两个算法都旨在解决中文实体和它的描述性短文本的关系抽取问题。其中, PNRE 挖掘短文本中的频繁语言模式, 然而对这些语言模式的语义没有进行深度理解；DNRE 进一步采用数据驱动的方法对短文本进行切分, 在切分的基础上生成中文实体与切分后的词组之间的关系。这两种算法都没有对短文本本身的语义进行理解。在本节中，我们特别关注短文本的**习语性**（Idiomaticity）问题，并且论述中文短文本的习语性分类是如何有利于语义关系的扩展。

4.5.1 习语性分类问题

习语性现象在自然语言中广泛存在，指词汇的含义不能从其构成部分的语义直接推出 [217]，因而具有引申或暗喻的含义，例如 “一窝蜂”、“拍马屁”、“哑巴吃黄连” 等。习语的广泛存在对语言的精准理解造成了很大的挑战，例如，机器翻译的准确性在习语较多的语料上明显下降 [218]。

习语性语言的检测和处理在计算语言学和 NLP 领域有很多研究。例如，**习语分类**的任务目标是根据上下位语境，将目标表达式的语境含义分类为习语或字面含义。典型的研究工作着眼于英语动名词短语的习语分类 [199, 201]。另一个紧密相关的任务为**复合名词**的**组合性分析** [207, 219]，其目标为预测构成复合名词的两个名词语义是否可分，这是由于复合名词的组合性与其习语性密切相关。例如在英语中，名词词组 “apple tree” 是可分解的，其词组的含义可以通过将这两个词的含义组合起来加以推断（“trees where apples grow”）；而 “cloud nine” 中的两个词是

完全不可分的，因为 “cloud nine” 的含义（指极乐心境、狂喜状态）与 “cloud” 和 “nine” 各自的字面含义无关。

与英语不同，中文短文本的习语性分析面临的挑战更大。这是由于在中文中，存在大量暗喻性的语言，这些语言的语义一般属于常识性知识，很难通过语言特征进行分析 [220]。因此，中文短文本的习语性分析与语义理解需要更多与中文语言学知识相关的数据建模。在本节中，我们特别研究中文复合名词的习语性问题，这是由于复合名词占中文描述性短文本的很大部分，蕴含丰富的语义知识。特别地，对于一个中文复合名词 N_1N_2，我们根据文献 [221] 提出的分类方法，按照中文复合名词的习语性程度分为四个等级，概述如下：[10]

1. **透明**（Transparent）：N_1 显示地修饰 N_2，描述了 N_2 的一种属性。我们也可以推断出 N_1N_2 是一种 N_2。例如，在 “固体燃料” 中，“固体” 是 “燃料” 的物理属性。

2. **部分模糊**（Partly Opaque）：N_1 不直接修饰 N_2，N_1 和 N_2 之间存在某种动词性关系。同样地，我们可以推断 N_1N_2 是一种 N_2。例如，在 “办公用品” 中，“办公” 不是 “用品” 的属性，“办公用品” 描述了一种用于 “办公” 的 “用品”，“用于” 是两者之间的动词性关系。

3. **部分习语性**（Partly Idiomatic）：N_1 和 N_2 语义可分，而 N_1 在 N_1N_2 的含义是暗喻性的，而非采用其字面含义。因此，N_1 和 N_2 之间无显示地语义关系，但我们仍然可以推断 N_1N_2 是一种 N_2。例如，在 “计划经济” 中，“计划” 指这一经济系统中最重要的特征，即为商品和其他资源的分配通过政府制定的计划来实施。

4. **完全习语性**（Completely Idiomatic）：N_1N_2 完全不可分，指既不是 N_1 也不是 N_2 的一种概念。例如，“夫妻肺片” 是中餐菜式的名称，既不是一种 “夫妻”，也不是一种 “肺片”。

这四类中文复合名词的习语性程度示例见表 4.17。对中文短文本按照 [221] 的框架进行习语性分类，我们可以进行自然语言推理，抽取出更多的语义关系 [65]。例如，如果我们预测 “固体燃料” 是透明的，我们可以推出如下两个语义关系：

[10]在语言学中，复合名词的这一性质一般被称为 “语义透明性”（Semantic Transparency）[182]，而在 NLP 的研究中，我们更多采用 “习语性” 来描述各种词汇单元的这种语言特性。在本研究中，我们统称为 “习语性”，不讨论这两种术语的细节区别。

表 4.17: 中文复合名词的四种习语性程度及其示例

习语性程度	示例	字面翻译	意译
类别 I：透明	固体 **燃料**	Solid **Fuel**	Solid fuel
	沿海 **地区**	Close to sea **Area**	Coastal area
类别 II：部分模糊	办公 **用品**	Office **Appliance**	Office supplies
	国家 **联盟**	Country **Alliance**	Coalition of nations
类别 III：部分习语性	计划 **经济**	Plan **Economy**	Planned economy
	纳米 **技术**	Nanometer **Technology**	Nanotechnology
类别 IV：完全习语性	夫妻 **肺片**	Husband and wife **Lung piece**	Mr and Mrs Smith, sliced beef and ox tongue in Chilli sauce
	意识 **形态**	Consciousness **Character**	Ideology

（固体燃料，具有属性，固体），（固体燃料，属于，燃料）

对于部分模糊的“办公用品”，我们也可以抽取出如下关系元组：

（办公用品，用于，办公），（办公用品，属于，用品）

其中，关系谓词“用于”可以通过**名词短语解释**技术生成 [180]。对于完全习语性的“夫妻肺片”，我们可以推知前述关系生成方法是不适用于这一名词短语的。

我们也可以将这一技术用于知识图谱的补全。例如在前述研究中，如果我们预测“无烟煤”是一种“固体燃料”，我们可以进一步推断：

（无烟煤，具有属性，固体），（无烟煤，属于，燃料）

同样地，如果我们抽取出“钢笔”是一种“办公用品”，我们也可以推断：

（钢笔，用于，办公），（钢笔，属于，用品）

在下文中，我们详细描述用于中文复合名词习语性分类的 RCRL 模型，对 RCRL 的实验效果进行详细评测，并对 RCRL 的多个应用场景进行研究分析。

4.5.2 算法模型

本节详细描述 RCRCL 的模型细节。根据中文复合名词的语言学特性，我们观察到以下两个现象：

- **现象 1**：部分 N_1 和 N_2 的语言模式与 N_1N_2 的习语性程度相关。

例如，给定包含中文复合名词 N_1N_2 的句子，如果核心词 N_2 也在同一句句子中单独出现，很可能其他与 N_2 相关的语言表达可以解释 N_1N_2。因此，N_1N_2 的两个名词含义可分割，N_1N_2 是完全习语性的概率很小。从“流行音乐是一种以盈利为主要目的而创作的音乐”句中，我们可以推断“流行音乐”不是完全习语性的，它的含义必然与“音乐”相关联。

- **现象 2**：具有相似组合性的中文复合名词有相似的习语性程度。

例如，两个中文复合名词“固体燃料”和“液体燃料”中的名词都是语义可分的。“固体”和“液体”，描述了“燃料”的物理性质，因此这两个中文复合名词在习语性程度上也是透明的。

设 f_i 和 $\tilde{f}_i$ 是中文复合名词 x_i 的真实和预测的习语性程度。令 L 和 U 为中文复合名词的训练和测试集。基于两个观察到的现象,RCRL 模型学习中文复合名词 $x_i \in L \cup U$ 的**关系性表示**（Relational Representation）$\vec{r}_i$ 和**组合性表示**（Compositional Representation）$\vec{c}_i$。关系性表示 $\vec{r}_i$ 建模中文复合名词中 N_1 和 N_2 的语义关系，组合性表示 $\vec{c}_i$ 描述 N_1 和 N_2 语义可分的程度。RCRL 模型的目标为最小化如下损失函数：[11]

$$\mathcal{J} = \sum_{x_i \in L} sl(f_i, \tilde{f}_i) + \lambda \sum_{x_i, x_j \in L \cup U} \mu_{i,j} ul(\tilde{f}_i, \tilde{f}_j)$$

其中，$sl(f_i, \tilde{f}_i)$ 是在训练集上的习语性程度预测损失，其特征根据现象 1 定义。$\mu_{i,j}$ 是两个中文复合性名词 x_i 和 x_j 的组合性相似度，$ul(\tilde{f}_i, \tilde{f}_j)$ 是非监督学习损失，使具有相似组合性的中文复合名词预测出的习语性程度相似（即现象 2）。λ 是监督和非监督学习的平衡性超参数。

关系性表示学习： 与通用词嵌入不同，关系性表示建模一个中文复合名词 N_1N_2 中，N_1 和 N_2 之间的语义关系。为了将现象 1 中的假设进行有效表示，我们在海量中文语料库上计算中文复合名词 x_i 的关系性特征 $\mathcal{F}_i$。关系性表示 $\vec{r}_i$ 可以通过下式进行计算：$\vec{r}_i = \mathbf{M}_r \mathcal{F}_i$，其中 $\mathbf{M}_r$ 是线性投影矩阵。关系性特征 $\mathcal{F}_i$ 定义如下：

助词特征. 在中文中，如果存在语言模式“N_1 的 N_2”，这表示 N_1 显示地修饰 N_2。因此，很有可能 N_1N_2 在语义上是透明的。因为这一语言模式在语料库中可能表达多次，并且模式的表达可能存在噪声，与文献 [222] 的方法相似，我们定义 r_a 为语言模式冗余性因子，通常设为较小的正整数。为了加速检索速度，我们在中文

[11]为了简单起见，我们省略了目标函数的正则项。

Algorithm 16 动词特征抽取算法

初始化 $V(N_1, N_2) = \emptyset$
for 每句句子 $s \in S^k_{q_{verb}}$ **do**
  if $N_1 \in s$ 且 $N_2 \in s$ **then**
    将 N_1N_2 的上下文动词加入 $V(N_1, N_2)$
  end if
end for
抽取 $V(N_1, N_2)$ 中的 Top-r_v 词频的动词集合 $V_{r_v}(N_1, N_2)$
return 动词特征 $f_{verb}(N_1, N_2)$

语料库上建立句子级别的倒排索引，并记 S^k_q 为语料库中查询 q 返回的 Top-k 句子集合。我们假设如果 "N_1 的 N_2" 至少出现了 r_a 次，则 N_1N_2 很可能为透明的。我们定义助词特征如下：[12]

$$f_{aux}(N_1, N_2) = \min\{1, \frac{1}{r_a} \sum_{s \in S^k_{q_{aux}}} I(q_{aux} \in s)\}$$

其中，查询 q_{aux} ="N_1 的 N_2"，$I(\cdot)$ 是指示函数。

动词特征. 这一特征建模两个名词之间有多大可能存在动词表示他们的关系。在中文中，由于动词关系的表达比较灵活 [34]，我们采用算法 16来抽取动词特征。令动词查询 q_{verb} 为 "(N_1 AND N_2) NOT N_1N_2", 对每句句子 $s \in S^k_{q_{verb}}$, 我们抽取 N_1 和 N_2 的依存解析路径中的动词作为 N_1 和 N_2 的上下文动词。令 $V(N_1, N_2)$ 为这一动词的集合，$c(v)$ 是动词 v 的计数。类似地，记 r_v 为动词冗余性因子，$V_{r_v}(N_1, N_2)$ 是 $V(N_1, N_2)$ 中有 Top-r_v 词频的动词子集。假设如果至少有 r_v 个动词出现了至少 r_v 次，这两个名词之间存在关系性动词，表达他们的关系。动词特征定义如下：

$$f_{verb}(N_1, N_2) = \min\{1, \frac{1}{r_v^2} \sum_{v \in V_{r_v}(N_1, N_2)} c(v)\}$$

核心词共现特征. 根据现象 1, 如果 N_1N_2 和 N_2 在同一句句子中共现, N_1N_2 不太可能是完全习语性的。令 r_c 为核心词共现因子，查询 q_{head} 为 "N_2 AND N_1N_2"。$I_c(s, N_1N_2)$ 为指示函数，当且仅当 $N_1N_2 \in s$ 且 $N_2 \in s \setminus \{N_1N_2\}$，函数返回 1。核

[12]我们用 $x \in y$ 来表示 x 是 y 的子串。

表 4.18: RCRL 的特征模板

特征名称	**数学定义**
助词特征	$f_{aux}(N_1, N_2) = \min\{1, \frac{1}{r_a}\sum_{s \in S^k_{q_{aux}}} I(q_{aux} \in s)\}$
动词特征	$f_{verb}(N_1, N_2) = \min\{1, \frac{1}{r_v^2}\sum_{v \in V_{r_v}(N_1,N_2)} c(v)\}$
核心词共现特征	$f_{head}(N_1, N_2) = \min\{1, \frac{1}{r_c}\sum_{s \in S^k_{q_{head}}} I_c(s, N_1N_2)\}$
修饰词扩展的助词特征	$f^m_{aux}(N_1, N_2) = \frac{1}{\tau}\sum_{n_1 \in C_p(N_1)} \text{top}(f_{aux}(n_1, N_2), \tau)$
核心词扩展的助词特征	$f^h_{aux}(N_1, N_2) = \frac{1}{\tau}\sum_{n_2 \in C_{p/2}(N_2)} \text{top}(f_{aux}(N_1, n_2), \tau)$
修饰词扩展的助词特征	$f^m_{verb}(N_1, N_2) = \frac{1}{\tau}\sum_{n_1 \in C_p(N_1)} \text{top}(f_{verb}(n_1, N_2), \tau)$
核心词扩展的助词特征	$f^h_{verb}(N_1, N_2) = \frac{1}{\tau}\sum_{n_2 \in C_{p/2}(N_2)} \text{top}(f_{verb}(N_1, n_2), \tau)$
修饰词扩展的核心词共现特征	$f^m_{head}(N_1, N_2) = \frac{1}{\tau}\sum_{n_1 \in C_p(N_1)} \text{top}(f_{head}(n_1, N_2), \tau)$
核心词扩展的核心词共现特征	$f^h_{head}(N_1, N_2) = \frac{1}{\tau}\sum_{n_2 \in C_{p/2}(N_2)} \text{top}(f_{head}(N_1, n_2), \tau)$

心词共现特征定义如下：[13]

$$f_{head}(N_1, N_2) = \min\{1, \frac{1}{r_c}\sum_{s \in S^k_{q_{head}}} I_c(s, N_1N_2)\}$$

