博士学位论文

基于分布表示的跨语言跨任务自然语言分析

DISTRIBUTED REPRESENTATIONS FOR CROSS-LINGUAL CROSS-TASK NATURAL LANGUAGE ANALYSIS

郭江

哈尔滨工业大学

2017 年 03 月

国内图书分类号：TP391.2　　学校代码：10213
国际图书分类号：681.324　　密级：公开

工学博士学位论文

基于分布表示的跨语言跨任务自然语言分析

博士研究生：郭江
导　　师：王海峰教授
副 导 师：车万翔副教授
申请学位：工学博士
学　　科：计算机应用技术
所在单位：计算机科学与技术
答辩日期：2017 年 03 月
授予学位单位：哈尔滨工业大学

Classified Index: TP391.2
U.D.C: 681.324

Dissertation for the Doctoral Degree in Engineering

DISTRIBUTED REPRESENTATIONS FOR CROSS-LINGUAL CROSS-TASK NATURAL LANGUAGE ANALYSIS

Candidate: Guo Jiang
Supervisor: Prof. WANG Haifeng
Associate Supervisor: Associate Prof. Che Wanxiang
Academic Degree Applied for: Doctor of Engineering
Specialty: Computer Applied Technology
Affiliation: School of Computer Science and Technology
Date of Defence: March, 2017
Degree-Conferring-Institution: Harbin Institute of Technology

摘 要

特征表示是统计机器学习的基础工作，也是影响机器学习系统性能的关键因素之一。在基于统计的自然语言处理研究中，最常见的特征表示是离散形式的符号表示，比如对于词的独热表示（One-Hot）以及对于文档的词袋表示（Bag-of-Words）等。这种表示方式直观简洁，易于计算，结合特征工程以及传统机器学习算法（如最大熵、支持向量机、条件随机场等），可以有效地应用于大部分自然语言处理的主流任务。另一种重要的特征表示机制称为分布表示，通常为连续、稠密、低维的向量表示，比如早期的潜在语义分析（Latent Semantic Analysis）以及近年来应用甚广的“特征嵌入”（Feature Embedding）方法等。

近年来，特征的分布表示被广泛应用在基于深度学习的自然语言处理模型中。与符号表示相比，分布表示可以更自然地与学习能力较强的深度神经网络模型相结合，并通过逐层抽象的表示学习来获得更适用于具体任务的高层语义表示。这也是填补自然语言处理语义鸿沟的一种有效手段。更重要的，分布表示提供了一种通用的语义表示空间，为不同任务、不同语言、不同模态数据之间的信息交互构建了一座桥梁。这种语义表示上的通用性使得多源训练信息能够相互融合，进而起到知识迁移的作用。比如，从无标注的生文本中训练神经网络语言模型而得到的词汇分布表示，被证明能够有效地提升大多数自然语言处理主流任务的性能。

本文正是利用分布表示的这些特点，尤其针对其在语义表示上的通用性，研究了分布表示在跨语言、跨数据类型以及跨任务知识迁移中的关键技术。主要包含以下几个方面：

1. **基于双语数据的词义分布表示学习**。针对前人提出的词汇分布表示无法刻画一词多义现象的问题，本文提出利用双语数据中所蕴含的词义对齐信息来学习词义级的分布表示。一方面能够更完整地刻画词义信息，另一方面可以结合循环神经网络对单语数据进行词义消歧，进而服务于上层应用。

2. **基于分布表示的跨语言依存句法分析**。对于世界上绝大多数自然语言，句法标注资源难以获取，且人工标注代价较高。因此，本文提出多语言分布表示学习的方法，将不同语言的词语表示在一个相同的向量空间之内，构成了句法知识在不同语言之间进行迁移的一座桥梁。进而利用资源丰富语言（如英语）的句法资源，来对资源稀缺语言进行依存句法分析。

3. **基于深度多任务学习的多类型树库迁移学习**。对于句法分析而言，现有的依存树库多种多样，或来自不同语言、或采用不同的标注规范。本文提出基于多层次分布表示共享的深度多任务学习结构，能够有效地从不同类型的源句法树库（不同语言、不同标规范）中进行知识萃取，从而提升句法模型在目标树库上的分析精度。

4. **面向语义角色标注与关系分类的统一框架**。不同任务之间往往存在一定的共性，比如语义角色标注与（实体）关系分类，它们都涉及对句子中的语义关系进行分析。本文提出一个统一的深度神经网络模型，将语义角色标注与（实体）关系分类任务进行融合，并采用深度多任务学习来提升目标任务上的性能。

总的来说，本论文利用分布表示在语义表示上的通用性，深入地研究了其在跨语言、跨任务与跨数据类型学习上的应用，在词汇、句法、语义层面上显著地提升了不同任务的性能。我们期待这些研究成果可以进一步延展至更多类型的数据以及任务，甚至应用于跨领域分析，以进一步推动自然语言处理领域的发展。

关键词： 自然语言处理；多语言；多任务；分布表示；迁移学习；神经网络

Abstract

Feature representation is the foundation work of statistical machine learning, and also one of the key factors that affect the performance of a machine learning system. In the field of statistical natural language processing (NLP), the most commonly-used feature representation is the discrete symbolic representation, such as the One-Hot representation of words and the Bag-of-Words representation of documents. This kind of feature representation is intuitive, concise and easy to calculate. It has been successfully employed by most of the mainstream NLP tasks, combined with feature engineering and conventional machine learning algorithms, e.g., Maximum Entropy Model, Support Vector Machine and Conditional Random Fields. Another important feature representation mechanism is known as distributed representation, mostly appearing as continuous, dense and low-dimensional vector representations of features. Typical distributed representation learning approaches include the early-stage latent semantic analysis and the "feature embedding" approach which has gained a lot of interests recently.

In recent years, distributed feature representations have been extensively used in deep learning models for NLP. Compared with symbolic representations, distributed representations can be much more naturally combined with deep neural network models for learning high-level task-specific semantic representations through layer-wise representation abstraction. This has been an effective way of bridging the semantic gap in NLP. More importantly, distributed representation provides a universal semantic representation space across different tasks, languages and data modalities, so that training signals from multiple sources can be incorporated effectively, which further promotes knowledge transfer in practical learning tasks. For example, the distributed word representations learned from the neural network language modeling task on plain texts have been proven highly beneficial for a variety of NLP mainstream tasks.

Inspired by these characteristics of distributed representation, especially its universality in semantic representation, this paper investigates the key technologies in distributed representation learning for knowedge transfer across languages, multi-typed data and tasks. To be more specific, our contributions include:

1. Learning sense-specific word embeddings by exploiting bilingual parallel data. Single word embeddings have been shown poor in representing polysemy. To address this

problem, we propose a novel and effective approach for learning sense-specific word embeddings by exploiting the sense-alignment information contained in bilingual resources. The proposed embeddings are expected to capture the multiple senses of polysemous words, and also benefit downstream applications.

2. Learning multilingual distributed representations for cross-lingual transfer parsing. The majority of languages in the world are low-resource for dependency parsing, and it's labor-intensive to annotate treebanks for every language. We present cross-lingual word representation learning, to map words from different languages into a common vector space, and thus build a bridge connecting different languages. Therefore, the large-scale treebanks of rich-resource languages can be exploited to induce parsers for low-resource languages through transfer parsing.

3. A deep multi-task learning framework for transfer parsing across multi-typed treebanks. Various treebanks have been released for dependency parsing, either belonging to different languages or annotated with different schemes. We propose a deep multi-task learning architecture with representation-level parameter sharing, in order to distill knowledge from multi-typed treebanks and benefit parsing of the target treebank.

4. A unified model for semantic role labeling and relation classification. It is common for different NLP tasks to be related in certain ways. For example, semantic role labeling and (entity) relation classification both involve categorizing the semantic relation between words in a sentence. We propose a unified neural architecture that ties together the task of semantic role labeling and relation classification, and further apply deep multi-task learning to leverage their potential mutual benefits.

Overall, this paper systematically and deeply investigates the application of distributed representation learning on knowledge transfer across languages, tasks and multi-typed data. We will show its effectiveness through substantial empirical studies on lexical, syntactic and semantic tasks respectively. In the future, we expect to apply our research achievements to more diverse data and tasks, even to domain adaptation, and finally make a significant difference in the NLP field.

Keywords: Natural Language Processing, Multilingual Learning, Multitask Learning, Distributed Representations, Transfer Learning, Neural Networks

目 录

Contents

第 1 章 绪论

1.1 课题背景及意义

1.1.1 课题背景

基于统计机器学习的方法是目前自然语言处理研究的主流。其中，有监督学习（Supervised Learning）的方式利用人工标注的训练数据进行归纳式学习，在大多数自然语言处理任务中取得了令人满意的结果。同时，也应注意到，对于自然语言处理中一些较为复杂的任务，如句法、语义分析等，由于存在标注难度及代价较高、规范性差等问题，大规模的标注资源往往不易获取。这也为有监督学习带来了困难。

针对这个难题，人们沿着不同的研究路线进行了探索，如从未标注数据中进行无监督学习（Unsupervised Learning）[1]，融合有标注和未标注数据的半监督学习（Semi-supervised Learning）[2]，或者采用众包的方式[3]等。与此同时，也有不少学者通过融合不同类型的数据资源来提升目标任务（或数据）上的分析精度，如不同语言[4, 5]、不同领域[6–8]、甚至不同任务的资源[9, 10]。这种方式使得不同的知识能够在模型中进行迁移或者融合，相应的学习方法可以认为是一种迁移学习（Transfer Learning）[11][12]，这也是本文的研究重心。以下从跨语言、跨数据类型以及跨任务三个方面简要阐述相应的研究背景。

1. 跨语言知识迁移。随着互联网数据（如：社交媒体数据）的爆炸性增长，对不同语言数据的自动分析在很多场合下的重要性日渐凸显。然而，目前国际上活跃的语言超过7000种，对于大多数自然语言处理任务而言，不同语言的标注资源规模却呈现出极严重的长尾现象。以标注难度较高的句法分析为例，目前只有50多种语言存在一定规模的句法树库[1]。于是，对于资源稀缺语言进行自动分析逐渐成为一个重要的研究课题。

2. 跨数据类型知识迁移。由于自然语言所具有的复杂性及非规范性，很多具体任务的数据标注规范并不统一。例如，中文分词任务中不同数据集的标注粒度不尽相同；依存句法分析的标注体系也存在CoNLL[13]、Stanford[14]及Universal Dependencies[15]等不同规范。然而，尽管标注规范存在差

[1] http://universaldependencies.org

异，这些数据所蕴含的领域知识却有较强的共性，在应用中存在互补的可能。如何有效地融合这些数据，也是研究者一直以来颇为关注的问题。

3. 跨任务知识迁移。不同任务之间的迁移学习广泛存在于人类的活动中。康奈尔大学的Caruana在他的博士论文中写道：“在这个世界上，我们需要学习很多事物。它们遵循同样的物理定律，产生于同样的人类文化，……，也许是因为这些任务之间的相似性，才使得我们只需要少量的经验就能够学会这么多不同的事物。”[16]。在生活中，我们学习走路，学习奔跑，学习跳跃，学习读、写、听、说等。它们虽是不同的任务，却是相关的。而正是因为它们之间的相关性，使得我们无需对于每个任务都拥有充分的训练数据，就能够学会它们。在自然语言处理中也是如此，词法、句法、语义分析等任务之间同样存在紧密的内在联系。在一个任务上的良好表现对于其在相关任务上的泛化能力有很大的促进作用。充分利用不同任务之间的信息交互，也是实现一个通用人工智能系统的必要条件。

另外，近年来，深度学习在自然语言处理的诸多任务上取得了较大的突破，并逐渐成为研究与应用的一大热点。我们知道，在深度学习模型的应用过程中，随着网络变得越来越复杂，其表示能力也越来越强。由此带来的问题是如果训练数据规模不够大，则很容易使模型陷入过拟合的状态。因此，深度学习对于数据资源的需求也更为强烈。在这样的背景之下，本文的研究重点——跨语言、跨任务知识迁移，提供了一种具有重要价值的研究思路，且具有非常广泛的应用前景。

1.1.2 课题意义

在统计自然语言处理中，一般的思路是先针对目标对象进行特征抽取，从而将每一个实例表示成特征向量并以此作为统计模型的输入；再使用统计机器学习的方法对模型中的参数进行估计。以命名实体识别任务为例，为了判断文本中某个词是否为命名实体以及相应的实体类型，我们通常会抽取出它在一定宽度窗口中的上下文、词性等相关信息，并构成相应的特征向量；然后再结合相应的统计模型（如最大熵[17]、感知机[18]、条件随机场[19]等），在含标注的训练数据上学习相应特征的权重。

可见，对于特征的表示是统计自然语言处理研究中的基础工作。最常见的特征表示是离散形式的符号表示（symbolic representation），比如对于词的独热表示（One-Hot）以及对于文档的词袋表示（Bag-of-Words）等。这种表示方式

直观简洁，易于计算，结合**特征工程**（feature engineering）以及传统机器学习算法，可以有效地应用于大部分自然语言处理的主流任务。然而，离散的符号表示只是孤立地表示特征（如：词）本身，而没有刻画其蕴含的语义信息，因而也无法充分地表达不同符号数据之间的语义关联。为了弥补这项缺陷，以往的大部分工作采用特征工程的方式，根据由丰富的专家经验所设计的特征模板，来抽取大量与具体任务相关的其他语义特征，从而对其语义维度进行补充。特征工程将人类宝贵的知识、经验以及智慧融入了机器学习模型，一度引领着统计自然语言处理研究的发展。然而，对于人力的过度依赖，却也使其成为了制约机器智能进一步发展的瓶颈。一方面，专家经验不易获得，常常需要积年累月的实验探索与尝试。另一方面，由这种方式所获得的特征向量往往维度很高（百万级），从而导致“维度灾难”问题（Curse of Dimensionality）[20]，使得其对模型复杂度有所限制。因此，在以往的统计自然语言处理研究中，最常用的依然是复杂度较低的模型，如（广义）线性模型（Generalized Linear Models）。