扩展特征. 在某些情况下，前述特征相关的语言模式对于人类是常识性知识，在语料库中出现频率仍然有限。我们采用基于词嵌入的查询扩展技术 [223] 提升文本检索的召回率，并且提出扩展特征。例如，模式 “液体的燃料” 可以用来预测 “固体燃料” 是透明的，因为 “液体” 和 “固体” 语义相似。在这一情况下，“固体” 和 “燃料” 甚至不需要在同一句句子中共现。令 $C_p(w)$ 为词 w 的词嵌入的 p 近邻。对于助词特征，我们将 N_1 替换为 $C_p(N_1)$ 中的词，并且定义修饰词扩展的助词特征为：

$$f^m_{aux}(N_1, N_2) = \frac{1}{\tau}\sum_{n_1 \in C_p(N_1)} \text{top}(f_{aux}(n_1, N_2), \tau)$$

其中，如果 $f_{aux}(n_1, N_2)$ 是 $f_{aux}(\tilde{n}_1, N_2)$ ($\tilde{n}_1 \in C_p(N_1)$) 中前 τ 大的值，在式中，$\text{top}(f_{aux}(n_1, N_2), \tau) = f_{aux}(n_1, N_2)$；否则 $\text{top}(f_{aux}(n_1, N_2), \tau) = 0$。类似地，核心词扩展的助词特征为：

$$f^h_{aux}(N_1, N_2) = \frac{1}{\tau}\sum_{n_2 \in C_{p/2}(N_2)} \text{top}(f_{aux}(N_1, n_2), \tau)$$

我们启发式地采用 $\frac{p}{2}$ 近邻而非 p 近邻，因为核心词的改变对中文复合名词的词义影响比修饰词更大。其他特征细节从略，RCRL 的所有特征模板参见表 4.18 。

[13]我们在这三个特征定义中都使用 $\min\{1, \cdot\}$ 保证了特征的归一化。

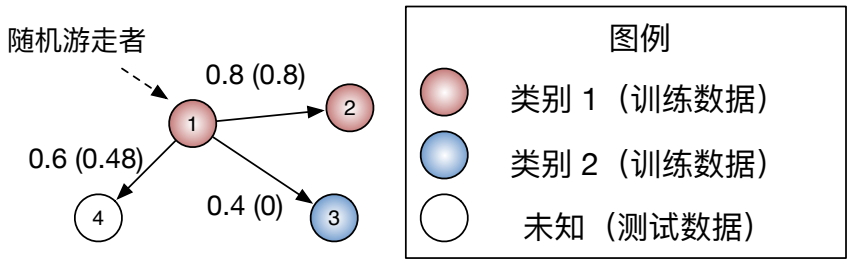


图 4.12: RCRL 的随机游走过程示例

组合性表示学习： 我们扩展了 Cordeiro 等人的工作 [207] 学习中文复合名词的组合性表示。我们将两个复合名词 x_i、x_j 表示为 N_1N_2 和 $N_1^{'}N_2^{'}$，他们的组合性相似度 $\mu_{i,j}$ 定义为：

$$\mu_{i,j} = \frac{1}{2}|\cos(\vec{v}(N_1N_2), \vec{v}(N_1 + N_2)) - \cos(\vec{v}(N_1^{'}N_2^{'}), \vec{v}(N_1^{'} + N_2^{'}))|$$

与 DNRE 所用的技术相似，$\vec{v}(N_1N_2)$ 为 N_1N_2 的复合词嵌入，$\vec{v}(N_1 + N_2)$ 是两个名词词嵌入之和。可以观察到 $\mu_{i,j} \in [0, 1]$，具有相似组合性的复合名词具有相似的 $\mu_{i,j}$ 分数。

为了最小化非监督式损失 $\sum_{x_i,x_j \in L \cup U} \mu_{i,j} ul(\tilde{f}_i, \tilde{f}_j)$，我们采用与 SphereRE 算法相似的图嵌入学习算法。令 $G(\Phi, \Psi, W)$ 为无向、边带权重的完全图，其中节点集合 Φ 表示所有中文复合名词，边集合 Ψ 描述了中文复合名词之间的组合性相似度。W 是边权重向量，对于每条边 $(x_i, x_j) \in \Psi$，我们定义权重 $w_{i,j}$：

$$w_{i,j} = \begin{cases} \mu_{i,j} & x_i \in L, x_j \in L, f_i = f_j \\ 0 & x_i \in L, x_j \in L, f_i \neq f_j \\ \mu_{i,j} \cdot \gamma & \text{其他} \end{cases}$$

其中，$\gamma \in (0, 1)$ 是衰减因子，降低了未标注数据的权重。在图 $G(\Phi, \Psi, W)$ 中，我们采用随机游走模型生成中文复合名字的序列。我们规定，随机游走者从 x_i 到 x_j 的概率正比于 $w_{i,j}$，这一随机游走的示例参见图 4.12。假设 $\mu_{1,2} = 0.8$、$\mu_{1,3} = 0.4$、$\mu_{1,4} = 0.6$、$\gamma = 0.8$，根据随机标签信息，我们有 $w_{1,2} = 0.8$、$w_{1,3} = 0$、$w_{1,4} = 0.48$。因此，随机游走的概率为 $\Pr(1 \to 2) = \frac{0.8}{0.8+0.48}$、$\Pr(1 \to 3) = 0$、$\Pr(1 \to 4) = \frac{0.48}{0.8+0.48}$。

令随机游走的序列为 $\mathcal{S} = \{x_1, \cdots, x_{|\mathcal{S}|}\}$，其中 l 为窗口大小。与 node2vec 模

Algorithm 17 RCRL 的组合性表示学习算法

for 每个中文复合名词 $x_i \in L \cup U$ **do**
    随机初始化组合性表示 $\vec{c}_i$
end for
构建图 $G(\Phi, \Psi, W)$
for $i = 1$ 到最大迭代数 **do**
    从 G 随机采样节点 $x^* \in \Phi$
    生成序列 $\mathcal{S} = \{x_1, \cdots, x_{|\mathcal{S}|}\}$，其中 $x_1 = x^*$
    最小化 $-\sum_{\mathcal{S}} \sum_{x_i \in \mathcal{S}} \sum_{j=i-l(j\neq i)}^{i+l} \log \Pr(x_j|\vec{c}_i)$
end for

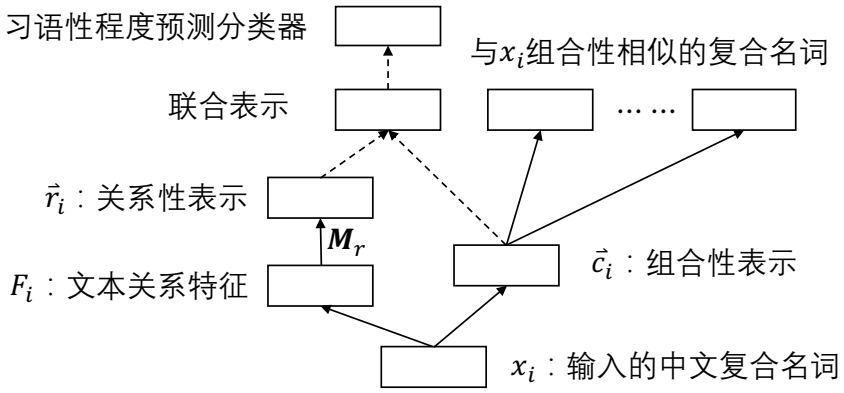


图 4.13: RCRL 的联合优化神经网络架构

型 [166] 相似，最小化损失 $\sum_{x_i, x_j \in L \cup U} \mu_{i,j} ul(\tilde{f}_i, \tilde{f}_j)$ 的目标可以改写为：

$$-\sum_{\mathcal{S}} \sum_{x_i \in \mathcal{S}} \sum_{j=i-l(j\neq i)}^{i+l} \log \Pr(x_j|\vec{c}_i)$$

因此，对于每个中文复合名词 $x_i \in L \cup U$，我们可以学习其组合性表示 $\vec{c}_i$ 使得组合性相似度 $\mu_{i,j}$ 高的中文复合名词具有类似的表示。

RCRL 的组合性表示学习算法见算法 17。在初始化阶段，算法随机初始化所有组合性表示 $\vec{c}_i$，并且构建图 $G(\Phi, \Psi, W)$。在算法的迭代阶段，它从随机选择的起点开始，通过随机游走生成序列 $\mathcal{S}$，并且通过最小化目标函数更新组合性表示 $\vec{c}_i$。

联合优化： 图 4.13中给出了基于多任务学习的 RCRL 联合优化神经网络，其中实线箭头指神经网络的直接连接，虚线箭头指神经网络的若干隐藏层（在 RCRL 的实现中，我们只采用一层隐藏层）。对每个中文复合名词 x_i，RCRL 从文本中抽取特征 $\mathcal{F}_i$，并且计算关系性表示 $\vec{r}_i = \mathbf{M}_r \mathcal{F}_i$。对于组合性表示，$x^{(i)}$ 映射到其表示向量 $\vec{c}_i$，并用其预测 $x^{(i)}$ 在图 $G(\Phi, \Psi, W)$ 中的 “邻居” 节点（即与 $x^{(i)}$ 组合性相似

Algorithm 18 RCRL 的联合优化学习算法

for 每个中文复合名词 $x_i \in L \cup U$ **do**
  随机初始化组合性表示 $\vec{c}_i$
  if $x_i \in L$ **then**
    计算文本关系特征 $\mathcal{F}_i$
  end if
end for
构建图 $G(\Phi, \Psi, W)$
while 算法不收敛 **do**
  for $i = 1$ 到最大迭代数 **do**
    从 G 随机采样节点 $x^* \in \Phi$
    生成序列 $\mathcal{S} = \{x_1, \cdots, x_{|\mathcal{S}|}\}$，其中 $x_1 = x^*$
    通过训练负采样训练器，更新组合性表示 $\vec{c}_i$
  end for
  利用组合性表示 $\vec{c}_i$ 和文本关系特征 $\mathcal{F}_i$ 训练习语性程度分类器
end while

的其他节点)。在网络的另一部分，模型预测习语性程度标签 $\tilde{f}_i$ 基于关系性表示 $\vec{r}_i$ 和组合性表示 $\vec{c}_i$。我们将 RCRL 的整体优化目标改写为：

$$\mathcal{J} = -\sum_{x_i \in L} \sum_{t \in T} I(t = f_i) \log \Pr(\tilde{f}_i = t | \mathcal{F}_i, \vec{c}_i) - \lambda \sum_{\mathcal{S}} \sum_{x_i \in \mathcal{S}} \sum_{j=i-l(j \neq i)}^{i+l} \log \Pr(x_j | \vec{c}_i)$$

其中，T 是所有标签的集合 (即四种习语性程度标签)。

在实际应用中，上式的优化复杂度仍然比较高。为了高效训练 RCRL 模型，我们采用 Skip-Gram 的负采样技术 [51] 进行近似优化。我们训练二分类逻辑斯蒂回归分类器预测任一中文复合名词 x_j 是否在中心复合名词 x_i 的采样序列 $\mathcal{S}$。当分类错误最小化时，组合性表示 $\vec{c}_i$ 也随之更新。将这一训练方法与习语性程度预测模型结合，我们给出了 RCRL 的整体迭代优化算法，参见算法 18。当联合损失 $\mathcal{J}$ 不明显下降时，模型训练终止。

4.5.3 实验分析

在本节中，我们对 RCRL 的实验效果在多个数据集上进行详细评测，并与其他的基线算法作对比。

数据集与实验设置： 在 RCRL 的实验中，我们采用与评测 DNRE 和 PNRE 相同的中文语料库和 FudanNLP [209] 工具，在本节中不再赘述。根据前期的调研，SemTransCNC [224] 是唯一关于中文复合名词的标注数据集。它描述了中文双字复

合词的语义透明性, 例如, “马” 和 “虎” 两个汉字是如何构成词语 “马虎” 的。这一数据集与 RCRL 关注的语言现象有显著的不同, 因此并不适合直接用于评测 RCRL 的效果。另一个相关的数据集由 Qi 等人 [208] 构建，包含了从中文语义词典 HowNet 中抽取的义原（Sememe）短文本。这一数据集用于评测义原短文本的语义相似度, 也不适合直接评测 RCRL。

在本研究中,我们构建两个数据集对 RCRL 进行评测。第一个数据集为 CNCBaike, 它包含从百度百科中抽取的概念类别子集。我们随机从中抽取 2500 个概念类别中的中文复合名词，并令中文母语使用者标注这些复合名词的习语性程度。如果不同的数据标注人员对同一复合名词有不同标注，我们舍弃这一名词，最终得到 1330 个具有人工标注的中文复合名词。第二个数据集为 CNCWeb，包括 815 个标注的中文复合名词，这些复合名词从前述中文语料库中采用基于 POS 的启发式规则抽取。数据标注过程同 CNCBaike。这两个数据集已在 GitHub 上开源 [14]。

在 RCRL 的特征抽取阶段, 我们设为超参数的默认值为：$k = 500$、$r_a = r_v = 3$、$\tau = 2$、$r_c = 20$ 以及 $p = 16$。我们随机将 CNCBaike 数据集分为训练集、验证集和测试集，比例为 70%:10%:20%。因为 CNCWeb 的数据量比较小，我们将 CNCWeb 中所有数据作为测试集，将 CNCBaike 的数据作为训练集。在组合性表示学习阶段, 我们运行算法 5000 个迭代，参数默认设为 $|\mathcal{S}| = 100$、$l = 5, \lambda = 0.1$、$\gamma = 0.8$，两种表示的维度为 $d = 50$。我们也在后续实验中用验证集评测这些超参数的取值对实验效果的影响。

整体实验比较： 由于没有先前的工作与 RCRL 解决同样的习语性程度分类任务，我们设立多个与 RCRL 相关的算法作为强基线算法：

- **词汇关系分类**：我们采用三个经典的分布式词汇关系分类模型用于习语性分类，其特征分别为：$\vec{N}_1 \oplus \vec{N}_2$、$\vec{N}_1 + \vec{N}_2$ 和 $\vec{N}_1 - \vec{N}_2$，分类算法为线性核 SVM 分类器。

- **习语分类**：采用基于神经网络的分类模型 [201], 结合中文复合名词的词嵌入及其上下文进行分类。与文献 [201] 中的原始实现不同，我们对中文复合名词进行习语性程度四分类，而非原始实现的二分类。

- **组合性分析**：我们考虑两种基于词嵌入的组合性分析模型 [205, 207] 作为基线算法。由于这两个算法的输入为连续值，我们在验证集上进行测试选择合

[14] https://chywang.github.io/data/access.zip

表 4.19: 不同习语性程度分类算法在 CNCBaike 和 CNCWeb 的实验效果

数据集	CNCBaike			CNCWeb		
方法	**精准度**	**召回率**	**F 值**	**精准度**	**召回率**	**F 值**
$\vec{N}_1 + \vec{N}_2$	0.622	0.631	0.626	0.512	0.508	0.510
$\vec{N}_1 \oplus \vec{N}_2$	0.663	0.657	0.660	0.508	0.472	0.489
$\vec{N}_1 - \vec{N}_2$	0.567	0.606	0.586	0.597	0.478	0.531
King 和 Cook [201]	0.664	0.691	0.682	0.563	0.582	0.572
Salehi 等人 [205]	0.675	0.663	0.669	0.705	0.648	0.675
Cordeiro 等人 [207]	0.704	0.693	0.698	0.723	0.652	0.686
Pattern	0.770	0.766	0.768	0.745	0.687	0.715
RRL	0.785	0.776	0.780	0.762	0.703	0.731
RCRL	**0.801**	**0.783**	**0.792**	**0.784**	**0.733**	**0.758**

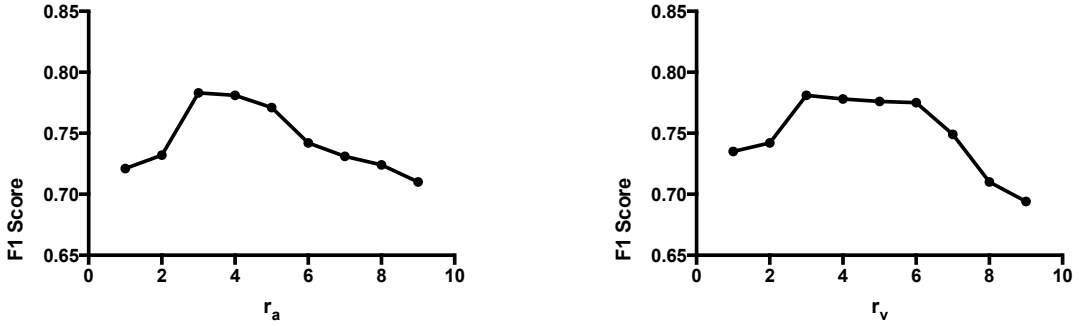
适的阈值，将算法的输出转换为四个习语性程度标签。

- **基于模式的方法**：我们采用基于 SVM 的分类模型直接在特征集 $\mathcal{F}_i$ 进行分类，将这个模型记为 Pattern。

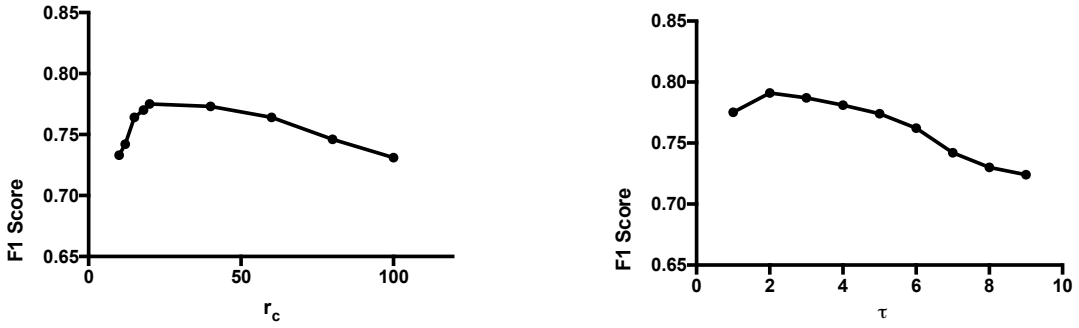
- **RCRL 的变体**：这一模型与 RCRL 类似，去除了组合性表示，记为 RRL。

在两个数据集上的实验结果见表 4.19。词汇关系分类的基线方法在本任务上的效果不佳，其 F 值的范围在 40% 到 60% 之间，这是因为算法没有学习一个中文复合名词中的两个名词之间的关系。King 和 Cook 的方法 [201] 与词汇关系分类方法的实验效果类似。组合性分析的两种基线模型 [205, 207] 与我们的任务最相关，但是其效果比 RCRL 略低，因为这两种算法没有考虑中文语言的特殊语言模式。比较 Pattern、RRL 和 RCRL，我们发现无论是学习关系性表示还是组合性表示，都对中文复合名词的习语性程度的预测有效。整体而言，提出的 RCRL 模型效果显著超越了先前的算法。

参数分析： 在特征抽取阶段，参数 k 和 p 的设置与语料库的大小和质量有关系。对于参数 k 的设置，我们将 k 的值在 $\{50, 100, 200, 500, 1000, 2000\}$ 中进行调整，我们发现当 $k \geq 500$ 时，我们可以抽取到足够多的与查询相关的句子。因此，我们将 k 默认设为 500。参数 p 的调整过程与 k 类似：当 p 过小时，扩展的语言模式与原模式过于相像，因此扩展特征的效果有限；当 p 过大时，扩展的语言模式 “语义漂移” 问题将会严重。在我们的语料库中，合适的设置为 $p = 16$。下一步，我们调整参数 r_a、r_v、r_c 和 τ 的值。在这一实验中，我们采用基于分类器的调参技巧调整上述参数的值。当某一组参数的值设置完毕后，我们在特征集 $\mathcal{F}_i$ 上训练逻辑斯蒂回归分类器，在验证集上进行习语性程度分类，并且采用宏观平均 F 值衡量参数设



(a) 调整参数 r_a　　(b) 调整参数 r_v



(c) 调整参数 r_c　　(d) 调整参数 τ

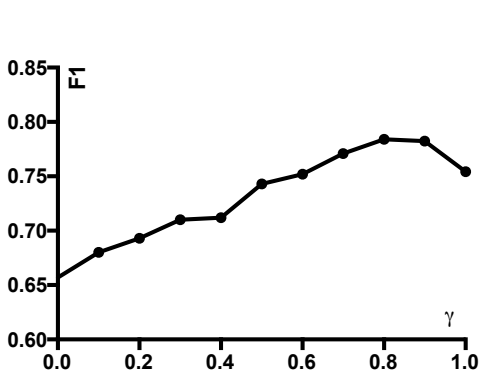
图 4.14: RCRL 的特征抽取中参数 r_a、r_v、r_c 和 τ 的调整实验结果

表 4.20: RCRL 的错误预测案例

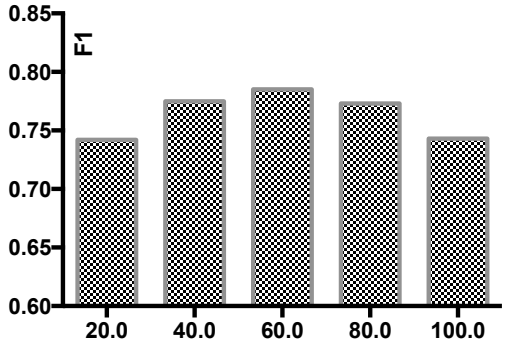
示例	字面翻译	意译	预测结果	真实结果
知识 **青年**	Knowledge **Youth**	Sent-down youth	I	III
青铜 **时代**	Bronze **Age**	The Bronze Age	II	III
邮政 **编码**	Postal service **Coding**	Postal code	IV	II
宗教 **仪式**	Religion **Ceremony**	Religious ceremony	III	II

置的优劣性。我们每次调整其中一个参数的值，将其他参数的值设为默认值，实验结果如图 4.14。

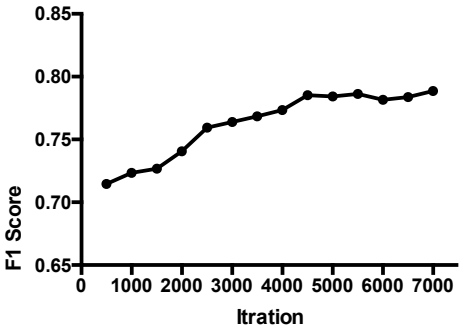
当特征 $\mathcal{F}_i$ 抽取完毕后，我们调整 RCRL 整体算法的参数，实验结果如图 4.15。我们设置参数 γ 和 d 的默认值为：$\gamma = 0.8$ 和 $d = 60$，并且每次调整一个参数的值。从实验结果，我们可以得出两个结论：i）在随机游走过程中采用参数 γ 增强了组合性表示学习的效果；ii）当表示学习向量的维度 d 的设置与词向量的设置相似时，模型的效果更精确。我们也调整了算法 18的迭代次数，每次运行 500 个迭代时，我们对模型的效果进行评测，并且汇报其 F 值。由结果可见，算法的 F 值随着迭代次数的增加而缓慢上升，当算法运行 4500 个迭代后，实验效果变得稳定。



(a) 调整参数 γ



(b) 调整参数 d



(c) 调整迭代次数

图 4.15: RCRL 整体参数调整实验结果

错误分析与案例研究： 我们对 RCRL 的错误案例进行分析，部分错误示例见表 4.20。我们分析了 RCRL 的 300 个错误案例，整体而言，一共有两种主要错误原因：暗喻检测错误（Metaphor Detection Error，缩写为 MDE）和模式缺失错误（Lack-of-Pattern Error，缩写为 LPE）。MDE 占到所有预测错误的 43.8%，因为在部分案例中，模型会受到部分语言模式的误导，低估了中文复合名词的习语性程度。例如，“知识青年” 指 20 世纪 50 年代至 “文化大革命” 结束后受过高等教育的、从城市到农村生活工作的年轻人，具有习语性含义。由于在特征抽取过程中，模型在中文语料库中检测出 “有知识的青年”、“没有知识的青年” 等语言模式，错误判断该复合名词是透明的。剩余的错误类别为 LPE，指和这些复合名词相关的语言模式在海量语料库中仍然比较稀疏，因此导致预测错误。

4.5.4 应用研究

在本节中，我们对 RCRL 的实际应用价值进行研究。我们在三个应用中讨论了中文复合名词的语义性程度预测与中文 NLP 的关系，并且研究 RCRL 是如何促

表 4.21: 中文网络语料库中复合名词的习语性程度分布

方法	类别 I	类别 II	类别 III	类别 IV
Pattern	47.2%	33.6%	15.0%	4.2%
RRL	53.1%	31.2%	14.2%	1.5%
RCRL	51.1%	34.6%	12.2%	2.1%
人工估计值	**(49.2%±1.4%)**	**(38.1%±1.5%)**	**(10.8%±0.4%)**	**(1.9%±1.4%)**

进中文自然语言理解相关研究的。

中文语言的整体习语性分析： 复合名词的习语性影响了自然语言的整体习语性。在本研究中，我们采用数据驱动的方法研究中文网络语料库中，中文复合名词习语性程度的分布。我们从中文网络语料库中随机采样 5000 个复合名词，使用先前实验中三种准确度较高的方法（即 Pattern、RRL 和 RCRL）预测这 5000 个复合名词的习语性程度。在表 4.21中，我们列出了三种模型预测出的中文复合名词习语性程度的分布。此外，我们也从这一数据集合中随机采样了 400 个复合名词三次，每一次人工标注其习语性程度，计算中其真实的分布。在这三次采样之后，利用标准的 t 检验估计分布的置信度区间，其置信度为 95%。由结果可见，与 Pattern 和 RRL 相比，由 RCRL 生成的概率分布与真实的人工估计分布是最接近的。从这一结果的另一个发现是，过半的中文复合名词（约 50.8%）的语义含义是不透明的，在不同程度上具有部分习语性意义。

从本研究得出的数据也部分反映出中文自然语言理解的困难度。我们的研究显示出，如果将中文高度习语性的表达与其他表达式区分处理，其他下游的中文 NLP 任务准确度会进一步提升。比如，当计算中文短文本的语义相似度时，如果这一短文本是高度习语性的，我们应当将这一表达视为一个整体学习其语义表达，而不是将中文短文本中的各个词的词向量加以平均，作为它的整体语义表达。这一做法也和部分其他研究（例如 Qi 等人 [208]）的结论相吻合。

习语性与机器翻译准确度的关系： 如文献 [218] 所示，习语性语言的存在给精准的机器翻译带来很大的技术挑战。在本研究中，我们特别研究中文复合名词的习语性程度与机器翻译准确性的关联性。我们利用两个工业界著名的机器翻译引擎 Google Translation[15]和 Bing Microsoft Translator[16]将 CNCBaike 数据集中的中文复合名词翻译成英语。因为经典的机器翻译评估指标（例如 BLEU）不适合评价中文复合名词的翻译准确性，我们人工标注翻译结果的正确性，汇总不同习语性程度的翻

[15]https://translate.google.com

[16]https://www.bing.com/translator

表 4.22: 不同习语性程度的中文复合名词的机器翻译准确度比较

准确度	类别 I	类别 II	类别 III	类别 IV
Google Translation	98.2%	92.6%	75.0%	64.2%
Microsoft Translator	97.4%	90.2%	78.2%	58.2%

表 4.23: Google Translation 和 Microsoft Translator 对中文复合名词的翻译结果

中文复合名词	Google Translation 结果	Microsoft Translator 结果
夫妻肺片	Couple lungs	Couple lung slices
正确翻译：*Mr and Mrs Smith (Sliced beef and ox organs in chili sauce)*		
竹书纪年	Bamboo book year	The Annals of Bamboo Books
正确翻译：*Bamboo Annals (A chronicle of ancient China)*		
民办教师	Private teacher	Becoming
正确翻译：*Citizen-managed teacher (teachers in rural schools who do not receive the normal remuneration from the government)*		

译准确性，结果见表 4.22。从中可以看见，机器翻译的准确性与习语性程度具有强关联性。语义透明的复合名词的翻译结果绝大部分是准确的，两个翻译引擎的结果分别为 98.2% 和 97.4%。中文复合名词的习语性程度越高，翻译质量越差，

表 4.23中给出了 3 个中文复合名词的正确英语翻译以及两个翻译引擎的翻译结果。由此可见，机器翻译错误源于模型将中文复合名词进行逐词直译，忽略了这些名词的习语性含义，例如，Microsoft Translator 将 “夫妻肺片” 翻译成 “Couple lung slices”。在某些例子中，由于不明原因，机器翻译引擎给出了人类无法解释的翻译，例如，Microsoft Translator 将 “民办教师” 翻译成 “Becoming”。这一现象显示出，目前的机器翻译技术仍然难以处理高习语性的输入。最有可能的原因是，机器翻译算法（无论是统计机器翻译还是采用注意力机制的神经机器翻译 [159]）都倾向于学习不同语言之间的词对齐，忽略了习语性表达的词间组合性。因此，我们提出的任务和算法可以与其他 NLP 任务进行结合，提升模型的自然语言理解能力。