另一种重要的特征表示方法被称为分布表示，通常为连续、稠密的低维度向量表示。比如早期应用于文档表示的潜在语义分析（Latent Semantic Analysis）[21]以及近年来应用甚广的“特征嵌入”方法（Feature Embedding）[22]等。尽管分布表示的思想提出较早[23, 24]，但其在大多数自然语言处理的主流任务中并没有得到广泛的应用。近年来，随着深度学习技术的发展，研究者们开始普遍认识到分布表示对于统计自然语言处理的重要性。我们知道，深度学习通常是指建立在含有多层非线性变换的神经网络结构之上，对数据的表示进行抽象和学习的一系列机器学习算法[25]，它最重要的性质在于“表示学习”（Representation Learning）。在自然语言处理的应用中，其“表示学习”作用主要表现在以下两个层面：

1. 对于基础特征的分布表示。前面提到，离散形式的符号表示无法充分地表达特征的潜在语义信息。例如，大多数自然语言处理的核心问题都以“词”作为最基础的研究对象，包括词性标注、句法分析等。因此，“词”特征也是最常用的基础特征类型之一。在基于符号表示的特征空间内，不同词的特征向量表示之间完全正交，难以有效地表达它们之间的相似性，从而也限制了统计模型的泛化能力。而在基于分布表示的特征空间中，每个词由多维度信息来联合表示，其每一维度反映了该词的一种潜在语义信息。利用分布表示，统计模型能够更好地利用特征之间的相似性，从而在一定程度上缓解数据稀疏的影响，获得更好的泛化能力。

2. 基于深度神经网络的特征学习。在大部分自然语言处理任务中，特征之

间需要相互组合或者交互才能够发挥作用。分布表示的引入虽然更好地表达了单特征的语义信息，但是对于特征之间的组合却不如传统的“特征工程”方式那么直接。幸运的是，含有多个隐含层的深度神经网络具有优异的特征抽象及学习能力。在基于分布表示的特征空间之内，使用循环、卷积、递归等神经网络模型能够有效地对特征进行组合，获得更抽象、更本质、更有利于具体任务（如分类、推理、生成）的特征表示，从而填补底层特征与高层语义之间的语义鸿沟。另一方面，从计算的角度来看，低维的分布表示特征空间也允许我们使用复杂度更高的机器学习模型对自然语言进行建模。

可以看出，分布表示在基于深度学习的自然语言处理技术中处于一个非常关键的位置。近年来，越来越多的自然语言处理学者投入到分布表示学习的研究中，其中既包括从语义表达角度出发的表示学习本身，即：如何学习更好的词、短语、句子甚至篇章分布表示；也不乏结合具体任务的表示学习研究，如基于句子分布表示以及循环神经网络的神经机器翻译（Neural Machine Translation）[26]、对话系统（Dialogue System）[27]等。可以说，结合分布表示的深度学习技术将人们一直以来所追求的目标——自然语言理解——往前推进了一大步。

本论文的主要研究内容是基于分布表示的跨语言、跨任务迁移学习在自然语言处理中的应用。我们将看到，分布表示为不同语言、不同任务之间构建了一座信息交互、知识迁移的桥梁，从而使得我们能够充分利用不同语言、不同标注规范的数据、甚至不同任务的标注资源，以提升目标语言或者目标任务上的分析精度。

1.2 研究现状与分析

本节首先介绍分布表示的两种定义以及典型的分布表示学习方法，然后介绍分布表示在多语言上的扩展。接着重点介绍目前分布表示在自然语言处理中的应用，最后介绍基于分布表示的迁移学习在自然语言处理中的研究现状。

1.2.1 分布表示

如前所述，分布表示指的是对于一个客观对象（如：词）的多维度向量表示，通常具有低维、连续、稠密的特性。事实上，在自然语言处理领域存在两种对于分布表示的定义。第一种称为“distributional representation”，可追溯到上世纪Firth提出的语义分布假设（Distributional Hypothesis）：“You shall know a

word by the company it keeps”[23, 24]。该假设指出，一个词的意义是由与其共现的上下文决定的。也就是说，上下文相似的词通常意义相近。第二种定义称为“distributed representation”，其概念最早由Hinton提出[28, 29]。Hinton将神经网络中隐含层神经元的激活向量，称为对输入数据的分布表示。Bengio等首次利用“词嵌入”的方法，将词映射为其分布表示向量，然后通过神经网络语言模型（Neural Network Language Model，简称NNLM）的学习，对其进行迭代更新[22]。因此，这种词汇分布表示也称为Word Embedding。

两种定义下的分布表示学习机制存在一定的差异。在第一种定义下，词汇分布表示的各个语义维度是显式建模的，一般是通过“计数”的方式获取。而第二种定义下的分布表示学习则是通过“预测”来获得，通常是先假定每个词的向量表示（可以是随机的），再将其作为模型的参数进行学习。需要注意的是，尽管两者在学习方法上有所不同，但实际上，两种方法都是基于语义分布假设，利用上下文信息来刻画词语之间的相似性。因此，两者殊途同归。本文不对两者进行概念上的区分。接下来，我们分别对“计数”模型以及“预测”模型进行详细阐述。

“计数”模型。潜在语义分析（Latent Semantic Analysis，简称LSA）[21]是一种典型的“计数”模型。在LSA中，我们首先构建一个词–上下文（word-context）的共现矩阵，然后使用矩阵分解技术对其进行降维处理，从而得到每个词的分布表示向量。其中有三个重要的技术细节：

1. 上下文的选取。不同的上下文决定了最终所得到的词表示所蕴含的性质。比较常用的上下文是“文档”[30]，一定窗口上下文中的“词”[31]，或者“*n*元短语”（n-gram）。一般的，假如上下文是文档，或者窗口很大的无序上下文，那么词表示更倾向于表现出主题或者语义层面上的性质；假如上下文是与目标词距离较近的词，或者含有词序信息，那么词表示将蕴含更多的句法性质。我们甚至可以直接使用与具体性质有关的上下文或者特征，比如根据词语在句法树上的依存关系所抽取出的依存上下文[32]。这样所得到的词表示则会蕴含更多的依存句法搭配的性质。
2. 共现矩阵中值的确定。一般需要对共现矩阵进行加权处理，以使得矩阵中每个元素能够更好地表达词与上下文之间的交互。常用的加权方式有tf-idf，点互信息（Pointwise Mutual Information, PMI），取对数等。
3. 降维。在共现矩阵中，每个词所对应的是一个高维且稀疏的向量表示，因此需要对其进行降维。最常用的降维方式有奇异值分解（Singular Value Decomposition, SVD），非负矩阵分解（Non-negative Matrix Factorization,

NMF），典型关联分析（Canonical Correlation Analysis, CCA），Hellinger PCA[33]等。除此之外，也可以使用一些非线性降维方式，比如自编码器（AutoEncoder）。降维能够消除一些原向量中所含有的噪声，同时也会损失一部分可能有价值的信息。因此，在实际应用中，需要根据具体情况来调节降维之后的向量维度。

值得一提的是Pennington等人提出的Glove模型[34]。Glove可以认为是一种对“词–词”矩阵进行分解的模型，它通过回归的方式来对加权之后的共现矩阵进行拟合。不同于其他矩阵分解方式，Glove只考虑矩阵中的非零值[2]。

“预测”模型。目前主流的预测模型主要是受到神经网络语言模型所启发。不同于传统的基于“计数+平滑”的非参数N元语法模型（N-gram model），神经网络语言模型是一种参数化的判别式模型，其基本思想是利用文本中上下文的词表示向量作为输入，通过神经网络计算出当前词的概率分布。Bengio等首次提出基于分布表示与前馈神经网络的语言模型（Feed-forward NNLM）[3]，其结构如图1-1所示。

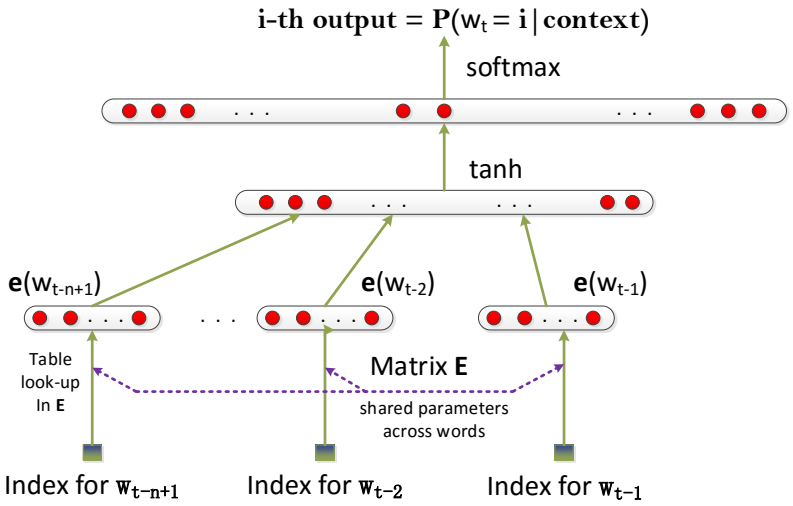


图 1-1 前馈神经网络语言模型[22]。

Figure 1-1 Feed-forward Neural Network Language Model.

该模型在普通的两层神经网络结构基础之上，增加了一层投射层（矩

[2]实际上，Glove首先假设每个词的向量表示，再将两两之间的点积去拟合词对在共现矩阵中的值，因此也可以认为是一种“预测”模型。但是，从模型所利用的信息角度来考虑，将Glove视为一种“计数”模型更加合适。

[3]实际上，Xu与Rudnicky在2000年最早提出基于神经网络的语言模型[35]，但是他们的模型使用词的One-Hot表示作为网络的输入，而非分布表示。

阵$\mathbf{E}_{V\times d}$，V为词表大小，d为分布表示向量维度），又称“词嵌入”层，将输入层的词投射为其分布表示向量。接下来，模型以句子中w_t的前$n-1$个词向量（历史）作为输入：$\boldsymbol{x} = [\boldsymbol{e}(w_{t-n+1}) \oplus ... \oplus \boldsymbol{e}(w_{t-2}) \oplus \boldsymbol{e}(w_{t-1})]$，再将其依次送至非线性隐含层以及输出层[4]。输出层维度为词表大小，模型采用`softmax`函数对输出层进行归一化，从而获得目标词w_t出现的概率。

最后，通过最大似然估计（Maximum Likelihood Estimation, MLE）或者其他合适的损失函数，即可对该网络进行训练（参数估计）。训练完成之后，投射矩阵$\mathbf{E}$的每一行则为对应词的分布表示向量。

固然，如果以语言模型为学习目标，那么该神经网络的输入一般限定为所预测词的前文；但是，如果以词的分布表示为学习目标，该神经网络的输入则更为自由，可以同时包含所预测词的前文及后文，甚至更丰富的信息。

同样的词嵌入思想可以扩展到使用其他网络结构的语言模型，比如循环神经网络语言模型（Recurrent Neural Network Language Model，简称RNNLM）[36]。RNNLM是一种时序模型，通过隐含层向输入层的反馈，使得在对语言进行建模时能够利用更长的历史信息。除此之外，Mnih与Hinton提出对数–双线性语言模型（Log-Bilinear Language Model，简称LBL）[37, 38]，去掉了隐含层的非线性变换操作，同时，输出层直接使用词向量矩阵作为权值矩阵。LBL的训练效率显著优于前馈神经网络语言模型以及循环神经网络语言模型。基于相似的思想，Mikolov等于2013年提出了word2vec[39]，在前馈神经网络语言模型的结构之上作了最大限度的简化，极大优化了模型的学习效率，使得在大规模数据上训练词语分布表示成为可能。word2vec中描述了两个重要的模型，分别是**连续词袋模型**（Continuous Bag-of-Words Model，简称CBOW）以及**跳跃文法模型**（Skip-gram Model），如图1-2所示。

相对于之前的模型，CBOW模型作了两方面简化：1. 去掉了非线性隐含层；2. 与LBL相比，CBOW不考虑上下文的词序信息，而是将输入的上下文词向量直接相加取平均，再计算与目标词向量之间的点积。形式化地，假设目标词为w_t，取其左右各n个词作为其上下文$C = \{w_{t-n}, ..., w_{t-1}, w_{t+1}..., w_{t+n}\}$。首先我们将上下文词向量的平均作为上下文表示：$\boldsymbol{x} = \frac{1}{|C|}\sum_{w_j\in C}\boldsymbol{e}(w_j)$，然后使用对数–线性模型对目标词进行预测：

$$P(w_t|C) = \frac{\exp(\boldsymbol{e}'(w_t)^\top \boldsymbol{x})}{\sum_{w'\in\mathbb{V}}\exp(\boldsymbol{e}'(w')^\top \boldsymbol{x})} \tag{1-1}$$

[4]在Bengio等人[22]的工作中，输入层还可以直接与输出层进行连接，相当于一个非线性神经网络模型与一个线性模型的集成。

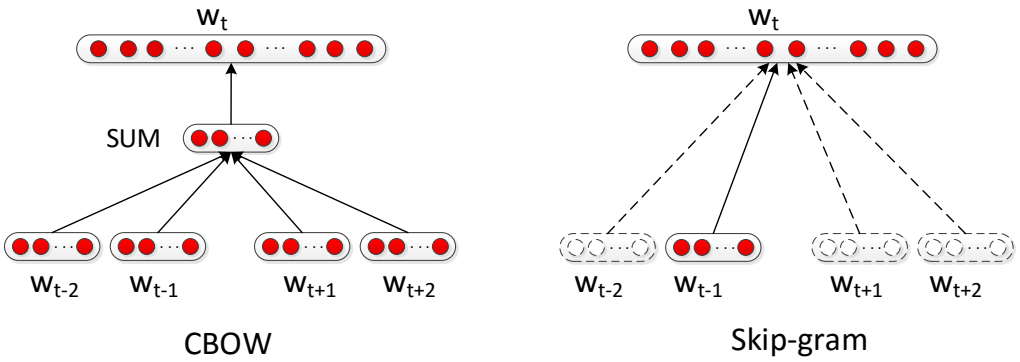


图 1-2 连续词袋模型（左）与跳跃文法模型（右）。

Figure 1-2 CBOW model (left) and Skip-gram model (right).

注意到word2vec采用了两个分布表示矩阵，以区分作为输入与输出的词向量（$\boldsymbol{e}$与$\boldsymbol{e}'$）。另外，word2vec实际上表示的是一个有向图，所有“上下文”到“词”的连接均为有向边，因此对输入、输出进行区分是合理的，正如在LSA中所得到的“词”与“上下文”低秩表示矩阵。实际上，与“计数”模型类似，“预测”模型也可以使用更丰富的上下文，比如依存关系的上下文[40]，甚至跨语言上下文（见1.2.2节）等。获得条件概率分布之后，即可以利用最大似然估计来训练CBOW模型。

Skip-gram模型可以视作CBOW模型的一个特例，见图1-2（右）。在Skip-gram模型中，我们每次只从上下文集合$\mathcal{C}$中选择一个词c，将其词向量作为模型的输入，直接对目标词w_t进行预测：

$$P(w_t|c) = \frac{\exp(\boldsymbol{e}'(w_t)^\top \boldsymbol{e}(c))}{\sum_{w' \in \mathbb{V}} \exp(\boldsymbol{e}'(w')^\top \boldsymbol{e}(c))} \tag{1-2}$$

类似的，可以利用最大似然估计进行训练。因此，在Skip-gram模型中，实际上是直接对“词”与“词”之间的上下文共现关系进行建模，也与LSA类似。

由于输出层对于整个词表进行遍历的代价较高，因此，为了提升训练效率，word2vec提供了两种方法，分别是层次化softmax[37]，以及负采样技术（negative sampling）。可以证明，使用负采样技术的Skip-gram模型实际上等价于一种隐式的矩阵分解[41]。

近几年，学术界也一直在讨论“计数”模型和“预测”模型孰优孰劣。譬如，Baroni等人在词汇语义相关任务上的实验证明，“预测”模型显著优于“计数”模型[42]。而Levy等人则证明，如果将word2vec中相似的技术细节迁移至“计数”模型中，两者的表现并无显著差异[43]。不过总体而言，由于“计数”模型涉及高内存开销的矩阵分解操作，从而限制了其对于数据规模的扩展

性。同时，“计数”模型一般需要对细节进行细致地调整，才能达到与“预测”模型相当的性能，因此，目前研究人员们使用更多的是“预测”模型，尤其是word2vec。

1.2.2 跨语言分布表示学习

我们已经知道，分布表示是以低维连续的数值向量来反映所表达对象的语义信息。因此，分布表示可以自然地联结表现形式不同但语义空间相同（或相似）的自然语言处理对象，最典型的莫过于跨语言词汇。譬如，中文与英文所用字符集完全不同，但是表达的却是相同的语义空间。近年来，越来越多的研究致力于跨语言分布表示学习：将不同语言的词汇嵌入至同一个向量空间之内，使得相似语义的词在该向量空间内距离接近，如图1-3所示。跨语言的分布表示对很多跨语言应用（如文本分类、句法分析、机器翻译等）都有显著的推动作用。

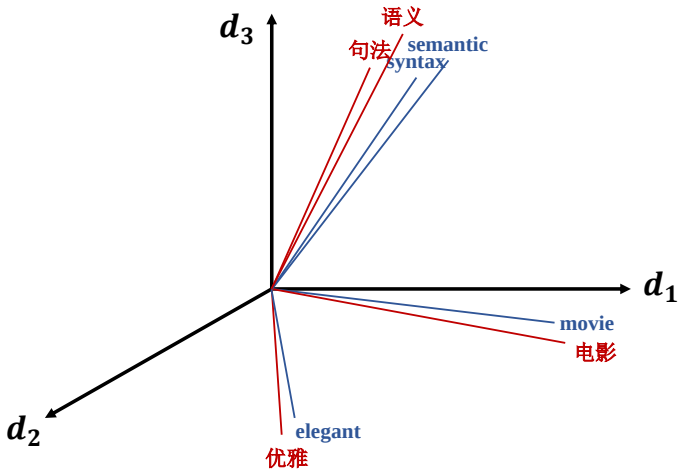


图 1-3 跨语言词汇分布表示（三维空间下的示意）。

Figure 1-3 Cross-lingual distributed word representation (3-dimensional illustration).

根据学习方式的不同，可以将现有的跨语言分布表示学习分为二类，分别是：1. 基于线下处理的方法；2. 基于联合学习的方法。下面对这两类进行详细阐述。

1. **线下处理方法**的基本思路是先独立学习各个语言的词汇分布表示，然后对两者进行对齐。Mikolov等发现，使用word2vec学习得到的不同语言的分布表示之间存在一定程度的线性映射关系，于是提出“翻译矩阵”学习的方法来实现跨语言分布表示的映射[44]。具体地，给定一个翻译词对的集合$\mathbb{D} = \{x_i, z_i\}_{i=1}^{n}$

（即双语词典，其中x_i为源语言中第i个词，z_i为目标语言中与x_i互为翻译的词），$\mathbb{D}$中词对所对应的分布表示矩阵分别记为$\boldsymbol{E}_{\mathbb{D}}^{src}$以及$\boldsymbol{E}_{\mathbb{D}}^{tgt}$。作者假设存在一个从源语言到目标语言的线性映射矩阵$\boldsymbol{W}$，使得平方误差损失最小：

$$\min_{\boldsymbol{W}} \| \boldsymbol{E}_{\mathbb{D}}^{src} \cdot \boldsymbol{W} - \boldsymbol{E}_{\mathbb{D}}^{tgt} \|^2 \tag{1-3}$$

再将$\boldsymbol{W}$应用于词典之外的其他词的跨语言映射。这种方法也能够很自然地扩展到多语言（≥ 2）的情形，只需为源语言与每种目标语言之间都学习一个相应的映射矩阵。

另外一种方法是典型关联分析（CCA）。CCA是一种度量两个多维变量之间线性相关性的技术。对于两个多维变量，CCA寻找两个映射矩阵，将原始的变量分别映射至一个新的子空间（一般情况下维度更低），使得两个变量之间的相关性最大[45]。如果把两种语言的词表示看作是两个多维变量，则可利用CCA对这两种语言的词表示矩阵进行变换[46]。我们仍然考虑双语词典$\mathbb{D}$，并将词表示矩阵记为$\boldsymbol{E}_{\mathbb{D}}^{l_1}, \boldsymbol{E}_{\mathbb{D}}^{l_2}$，CCA优化以下目标：

$$\max_{\boldsymbol{W},\boldsymbol{V}} \mathtt{Corr}(\boldsymbol{E}_{\mathbb{D}}^{l_1} \cdot \boldsymbol{W}, \boldsymbol{E}_{\mathbb{D}}^{l_2} \cdot \boldsymbol{V}) \tag{1-4}$$

Corr表示相关系数。$\boldsymbol{W}$与$\boldsymbol{V}$则可用于对两种语言的词表示矩阵进行映射。可见，CCA仍然是立足于线性变换的基础之上的。然而，由于语言的复杂性，线性变换对于词表示向量之间的映射关系刻画得并不理想，鉴于此，Lu等人提出深度典型关联分析（Deep CCA，DCCA），先对原始词向量进行多层非线性变换，再使用CCA处理[47]。

2. **联合学习方法**的目标则是同时学习多语言的词汇分布表示。Klementiev等人最早提出基于多任务学习（Multi-task Learning, MTL）来同时更新不同语言的词汇分布表示矩阵[48]。以英文-法文为例，该方法首先从英-法双语平行数据中根据词对齐信息获得双语词汇相似度矩阵（对齐频率）。分布表示的学习仍然依赖单语数据下的神经网络语言模型，每当计算英语中w_i^{en}的梯度时，都会根据相似度矩阵为相应法语词汇赋予梯度（根据相似度进行加权），从而达到联合学习的目的。Zou等人则借鉴前文所述“翻译矩阵”的思想，将双语词汇相似度矩阵作为跨语言映射矩阵，而以相应的平方误差损失作为单语分布表示学习目标的正则项[49]。另一个比较独特的思路是使用跨语言AutoEncoder[50]，也称相关性网络（Correlational Neural Network）。该方法将AutoEncoder的重构思想应用于双语平行数据之间的语义重构，基本思路是将一种语言的句子表示（Bag-of-Words）作为输入，期望重构出另一种语言的句子表示。对应的AutoEncoder权值矩阵则为所学到的跨语言分布表示。

受跨语言AutoEncoder所启发，我们可以通过优化双语平行数据中两种语言句子表示之间的距离，来实现跨语言分布表示的学习。在词分布表示的基础之上组合得到句子表示的方法有很多，为了保证分布表示学习的效率，一般使用较为简单的语义组合模型。Hermann与Blunsom提出双语组合语义向量模型（Bilingual compositional vector model, BiCVM）[51]，考虑了两种基本的语义组合方式，分别是ADD模型：$f(x) = \sum_{i=1}^{n} \boldsymbol{x}_i$，与BI模型：$f(x) = \sum_{i=1}^{n} \tanh(\boldsymbol{x}_{i-1} + \boldsymbol{x}_i)$ (考虑Bigram信息)。注意到，双语平行数据对于很多语言对而言都是非常有限的，因此，Gouws等人进一步提出Bilbowa模型[52]，将单语分布表示学习的优化目标与双语平行约束进行融合，使得模型能够同时利用大规模单语数据以及有限的双语数据。

线下处理与联合学习的方法各有优劣。从资源依赖的角度，线下处理方法使用的是双语词典，而联合学习方法依赖双语平行数据。双语词典既可以从平行数据中通过自动词对齐来获得，也可以从一些在线词典资源中获取，比如PanLex[5], Wikitionary[6]等；相对而言，大多数语言对之间的高质量双语平行数据较难获取。因此，线下处理方法的可扩展性更强。然而也应该注意到，线下处理方法对于跨语言映射的线性变换假设不尽合理，这在很大程度上制约了其分布表示学习的质量。联合学习方法则通常不对跨语言分布表示之间的映射关系进行约束，因此更为自由。

1.2.3 基于分布表示的自然语言处理

分布表示在自然语言处理任务中有着广泛的应用。首先，分布表示旨在对自然语言处理对象的语义进行较为全面的表达，因此能够方便地用于语义相似性的计算；其次，从未标注文本中学习得到的分布表示能够为传统的统计自然语言处理模型提供额外的特征，在一定程度上缓解由离散的符号表示所带来的数据稀疏问题；最后，分布表示可以与深度神经网络等非线性模型进行联合学习，利用深度神经网络强大的特征组合与学习能力，在基础特征的分布表示之上获得更高层次、更抽象、与目标任务更为相关的的分布表示。接下来对这三个角度进行详细介绍。

（1）语义相似性

分布表示的低维、连续向量表达形式，使其非常适用于对象之间语义相似

[5]https://panlex.org

[6]https://en.wiktionary.org

度的计算（如最常用的cosine距离）。事实上，在早期词汇分布表示学习的研究中，就已经采用语义相似性来对分布表示的质量进行评价[31, 53]。后续研究中常用的评测集有WordSim-353（英文）[54]以及相应的中文版本[55]等等。同时，词汇的语义相似度计算能够对诸多上层应用提供支持，比如语言模型、词义消歧、信息检索、复述识别等[31, 53, 56]。

由语义相似性衍生出来的一个有趣的性质是分布表示对于语义关系的表达。Mikolov等首先发现，满足相同语义关系的两个词对，其分布表示的差向量也非常接近[57]。一个典型的例子是：

$$\mathbf{e}(woman) - \mathbf{e}(man) \approx \mathbf{e}(queen) - \mathbf{e}(king)$$

根据这种“类比”性质，我们能够回答类似：“北京之于中国，正如华盛顿之于?”的问题。同时也启发了后续基于分布表示的关系学习（Relational Learning）研究。比如，Fu等人发现，词分布表示在一定程度上蕴含了实体与概念之间的上下位关系，从而能够有效地用于上位词识别[58]。而关系分布表示学习也逐渐成为知识库补全（Knowledge Base Completion）任务的典型方法之一[59, 60]。

（2）线性模型下的半指导学习——作为特征

根据1.2.1与1.2.2节中对于分布表示学习的描述与分析，可以发现，词汇分布表示通常是从未标注文本中进行训练的，对于标注资源的依赖非常小。这种特性使得半指导学习变得代价极低而且异常简单。人们很早就发现，从大规模未标注数据中学习得到的词聚类（如布朗聚类[61]）、互信息等特征能够显著提升多项自然语言处理任务的性能，如序列标注（中文分词、词性标注、命名实体识别等）[2, 62–64]以及依存句法分析[65]等等。词聚类的基本思想是给上下文相似的词打上一个离散的标签，相当于一种细粒度的词性，因此能够表达的语义信息较为有限。而分布表示的提出与应用则为这种半指导学习方法提供了一个新的思路。Turian等首次在序列标注模型中融合了基于神经网络语言模型的词分布表示特征，在命名实体识别以及组块分析（Chunking）任务中取得了显著的性能提升[66]。他们的做法是直接将原有的符号表示特征向量（离散、稀疏、高维）与分布表示特征向量（连续、稠密、低维）拼接在一起，从而构成新的特征向量，再统一由线性模型处理（见图1-4（左））。

然而，这种简单直接的拼接方式似乎显得过于朴素，使我们不禁考虑一个问题：**分布表示特征空间是否适合由线性模型来处理?** 在基于符号表示的特征空间下之所以多用线性模型，原因之一是特征维度太高，对模型的

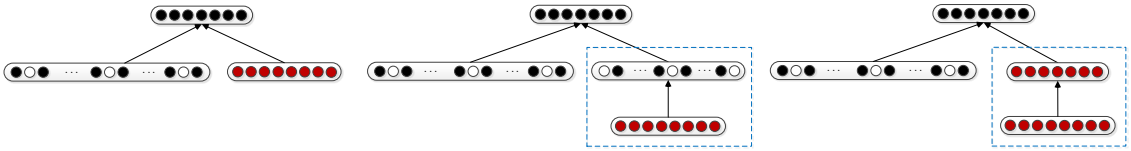


图 1-4 线性模型下引入分布表示特征的方式：Turian（左），Guo（中）以及Zhang（右）。

Figure 1-4 Methods of augmenting distributed-represented features in (generalized) linear models: Turian et al., (left), Guo et al., (middle) and Zhang et al., (right).