RCRL 对英语组合性预测任务的应用： 尽管 RCRL 解决的是中文复合名词的习语性问题，我们研究 RCRL 能否应用于英语语言。在英语中与 RCRL 解决的任务最接近的任务是复合名词的组合性分析。我们在英语语言上实现了 RCRL 的变体，在广泛应用的 Reddy 数据集 [203] 上评测它在英语复合名词的组合性分析任务的实验效果。在实现中，因为我们使用中文助词模式进行特征抽取，我们将这些模式翻译成英语，例如 “[...] of [...]”、“[...]’s [...]”、“[...] that is [...]”、“[...] which is [...]” 等。RCRL 中的组合性表示是语言独立的，所以在英语语言中不需要改变。由于 Reddy 数据集的任务是对组合性分数进行预测，我们将 RCRL 的分类错误损失

表 4.24: 英语复合名词的组合性预测结果

方法	**斯皮尔曼相关系数 ρ**
Reddy 等人 [203]	0.71
Salehi 等人 [205]	0.80
Cordeiro 等人 [207]	0.82
RCRL	0.81

改为回归损失，定义如下：

$$\mathcal{J} = \sum_{x_i \in L} (\tilde{s}(x_i; \mathcal{F}_i, \vec{c}_i) - s_i)^2 - \lambda \sum_{\mathcal{S}} \sum_{x_i \in \mathcal{S}} \sum_{j=i-l(j\neq i)}^{i+l} \log \Pr(x_j | \vec{c}_i)$$

其中，$\tilde{s}(x_i; \mathcal{F}_i, \vec{c}_i)$ 是英语复合名词 x_i 的预测组合性分数，s_i 为其人工标注分数。

在测试阶段，我们将预测与真实得分相对比，以斯皮尔曼相关系数 ρ 作为评测指标。实验设置与英语语料库的设置同文献 [207]，实验结果见表 4.24。由此可见，我们的方法 RCRL 的实验结果为 $\rho = 0.81$，比 Reddy 等人 [203] 提出的模型更加精确，也与两个近年提出的 SOTA 方法 [205, 207] 结果相似。因此，RCRL 不完全是针对中文特定语言的，也可以在英语数据集上应用。

4.6 小结

在本章中，我们分别从频繁模式挖掘和数据驱动两个不同的角度，提出了 PNRE 和 DNRE 两种算法，从中文短文本中抽取多种类别的语义关系。其中，PNRE 采用非监督的图挖掘技术，从中文短文本中挖掘出表达语义关系的频繁语言模式，并且自动抽取出相应的关系三元组；DNRE 克服了 PNRE 只能抽取频繁语言模式对应关系这一弱点，提出了三阶段的数据驱动算法，在不依靠训练数据的情况下，完成从短文本切分到关系生成的完整算法流程，抽取出更多数量的关系三元组。在中文维基百科数据上的实验表明了上述算法可以在不预先定义待抽取关系类别的情况下，获得多种高精度关系。在此基础上，我们进一步注意到，通过利用习语性语义理解技术，我们可以从中文关系类别短文本中推导出更多关系，并且提出 RCRL 的表示学习框架，实现中文复合名词的习语性程度预测。与第二章、第三章的研究工作相比，本章提出的算法在基于中文短文本的关系抽取过程中，不受预定义关系类别限制，取得了较好的效果。

第五章　总结与展望

本章对本文中面向中文短文本的关系抽取研究工作进行简要总结，并且针对自然语言处理的发展趋势，提出对未来工作的研究展望。

5.1　总结

互联网中的海量、异构、碎片化数据给人们快速获得所需的知识带来了挑战。关系抽取作为 NLP 中最基础的任务之一，从无结构化的数据源中抽取出结构化的知识，为大规模知识图谱的构建和补全提出了技术基础，支撑语义检索、智能问答、机器阅读理解等多个 NLP 任务。然而目前关系抽取的研究主要面向句子和文档级，不能有效从中文短文本获取上下文高度稀疏、表达方式各异的知识。因此，本文研究基于海量互联网中的中文短文本，深入探索如何利用深度学习和文本挖掘技术，同时考虑中文自然语言的特性，从多个角度解决中文短文本的关系抽取问题，给出了较为完整的解决方案。下文简要总结本文面向中文短文本的关系抽取框架中的主要技术贡献。

1. **基于词嵌入的上下位关系抽取**：分类体系是大规模知识图谱中概念的层次化表示和组织的重要形式，由大量上下位关系构成。与英语语言相比，由于中文语言表述具有高度灵活性，中文上下位关系抽取不能简单利用文本匹配的方法来实现。我们结合深度神经语言模型的最新研究成果和中文语言本身的特点，采用词嵌入作为中文术语的特征表示，建模中文上下位关系在词嵌入空间的表示，即学习中文下位词的词向量是如何映射其上位词的词向量的。这一部分的研究工作包括了三个模型：**半监督式上下位关系扩展模型**（IPM）、**基于转导学习的上下位关系分类模型**（TPM）和**基于模糊正交投影的上下位关系分类模型**（FOPM）。实验结果表明，IPM 对于中文上下位关系的扩展、TPM 和 FOPM 对于中文上下位关系分类的效果明显，超过了现有最佳方法。

2. **知识增强的语义关系抽取**：上述基于词嵌入的上下位关系抽取模型对人工标注的训练集高度依赖，对外部知识和其他辅助任务没有加以良好运用。我们以词嵌入投影模型作为基础，深入地探索知识增强的语义关系抽取算法，从

多知识源、多语言、多词汇关系三个角度，扩展了这一类算法的应用空间。其中，**分类体系增强的对抗学习框架**（TEAL）利用深度对抗学习机制，将大规模分类体系中的上下位关系知识融入基于训练集的基础词嵌入投影神经网络中；**迁移模糊正交投影模型**（TFOPM），及其扩展算法**迭代迁移模糊正交投影模型**（ITFOPM）结合了深度迁移学习和双语术语对齐技术，实现了面向小语种的跨语言上下位关系预测；**超球关系嵌入模型**（SphereRE）将多种类别的词汇关系分别进行语义建模，学习这些词汇关系的超球嵌入表示，使模型可以对多种词汇关系进行分类。相应 NLP 任务的实验效果证明了这三种模型的有效性。

3. **非上下位关系抽取与语义理解**：中文短文本中往往存在多种类别的非上下位关系，前述模型预测的关系类别是人工预先定义的，难以扩展至开放领域，而且对中文短文本的语义缺乏深度理解。我们首先提出两种非监督的关系挖掘算法，从中文短文本中抽取多种类别的语义关系。其中，**基于模式的非上下位关系抽取算法**（PNRE）采用图挖掘算法，从数据源挖掘出表达语义关系的频繁语言模式，自动抽取出与这些模式相对应的非上下位关系三元组；**数据驱动的非上下位关系抽取算法**（DNRE）为三阶段的数据驱动算法，克服了 PNRE 只能抽取频繁模式对应关系的缺点，实现从中文短文本切分到关系生成的完整算法流程，提升了关系抽取的覆盖率。我们进一步观察到，习语性语义理解技术有助于从中文短文本中推导出更多关系，并且据此提出**关系性与组合性表示学习框架**（RCRL），对中文复合名词的习语性程度进行预测。实验结果表明，这些算法在中文短文本的关系抽取过程中，不受人工定义关系类别约束，取得较为精确的实验结果。

这三大模块在主题上有机统一，都旨在解决中文短文的关系抽取问题；在技术路线上，互相关联、层层递进、逐渐深入。其中，**基于词嵌入的上下位关系抽取**主要面向中文上下位关系，构建了基于词嵌入的关系预测模型的技术基础；**知识增强的语义关系抽取**将目标扩展至多知识源、多语言、多词汇关系的场景，并且扩展了基于词嵌入的关系预测模型的应用空间；**非上下位关系抽取与语义理解**则从封闭域下的关系抽取拓展至开放域，在采用词嵌入模型进行语义计算的基础上，挖掘出种类繁多的中文非上下位关系。无论是理论分析，还是实验结果，都证明了上述研究框架具有高度有效性。

5.2 未来工作展望

我们将继续研究面向中文短文本的关系抽取中的几个重要方向，包括融合异构知识源的中文关系抽取、基于神经网络的复杂语义关系自动推理、常识性知识的表示学习与关系补全、编码中文语言学知识的神经网络模型，以及基于深度语言模型的关系抽取与理解等问题，以促进中文语言的机器智能认知和理解。

1. **融合异构知识源的中文关系抽取**：本文提出的中文关系抽取算法主要关注短文本，与句子级别的关系抽取神经网络模型 [35, 171, 172, 174, 176] 可以构成互补关系。因此，可以通过多视图学习 [62]、对抗学习 [29, 120] 等机制，融合各种结构的中文文本（例如短文本、完整语句（句子级别或文档级别）、属性-值对等形式的半结构化数据等）作为知识源，协同训练关系抽取模型，达到互相增强、互相补充的目的。

2. **基于神经网络的复杂语义关系自动推理**：语义关系本身的性质与其表示学习与抽取关系密切，例如双曲正切嵌入空间适合建模上下位关系的传递性 [75, 76]，部分逻辑表达式的应用可以提升知识图谱表示学习的效果 [225]。在我们的研究中，IPM、TPM、FOPM 和 SphereRE 等模型都在不同的向量空间建模了特定语义关系的性质。然而，这些方法仍然需要大量人类观察的结果作为建模的依据，难以适应任意关系的复杂语义。在未来的研究工作中，我们计划探索利用神经网络进行复杂语义关系的自动推理，从而更好地学习不同类别关系在词嵌入空间的表示。

3. **常识性知识的表示学习与关系补全**：基于深度学习的 NLP 模型的缺陷之一在于他们对于常识性知识的表示和推理能力较弱，阻碍了模型对理解自然语言能力的提升。在本文对 DNRE 和 RCRL 算法的研究中，我们都发现常识性知识的挖掘和应用对中文短文本的关系抽取和语义理解至关重要。我们计划以现有常识性知识图谱（例如 [179]）作为基础训练数据，利用深度语言模型作为常识性知识的表示学习算法 [196]，联合学习面向海量语料库的文本解释模型 [184, 185, 187, 189]，探索迭代式常识性知识表示学习和知识挖掘框架，以提升本研究中关系抽取的覆盖率。

4. **编码中文语言学知识的神经网络模型**：本文提出的 IPM、TPM、DNRE 和 RCRL 四种模型都在不同程度上融入了中文语言学知识或语法规则，取得了

较好的实验效果。在以后的研究中，我们将重点考虑如何在深度神经网络中编码中文语言学知识（例如参考 [226, 227] 等），使得神经网络在学习文本特征的同时，不违反现有规则并且试图挖掘更多语义规则，使得模型预测的结果更符合自然语言的规则，提升模型的可解释性。

5. **基于深度语言模型的关系抽取与理解**：随着深度学习技术在 NLP 的进一步发展，深度语言模型（Deep Language Models）可以对自然语言中的各种元素进行更加精确的深度表示学习。这些模型包括 ELMo [228]、BERT [229]、Transformer-XL [230]、XLNet [231] 等。由于本论文研究的对象是中文短文本，缺乏足够的上下文信息，直接将深度语言模型应用于本文的研究具有一定挑战性。在未来，我们可以将深度语言模型的表示学习能力与现有方法结合，提升现有方法的准确度。

参考文献

[1] 中国互联网络发展状况统计报告[EB/OL]. 2019. http://www.cac.gov.cn/pdf/20190829/44.pdf.

[2] Idc white paper[EB/OL]. 2018. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

[3] LU R, JIN X, ZHANG S, et al. A study on big knowledge and its engineering issues [J]. IEEE Trans. Knowl. Data Eng., 2019, 31(9): 1630–1644.

[4] Google knowledge graph[EB/OL]. 2012. https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html.

[5] Satori[EB/OL]. 2013. https://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/.

[6] 百度知心[EB/OL]. 2015. https://tupu.baidu.com/xiaoyuan/.

[7] 搜狗知识搜索[EB/OL]. 2012. http://zhishi.sogou.com/.

[8] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: a core of semantic knowledge[C]//Proceedings of the 16th International Conference on World Wide Web (WWW). Banff, Alberta, Canada: ACM, 2007: 697–706.

[9] CARLSON A, BETTERIDGE J, KISIEL B, et al. Toward an architecture for never-ending language learning[C]//Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI). Atlanta, Georgia, USA: AAAI, 2010: 1306–1313.

[10] WU W, LI H, WANG H, et al. Probase: a probabilistic taxonomy for text understanding[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD). Scottsdale, AZ, USA: ACM, 2012: 481–492.

[11] XU B, XU Y, LIANG J, et al. Cn-dbpedia: A never-ending chinese knowledge extraction system[C]//Advances in Artificial Intelligence: From Theory to Practice - 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE). Arras, France: Springer, 2017: 428–438.

[12] NIU X, SUN X, WANG H, et al. Zhishi.me - weaving chinese linking open data[C]//Proceedings of 10th International Conference on Semantic Web (ISWC). Bonn, Germany: Springer, 2011: 205–220.

[13] WANG Z, LI J, WANG Z, et al. Xlore: A large-scale english-chinese bilingual knowledge graph[C]//Proceedings of the ISWC 2013 Posters & Demonstrations Track. Sydney, Australia: Springer, 2013: 121–124.

[14] SHEN W, WANG J, HAN J. Entity linking with a knowledge base: Issues, techniques, and solutions[J]. IEEE Trans. Knowl. Data Eng., 2015, 27(2): 443–460.

[15] WANG Z, ZHAO K, WANG H, et al. Query understanding through knowledge-based conceptualization[C]//Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI). Buenos Aires, Argentina: IJCAI Organization, 2015: 3264–3270.

[16] MANTLE M, BATSAKIS S, ANTONIOU G. Large scale distributed spatio-temporal reasoning using real-world knowledge graphs[J]. Knowl.-Based Syst., 2019, 163: 214–226.

[17] MILLER G A. Wordnet: A lexical database for english[J]. Commun. ACM, 1995, 38(11): 39–41.

[18] 知网[EB/OL]. 2013. http://www.keenage.com/.

[19] PONZETTO S P, STRUBE M. Deriving a large-scale taxonomy from wikipedia [C]//Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence (AAAI). Vancouver, British Columbia, Canada: AAAI, 2007: 1440–1445.

[20] CHEN J, WANG A, CHEN J, et al. Cn-probase: A data-driven approach for large-scale chinese taxonomy construction[C]//Proceedings of the 35th IEEE International Conference on Data Engineering. Macao, China: IEEE, 2019: 1706–1709.

[21] LI J, WANG C, HE X, et al. User generated content oriented chinese taxonomy construction[C]//Web Technologies and Applications - 17th Asia-Pacific Web Conference (APWeb). Guangzhou, China: Springer, 2015: 623–634.

[22] BARZEGAR S, DAVIS B, HANDSCHUH S, et al. Classification of composite semantic relations by a distributional-relational model[J]. Data Knowl. Eng., 2018, 117: 319–335.

[23] ROLLER S, ERK K. Relations such as hypernymy: Identifying and exploiting hearst patterns in distributional vectors for lexical entailment[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP). Austin, Texas, USA: ACL, 2016: 2163–2172.

[24] SANTUS E, SHWARTZ V, SCHLECHTWEG D. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Valencia, Spain: ACL, 2017: 65–75.

[25] FU R, GUO J, QIN B, et al. Learning semantic hierarchies via word embeddings [C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL). Baltimore, MD, USA: ACL, 2014: 1199–1209.

[26] CAI R, ZHANG X, WANG H. Bidirectional recurrent convolutional neural network for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL). Berlin, Germany: ACL, 2016: 756–765.

[27] ETZIONI O, FADER A, CHRISTENSEN J, et al. Open information extraction: The second generation[C]//Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI). Barcelona, Catalonia, Spain: IJCAI Organization, 2011: 3–10.

[28] MAUSAM, SCHMITZ M, SODERLAND S, et al. Open language learning for information extraction[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). Jeju Island, Korea: ACL, 2012: 523–534.

[29] QIN P, XU W, WANG W Y. DSGAN: generative adversarial training for distant supervision relation extraction[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne, Australia: ACL, 2018: 496–505.

[30] TAKANOBU R, ZHANG T, LIU J, et al. A hierarchical framework for relation extraction with reinforcement learning[C]//Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI). Honolulu, Hawaii, USA: AAAI, 2019: 7072–7079.

[31] YAHYA M, WHANG S, GUPTA R, et al. Renoun: Fact extraction for nominal attributes[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: ACL, 2014: 325–335.

[32] PASUPAT P, HAKKANI-TÜR D. Unsupervised relation detection using automatic alignment of query patterns extracted from knowledge graphs and query click logs [C]//Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH). Dresden, Germany: ISCA-Speech, 2015: 2714–2718.

[33] NASTASE V, STRUBE M. Decoding wikipedia categories for knowledge acquisition[C]//Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI). Chicago, Illinois, USA: AAAI, 2008: 1219–1224.

[34] QIU L, ZHANG Y. ZORE: A syntax-based system for chinese open relation extraction[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: ACL, 2014: 1870–1880.

[35] WEN J, SUN X, REN X, et al. Structure regularized neural network for entity relation classification for chinese literature text[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). New Orleans, Louisiana, USA: ACL, 2018: 365–370.

[36] LI Z, DING N, LIU Z, et al. Chinese relation extraction with multi-grained information and external linguistic knowledge[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL). Florence, Italy: ACL, 2019: 4377–4386.

[37] LI H, WU X, LI Z, et al. A relation extraction method of chinese named entities based on location and semantic features[J]. Appl. Intell., 2013, 38(1): 1–15.

[38] HEARST M A. Automatic acquisition of hyponyms from large text corpora[C]// Proceedings of the 14th International Conference on Computational Linguistics (COLING). Nantes, France: ACL, 1992: 539–545.

[39] WASHIO K, KATO T. Neural latent relational analysis to capture lexical semantic relations in a vector space[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium: ACL, 2018: 594–600.

[40] WANG C, GAO M, HE X, et al. Challenges in chinese knowledge graph construction[C]//Proceedings of the 31st IEEE International Conference on Data Engineering Workshops (ICDE Workshops). Seoul, South Korea: IEEE, 2015: 59–61.

[41] JIANG J, ZHAI C. A systematic exploration of the feature space for relation extraction[C]//Proceedings of the 2007 Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (NAACL-HLT). Rochester, New York, USA: ACL, 2007: 113–120.

[42] XU J, WEN J, SUN X, et al. A discourse-level named entity recognition and relation extraction dataset for chinese literature text[J/OL]. CoRR, 2017, abs/1711.07010. http://arxiv.org/abs/1711.07010.

[43] EL-KISHKY A, SONG Y, WANG C, et al. Scalable topical phrase mining from text corpora[J]. PVLDB, 2014, 8(3): 305–316.

[44] LIU J, SHANG J, WANG C, et al. Mining quality phrases from massive text corpora [C]//Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015. Melbourne, Victoria, Australia: ACM, 2015: 1729–1744.

[45] LI P, ZHU Q, ZHOU G, et al. Global inference to chinese temporal relation extraction[C]//Proceddings of the 26th International Conference on Computational Linguistics. Osaka, Japan: ACL, 2016: 1451–1460.

[46] ZHANG N, DENG S, SUN Z, et al. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Minneapolis, MN, USA: ACL, 2019: 3016–3025.

[47] MAUSAM. Open information extraction systems and downstream applications [C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI). New York, NY, USA: IJCAI Organization, 2016: 4074–4077.

[48] YAO Y, YE D, LI P, et al. Docred: A large-scale document-level relation extraction dataset[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics. Florence, Italy: ACL, 2019: 764–777.

[49] SHWARTZ V, DAGAN I. Still a pain in the neck: Evaluating text representations on lexical composition[J]. Trans. Assoc. Comput. Linguistics, 2019, 7: 403–419.

[50] SHARP B, SEDES F, LUBASZEWSKI W. Cognitive approach to natural language processing[M]. [S.l.]: Elsevier, 2017.

[51] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C]//Proceedings of the 1st International Conference on Learning Representations Workshop Track (ICLR). Scottsdale, Arizona, USA: ICLR, 2013.

[52] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: ACL, 2014: 1532–1543.

[53] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[J]. TACL, 2017, 5: 135–146.

[54] SHWARTZ V, GOLDBERG Y, DAGAN I. Improving hypernymy detection with an integrated path-based and distributional method[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL). Berlin, Germany: ACL, 2016: 2389–2398.

[55] SEITNER J, BIZER C, ECKERT K, et al. A large database of hypernymy relations extracted from the web[C]//Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC). Portorož, Slovenia: ELRA, 2016: 360–367.

[56] LUU A T, KIM J, NG S. Taxonomy construction using syntactic contextual evidence[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: ACL, 2014: 810–819.

[57] SNOW R, JURAFSKY D, NG A Y. Learning syntactic patterns for automatic hypernym discovery[C]//Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems (NIPS). Vancouver, British Columbia, Canada: NIPS, 2004: 1297–1304.

[58] BANSAL M, BURKETT D, DE MELO G, et al. Structured learning for taxonomy induction with belief propagation[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL). Baltimore, MD, USA: ACL, 2014: 1041–1051.

[59] NAVIGLI R, VELARDI P. Learning word-class lattices for definition and hypernym extraction[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL). Uppsala, Sweden: ACL, 2010: 1318–1327.

[60] NAKASHOLE N, WEIKUM G, SUCHANEK F M. PATTY: A taxonomy of relational patterns with semantic types[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). Jeju Island, Korea: ACL, 2012: 1135–1145.

[61] KOZAREVA Z, HOVY E H. A semi-supervised method to learn and construct taxonomies using the web[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP). MIT State Center, Massachusetts, USA: ACL, 2010: 1110–1118.

[62] CARLSON A, BETTERIDGE J, WANG R C, et al. Coupled semi-supervised learning for information extraction[C]//Proceedings of the Third International Conference on Web Search and Web Data Mining (WSDM). New York, NY, USA: ACM, 2010: 101–110.

[63] FU R, QIN B, LIU T. Exploiting multiple sources for open-domain hypernym discovery[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP). Seattle, Washington, USA: ACL, 2013: 1224–1234.

[64] ALFARONE D, DAVIS J. Unsupervised learning of an IS-A taxonomy from a limited domain-specific corpus[C]//Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI). Buenos Aires, Argentina: IJCAI Organization, 2015: 1434–1441.

[65] GUPTA A, LEBRET R, HARKOUS H, et al. Taxonomy induction using hypernym subsequences[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM). Singapore: ACM, 2017: 1329–1338.

[66] ROLLER S, KIELA D, NICKEL M. Hearst patterns revisited: Automatic hypernym detection from large text corpora[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne, Australia: ACL, 2018: 358–363.

[67] WEEDS J, WEIR D J. A general framework for distributional similarity[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Sapporo, Japan: ACL, 2003: 81–88.

[68] LENCI A, BENOTTO G. Identifying hypernyms in distributional semantic spaces [C]//Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM). Montréal, Canada: ACL, 2012: 75–79.

[69] CLARKE D. Context-theoretic semantics for natural language: an overview[C]// Proceedings of the Workshop on Geometrical Models of Natural Language Semantics (GEMS). Athens, Greece: ACL, 2009: 112—119.

[70] SANTUS E, LENCI A, LU Q, et al. Chasing hypernyms in vector spaces with entropy[C]//Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Gothenburg, Sweden: ACL, 2014: 38–42.

[71] ROLLER S, ERK K, BOLEDA G. Inclusive yet selective: Supervised distributional hypernymy detection[C]//Proceedings of the 25th International Conference on Computational Linguistics (COLING). Dublin, Ireland: ACL, 2014: 1025–1036.

[72] NGUYEN K A, KÖPER M, Schulte im Walde S, et al. Hierarchical embeddings for hypernymy detection and directionality[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). Copenhagen, Denmark: ACL, 2017: 233–243.

[73] CHANG H, WANG Z, VILNIS L, et al. Distributional inclusion vector embedding for unsupervised hypernymy detection[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). New Orleans, Louisiana, USA: ACL, 2018: 485–495.

[74] LIANG J, ZHANG Y, XIAO Y, et al. On the transitivity of hypernym-hyponym relations in data-driven lexical taxonomies[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI). San Francisco, California, USA: AAAI, 2017: 1185–1191.

[75] NICKEL M, KIELA D. Learning continuous hierarchies in the lorentz model of hyperbolic geometry[C]//Proceedings of the 35th International Conference on Machine Learning (ICML). Stockholmsmässan, Stockholm, Sweden: ICML, 2018: 3776–3785.

[76] GANEA O, BÉCIGNEUL G, HOFMANN T. Hyperbolic entailment cones for learning hierarchical embeddings[C]//Proceedings of the 35th International Conference on Machine Learning (ICML). Stockholmsmässan, Stockholm, Sweden: ICML, 2018: 1632–1641.

[77] BARONI M, BERNARDI R, DO N, et al. Entailment above the word level in distributional semantics[C]//Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Avignon, France: ACL, 2012: 23–32.

[78] WEEDS J, CLARKE D, REFFIN J, et al. Learning to distinguish hypernyms and co-hyponyms[C]//Proceedings of the 25th International Conference on Computational Linguistics (COLING). Dublin, Ireland: ACL, 2014: 2249–2259.

[79] TURNEY P D, MOHAMMAD S M. Experiments with three approaches to recognizing lexical entailment[J]. Natural Language Engineering, 2015, 21(3): 437–476.

[80] YU Z, WANG H, LIN X, et al. Learning term embeddings for hypernymy identification[C]//Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI). Buenos Aires, Argentina: IJCAI Organization, 2015: 1390–1397.

[81] LUU A T, TAY Y, HUI S C, et al. Learning term embeddings for taxonomic relation identification using dynamic weighting neural network[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP). Austin, Texas, USA: ACL, 2016: 403–413.

[82] LEVY O, REMUS S, BIEMANN C, et al. Do supervised distributional methods really learn lexical inference relations?[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Denver, Colorado, USA: ACL, 2015: 970–976.

[83] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735–1780.

[84] LE M, ROLLER S, PAPAXANTHOS L, et al. Inferring concept hierarchies from text corpora via hyperbolic embeddings[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL). Florence, Italy: ACL, 2019: 3231–3241.

[85] HELD W, HABASH N. The effectiveness of simple hybrid systems for hypernym discovery[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL). Florence, Italy: ACL, 2019: 3362–3367.

[86] YAMANE J, TAKATANI T, YAMADA H, et al. Distributional hypernym generation by jointly learning clusters and projections[C]//Proceedings of the 26th International Conference on Computational Linguistics (COLING). Osaka, Japan: ACL, 2016: 1871–1879.

[87] BIEMANN C, USTALOV D, PANCHENKO A, et al. Negative sampling improves hypernymy extraction based on projection learning[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Valencia, Spain: ACL, 2017: 543–550.

[88] KHULLER S, MOSS A, NAOR J. The budgeted maximum coverage problem[J]. Inf. Process. Lett., 1999, 70(1): 39–45.

[89] MENON A, MEHROTRA K, MOHAN C K, et al. Characterization of a class of sigmoid functions with applications to neural networks[J]. Neural Networks, 1996, 9(5): 819–835.

[90] LIU H, YANG Y. Bipartite edge prediction via transductive learning over product graphs[C]//Proceedings of the 32nd International Conference on Machine Learning (ICML). Lille, France: ICML, 2015: 1880–1888.

[91] XU R, YANG Y, LIU H, et al. Cross-lingual text classification via model translation with limited dictionaries[C]//Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM). Indianapolis, IN, USA: ACM, 2016: 95–104.

[92] MIRZA P, TONELLI S. On the contribution of word embeddings to temporal relation classification[C]//Proceedings of the 26th International Conference on Computational Linguistics (COLING). Osaka, Japan: ACL, 2016: 2818–2828.

[93] BARONI M, LENCI A. How we BLESSed distributional semantic evaluation[C]// Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics. Edinburgh, UK: ACL, 2011: 1–10.

[94] XING C, WANG D, LIU C, et al. Normalized word embedding and orthogonal transform for bilingual word translation[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Denver, Colorado, USA: ACL, 2015: 1006–1011.

[95] MARKLEY F L, CRASSIDIS J L. Fundamentals of spacecraft attitude determination and control: volume 33[M]. [S.l.]: Springer, 2014.