复杂度有限制；另一个原因是其本身的高维、稀疏性质使得其线性可分性较强。而低维、稠密的分布表示特征空间则不然。针对这个问题，Wang等人在NER与Chunking任务上，研究了线性模型与非线性模型分别在符号表示特征空间与分布表示特征空间下的应用效果[67]。他们的实验表明，在符号表示特征空间下，线性模型显著优于非线性模型；而在分布表示特征空间下，非线性模型则远远优于线性模型。

受该研究的启发，Guo等人提出将分布表示特征进行离散化，再应用于线性模型[68]（见图1-4（中））。我们研究了三种离散化方式，分别是二值化、聚类以及分布原型法（distributional prototype）。在NER上的实验表明，离散化之后的分布表示特征能够带来更为显著的提升。同时，我们也设计实验对不同特征表示下数据的线性可分性（linear separability）进行了分析，发现在离散化之后，特征空间的线性可分性确实明显优于原始的分布表示。

然而，离散化虽然证明有效，但是不可避免地会损失一部分原始分布表示中的信息。另一种思路是：能否对分布表示特征进行某种非线性变换，使得变换之后的特征空间线性可分性更强？Zhang等人尝试对分布表示进行一次非线性变换，再与原有的离散表示特征进行拼接（见图1-4（右）），在依存句法分析任务中取得了较好的结果[69]。事实上，使用非线性模型（尤其是深度神经网络）进行特征学习的研究由来已久，接下来我们将进行详细介绍。

（3）非线性模型下的特征学习

使用非线性模型进行特征学习的基本出发点是对组合特征的捕捉。对于自然语言处理大部分核心任务，尤其是序列标注、句法分析等结构预测问题而言，其假设空间（hypothesis space）通常过于庞大。因此，一个好的特征表示对于统计模型的学习至关重要。在传统方法中，由于线性模型无法有效地刻画特征之间的组合性质，因此，人们根据自己对于特定任务的先验知识以及大量工程经验，精心设计了复杂的组合特征模板。以语义角色标注任务（Semantic Role Labeling, SRL）为例，CoNLL 2009评测中最好的系统使用了超过50个特征

模板[70]。这种严重依赖于专家知识的“特征工程”不仅代价较高，而且受到特征组合不完全、特征空间维度高以及特征抽取时间代价高等问题的影响[71]。在这样的背景之下，基于非线性模型的特征自动学习提供了一个优雅的解决方案，其中尤以深度神经网络为代表。

回顾1.2.1中所介绍的神经网络语言模型，如NNLM，模型以上下文分布表示向量作为输入，经过一个非线性隐含层（tanh激活函数）得到中间表示，再由输出层（softmax，一种广义线性模型）计算目标词的概率分布。在这个简单的前馈网络中，隐含层所得到的分布表示实际上就是对于作为输入的上下文分布表示的一种组合与抽象。同样的思想也可以用于其他任务，比如句法分析。Chen与Manning在基于转移的依存句法分析系统中使用非线性隐含层（cube激活函数）来计算转移状态的分布表示，再使用softmax进行转移动作的预测[71]。他们的模型只需使用少量基础特征的分布表示作为输入，而在隐含层实现特征之间的组合，从而大大提高了依存句法分析的效率（每秒钟处理1,000个句子）。

除了结构相对简单的前馈神经网络，在自然语言处理中常用的网络结构还有循环神经网络、卷积神经网络和递归神经网络等。

1. 循环神经网络（RNN）是一种时序模型，与前馈神经网络的主要区别在于引入了回馈机制。Jordan在1986年最早提出在神经网络中引入由输出层到输入层的回馈连接，以获取时序行为中的动态记忆（dynamic memory）[72]；Elman对该结构进行了修改，将回馈连接改为由隐含层指向输入层[73]。因此，在Elman循环神经网络中，下一时刻所看到的不是上一时刻的输出结果，而是其中间表示。Elman结构是现在最为常用的循环神经网络结构，也称为精简循环网络（Simple Recurrent Network），其基本模型结构如图1-5所示。对于一个输入序列（比如由词构成的句子）：$\mathbf{x} = [x_1, x_2, ..., x_n]$，RNN从左至右（或相反顺序）依次处理每个时刻的输入。同时，当前时刻的隐含层输出也作为下一时刻的输入：

$$\boldsymbol{h}_t = f(\boldsymbol{W} \cdot [\boldsymbol{x}_t \oplus \boldsymbol{h}_{t-1}] + \boldsymbol{b}) \tag{1-5}$$

因此，在每个时刻，RNN计算的是当前输入与（理论上）所有历史信息的组合分布表示。当模型训练完成之后，我们可以将最后时刻的隐层输出$\boldsymbol{h}_n$作为对于该序列$\mathbf{x}$的特征表示。另外，我们也可以综合考虑历史上所有隐含层表示：$\boldsymbol{h} = g(\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_n)$。其中$g$的设定较为自由，可以简单的取平均操作，也可以是更复杂的加权求和，甚至是动态权重——即注意力机制[74]。

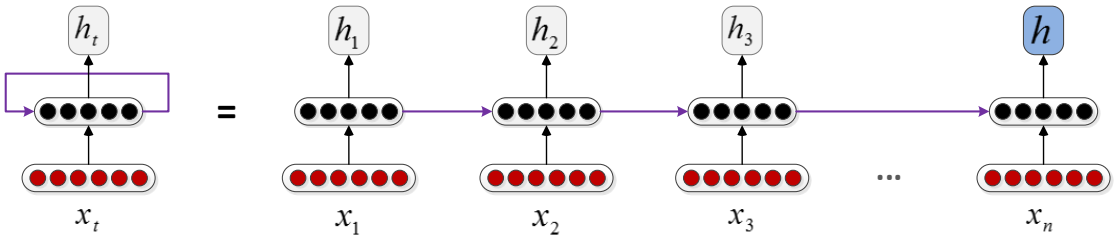


图 1-5 循环神经网络结构。

Figure 1-5 Architecture of recurrent neural network (RNN).

RNN在自然语言处理中取得了极大的成功。早在上世纪RNN刚被提出时，研究者们就已经成功地将其应用于词类标注、口语分析等自然语言处理任务[75–77]。但是受限于当时的计算能力，RNN并没有引起广泛的重视。近年来，随着RNN在语言模型任务上的出色表现[36, 78]，人们开始普遍意识到RNN对于文本处理这一对于长距离依赖较为敏感的领域具有很强的适用性。目前，RNN几乎被应用于绝大多数自然语言处理问题之中，包括传统的结构预测任务（词法、句法、语义分析等）以及NLP最核心的问题之一——机器翻译。同时，RNN也为一些传统技术很难实现的目标带来了一线曙光，如文本摘要、诗歌生成、对话系统、图片字幕生成等等。在这些令人眼花瞭乱的成果背后，离不开两大里程碑式的贡献。一是长短时记忆（Long Short-Term Memory, LSTM）的提出[79]。在上个世纪人们就已经发现，RNN这种极深的网络在训练过程中存在严重的梯度消失（gradient vanish）或者梯度爆炸（gradient explode）问题[80]，使得RNN在实际中对长距离依赖的捕捉效果并不理想[81]。以LSTM为代表的带门循环神经网络（gated RNN）通过对网络中门的控制，来强调或者忘记某些输入，使得RNN中的信息流能够有效地进行长距离传递。其次是编码–解码（encoder-decoder）框架的提出[26]。该框架为序列–序列（sequence to sequence）这一大类问题提供了一套简单有效的端到端（end to end）解决方案。

2. 卷积神经网络（Convolutional Neural Network, CNN）主要考虑了生物神经网络中的局部接收域（reception field）性质[82]，即隐含层神经元只与部分输入层神经元连接，同时不同隐含层神经元的局部连接权值是共享的。CNN主要由卷积层（convolutional layer）与池化层（pooling layer）构成。其模型结构如图1-6所示。可以看出，我们记$\boldsymbol{x}_{1:n} = \boldsymbol{x}_1 \oplus \boldsymbol{x}_2 \oplus ... \oplus \boldsymbol{x}_n$，表示输入特征向量按顺序拼接。卷积层计算方式如下：

$$\boldsymbol{h}_i = f(\boldsymbol{W} \cdot \boldsymbol{x}_{i:i+k-1} + \boldsymbol{b}) \tag{1-6}$$

其中k为预设的卷积窗口大小。池化层常用的操作有最大化池化（Max Pool-

ing）、平均池化（Mean Pooling）等，如：

$$\boldsymbol{h} = \mathtt{max\_pooling}(\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_{n-k+1}) \tag{1-7}$$

即为最终对于序列的特征表示。在实际应用中，通常可以使用不同窗口大小的卷积组合，以捕捉不同粒度的局部特征组合。

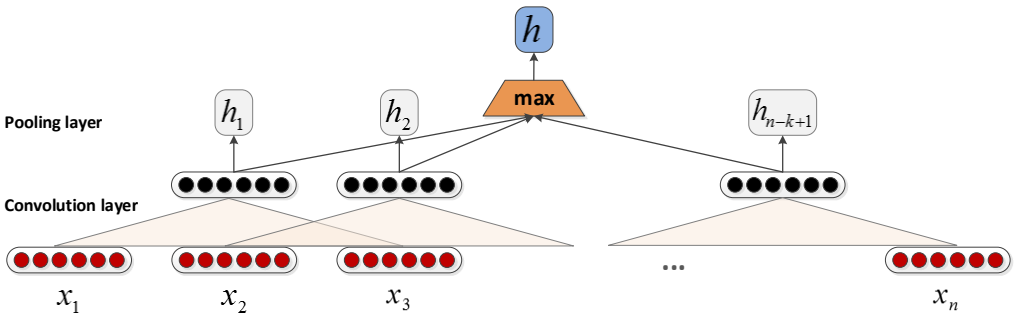


图 1-6 卷积神经网络结构。

Figure 1-6 Architecture of convolutional neural network (CNN).

CNN的局部连接性质使得它对于数据的平移、比例缩放、倾斜等形式的变化具有高度不变性，因而目前被广泛应用于图像处理领域[83]。同时，这一性质在很多自然语言处理的任务中也有所体现，如对评论文本进行分类，最终的褒贬性往往由局部的一些短语决定，同时不需要顾及这些短语在文本中的位置信息。因此，CNN近年来也被大量应用于文本分类相关的任务并取得了非常理想的效果[84–86]。事实上，CNN在自然语言处理中最早的应用是由Collobert等提出的多任务自然语言处理框架，也被称为C&W模型[9, 10]。C&W模型主要面向序列标注任务，主要思想是将分布表示特征（词、词缀、类型等基本特征）作为多任务深度神经网络的输入，通过卷积网络自动学习特征之间的组合与交互，再经过一层全连接的非线性隐含层，最终在输出端进行标签预测。不过由于序列标注任务对于词序等信息更为敏感，所以CNN不一定是最理想的特征学习方案。从近年来的研究进展来看，一个更为理想的解决方案是使用循环神经网络。

3. 递归神经网络（Recursive Neural Network，RecNN）是循环神经网络针对自然语言句法结构所作的一种变形。如前所述，RNN是以从左至右（或者相反）的线性顺序对输入序列进行组合，而CNN的组合方式也较为单一。相对而言，RecNN中的语义组合方式则比较自由，可以完全根据自然语言的语法结构进行自底向上的递归组合，如图1-7所示：在对输入序列进行组合时，首先组合x_1与x_2，生成隐层表示$\boldsymbol{h}_1$；再组合x_3与x_4，获得$\boldsymbol{h}_2$；$\boldsymbol{h}_1$与$\boldsymbol{h}_2$再进行组合得到$\boldsymbol{h}_3$，

最后与x_5组合得到最终的序列表示$\boldsymbol{h}$。

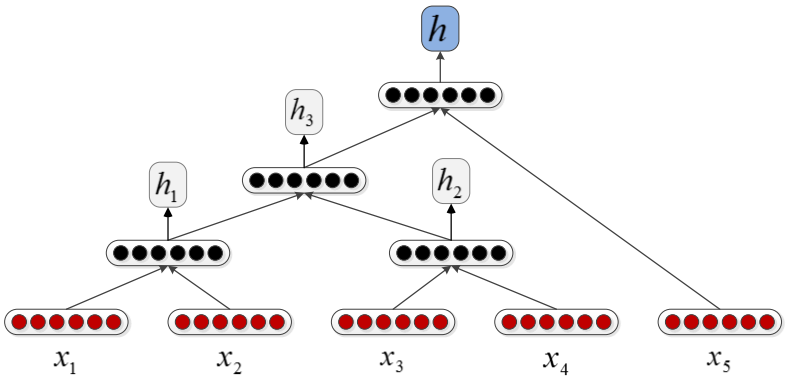


图 1-7 基于给定句法结构的递归神经网络结构。

Figure 1-7 Architecture of recursive neural network (RecNN) based on syntactic structure.