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

[97] VELARDI P, FARALLI S, NAVIGLI R. Ontolearn reloaded: A graph-based algorithm for taxonomy induction[J]. Computational Linguistics, 2013, 39(3): 665–707.

[98] ZHANG M, LIU Y, LUAN H, et al. Adversarial training for unsupervised bilingual lexicon induction[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). Vancouver, Canada: ACL, 2017: 1959–1970.

[99] CHEN X, SHI Z, QIU X, et al. Adversarial multi-criteria learning for chinese word segmentation[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). Vancouver, Canada: ACL, 2017: 1193–1203.

[100] FLATI T, VANNELLA D, PASINI T, et al. Two is bigger (and better) than one: the wikipedia bitaxonomy project[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL). Baltimore, MD, USA: ACL, 2014: 945–955.

[101] LIANG J, XIAO Y, WANG H, et al. Probase+: Inferring missing links in conceptual taxonomies[J]. IEEE Trans. Knowl. Data Eng., 2017, 29(6): 1281–1295.

[102] LAMPLE G, CONNEAU A, RANZATO M, et al. Word translation without parallel data[C]//Proceedings of the 6th International Conference on Learning Representations (ICLR). Vancouver, BC, Canada: ICLR, 2018.

[103] NECSULESCU S, MENDES S, JURGENS D, et al. Reading between the lines: Overcoming data sparsity for accurate classification of lexical relationships[C]// Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM). Denver, Colorado, USA: ACL, 2015: 182–192.

[104] SHWARTZ V, DAGAN I. Path-based vs. distributional information in recognizing lexical semantic relations[C]//Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex@COLING). Osaka, Japan: ACL, 2016: 24–29.

[105] WASHIO K, KATO T. Filling missing paths: Modeling co-occurrences of word pairs and dependency paths for recognizing lexical semantic relations[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT). New Orleans, Louisiana, USA: ACL, 2018: 1123–1133.

[106] GLAVAS G, VULIC I. Discriminating between lexico-semantic relations with the specialization tensor model[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). New Orleans, Louisiana, USA: ACL, 2018: 181–187.

[107] MCCRAE J P, QUATTRI F, UNGER C, et al. Modelling the semantics of adjectives in the ontology-lexicon interface[C]//Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex@COLING). Dublin, Ireland: ACL, 2014: 198–209.

[108] CHEN M, TIAN Y, CHEN X, et al. On2vec: Embedding-based relation prediction for ontology population[C]//Proceedings of the 2018 SIAM International Conference on Data Mining (SDM). SanDiego, CA, USA: SIAM, 2018: 315–323.

[109] LIU W, ZHANG Y, LI X, et al. Deep hyperspherical learning[C]//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017 (NIPS). Long Beach, CA, USA: NIPS, 2017: 3950–3960.

[110] DAVIDSON T R, FALORSI L, CAO N D, et al. Hyperspherical variational auto-encoders[C]//Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI). Monterey, California, USA: AUAI, 2018: 856–865.

[111] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014 (NIPS). Montreal, Quebec, Canada: NIPS, 2014: 2672–2680.

[112] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[C]//Proceedings of the 4th International Conference on Learning Representations (ICLR). San Juan, Puerto Rico: ICLR, 2016.

[113] WANG K, ZHAO R, SU H, et al. Generalizing eye tracking with bayesian adversarial learning[C]//Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 11907–11916.

[114] XIONG W, LUO W, MA L, et al. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE, 2018: 2364–2373.

[115] YU L, ZHANG W, WANG J, et al. Seqgan: Sequence generative adversarial nets with policy gradient[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI). San Francisco, California, USA: AAAI, 2017: 2852–2858.

[116] DE MASSON D'AUTUME C, ROSCA M, RAE J W, et al. Training language gans from scratch[J/OL]. CoRR, 2019, abs/1905.09922. http://arxiv.org/abs/1905.09922.

[117] ALZANTOT M, SHARMA Y, ELGOHARY A, et al. Generating natural language adversarial examples[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium: ACL, 2018: 2890–2896.

[118] ZHANG H, ZHOU H, MIAO N, et al. Generating fluent adversarial examples for natural languages[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL). Florence, Italy: ACL, 2019: 5564–5569.

[119] REN S, DENG Y, HE K, et al. Generating natural language adversarial examples through probability weighted word saliency[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL). Florence, Italy: ACL, 2019: 1085–1097.

[120] WU Y, BAMMAN D, RUSSELL S J. Adversarial training for relation extraction [C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). Copenhagen, Denmark: ACL, 2017: 1778–1783.

[121] SHI G, FENG C, HUANG L, et al. Genre separation network with adversarial training for cross-genre relation extraction[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium: ACL, 2018: 1018–1023.

[122] ZHOU J T, ZHANG H, JIN D, et al. Dual adversarial neural transfer for low-resource named entity recognition[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL). Florence, Italy: ACL, 2019: 3461–3471.

[123] WANG Y, LEE H. Learning to encode text as human-readable summaries using generative adversarial networks[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium: ACL, 2018: 4187–4195.

[124] QIN L, ZHANG Z, ZHAO H, et al. Adversarial connective-exploiting networks for implicit discourse relation classification[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). Vancouver, Canada: ACL, 2017: 1006–1017.

[125] CAI L, WANG W Y. KBGAN: adversarial learning for knowledge graph embeddings[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). New Orleans, Louisiana, USA: ACL, 2018: 1470–1480.

[126] YASUNAGA M, KASAI J, RADEV D R. Robust multilingual part-of-speech tagging via adversarial training[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). New Orleans, Louisiana, USA: ACL, 2018: 976–986.

[127] TRAN V, NGUYEN L. Adversarial domain adaptation for variational neural language generation in dialogue systems[C]//Proceedings of the 27th International Conference on Computational Linguistics (COLING). Santa Fe, New Mexico, USA: ACL, 2018: 1205–1217.

[128] PAN S J, YANG Q. A survey on transfer learning[J]. IEEE Trans. Knowl. Data Eng., 2010, 22(10): 1345–1359.

[129] WANG M, MANNING C D. Cross-lingual projected expectation regularization for weakly supervised learning[J]. TACL, 2014, 2: 55–66.

[130] PRETTENHOFER P, STEIN B. Cross-language text classification using structural correspondence learning[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL). Uppsala, Sweden: ACL, 2010: 1118–1127.

[131] ZHOU X, WAN X, XIAO J. Cross-lingual sentiment classification with bilingual document representation learning[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL). Berlin, Germany: ACL, 2016: 1403–1412.

[132] EGER S, DAXENBERGER J, STAB C, et al. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need![C]//Proceedings of the 27th International Conference on Computational Linguistics (COLIMG). Santa Fe, New Mexico, USA: ACL, 2018: 831–844.

[133] OTANI N, KIYOMARU H, KAWAHARA D, et al. Cross-lingual knowledge projection using machine translation and target-side knowledge base completion[C]// Proceedings of the 27th International Conference on Computational Linguistics (COLIMG). Santa Fe, New Mexico, USA: ACL, 2018: 1508–1520.

[134] MCDONALD R T, PETROV S, HALL K B. Multi-source transfer of delexicalized dependency parsers[C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)). Edinburgh, UK: ACL, 2011: 62–72.

[135] YU Z, MARECEK D, ZABOKRTSKÝ Z, et al. If you even don't have a bit of bible: Learning delexicalized POS taggers[C]//Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC). Portorož, Slovenia: ELRA, 2016: 96–103.

[136] KUNDU G, SIL A, FLORIAN R, et al. Neural cross-lingual coreference resolution and its application to entity linking[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne, Australia: ACL, 2018: 395–400.

[137] XIE J, YANG Z, NEUBIG G, et al. Neural cross-lingual named entity recognition with minimal resources[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium: ACL, 2018: 369–379.

[138] XIONG F, GAO J. Entity alignment for cross-lingual knowledge graph with graph convolutional networks[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI). Macao, China: IJCAI Organization, 2019: 6480–6481.

[139] CAO P, CHEN Y, LIU K, et al. Adversarial transfer learning for chinese named entity recognition with self-attention mechanism[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium: ACL, 2018: 182–192.

[140] LI J, YE D, SHANG S. Adversarial transfer for named entity boundary detection with pointer networks[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI). Macao, China: IJCAI Organization, 2019: 5053–5059.

[141] WANG X, HAN X, LIN Y, et al. Adversarial multi-lingual neural relation extraction[C]//Proceedings of the 27th International Conference on Computational Linguistics (COLING). Santa Fe, New Mexico, USA: ACL, 2018: 1156–1166.

[142] ZOU B, XU Z, HONG Y, et al. Adversarial feature adaptation for cross-lingual relation classification[C]//Proceedings of the 27th International Conference on Computational Linguistics (COLING). Santa Fe, New Mexico, USA: ACL, 2018: 437–448.

[143] CHEN X, CARDIE C. Unsupervised multilingual word embeddings[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium: ACL, 2018: 261–270.

[144] WADA T, IWATA T, MATSUMOTO Y. Unsupervised multilingual word embedding with limited resources using neural language models[C]//Proceedings of the

57th Conference of the Association for Computational Linguistics (ACL). Florence, Italy: ACL, 2019: 3113–3124.

[145] NGUYEN K A, Schulte im Walde S, VU N T. Distinguishing antonyms and synonyms in a pattern-based neural network[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Valencia, Spain: ACL, 2017: 76–85.

[146] VYLOMOVA E, RIMELL L, COHN T, et al. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL). Berlin, Germany: ACL, 2016: 1671–1682.

[147] ATTIA M, MAHARJAN S, SAMIH Y, et al. Cogalex-v shared task: GHHH - detecting semantic relations via word embeddings[C]//Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex@COLING). Osaka, Japan: ACL, 2016: 86–91.

[148] BOURAOUI Z, JAMEEL S, SCHOCKAERT S. Relation induction in word embeddings revisited[C]//Proceedings of the 27th International Conference on Computational Linguistics (COLING). Santa Fe, New Mexico, USA: ACL, 2018: 1627–1637.

[149] NGUYEN K A, Schulte im Walde S, VU N T. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL). Berlin, Germany: ACL, 2016: 454–459.

[150] HASHIMOTO K, STENETORP P, MIWA M, et al. Task-oriented learning of word embeddings for semantic relation classification[C]//Proceedings of the 19th Conference on Computational Natural Language Learning (CoNLL). Beijing, China: ACL, 2015: 268–278.

[151] MRKSIC N, VULIC I, SÉAGHDHA D Ó, et al. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints[J]. TACL, 2017, 5: 309–324.

[152] CHEN H, LEE C, LIAO K, et al. Word relation autoencoder for unseen hypernym extraction using word embeddings[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium: ACL, 2018: 4834–4839.

[153] CAMACHO-COLLADOS J, ANKE L E, SCHOCKAERT S. Relational word embeddings[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL). Florence, Italy: ACL, 2019: 3286–3296.

[154] MASUMURA R, ASAMI T, MASATAKI H, et al. Hyperspherical query likelihood models with word embeddings[C]//Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP). Taipei, Taiwan: ACL, 2017: 210–216.

[155] LV X, HOU L, LI J, et al. Differentiating concepts and instances for knowledge graph embedding[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium: ACL, 2018: 1971–1979.

[156] PHAM N T, LAZARIDOU A, BARONI M. A multitask objective to inject lexical contrast into distributional semantics[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP). Beijing, China: ACL, 2015: 21–26.

[157] DENTON E L, GROSS S, FERGUS R. Semi-supervised learning with context-conditional generative adversarial networks[J/OL]. CoRR, 2016, abs/1611.06430. http://arxiv.org/abs/1611.06430.

[158] KINGMA D P, BA J. Adam: A method for stochastic optimization[C]//Proceedings of the 3rd International Conference on Learning Representations (ICLR). San Diego, CA, USA: ICLR, 2015.

[159] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017 (NIPS). Long Beach, CA, USA: NIPS, 2017: 5998–6008.

[160] ZAREMOODI P, BUNTINE W L, HAFFARI G. Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne, Australia: ACL, 2018: 656–661.

[161] GUO C, PLEISS G, SUN Y, et al. On calibration of modern neural networks[C]// Proceedings of the 34th International Conference on Machine Learning, (ICML). Sydney, NSW, Australia: ICML, 2017: 1321–1330.

[162] KOTLERMAN L, DAGAN I, SZPEKTOR I, et al. Directional distributional similarity for lexical inference[J]. Natural Language Engineering, 2010, 16(4): 359–389.

[163] BOND F, FOSTER R. Linking and extending an open multilingual wordnet[C]// Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL). Sofia, Bulgaria: ACL, 2013: 1352–1362.

[164] KIELA D, RIMELL L, VULIC I, et al. Exploiting image generality for lexical entailment detection[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP). Beijing, China: ACL, 2015: 119–124.

[165] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). New York, NY, USA: ACM, 2014: 701–710.

[166] GROVER A, LESKOVEC J. node2vec: Scalable feature learning for networks[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). San Francisco, CA, USA: ACM, 2016: 855–864.

[167] SANTUS E, LENCI A, CHIU T, et al. Nine features in a random forest to learn taxonomical semantic relations[C]//Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC). Portorož, Slovenia: ELRA, 2016: 4557–4564.

[168] SANTUS E, YUNG F, LENCI A, et al. Evalution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models[C]//Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications (LDL@IJCNLP). Beijing, China: ACL, 2015: 64–69.

[169] SANTUS E, GLADKOVA A, EVERT S, et al. The cogalex-v shared task on the corpus-based identification of semantic relations[C]//Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex@COLING). Osaka, Japan: ACL, 2016: 69–79.

[170] MAATEN L V D, HINTON G. Visualizing data using t-sne[J]. Journal of machine learning research, 2008, 9(Nov): 2579–2605.

[171] ZHOU P, SHI W, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL). Berlin, Germany: ACL, 2016: 207–212.

[172] WU S, HE Y. Enriching pre-trained language model with entity information for relation classification[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM). Beijing, China: ACM, 2019: 2361–2364.

[173] LUO B, FENG Y, WANG Z, et al. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). Vancouver, Canada: ACL, 2017: 430–439.

[174] WU S, FAN K, ZHANG Q. Improving distantly supervised relation extraction with neural noise converter and conditional optimal selector[C]//Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI). Honolulu, Hawaii, USA: AAAI, 2019: 7273–7280.

[175] BANKO M, CAFARELLA M J, SODERLAND S, et al. Open information extraction from the web[C]//Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI). Hyderabad, India: IJCAI Organization, 2007: 2670–2676.