递归神经网络最早藉由Socher等人在句法分析、情感分类、复述识别等任务上的一系列研究工作[56, 87–91]而引起自然语言处理领域学者的广泛关注。人们普遍认为，对于自然语言这种内在递归性质较强的研究对象，融合了句法结构的RecNN应比RNN/CNN更为适用。

与RNN类似，RecNN在训练时也会受到梯度消失或者梯度爆炸的影响。因此，人们将长短时记忆（LSTM）的思想引入RecNN，提出了树状结构的LSTM（Tree-LSTM）[92, 93]。

1.2.4 基于分布表示的迁移学习

自然界中万事万物，无不呈现着各自不同的形态。比如我们眼中看到的景致、听到的声音、读到的文字等。即使是同一种形态的信号，其表现形式也可能存在极大的差异，比如不同语言的文字。然而，在语义层面，这些信息的表示却是统一的。当我们读到“细雨鱼儿出”的时候，大脑中会浮现出一幅“鱼儿在细雨中摇曳着身躯，吐着水泡儿”的自然图景；当我们阅读一篇英文文章的时候，同样会联想其对应的中文翻译。这些不同的信号在大脑中相互转化与融合，不断地帮助我们更加深刻地理解和感受这个美妙的世界。分布表示提供了一种具有**通用性**的语义表示方法，为不同语言、不同模态数据、不同任务之间建立了一座信息交互、知识迁移的桥梁。

1. 跨语言迁移学习。来自不同语言的数据是一种重要的知识来源，尤其是对于资源稀缺语言而言。Yarowsky等首次提出使用跨语言映射的方法，利用源语言（英语）端丰富的标注资源自动构建目标语言的训练数据[4]。这种方法

又称为“数据迁移”（data transfer），类似的方法也被应用于跨语言依存句法分析[5, 94–96]和语义角色标注[97]。句法、语义分析资源的构建难度远远大于词法分析，因此，跨语言资源显得更为重要和宝贵。

“数据迁移”方法有诸多限制，比如对双语平行数据的需求、容易受到词对齐错误的影响、需要人工设计跨语言映射的规则等等。因而，近年来另一种跨语言迁移方法——“模型迁移”（model transfer）开始受到广泛关注。比如，McDonald等人构建的“去词汇化”（delexicalized）依存句法分析器能够直接在源语言句法树库上进行训练并对目标语言文本进行句法分析[98]。之所以选择“去词汇化”，主要原因是符号化表示的词汇化特征无法在不同语言之间进行迁移。幸运的是，1.2.2节中所介绍的跨语言分布表示学习为此提供了一种有效的解决方案。其基本思路是在源语言端构建一个基于分布表示的模型，并使用源语言的特征分布表示进行训练；这样所得到的模型便可以直接应用于目标语言数据。基于跨语言分布表示的“模型迁移”已经被成功应用于文本分类[48, 50, 51]、情感分析[99]等任务中。

2. 跨模态数据迁移学习。不同模态的数据，如图像、声音、文字等，反映了语义的不同表现形式。在很多场景中，单独的某种模态表现能力不足。所以当人类在相互交流的过程中，会频繁地借助语气、文字、图像、肢体语言等丰富的表达方式。在自然语言处理中，借助跨模态分布表示学习，可以将不同模态的数据表示在一个统一的语义空间内，从而实现知识的融合与迁移。比如，通过图片与词语之间的跨模态分布表示学习，能够使得图片分类模型对于训练数据中未出现过的图片类别进行有效的识别（zero-shot learning）[100]；Socher等进一步对图片及图片描述（组合语义）进行联合分布表示学习，显著提升了“以句子搜图”等检索任务的效果[101]。最近备受关注的“图片字幕生成”（Image Caption Generation）任务主要采用Encoder-Decoder框架来实现图片与文本之间分布语义表示的映射[102]。

3. 跨任务迁移学习。相似（或相关）的任务之间往往蕴含着能够相互利用的知识。回顾1.2.3节中所介绍的，使用由神经网络语言模型所训练出来的词表示作为NER等任务的特征[66]或者深度神经网络的初始化参数[10]，能够显著提升模型在目标任务上的效果。这意味着，不同任务之间可以通过参数共享来实现知识迁移。概括而言，跨任务的迁移学习主要有两类方法，一类是将任务A中训练得到的模型参数（或者部分）直接应用于目标任务B的模型，或者作为相应参数的初始值。这种方法比较适用于目标任务的训练数据较少（甚至没有）的情形。另一类是多任务学习（Multi-task Learning，MTL），即对多个任务进行

联合训练，而两者共享一部分模型参数。多任务学习是一种归纳式的迁移学习（Inductive Transfer Learning），适用于目标任务训练数据也较为充分的情形。在多任务学习中，通常我们要求不同任务之间采用相似的模型结构。

分布表示与神经网络为多任务学习提供了一个简单有效的框架[16]。在深度神经网络中，越接近输出的网络层所学到的分布表示与目标任务越相关，而越底层的分布表示通用性越强[103]。这个性质可以为多任务学习中的参数共享机制提供有效的指导。同时，多任务学习是否有效的另一个重要的影响因素是各个任务之间的关联性（相似、互补）如何。Collobert等所提出的C&W模型[10]对词性标注、命名实体识别、语义角色标注进行联合学习，考虑的正是词法、语义任务之间内在的相似性与互补性。

从统计学习的角度，迁移学习一方面为目标任务提供了额外的训练数据，另一方面也引入了来自不同任务的归纳偏置（Inductive bias），从而缓解目标任务上的过拟合问题。从认知的角度，我们认为迁移学习在一定程度上反映了人类成长和学习的过程。事实上，从Hinton的观点出发，我们可以将迁移学习理解为一种广义的知识萃取（knowledge distillation）[104]，即对模型A中的知识进行进一步加工，再与模型B进行融合。然而，一直以来，由于自然语言的复杂性以及传统符号化的局部特征表示所带来的限制，这方面的研究并没有引起广泛的关注。本文将深入探讨分布表示学习的引入为不同语言、不同数据类型以及不同任务之间的知识迁移所带来的机遇。

1.3 本文的研究内容及章节安排

本文从自然语言处理的两个主要研究层面：语法（包含词法和句法）以及语义分析入手，系统深入地研究了基于分布表示的跨语言、跨任务、跨数据类型迁移学习在不同任务中的关键技术。总的来说，我们做了四方面的工作：（1）在词法层面，我们研究如何利用跨语言信息学习单语词义级的分布表示，旨在更准确地刻画自然语言中“一词多义”现象。同时，我们将词义分布表示以特征形式应用于序列标注任务，在命名实体识别上的效果显著优于传统分布表示特征；（2）在句法层面，我们提出三种面向资源稀缺语言依存句法分析的跨语言分布表示学习方法，填补了不同语言之间句法分析的“词汇化特征鸿沟”，极大提升了跨语言依存句法分析的性能；（3）我们也系统深入地研究了目标树库分别在稀缺与丰富的情况下，如何更有效地利用已有的树库资源，并提出了多层次参数共享的多任务学习机制，使得多类型树库能够有效地融

合；（4）在语义层面，我们提出了一个统一的深度神经网络框架，使得语义角色标注与语义关系分类这两个看似不同却有着较强相关性的任务能够进行联合学习，并在语义角色标注任务上取得了当前最好结果。

从知识迁移的角度出发，第（1）部分的工作使用的是“数据迁移”（也称标注映射）的方法来进行跨语言知识映射，从而自动构建目标语言上的词义消歧数据以用于词义分布表示的学习。第（2）、（3）、（4）部分工作则是采用迁移学习的方法，从模型层面上实现多语言多任务之间的知识迁移。其中，第（2）部分迁移学习工作是利用基础特征（词汇）的跨语言分布表示来实现迁移学习；而第（3）、（4）部分工作则进一步地利用深度神经网络中抽象层次更高的分布表示来进行多语言多任务的知识迁移。因此，这四部分工作以分布表示为核心，由浅入深地对语法及语义分析中的知识迁移方法进行了系统地研究。整篇论文结构框架如图1-8所示：

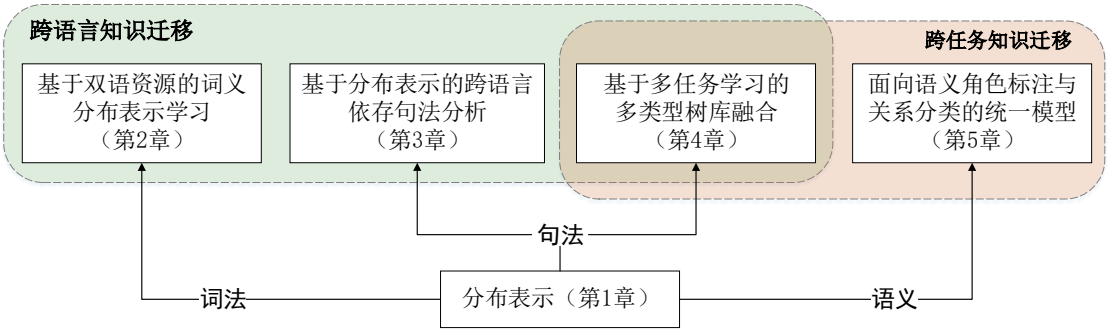


图 1-8 论文总体框架。

Figure 1-8 Structure of the thesis.

具体的，本文含5章，各章内容组织如下：

第1章，介绍了本文课题的研究背景、意义，并对分布表示的研究现状进行了概述与分析，最后对本文主要内容进行了规划。

第2章，针对词语的单向量分布表示无法有效表达自然语言中“一词多义”现象的问题，提出了基于双语资源的词义级分布表示学习方法。同时，构建了一个中文多义词相似度数据集。在该数据集上的评价结果表明，相比于单向量词分布表示，词义级分布表示能够更准确地表达多义词的词义。我们进一步将词义级表示特征应用于序列标注任务，取得了比单向量词分布表示更为显著的性能提升。

第3章，针对跨语言依存句法分析中的“词汇化特征鸿沟”问题，我们首先提出两种双语分布表示学习方法，分别是跨语言映射与典型关联分析。进一步

的，针对多源语言的情形提出了跨语言跳跃文法模型，并扩展了双语情况下的跨语言映射方法，使其能够对多种语言的分布表示进行联合学习。实验证明，跨语言分布表示能够显著提升稀缺资源语言的句法分析效果。

第4章，提出了基于多层次参数共享的深度多任务学习框架，使得不同任务能够在多个抽象表示层之间实现信息传递与交互。我们将此框架应用于多类型句法树库的融合，具体包括多语言同构树库以及同种语言的异构树库。实验证明，该框架能够有效地实现树库之间的迁移学习，显著提升目标树库上的句法分析效果。

第5章，提出了面向语义角色标注与关系分类的统一深度神经网络模型，该模型能够充分捕捉对于这两个任务最为重要的三类特征：全局上下文特征、句法特征以及词汇语义特征。在此基础之上，我们利用深度多任务学习有效地实现了语义角色标注与关系分类任务之间的知识迁移。实验表明，该框架在这两个任务上都取得了当前最好（或者可比较）的结果，而多任务学习则进一步提升了语义角色标注的效果，达到了目前最好的水平。

第 2 章 基于双语资源的词义表示学习

2.1 引言

近年来，分布表示学习的思想在自然语言处理中占据越来越重要的位置。词作为自然语言的基础研究单元之一，其分布表示学习获得了格外的关注。根据第1.2.1节的介绍，目前对于词的分布表示学习，主要有基于“计数”的模型以及基于“预测”的模型。而无论哪种模型，其遵循的基本假设都是Firth所提出的分布语义假设[24]，即：一个词的含义是由与其共现的上下文来确定的。在本文中，我们通过研究观察发现，该假设只在单义语境下适用，而不能准确地表达自然语言中常见的“一词多义”现象（polysemy）。对于多义词而言，如“制服”、“苹果”等，它们在表达不同含义时所搭配的上下文分布存在较大的差异，现有的绝大部分词表示学习方法只是简单地将每个词当作一个概念的集合，利用语料中所有出现过的上下文来进行学习。由这种方式所得到的多义词向量表示，实际上是该词的多种词义表示在向量空间中的某种线性组合[105]。然而，在没有有效的方法对其进行词义解析或者分离的条件下，这种组合只会为上层应用（词相似性的度量、作为特征等）带来干扰。

为了更好地表达词的不同意义或者用法，研究者提出了“多向量”模型（Multi-prototype）——对于词表中的每个词，分别学习K个向量表示，其中K是预先设定的一个值[106, 107]。相对于传统的单向量模型，这种方法固然能够对一个词的意义及用法进行更细粒度地划分，但是考虑到不同词的词义数目不尽一致，使用统一的K向量表示依然无法准确刻画多义词的词义信息。

在本文中，我们提出一种简单有效的方法，利用双语平行数据来学习词义级别的分布表示。该方法的核心思想是首先对数据进行词义归纳（Word Sense Induction, WSI），然后在标记了词义类别的数据上使用神经网络语言模型进行分布表示学习。这里，不同于大部分基于上下文聚类的词义归纳方法，我们利用双语资源对于词义的消歧作用来进行词义归纳。例如，“制服”所对应的英文翻译有：*investment*, *overpower*, *subdue*, *subjugate*, *uniform*等。其中，*overpower*, *subdue*, *subjugate*表达同一种含义，而*investment*, *uniform*表达“制服”的另一种含义。很自然地，我们可以通过对翻译词进行聚类来对“制服”进行词义归纳。接下来，我们将所获得的词义聚类通过双语词对齐映射回目标语言，从而得到