[176] CUI L, WEI F, ZHOU M. Neural open information extraction[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne, Australia: ACL, 2018: 407–413.

[177] PAL H, MAUSAM. Demonyms and compound relational nouns in nominal open IE [C]//Proceedings of the 5th Workshop on Automated Knowledge Base Construction (AKBC@NAACL-HLT). San Diego, CA, USA: ACL, 2016: 35–39.

[178] TANDON N, DE MELO G, SUCHANEK F M, et al. Webchild: harvesting and organizing commonsense knowledge from the web[C]//Proceedings of the Seventh ACM International Conference on Web Search and Data Mining (WSDM). New York, NY, USA: ACM, 2014: 523–532.

[179] SPEER R, CHIN J, HAVASI C. Conceptnet 5.5: An open multilingual graph of general knowledge[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI). San Francisco, California, USA: AAAI, 2017: 4444–4451.

[180] SHWARTZ V, DAGAN I. Paraphrase to explicate: Revealing implicit noun-compound relations[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne, Australia: ACL, 2018: 1200–1211.

[181] PASCA M. German typographers vs. german grammar: Decomposition of wikipedia category labels into attribute-value pairs[C]//Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM). Cambridge, United Kingdom: ACM, 2017: 315–324.

[182] NAKOV P. On the interpretation of noun compounds: Syntax, semantics, and entailment[J]. Natural Language Engineering, 2013, 19(3): 291–330.

[183] SÉAGHDHA D Ó, COPESTAKE A A. Interpreting compound nouns with kernel methods[J]. Natural Language Engineering, 2013, 19(3): 331–356.

[184] SHWARTZ V, WATERSON C. Olive oil is made of olives, baby oil is made for babies: Interpreting noun compounds using paraphrases in a neural model[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). New Orleans, Louisiana, USA: ACL, 2018: 218–224.

[185] FARES M, OEPEN S, VELLDAL E. Transfer and multi-task learning for noun-noun compound interpretation[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium: ACL, 2018: 1488–1498.

[186] DE CRUYS T V, AFANTENOS S D, MULLER P. MELODI: A supervised distributional approach for free paraphrasing of noun compounds[C]//Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT). Atlanta, Georgia, USA: ACL, 2013: 144–147.

[187] GRYCNER A, WEIKUM G. POLY: mining relational paraphrases from multilingual sentences[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP). Austin, Texas, USA: ACL, 2016: 2183–2192.

[188] HENDRICKX I, KOZAREVA Z, NAKOV P, et al. Semeval-2013 task 4: Free paraphrases of noun compounds[C]//Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT). Atlanta, Georgia, USA: ACL, 2013: 138–143.

[189] XAVIER C C, DE LIMA V L S. Boosting open information extraction with noun-based relations[C]//Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC). Reykjavik, Iceland: ELRA, 2014: 96–100.

[190] JIA S, E S, LI M, et al. Chinese open relation extraction and knowledge base establishment[J]. ACM Trans. Asian & Low-Resource Lang. Inf. Process., 2018, 17(3): 15:1–15:22.

[191] WEI X, YUAN Y. To construct the interpretation templates for the chinese noun compounds based on semantic classes and qualia structures[C]//Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation (PACLIC). Bali, Indonesia: ACL, 2012: 609–619.

[192] LENAT D B. CYC: A large-scale investment in knowledge infrastructure[J]. Commun. ACM, 1995, 38(11): 32–38.

[193] NARISAWA K, WATANABE Y, MIZUNO J, et al. Is a 204 cm man tall or small ? acquisition of numerical common sense from the web[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL). Sofia, Bulgaria: ACL, 2013: 382–391.

[194] COLLELL G, GOOL L V, MOENS M. Acquiring common sense spatial knowledge through implicit spatial templates[C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI). New Orleans, Louisiana, USA: AAAI, 2018: 6765–6772.

[195] XU F F, LIN B Y, ZHU K Q. Automatic extraction of commonsense located-near knowledge[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne, Australia: ACL, 2018: 96–101.

[196] BOSSELUT A, RASHKIN H, SAP M, et al. COMET: commonsense transformers for automatic knowledge graph construction[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL). Florence, Italy: ACL, 2019: 4762–4779.

[197] HASHIMOTO C, KAWAHARA D. Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features [C]//Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP). Honolulu, Hawaii, USA: ACL, 2008: 992–1001.

[198] PENG J, FELDMAN A, VYLOMOVA E. Classifying idiomatic and literal expressions using topic models and intensity of emotions[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: ACL, 2014: 2019–2027.

[199] SALTON G, ROSS R J, KELLEHER J D. Idiom token classification using sentential distributed semantics[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL). Berlin, Germany: ACL, 2016: 194–204.

[200] GHARBIEH W, BHAVSAR V, COOK P. A word embedding approach to identifying verb-noun idiomatic combinations[C]//Proceedings of the 12th Workshop on Multiword Expressions (MWE@ACL). Berlin, Germany: ACL, 2016: 112–118.

[201] KING M, COOK P. Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of english verb-noun combinations[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne, Australia: ACL, 2018: 345–350.

[202] LIU C, HWA R. Heuristically informed unsupervised idiom usage recognition[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels (EMNLP). Brussels, Belgium: ACL, 2018: 1723–1731.

[203] REDDY S, MCCARTHY D, MANANDHAR S. An empirical study on compositionality in compound nouns[C]//Proceedings of the Fifth International Joint Conference on Natural Language Processing (IJCNLP). Chiang Mai, Thailand: ACL, 2011: 210–218.

[204] KIELA D, CLARK S. Detecting compositionality of multi-word expressions using nearest neighbours in vector space models[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP). Seattle, Washington, USA: ACL, 2013: 1427–1432.

[205] SALEHI B, COOK P, BALDWIN T. A word embedding approach to predicting the compositionality of multiword expressions[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Denver, Colorado, USA: ACL, 2015: 977–983.

[206] YAZDANI M, FARAHMAND M, HENDERSON J. Learning semantic composition to detect non-compositionality of multiword expressions[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). Lisbon, Portugal: ACL, 2015: 1733–1742.

[207] CORDEIRO S, RAMISCH C, IDIART M, et al. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL). Berlin, Germany: ACL, 2016: 1986–1997.

[208] QI F, HUANG J, YANG C, et al. Modeling semantic compositionality with sememe knowledge[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL). Florence, Italy: ACL, 2019: 5706–5715.

[209] QIU X, ZHANG Q, HUANG X. Fudannlp: A toolkit for chinese natural language processing[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), System Demonstrations. Sofia, Bulgaria: ACL, 2013: 49–54.

[210] ALIDAEE B, GLOVER F W, KOCHENBERGER G A, et al. Solving the maximum edge weight clique problem via unconstrained quadratic programming[J]. European Journal of Operational Research, 2007, 181(2): 592–597.

[211] FANG Z, WANG H, GRACIA J, et al. Zhishi.lemon: On publishing zhishi.me as linguistic linked open data[C]//Proceedings of the 15th International Semantic Web Conference (ISWC). Kobe, Japan: Springer, 2016: 47–55.

[212] PASCA M. Interpreting compound noun phrases using web search queries[C]// Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Denver, Colorado, USA: ACL, 2015: 335–344.

[213] ZHANG D, YUAN J, WANG X, et al. Probabilistic verb selection for data-to-text generation[J]. TACL, 2018, 6: 511–527.

[214] ZHAI C, LAFFERTY J D. A study of smoothing methods for language models applied to ad hoc information retrieval[J]. SIGIR Forum, 2017, 51(2): 268–276.

[215] GASHTEOVSKI K, GEMULLA R, CORRO L D. Minie: Minimizing facts in open information extraction[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). Copenhagen, Denmark: ACL, 2017: 2630–2640.

[216] NIKLAUS C, CETTO M, FREITAS A, et al. A survey on open information extraction[C]//Proceedings of the 27th International Conference on Computational Linguistics (COLING). Santa Fe, New Mexico, USA: ACL, 2018: 3866–3878.

[217] SPORLEDER C, LI L, GORINSKI P, et al. Idioms in context: The IDIX corpus[C]//Proceedings of the International Conference on Language Resources and Evaluation (LREC). Valletta, Malta: ELRA, 2010: 639–646.

[218] SHAO Y, SENNRICH R, WEBBER B L, et al. Evaluating machine translation performance on chinese idioms with a blacklist method[C]//Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC). Miyazaki, Japan: ELRA, 2018: 31–38.

[219] Schulte im Walde S, HÄTTY A, BOTT S. The role of modifier and head properties in predicting the compositionality of english and german noun-noun compounds: A vector-space perspective[C]//Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM@ACL). Berlin, Germany: ACL, 2016: 148–158.

[220] LU X, WANG B P. Towards a metaphor-annotated corpus of mandarin chinese[J]. Language Resources and Evaluation, 2017, 51(3): 663–694.

[221] WANG L, WANG M. A study on the taxonomy of chinese noun compounds[C]// Proceedings of the 16th Workshop on Chinese Lexical Semantics (CLSW). Beijing, China: Springer, 2015: 262–269.

[222] BETTERIDGE J, RITTER A, MITCHELL T M. Assuming facts are expressed more than once[C]//Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference (FLAIRS). Pensacola Beach, Florida, USA: AAAI, 2014: 431–436.

[223] KUZI S, SHTOK A, KURLAND O. Query expansion using word embeddings [C]//Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM). Indianapolis, IN, USA: ACM, 2016: 1929–1932.

[224] WANG S, HUANG C, YAO Y, et al. Building a semantic transparency dataset of chinese nominal compounds: A practice of crowdsourcing methodology[C]// Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP@COLING). Dublin, Ireland: ACL, 2014: 147–156.

[225] ZHANG W, PAUDEL B, WANG L, et al. Iteratively learning embeddings and rules for knowledge graph reasoning[C]//Proceedings of the 2019 World Wide Web Conference (WWW). San Francisco, CA, USA: ACM, 2019: 2366–2377.

[226] XIONG W, DU B, ZHANG L, et al. Regularizing deep convolutional neural networks with a structured decorrelation constraint[C]//Proceedings of the IEEE 16th International Conference on Data Mining (ICDM). Barcelona, Spain: IEEE, 2016: 519–528.

[227] OKAJIMA Y, SADAMASA K. Deep neural networks constrained by decision rules [C]//Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI). Honolulu, Hawaii, USA,: AAAI, 2019: 2496–2505.

[228] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). New Orleans, Louisiana, USA: ACL, 2018: 2227–2237.

[229] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Minneapolis, MN, USA: ACL, 2019: 4171–4186.

[230] DAI Z, YANG Z, YANG Y, et al. Transformer-xl: Attentive language models beyond a fixed-length context[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL). Florence, Italy: ACL, 2019: 2978–2988.

[231] YANG Z, DAI Z, YANG Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[C]//Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS). Vancouver, BC, Canada: NeurIPS, 2019: 5754–5764.

附录

为便于读者查阅，附录汇总了与本文主题密切相关的数据集和源代码。值得注意的是，我们只列出作者及课题组成员构造的相关资源，以供学术研究使用，其他资源请联系相关论文作者获取。

公开数据集

- **中文分类体系数据集**[1]：从中文维基百科的类别标签系统中构建的中文分类体系，包括约 58.2 万个中文实体、7.9 万个类别和 131.8 万个上下位关系，整体精确度约为 95%。在本文中，我们用该数据集构建 IPM 算法的训练集。

- **中文上下位关系标注数据集**[2]：从百度百科的词条类别进行采样，生成 “实体-类别” 对的集合，并且进行上下位关系或非上下位关系的二分类人工标注。这一数据集包括 3870 个上下位关系元组和 3582 个非上下位关系元组，用于评测 IPM、TPM、FOPM 等模型的准确性。

- **中文维基类别抽取关系数据集**[3]：从中文维基类别网络中抽取的关系数据集，同时包含上下位关系和非上下位关系，用于中文知识图谱的补全。

- **中文短语习语性标注数据集**[4]：这一数据集包括两个部分。第一个部分为 CNCBaike，包括 1330 个具有人工标注习语性程度的中文复合名词，这些中文复合名词从百度百科中抽取的概念类别中抽取出来。第二部分为 CNCWeb，包括 815 个标注的中文复合名词，这些名词从中文语料库中采用基于 POS 的启发式规则抽取。CNCBaike 和 CNCWeb 用于评测中文复合名词习语性程度分类算法的有效性。

- **小语种上下位关系数据集**[5]：从 Open Multilingual Wordnet 项目中生成七种

[1] https://chywang.github.io/data/apweb2015.zip
[2] https://chywang.github.io/data/acl2017.zip
[3] https://chywang.github.io/data/tkde.zip
[4] https://chywang.github.io/data/access.zip
[5] https://chywang.github.io/data/www2019.zip

非英语语种的上下位关系分类训练集合测试集。这七种语言包括法语、中文、日语、意大利语、泰语、芬兰语和希腊语。上下位关系元组从对应的 Wordnet 中的概念层次类别中随机采样得到, 非上下位关系元组合并了对应的 Wordnet 中的其他多种关系（包括整体-部分关系、同义词关系等）。这一数据集用于评测跨语言上下位关系预测算法的准确性。

开源软件

- **CN-IterativeIsALearner** [6]：自动从中文实体-类别对集合中抽取上下位关系, 包含 IPM 算法的实现。
- **CN-TransductIsALearner** [7]：自动进行中文上下位关系和非上下位关系的分类，包含 TPM 算法的实现。
- **TEAL** [8]：分类体系增强的对抗学习框架（TEAL）的算法实现，用于上下位关系分类的模型效果增强。
- **FOP** [9]：包括基于模糊正交投影模型的三种算法实现（即 FOPM、TFOPM 和 ITFOPM），用于单语言和跨语言的上下位关系分类。
- **SphereRE** [10]：超球关系嵌入（SphereRE）的算法实现，用于词汇关系分类。
- **DRCLib** [11]：自动集成本文提出的多种投影模型，支持这些模型的组合和自动调用，用于词汇关系分类。
- **CN-WikiCatReader** [12]：从中文维基百科的类别系统中抽取上下位关系和多种非上下位关系，结合了 PNRE 和 DNRE 等算法。

[6] https://github.com/chywang/CN-IterativeIsALearner
[7] https://github.com/chywang/CN-TransductIsALearner
[8] https://github.com/chywang/TEAL
[9] https://github.com/chywang/FOP
[10] https://github.com/chywang/SphereRE
[11] https://github.com/chywang/DRCLib
[12] https://github.com/chywang/CN-WikiCatReader

致谢

在华东师范大学的五年博士学习生涯即将结束。这五年是我毕生难忘的五年，无论是在生活上还是科研上，我都经历了无数的喜怒哀乐。我学到了很多东西，我相信这些东西能使我受益终身。值此论文完成之际，我的心中感慨良多，体会到辛勤劳动后的喜悦和激动。同时，这篇论文的完成也与老师、同学、家人、朋友们的帮助和支持是分不开的，我想借此机会向他们由衷地致以感谢。

首先，我想深深地感谢我的指导老师何晓丰老师。从 2013 年本科学习起，他就对我的学业和科研工作谆谆教诲、耳提面命，把我这一个对于计算机科研领域一无所知的初学者，带入了数据挖掘和机器学习等领域，进入了科研的大门。他渊博的知识、勤恳的作风和一丝不苟的态度是我学习的楷模。他在科研中给我充分发挥的空间，使我收获了知识，提高了能力。感谢他一直鼓励我进行自然语言处理方面的研究，让我在自己感兴趣的领域里自由探索，发挥自己的优势；感谢他多次资助我远赴海外，参加各大国际学术会议，提供充足的学习交流机会，同时饱览亚欧美各大洲的壮美风景。感谢周傲英老师，周老师开阔的学术视野、敏锐的思维和严谨的学风深深地使我感动。他对我的鼓励、信任和支持是我不断前进的动力。感谢信息学部的其他老师给我的帮助和支持，包括钱卫宁老师、查宏远老师、王长波老师、王晓玲老师、宫学庆老师、张蓉老师、张伟老师、张召老师等。感谢学院为我提供优越的学习和科研氛围，能让我在良好的环境下尽自己所能。感谢学校研究生院的各位老师提供的帮助和科研资助。感谢阿里巴巴集团和蚂蚁金服集团的临在、岑鸣、岑尘等师兄师姐对我在迁移学习和深度语言模型方面的指导，让我进一步拓宽了眼界。

从我进入数学馆西 105 实验室的第一天以来，我就把实验室的格言（“以云水趣看成败，以木石心图将来”）当成我自己的座右铭，用以鼓励我在这里生活和战斗，在实验结果不理想的情况下仍然坚持研究下去。从某种意义上来说，数学馆西 105 就是我博士生活的家。我想感谢日日夜夜奋战在实验室学长学姐们、同学们、战友们和朋友们。他们在平时学习生活中给了我帮助和照顾。我们一起讨论问题，一起看论文，一起完成项目，他们的陪伴和帮助使我收获良多。特别感谢宋乐怡学

姐在我科研起步阶段对我细致入微的指导。感谢曾经和我一起奋战在 105 实验室的学长学姐和同学们：程文亮、苏永浩、张驰、肖冰、李金洋、王燕华、张伟佳、潘松松、雍若兰等。还有给实验室生活带来莫大欢乐的学弟学妹们：王露、徐国海、王坤、丁雨琦、孙阳、吴盈娇、张涛林、张君瑞、蔡烩、张艳丽、郑巧、胡楠、农伟、杨双吉、蔡泽锐等。

最后，我想把我最真挚的感谢献给我最好的朋友、最贴心的伙伴、生活上的导师、工作上的战友熊然同学。感谢她对我的认同，感谢她对我无条件的信任，感谢她为我做的一切。感谢辛苦养育和照顾我的父母和家人，他们无私的付出使我闯过生命中的一道道难关，在各方面不断成长。谢谢你们！

汪诚愚

二零二零年五月

简历

■ 基本信息

汪诚愚，男，1991 年 12 月出生于江苏省苏州市，华东师范大学软件工程学院在读博士生。

■ 教育经历

2011 年 9 月至 2015 年 7 月：华东师范大学软件学院，软件工程专业，获得工学学士学位；

2015 年 9 月至今：华东师范大学软件工程学院，软件工程专业（研究方向：数据科学与工程），本科直博。

■ 研究兴趣

知识抽取、知识图谱、计算语言学、自然语言理解

攻读博士学位期间发表的学术论文和科研情况

■ 发表的期刊论文

[1] **Chengyu Wang**, Xiaofeng He, Aoying Zhou: Open Relation Extraction for Chinese Noun Phrases[J]. IEEE Transactions on Knowledge and Data Engineering (TKDE) (In Press).

[2] **Chengyu Wang**, Xiaofeng He, Aoying Zhou: HEEL: Exploratory Entity Linking for Heterogeneous Information Networks[J]. Knowledge and Information Systems (KAIS), 2020, 62(2): 485–506.

[3] **Chengyu Wang**, Yan Fan, Xiaofeng He, Hongyuan Zha, Aoying Zhou: Idiomaticity Prediction of Chinese Noun Compounds and Its Applications[J]. IEEE Access, 2019,7: 142866-142878.

[4] **Chengyu Wang**, Yan Fan, Xiaofeng He, Aoying Zhou: Decoding Chinese User Generated Categories for Fine-grained Knowledge Harvesting[J]. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2019, 31(8): 1491–1505.

[5] **Chengyu Wang**, Yan Fan, Xiaofeng He, Aoying Zhou: Predicting Hypernym-Hyponym Relations for Chinese Taxonomy Learning[J]. Knowledge and Information Systems (KAIS), 2019, 58(3): 585–610.

[6] **Chengyu Wang**, Guomin Zhou, Xiaofeng He, Aoying Zhou: NERank+: A Graph-based Approach for Entity Ranking in Document Collections[J]. Frontiers of Computer Science (FCS), 2018, 12(3): 504–517.

[7] Jihong Yan, **Chengyu Wang**, Wenliang Cheng, Ming Gao, Aoying Zhou: A Retrospective of Knowledge Graphs[J]. Frontiers of Computer Science (FCS), 2018, 12(1): 55–74.

[8] **汪诚愚**, 何晓丰, 宫学庆, 周傲英. 面向上下位关系预测的词向量投影模型 [J]. 计算机学报, 2020, 43(5): 868-883.

[9] 李金洋, 王燕华, 樊艳, **汪诚愚**, 张蓉, 何晓丰. 中文分类体系的构建与查询系统 [J]. 计算机应用, 2016, 36(S1): 207-209, 227.

■ 发表的会议论文

[1] **Chengyu Wang**, Xiaofeng He: BiRRE: Learning Bidirectional Residual Relation Embeddings for Supervised Hypernymy Detection[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). Seattle, Washington, USA: ACL, 2020 (Accepted).

[2] **Chengyu Wang**, Xiaofeng He, Aoying Zhou: SphereRE: Distinguishing Lexical Relations with Hyperspherical Relation Embeddings[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy: ACL, 2019: 1727–1737.

[3] **Chengyu Wang**, Yan Fan, Xiaofeng He, Aoying Zhou: A Family of Fuzzy Orthogonal Projection Models for Monolingual and Cross-lingual Hypernymy Prediction[C]// Proceedings of the 2019 World Wide Web Conference (WWW). San Francisco, California, USA: ACM, 2019: 1965–1976.

[4] **Chengyu Wang**, Xiaofeng He, Aoying Zhou: Improving Hypernymy Prediction via Taxonomy Enhanced Adversarial Learning[C]// Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). Honolulu, Hawaii, USA: AAAI, 2019: 7128–7135.

[5] **Chengyu Wang**, Yan Fan, Xiaofeng He, Aoying Zhou: Learning Fine-grained Relations from Chinese User Generated Categories[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). Copenhagen, Denmark: ACL, 2017: 2577–2587.

[6] **Chengyu Wang**, Xiaofeng He, Aoying Zhou: A Short Survey on Taxonomy Learning from Text Corpora: Issues, Resources and Recent Advances[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). Copenhagen, Denmark: ACL, 2017: 1190–1203.

[7] **Chengyu Wang**, Junchi Yan, Aoying Zhou, Xiaofeng He: Transductive Non-linear Learning for Chinese Hypernym Prediction[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). Vancouver, British Columbia, Canada: ACL, 2017: 1394–1404.

[8] **Chengyu Wang**, Xiaofeng He: Chinese Hypernym-Hyponym Extraction from User Generated Categories[C]// Proceedings of the 26th International Conference on Computational Linguistics (COLING). Osaka, Japan: ACL, 2016: 1350–1361.

[9] **Chengyu Wang**, Rong Zhang, Xiaofeng He, Aoying Zhou: Error Link Detection and Correction in Wikipedia[C]// Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM). Indianapolis, Indiana, USA: ACM, 2016: 307–316.

[10] **Chengyu Wang**, Rong Zhang, Xiaofeng He, Guomin Zhou, Aoying Zhou: Event Phase Extraction and Summarization[C]// Proceedings of the 17th International Conference on Web Information Systems Engineering (WISE). Shanghai, China: Springer, 2016: 473–488.

[11] **Chengyu Wang**, Rong Zhang, Xiaofeng He, Guomin Zhou, Aoying Zhou: NERank: Bringing Order to Named Entities from Texts[C]// Proceedings of the 18th Asia Pacific Web Conference (APWeb). Suzhou, China: Springer, 2016: 15–27.

[12] **Chengyu Wang**, Rong Zhang, Xiaofeng He, Aoying Zhou: NERank: Ranking Named Entities in Document Collections[C]// Proceedings of the 25th World Wide Web Conference (WWW). Montreal, Quebec, Canada: ACM, 2016: 123–124.

[13] Hui Cai, **Chengyu Wang**, Xiaofeng He: Debiasing Learning to Rank Models with Generative Adversarial Networks[C]// Proceedings of the 4th APWeb-WAIM Joint Conference on Web and Big Data (APWeb-WAIM). Tianjin, China: Springer, 2020 (Accepted).

[14] Yan Fan, **Chengyu Wang**, Boxing Chen, Zhongkai Hu, Xiaofeng He: SPMM: A Soft Piecewise Mapping Model for Bilingual Lexicon Induction[C]// Proceedings of the 19th SIAM International Conference on Data Mining (SDM). Calgary, Alberta, Canada: SIAM, 2019: 244-252.

[15] Yan Fan, **Chengyu Wang**, Xiaofeng He: Exploratory Neural Relation Classification for Domain Knowledge Acquisition[C]// Proceedings of the 27th International Conference on Computational Linguistics (COLING). Santa Fe, New Mexico, USA: ACL, 2018: 265–2276.

[16] Guohai Xu, **Chengyu Wang**, Xiaofeng He: Improving Clinical Named Entity Recognition with Global Neural Attention[C]// Proceedings of the 2nd APWeb-WAIM Joint Conference on Web and Big Data (APWeb-WAIM). Macao, China: Springer, 2018: 264–279.

[17] Lu Wang, **Chengyu Wang**, Keqiang Wang, Xiaofeng He: BiUCB: A Contextual Bandit Algorithm for Cold-Start and Diversified Recommendation[C]// Proceedings of the 8th IEEE International Conference on Big Knowledge (ICBK). Hefei, China: IEEE, 2017: 248–253.

[18] Ruolan Yong, **Chengyu Wang**, Xiaofeng He: A Transfer Learning based Boosting Model for Emotion Analysis[C]// Proceedings of the 8th IEEE International Conference on Big Knowledge (ICBK). Hefei, China: IEEE, 2017: 264–269.

[19] Yan Fan, **Chengyu Wang**, Guomin Zhou, Xiaofeng He: DKGBuilder: An Architecture for Building a Domain Knowledge Graph from Scratch[C]// Proceedings of the 22nd International Conference on Database Systems for Advanced Applications (DASFAA). Suzhou, China: Springer, 2017: 663–667.

[20] Jinyang Li, **Chengyu Wang**, Xiaofeng He, Rong Zhang, Ming Gao: User Generated Content Oriented Chinese Taxonomy Construction[C]// Proceedings of the 17th Asia Pacific Web Conference (APWeb). Guangzhou, China: Springer, 2015: 623–634.

[21] Wenliang Cheng, **Chengyu Wang**, Bing Xiao, Weining Qian, Aoying Zhou: On Statistical Characteristics of Real-life Knowledge Graphs[C]// Proceedings of the 6th International Workshop on Big data benchmarks, Performance, Optimization and Emerging hardware (BPOE). Kohala, Hawaii, USA: Springer, 2015: 37–49.

[22] Kun Wang, Guohai Xu, **Chengyu Wang**, Xiaofeng He: A Hybrid Abnormal Advertising Traffic Detection Method[C]// Proceedings of the 8th IEEE International Conference on Big Knowledge (ICBK). Hefei, China: IEEE, 2017: 236–241.

[23] Chi Zhang, Yanhua Wang, **Chengyu Wang**, Wenliang Cheng, Xiaofeng He: SNExtractor: A Prototype for Extracting Semantic Networks from Web Documents[C]// Proceedings of the 17th International Conference on Web-Age Information Management (WAIM). Nanchang, China: Springer, 2016: 527–530.

[24] Yonghao Su, Chi Zhang, Jinyang Li, **Chengyu Wang**, Weining Qian, Aoying Zhou: Cross-Lingual Entity Query from Large-Scale Knowledge Graphs[C]// Proceedings of the 2nd International Workshop on Web Data Mining and Applications (WDMA). Guangzhou, China: Springer, 2015: 139–150.

■ 已投稿论文

[1] **Chengyu Wang**, Minghui Qiu, Jun Huang, Xiaofeng He: KEML: A Knowledge-Enriched Meta-Learning Framework for Lexical Relation Classification[C]// submitted to the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[2] **Chengyu Wang**, Minghui Qiu, Jun Huang, Xiaofeng He: Meta Fine-Tuning Neural Language Models for Multi-Domain Text Mining[C]// submitted to the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[3] Taolin Zhang, **Chengyu Wang**, Xiaofeng He: Knowledge-Empowered Representation Learning for Chinese Medical Reading Comprehension: Task, Model and Resources[C]// submitted to the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[4] Cen Chen, **Chengyu Wang**, Minghui Qiu, Dehong Gao, Linbo Jin, Li Wang, Jun Zhou: Domain-aware Transfer via Multi-teacher Knowledge Distillation for Retrieval-based Question Answering Systems[J]. submitted to IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP).

■ 攻读博士学位期间参加的科研项目及获得资助项目

[1] 国家重点研发计划 “大数据知识工程基础理论及其应用研究” 华东师范大学课题 “基于情景感知的知识导航”（2016YFB1000904）. 2016-2020.

[2] 华东师范大学 “优秀博士学位论文培育资助项目”（YB2016040）. 2016.12-2018.12.

[3] 华东师范大学教育学部专项经费 “教育网络数据平台”（14000-5154A5-15001/002）. 2015.