了一个标注了词义的目标语言数据。在该数据上，我们使用循环神经网络语言模型进行训练，便可得到词义级别的分布表示。

为了对词义级分布表示进行评价，我们构建了一个中文多义词相似性评价数据集，其中对401个词对进行了相似度标注。实验表明，与单向量表示以及K向量表示相比，我们的词义级分布表示对于多义词相似度的估计更为准确。我们进一步在上层任务——命名实体识别（NER）中，采用半监督特征的方式[66]（1.2.3节）对词义级分布表示进行评价。为此，我们进一步提出一种新的基于循环神经网络的词义消歧模型，以确定词的词义类别。实验结果表明，使用词义级别分布表示特征显著优于单向量词分布表示特征。

2.2 背景与相关工作

2.2.1 基于循环神经网络语言模型的分布表示学习

我们知道，神经网络语言模型是目前词分布表示学习的主要方法之一（见第1.2.1节）。自从Bengio等人2003年首次提出“词嵌入”的思想[22]并成功应用于前馈神经网络语言模型之后，出现了一系列的改进工作，如对数–双线性语言模型[38]、循环神经网络语言模型（RNNLM）[36]等。其中，循环神经网络语言模型由于其在时序建模方面的良好性质，近年来逐渐受到自然语言处理领域研究人员的亲睐。

本研究中，我们将以Mikolov等人所采用的循环神经网络语言模型[108]为基础，该模型结构如图 2-1所示。输入层包含两部分，分别是当前词的One-Hot表示输入$\boldsymbol{s}(w_t) = [0, 0, ..., 1, ..., 0]$，以及上一时刻隐含层分布表示向量$\boldsymbol{h}_{t-1}$。接下来，隐含层表示以及输出层可由下式进行计算：

$$\boldsymbol{h}_t = f(\boldsymbol{U} \cdot \boldsymbol{s}(w_t) + \boldsymbol{W} \cdot \boldsymbol{h}_{t-1} + \boldsymbol{b}) \tag{2-1}$$

$$\boldsymbol{y}_t = \mathtt{softmax}(\boldsymbol{V} \cdot \boldsymbol{h}_t) \tag{2-2}$$

其中f为非线性激活函数，本研究中使用的是`sigmoid`函数。

通常采用BPTT（Back Propagation Through Time）算法[109]来计算循环神经网络中参数的梯度，并辅助随机梯度下降（SGD）等优化方法对该网络进行训练。与Bengio等人的前馈神经网络语言模型不同，该模型的输入层中对于当前词采用的是One-Hot向量表示，而不是显式的“词嵌入”。因此，当模型训练完

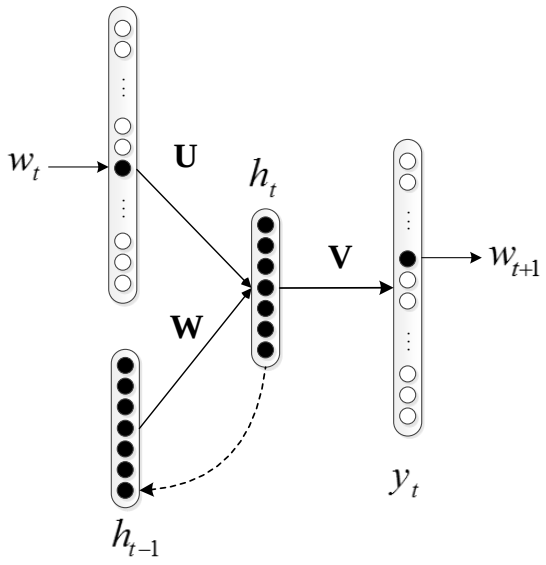


图 2-1 循环神经网络语言模型。

Figure 2-1 Recurrent neural network language models.

成之后，我们将$\boldsymbol{U} \cdot \boldsymbol{s}(w_t)$——即权值矩阵$\boldsymbol{U}$中对应$w_t$的列向量——作为$w_t$的分布表示向量。

2.2.2 面向多义词的分布表示学习

一词多义是自然语言处理中的一种常见现象，也是自然语言歧义性的一个重要来源。在传统的单向量分布表示学习模型中，多义词各个义项下的上下文被当成一个整体进行建模。为了研究这种单向量表示的性质，最近，普林斯顿大学的Arona等设计了一个实验[105]，主要做法是将语料中两个不同的词（u和v）合并为一个新的人造词（w），从而模拟“u与v是w的两种词义”这一情形。作者观察合并前u与v的向量表示：$\boldsymbol{e}(u)$，$\boldsymbol{e}(v)$与合并后w的向量表示$\boldsymbol{e}(w)$之间的联系。实验结果表明：$\boldsymbol{e}(w) \approx \alpha \cdot \boldsymbol{e}(u) + \beta \cdot \boldsymbol{e}(v)$，即$w$的向量表示实际上是它在不同词义（$u$与$v$）下向量表示的某种线性组合。这种组合为词语多义性的表达带来了困难。

为了在分布表示学习中更好地表达词语的多义性，直觉上最容易想到的是将分布表示学习与词义消歧模型结合起来：首先利用词义消歧模型对文本进行词义标记，然后在标记后的语料上训练词语在不同词义下的分布表示。事实上，大部分工作，包括本文所提出的方法，都是基于这种思路。大体上这方面的工作可以分为两类，分别是**基于知识库的方法**以及**无监督的方法**。

基于知识库的方法主要根据人工构建的知识库（如WordNet[110]等）来确定词的词义，由这种方式所获得的词义表示是“接地”的（grounded）。如Iacobacci等[111]使用BabelNet[112]作为词义知识库并使用Babelfy[113]对语料进行词义消歧。类似的，Chen等人[114]使用WordNet中所定义的词义及描述，提出一种简单的基于向量相似性排序的词义消歧方法，进而用于词义表示学习。此外，Rothe与Schütze则进一步利用WordNet中所定义的词义以及语义结构信息（如：上下位关系等）来学习词义表示[115]。

无监督方法则不依赖知识库，而是从数据中进行无监督的词义归纳（WSI）。Reisinger与Mooney首次提出多向量表示模型（Multi-prototype）[106]，采用上下文聚类的方式进行词义归纳，然后在标注了词义类别的数据上进行分布表示学习。Huang等采取了类似的做法[107]，不同于Reisinger与Mooney使用离散上下文特征向量作为聚类的基础，他们使用上下文词的分布表示向量平均作为词项的表示。除此之外，Tian等对word2vec中的Skip-gram模型进行了扩展，从而提出一种更为简单优雅的多向量表示学习模型[116]。他们将词义作为一个隐变量，对Skip-gram模型中词──上下文的预测概率进行分解，从而在不显式进行词义归纳的情况下获得每个词的多向量表示。

总的来说，基于知识库的方法对词义表示的学习更加准确，但是对资源的要求也较高，无法扩展到知识库不完善的语言。同时，它在实际中的应用效果受限于词义消歧系统的准确率。无监督方法通用性更强，但是由于缺少知识库的指导，绝大部分模型只能对词语的词义类别数进行粗糙的一致假设，严重限制了其对于词义的表达作用。同时，基于上下文聚类的词义归纳方法较为低效，尤其在语料规模较大的情况下。

本文所提出的方法属于一种无监督的方法。我们利用双语资源克服上下文聚类所带来的性能瓶颈，并使用仿射传播聚类算法来克服词义类别数的一致假设问题。

2.3 基于双语数据的词义表示学习方法

本节将详细介绍基于双语数据的词义表示学习方法。在该方法中，我们首先借鉴Gale等人[117]以及Chan和Ng等人[118]利用双语数据自动获取词义消歧（WSD）训练数据的思想，提出了一种基于双语资源的无监督词义归纳方法，然后使用循环神经网络语言模型在含有词义类别标记的数据上进行分布表示学习。如图2-2所示，我们的方法主要包括以下四个步骤：

1. **翻译词抽取**。从双语平行数据中抽取目标语言中每个词在源语言中的翻译词集合[1]。
2. **翻译词聚类**。对每个词的翻译词集合分别进行聚类，使得每个簇表示一种词义。
3. **跨语言词义映射**。将词义聚类映射至目标语言中每个词项（token），从而得到一个包含词义标记的语料。
4. **分布表示学习**。使用循环神经网络语言模型对该语料进行训练，进而获得词义级别的分布表示。

接下来对前三个步骤进行主要介绍。

2.3.1 翻译词抽取

本文提出一种基于双向词对齐概率的翻译词抽取方法。首先简要介绍一下词对齐的基本模型。记$c = (c_1, ..., c_J)$为一个中文句子，记$e = (e_1, ..., e_I)$为与其平行的英文句子。词对齐模型可以表示为：

$$p(c|e) = \sum_a p(a, c|e) \tag{2-3}$$

$$p(a, c|e) = \prod_{j=1}^{J} p_d(a_j|a_{j-}, j) p_t(c_j|e_{a_j}) \tag{2-4}$$

其中a为句子e到c的对齐模式，a_j表示e中第a_j个词（e_{a_j}）与c中第j个词（c_j）对齐，$p_t(c_j|e_{a_j})$即为e_{a_j}到c_j的翻译概率。通过对句子翻译概率在整体数据上的似然函数进行优化，便可以获得每个句对之间最优的词对齐，以及双语词对之间的翻译概率。

我们采用Liang所提出的基于双向一致性的词对齐模型[119]，并综合考虑中英双向翻译概率来进行翻译词抽取。具体的，给定目标词w_c，对于一个源语言词候选：w_e，只有在$p_t(w_e|w_c)$与$p_t(w_c|w_e)$均大于某个阈值$\delta \in [0, 1]$的情况下，才被认为是w_c的一个翻译词。

2.3.2 翻译词聚类

接下来，我们对目标语言中每个词的翻译词集合进行聚类，使得不同簇表达该词的不同含义。在聚类过程中，为了对词与词之间的距离进行度量，我们首先需要将每一个翻译词表示成一个实数特征向量。一种直观的方法是抽取该词在源语言中的上下文分布，并以此作为其向量表示[120]，然而这种方式的计

[1]在本文中，目标语言为中文，源语言为英文。

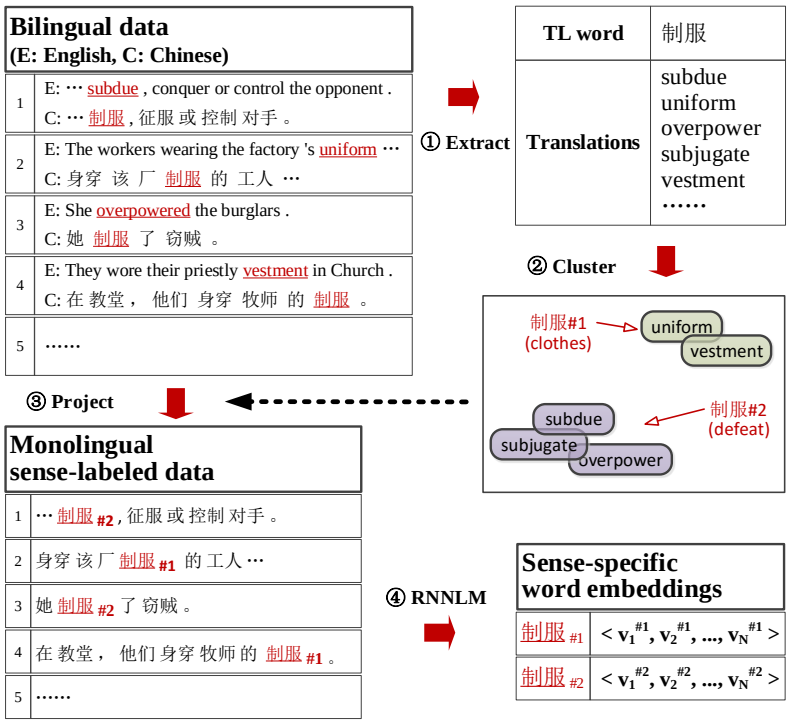


图 2-2 模型流程图。其中TL（Target Language）表示目标语言.

Figure 2-2 An illustration of the proposed method. TL stands for *target language*.

算代价太高，对于大规模语料的扩展性较差。因此，我们同样借鉴分布表示的思想，将源语言端词的低维分布表示向量作为其用于聚类的特征向量[2]。

词义聚类过程中的另一个关键问题是聚类数目的确定。不同词蕴含的词义数目不尽相同，比如“打”字常用的词义远比“制服”丰富。因此，常用的K均值聚类（k-means）以及K中心点聚类（k-medians）算法并不适用。在本文中，我们采用Frey和Dueck发表在《科学》杂志上的仿射传播算法（Affinity Propagation, AP）[121]对翻译词集合进行聚类。AP聚类的主要思想是通过在各个样本点之间不断传递信息（message passing）来选出各个簇的代表样本（examplar），以完成聚类。与k均值或k中心点聚类算法不同，AP聚类不需要人为设定聚类簇的数目，而是在迭代过程中根据样本点的分布来动态确定。

表2-1中给出了一些示例，其中第二列为翻译词抽取的结果，第三列为对翻译词进行AP聚类的结果。可以看出，聚类的结果很好地区分了目标词的常用词义。与K中心点聚类算法相似，在AP聚类的结果中，每个类别的中心（即examplar）为样本之一，即表中加粗的词。

[2]本文采用的是Collobert等在Wikipedia数据上训练得到的公开词向量（http://ml.nec-labs.com/senna/）。

表 2-1 翻译词抽取及聚类结果示意。第二列为对目标词的翻译词抽取结果（第 2.3.1节），第三列为AP聚类结果（第 2.3.2节）。

Table 2-1 Results of our approach on a sample of polysemous words. The second column lists the extracted translation words of the target language word (Section 2.3.1). The third column lists the clustering results using affinity propagation (Section 2.3.2).

TL Word	Translation Words	Translation Word Clusters
制服	investment, overpower, subdue, subjugate, uniform	#1: investment, **uniform** #2: **subdue**, subjugate, overpower
花	blossom, cost, flower, spend, take, took	#1: **flower**, blossom #2: take, cost, **spend**
法	act, code, France, French, law, method	#1: **France**, French #2: **law**, act, code #3: **method**
领导	lead, leader, leadership	#1: **leader**, leadership #2: **lead**

2.3.3 跨语言词义映射

接下来，我们将目标词在源语言端的词义聚类结果映射回目标语言，以完成词义归纳过程。对于目标词w的每一个词项$\hat{w}$，我们首先选择源语言端与之对齐概率最高的词作为其在当前语境下的翻译词，记为v。然后，我们计算v与每个词义聚类中心词之间的相似度，并选择相似度最高的类别作为$\hat{w}$的词义类别，并对$\hat{w}$进行标记。如果$\hat{w}$不与源语言端任何词对齐，那么我们将$\hat{w}$标记为w最常出现的词义类别。这样一来，我们为目标语言端所有词都附上了词义类别标签，从而获得了一个标注了词义的语料。我们进一步在该数据上使用循环神经网络语言模型进行训练，学习词义空间下的分布表示。

2.4 词义分布表示的应用

词义信息对于很多自然语言处理任务都非常有帮助。以中文命名实体识别为例，“美”在表达国家的概念时通常应标注为**地名**（Location），而在表达美丽的含义时，则不是一个命名实体。类似的，“华盛顿”既可以是人名实体（Person），也可为地名实体。因此，如果能够恰当地在NER模型中引入词义信息，将会对NER中的歧义消歧有所裨益。

我们考虑Turian等所采用的方法[66]，尝试将词义分布表示以半监督特征的

形式应用于NER任务。然而，对于词义分布表示而言，由于一个词可能包含多个分布表示向量且数目不定，其应用方式并不像词表示特征那么简单直接。为了使用NER中每个词的词义表示特征，我们需要首先对每个词进行词义消歧，然后才能使用对应词义下的向量表示。这里，我们将词义消歧看作是一个序列标注问题，并提出一种基于转移的词义消歧算法。在该算法中，我们使用循环神经网络语言模型来进行“转移动作”的预测。

我们知道，语言模型的基本思想是在给定历史上下文的条件下，预测当前词出现的概率：$p(w_t|w_{t-n+1},...,w_{t-1})$。自然的，我们可以根据此概率分布来预测最有可能的$w_t$，从而将语言模型应用于文本生成任务，比如生成式的机器翻译[26]。我们将这种思想用于确定当前词w_t最有可能的词义类别。记$w_t^{s_k}$为带词义标记的一个词项，其中s_k表示w_t的第k种词义，那么可以通过下式来预测w_t的词义类别s：

$$s = \arg\max_{s_k} p(w_t^{s_k}|w_{t-n+1}^{s_*},...,w_{t-1}^{s_*}) \tag{2-5}$$

注意，此时上下文$w_{t-n+1},...,w_{t-1}$均已被标记了词义，p为我们在上一节中学习词义分布表示的循环神经网络语言模型，可见，我们并不需要额外的训练代价。因此，最直接的解码方法是自左向右执行确定性地**贪心搜索**，每一步决策过程如图2-3所示。由于在预测$t+1$时刻词的词义时，词的原型是确定的（w_{t+1}），因此我们不需要在整个词表上进行概率归一化的计算，而只需要计算w_{t+1}在不同词义标记下的概率分布，并选择概率最高的词义作为预测的结果。算法 2-1描述了贪心解码的具体过程。

贪心搜索的一个缺点是在对某时刻的词进行预测时，只考虑上文的历史信息，而没有考虑下文信息。因此，对于句首词的预测通常是没有任何证据可参考的（只有句首符⟨s⟩能够提供极少量的上文信息）。同时，贪心搜索容易受到错误传播（error propagation）的影响。为了减少这些错误，我们在每一步决策时保留k个最可能的词义序列（top-k），并采用**柱搜索**（beam search）进行解码，以获得一个全局更优的词义序列。假设对于句子x，我们通过柱搜索得到的词义序列集合为$GEN(x)$，其中每个词义序列的分值由下式确定：

$$score(\mathbf{q}) = \sum_i \log p(\mathbf{q}_i|\mathbf{q}_{1:i-1}) \tag{2-6}$$

最优词义序列则为：

$$\mathbf{y} = \arg\max_{\mathbf{q}\in GEN(x)} score(\mathbf{q}) \tag{2-7}$$

算法 2-2描述了柱搜索解码的具体细节。 注意到在使用RNN的情况下，我

Input: Sentence $\mathbf{x} = \langle s \rangle, w_1, ..., w_n, \langle /s \rangle$
Word sense clusters: $\mathbb{SC}$
Recurrent neural network language model: `RNNLM`
Output: Word sense sequence: $\mathbf{y}$

for $i = 1..\texttt{len}(\mathbf{x})$ **do**
  $\mathcal{S} = get\_sense(\mathbb{SC}, w_i)$
  Select the best sense $s \in \mathcal{S}$ for w_i according to Equation 2-5
  $\mathbf{y}_i = s$
end
return y

算法 2-1 基于转移的词义消歧贪心解码。

Algo. 2-1 Transition-based greedy decoding for word sense disambiguation.

Input: Sentence $\mathbf{x} = \langle s \rangle, w_1, ..., w_n, \langle /s \rangle$
Word sense clusters: $\mathbb{SC}$
Recurrent neural network language model: `RNNLM`
Beam size: B
Output: Word sense sequence: $\mathbf{y}$

$agenda = []$; $GEN = []$
for $i = 1..\texttt{len}(\mathbf{x})$ **do**
  $\mathcal{S} = get\_sense(\mathbb{SC}, w_i)$
  for $\mathbf{q} \in GEN$ **do**
    for $s \in \mathcal{S}$ **do**
      $\texttt{insert}(agenda, \texttt{expand}(\mathbf{q}, w_i^s))$
    end
  end
  $GEN = \texttt{top\_B}(agenda, B)$ ▷ *ranking according to Equation 2-6*
  $agenda = []$
end
Determine $\mathbf{y}$ according to Equation 2-7 and **return**.

算法 2-2 基于转移的词义消歧柱搜索解码。

Algo. 2-2 Transition-based beam-search decoding for word sense disambiguation.

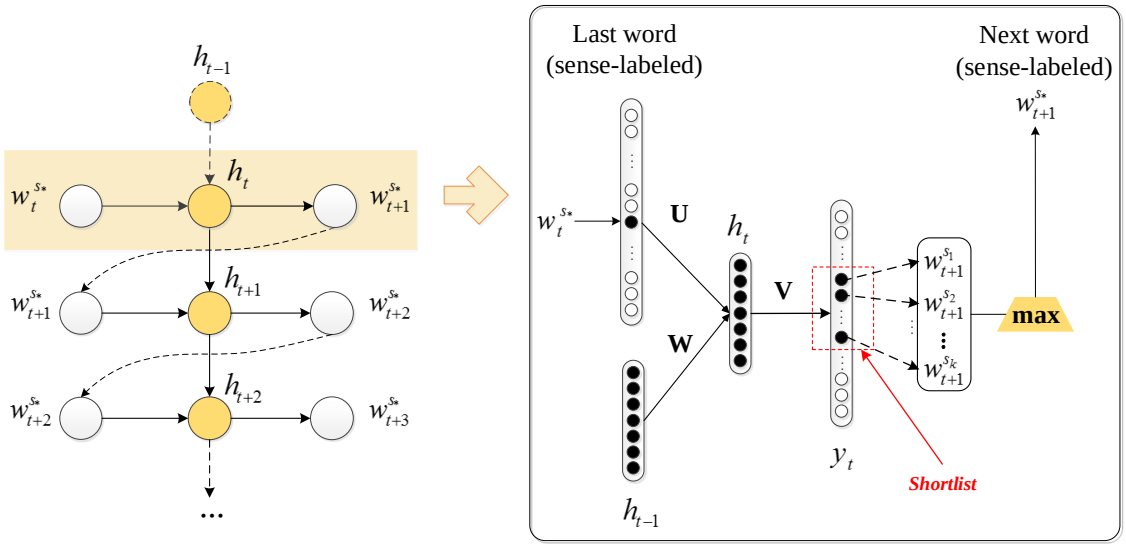


图 2-3 基于RNNLM的词义消歧（左）以及单步决策过程（右）。

Figure 2-3 Using RNNLM for WSD by sequential labeling (left). Decision at each step of the RNNLM-based WSD algorithm (right).

们的序列预测模型不满足马尔可夫性质，因此不存在多项式时间内的精确全局解码。这也是我们选择柱搜索解码而非维特比（Viterbi）解码的主要原因。

2.5 实验与分析

在这一节中，我们通过实验从词义相似度以及上层应用两个层面对词义分布表示进行评价。我们先介绍实验中所用的数据以及相关的实验设置，再介绍实验结果与分析。

2.5.1 实验设置

本实验中使用的双语平行语料库如表2-2所示。在过滤掉太长（≥ 40）或太短（≤ 10）的句子之后，我们获得918,681个双语句对（约21.7M个词）。我们使用BerkeleyAligner[119]进行词对齐[3]，并获得每个双语句对中双向词——词对齐概率，以及整体双语词表的翻译概率。我们使用scikit-learn[122]实现AP聚类算法[4]。需要注意的是，在AP聚类中仍然有一个可调的超参数称为“偏好”（preference）会对最终聚类的数目产生影响。较大的“偏好”值会鼓励算法产生更多的聚类

[3]code.google.com/p/berkeleyaligner/

[4]scikit-learn.org

数目。在本实验中，由于我们没有对于聚类数目倾向的先验或直觉，因此我们将AP算法的“偏好”值设定为样本相似度矩阵的中位数，从而使得聚类数目适中。我们使用Mikolov等开发的rnnlm工具[5]来训练循环神经网络语言模型并获得词/词义分布表示，分布表示的维度设为50。最后，我们一共获得了21.7K个词的分布表示，在完成词义归纳之后，8.4%的词存在多个词义类别，即多义词。

表 2-2 双语平行语料库数据统计信息。对于LDC04E12，我们采用1998年的一个子集。

Table 2-2 Statistics of the bilingual parallel datasets. For LDC04E12, the subset from 1998 is used.

Dataset	**Source**	**#Sent-pairs**
LDC03E14	news of FBIS	253,202
LDC04E12	United Nations	624,954
IWSLT08	IWSLT 2008 Evaluation campaign	39,947
PKU-863	Peking University	200,101

2.5.2 中文多义词相似度评测集以及评价结果

由于一个词可能具有多个分布表示向量，因此，对于两个词u和v，我们采用最大相似度（$MaxSim$）与平均相似度（$AvgSim$）两种度量标准[106]：

$$MaxSim(u,v) = \max_{1\le i\le k_u, 1\le j\le k_v} s(u^{s_i}, v^{s_j}) \tag{2-8}$$

$$AvgSim(u,v) = \frac{1}{k_u \times k_v} \sum_{i=1}^{k_u} \sum_{j=1}^{k_v} s(u^{s_i}, v^{s_j}) \tag{2-9}$$

其中k_u与k_v为u与v的词义类别数目。$s(\cdot,\cdot)$可以是任意一种相似性度量，这里我们使用的是$cosine$相似度。

在词汇相似性研究中，人们通常使用的数据集是WordSim-353（英文）[123]或者与其对应的中文版本[55]。这些数据集中几乎不含有多义词，因此不适用于本文的研究。而目前已发表的工作或者公开的数据集中，尚没有一个针对多义词词义的相似性评价集。为了填补这一空白，我们人工标注了一个中文多义词相似性数据集。以下是我们的具体标注过程。

我们借助HowNet（知网）[124, 125]来构建此数据集[6]。HowNet是一个以汉语和英语的词语所代表的概念为描述对象，以提示概念与概念之间以及概念所具

[5] www.fit.vutbr.cz/~imikolov/rnnlm/

[6] HowNet详细信息见：http://www.keenage.com

有的属性之间的关系为基本内容的常识知识库。对于汉语中的词汇，HowNet以“义原”为最小语义单位，记录了每个词语的每种义项。基于HowNet，我们的数据构建过程主要分为以下三个步骤：

1. 常用多义词抽取。我们首先依据HowNet所定义的每个词的不同义项来抽取多义词。然而，HowNet中所定义的词义粒度往往过细，对于非语言学研究人员而言不易区分。因此，我们通过人工观察对抽取出的多义词进行过滤，去掉了那些词义边界较为模糊的多义词。
2. 通过采样构成词对。对于抽取出来的每个多义词w，我们采样出数个其他的词与之构成词对。采样出的词可以分为两类：与w相关或者无关。相关词由人工进行采样，可以是该多义词某种词义的上位词（hypernym）、下位词（hyponym）、同级词（sibling）、同义或近义词（near-synonym）、反义词（antonym）或者只是主题相关的词。无关词则为随机采样。
3. 相似度标注。在收集好词对之后，我们请六位自然语言处理专业的研究生为每个词对进行相似度标记，参考WordSim-353，我们将相似度分值限制在区间$(0.0, 10.0)$。同时，为了保证标注的一致性，我们只保留了标注结果方差小于1.0的词对。最终我们一共获取了401个标注一致性较好的词对。与WordSim-353相比，我们的数据集词性更丰富，包含名词、名词、形容词等，而WordSim-353中则主要为名词。

表2-3中提供了一些标注示例：

表 2-3 多义词相似度评价数据示例。Mean.Sim为相似度标注的均值，Std.Dev为标准差。

Table 2-3 Sample word pairs of our dataset. Mean.Sim represents the mean similarity of the annotations, Std.Dev represents the standard deviation.

Word	Paired word	Category	Mean.Sim	Std.Dev
制服	征服$_{conquer}$	synonym	8.60	0.29
	重点$_{key\ point}$	unrelated	0.12	0.19
出	进$_{enter}$	autonym	7.90	0.97
	发表$_{publish}$	near-synonym	7.86	0.76
花	茎$_{plant\ stem}$	sibling	7.80	0.12
	费用$_{cost}$	topic-related	5.86	0.90
面	食物$_{food}$	hypernym	6.50	0.71

我们在该数据集上使用Spearman相关系数ρ以及Kendall相关系数τ进行评价，结果如表2-4所示。可以看出，词义分布表示对于多义词相似度的刻画能力显著

优于单向量表示。进一步地，我们也与多向量表示学习方法（Multi-prototype）进行了对比。借鉴Huang等人的做法[107]，对于一个词w，它在语料中的每个词项都由其上下文的平均词向量来表示。我们将上下文窗口设为10。接下来，对所有词项进行k平均聚类，我们在开发集上对k的值进行调优，最终设为2。

令我们感到惊讶的是，多向量表示没有取得预期的效果，其性能甚至稍低于单向量表示。这意味着为每个词学习相同数目的向量表示这种“一致假设”是不恰当的。

表 2-4 在多义词相似度数据集上的Spearman相关系数及Kendall相关系数评价结果。

Table 2-4 Spearman's ρ correlation and Kendall's τ correlation evaluated on the polysemous dataset.

System	MaxSim		AvgSim	
	$\rho \times 100$	$\tau \times 100$	$\rho \times 100$	$\tau \times 100$
Ours	**55.4**	**40.9**	49.3	35.2
SingleEmb	42.8	30.6	42.8	30.6
Multi-prototype	40.7	29.1	38.3	27.4

为了对所学到的词义分布表示有一个直观的印象，我们从评测集中选择了一些词，并根据词义分布表示计算它在各个词义下的近邻。结果如表2-5所示。

表 2-5 根据多义词的词义分布表示计算得到的在不同词义下的近邻。

Table 2-5 Nearest neighborhoods on a sample of polysemous words.

TL Word	Sense Cluster	Nearest Neighbours
制服	#1: uniform	穿着$_{dress}$, 警服$_{policeman\ uniform}$
	#2: subdue	打败$_{defeat}$, 击败$_{beat}$, 征服$_{conquer}$
花	#1: spend	花费$_{cost}$, 节省$_{save}$, 剩下$_{rest}$
	#2: flower	菜$_{greens}$, 叶$_{leaf}$, 果实$_{fruit}$
法	#1: law	法令$_{ordinance}$, 法案$_{bill}$, 法规$_{rule}$
	#2: method	概念$_{concept}$, 方案$_{scheme}$
	#3: French	德$_{Germany}$, 俄$_{Russia}$, 英$_{Britain}$
领导	#1: lead	监督$_{supervise}$, 决策$_{decision}$
	#2: leader	主管$_{chief}$, 上司$_{boss}$

2.5.3 中文命名实体识别上的实验结果

我们进一步在中文命名实体识别任务上对词义分布表示进行评价。命名

实体识别任务通常被建模为一个序列标注问题，即对序列中每个词进行标签预测。对应的标签可表示为“位置+类别”的形式，比如“B-PER”表示人名实体的开始，“I-ORG”则表示该词处于某机构名实体之内或者结束的位置。我们在人民日报1998年1月份与6月份的数据上进行实验。原始数据中标注了7种命名实体，分别为人名（Person），地名（Location），机构名（Organization），日期（Date），时间（Time），数字（Number）以及其他（Miscellany）。本实验中只研究其中最常用的三种实体类型：人名，地名和机构名。我们使用1月份的数据作为训练集（共计37,426个句子），6月份数据中的前2,000句作为开发集，后8,000句作为测试集。

为了在NER任务中使用词义分布表示特征，我们首先需要对NER数据进行词义消歧。我们分别使用2.4节中所介绍的两种消歧算法（贪心搜索与柱搜索）确定NER数据中每个词的词义类别。对于柱搜索算法，我们在开发集上对柱的宽度（beam size）进行调优。

本实验采用CRF模型进行序列标注，优化方法则使用L2正则的随机梯度下降（Stochastic Gradient Descent, SGD）。特征模板如表2-6所示。对于单向量词分布表示特征或者词义分布表示特征，我们均使用宽度为5的上下文窗口。

表 2-6 NER特征模板。其中Prefix/Suffix分别提取长度为l的前缀/后缀；Shape表示词的类型，如数字或字母；SingleEmb为单向量词分布表示；SenseEmb为词义分布表示。

Table 2-6 Features used in the Chinese NER system. Prefix and Suffix are the first and last l characters of a word. Shape indicates the shape of w_{i+k}, such as number or alphabet. SingleEmb represents the single word embeddings of w_{i+k}, SenseEmb represents the sense-specific embedding of w_{i+k} in the sense of s_{i+k}.

Baseline Features	00: $w_{i+k}, -2 \le k \le 2$
	01: $w_{i+k} \circ w_{i+k+1}, -2 \le k \le 1$
	02: $\texttt{Prefix}(w_{i+k}, l), -2 \le k \le 2, 1 \le l \le 4$
	03: $\texttt{Suffix}(w_{i+k}, l), -2 \le k \le 2, 1 \le l \le 4$
	04: $\texttt{Shape}(w_{i+k}), -2 \le k \le 2$
Embedding Features	$\texttt{SingleEmb}(w_{i+k}), -2 \le k \le 2$
	$\texttt{SenseEmb}(w_{i+k}, s_{i+k}), -2 \le k \le 2$

我们首先考察词义消歧过程中柱搜索宽度对于NER结果的影响。图2-4显示了在NER开发集上F值随着柱搜索宽度的变化。可以看出，当柱搜索宽度为4时，NER性能有显著的提升；当宽度继续增加时，NER性能的提升趋于平缓。同时，单向量词分布表示特征对于基准系统有显著的提升作用，这一点

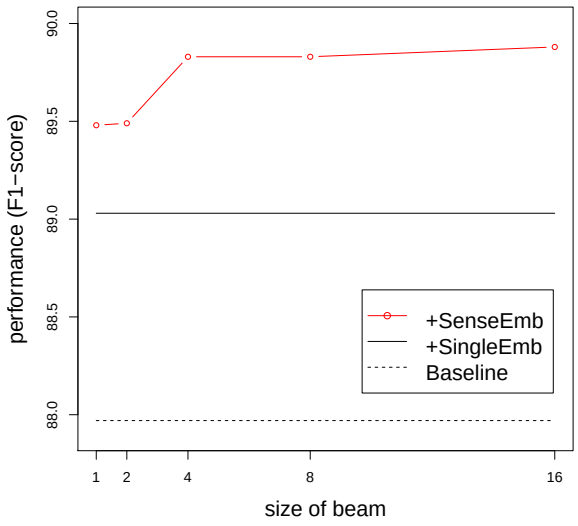


图 2-4 消歧算法中柱搜索宽度对于NER性能的影响。

Figure 2-4 Performance of NER w.r.t. the beam-size *B* of RNNLM-based WSD on development data.

与Turian等的实验发现一致[66]。而使用词义分布表示则取得了更大的提升。在测试集上的实验结果如表2-7所示，使用词义分布表示的模型相比单向量表示提升近1%（88.56 *vs.* 87.58）。使用`t-test`对两者预测结果进行显著性检验显示$p-value < 0.01$。

表 2-7 NER测试集上的实验结果。

Table 2-7 Performance of NER on test data.

System	P	R	F
Baseline	93.27	81.46	86.97
+SingleEmb	93.55	82.32	87.58
+SenseEmb (greedy)	93.38	83.56	88.20
+SenseEmb (beam search)	**93.59**	**84.05**	**88.56**

根据我们的假设，词义分布表示特征理应对于NER中多义词的预测准确性有所帮助。为了验证该假设，我们进一步评价NER测试结果中多义词的预测准确率。这里我们采用基于词项的准确率（per-token），而非基于实体。对于多义词的识别仍然使用HowNet。图2-5表明词义分布表示特征的应用的确对于多义词的预测有显著的提升作用，而单向量表示特征甚至对于多义词的预测准确率有所降低。同时，由于对上下文进行了消歧，词义分布表示特征对于单义词的预测也带来了更大的提升。

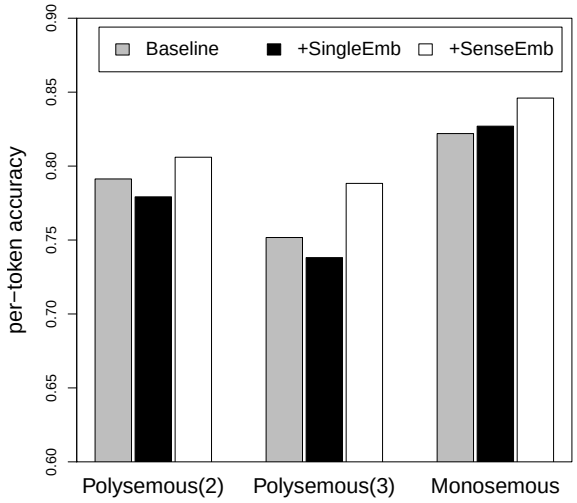


图 2-5 NER测试集中多义词及单义词的词项级预测准确率。Polysemous(k)表示HowNet中所定义的词义数大于等于k的词集。

Figure 2-5 Per-token accuracy on the polysemous and monosemous words in the NER test data. Polysemous(k) represents the set of words that have $\geq k$ senses defined in HowNet.

2.6 本章小结

本章针对自然语言处理中常见的“一词多义”现象提出一种基于双语资源的词义分布表示学习方法。与传统的单向量词表示相比，词义表示能够更好地捕捉多义词的不同含义。为了验证词义表示的有效性，我们构建了一套中文多义词相似度评价数据集，在该数据集上的评价结果显示，词义分布表示显著优于单向量词分布表示以及基于上下文聚类的多向量表示。

本章的另外一个重要的贡献是提出了一种基于循环神经网络语言模型的单语词义消歧算法，进而使得词义分布表示能够方便地应用于上层任务中。在中文命名实体识别中的实验结果显示，相比于单向量表示特征，词义分布表示特征能够带来更显著的性能提升，尤其是对于多义词的预测准确性。

在本章的工作中，为了与上层应用之间建立直接的联结，我们采用的是循环神经网络语言模型（RNNLM）来学习分布表示。然而RNNLM训练速度较慢，对于大规模数据的处理能力有限。因此，在未来的工作中，我们希望借助更高效的模型，如word2vec[39]，在更大规模数据上进行分布表示学习。与此同时，我们也需要设计新的效率更高的词义消歧算法。

第 3 章 基于分布表示的跨语言依存句法分析

3.1 引言

依存句法分析（dependency parsing）是自然语言处理中的核心任务之一，旨在根据依存文法对自然语言句子的内在语法结构进行解析，将输入句子从序列形式变为树状结构，从而可以捕捉到句子内部词语之间的远距离搭配或修饰关系[126]。绝大部分依存句法分析研究集中在资源丰富语言上，如英文、中文等。对于这些语言，人们已经标注了丰富的树库资源，可以方便地用来进行有监督学习。然而，在世界上现存的7,000多种语言中，绝大部分语言并不存在（或者存在极少量）可利用的标注树库。那么我们面临的一个关键问题就是：如何自动对这些语言的文本进行句法分析？考虑到句法树库的标注困难，人们开始探索无监督方法[1]、跨语言标注映射方法[5]以及模型迁移方法[98]来对资源稀缺语言进行句法分析。从目前的研究现状来看，无监督方法性能还远远没有达到基于跨语言迁移的句法分析性能。因此，本文主要关注模型迁移方法，目标是利用资源丰富的源语言树库资源来构建直接可用于资源稀缺语言的句法分析器。

目前，模型迁移方法主要面临两方面的挑战：1. 不同语言之间的“词汇化特征”鸿沟；2. 不同语言之间由于词序不同所带来的句法结构差异。

在句法分析模型中，词汇化特征（如词特征及组合特征）起到了很关键的作用，且对于弧上关系的判别尤其重要。而不同语言之间的词表通常存在较大的差异，导致词汇化特征无法进行跨语言迁移。为了规避这个问题，McDonald等人[98]采取了“去词汇化”（delexicalized）策略，只采用词性、弧上关系等非词汇化特征来学习跨语言模型。这种方式固然可行，但是由于损失了词汇化特征，使得句法分析性能较低，尤其是LAS（Labeled Attachment Score）值。Täckström等人[127]进一步提出使用跨语言词聚类特征来弥补词汇化特征的缺失。词聚类可以认为是一种粗粒度的词特征或者细粒度的词性特征，虽然在一定程度上对词汇化特征进行了补充，但是仍然损失了更细粒度的词汇化信息。

在本研究中，我们提出基于分布表示的跨语言依存句法分析框架，将不同语言中离散的“词汇化”特征映射至统一的分布表示空间，从而实现“词汇化”特征的跨语言迁移。具体的，我们的框架包含两个主要部分：

1. **基于神经网络的依存句法分析系统**。根据第1.2.3节的介绍，分布表示特征更适用于非线性模型[67]。Chen和Manning[71]提出的依存句法分析系统使用神经网络对每一步转移动作进行预测，非常显著地提升了在局部学习以及贪心解码条件下的基于转移的依存句法分析器性能。我们将在第3.3节中详细描述该模型。
2. **跨语言分布表示学习**。填补“词汇化特征”鸿沟的关键在于将不同语言的词汇化特征表示在同一向量空间之内，并使得语义相似的特征距离接近，也就是第1.2.2节中所介绍的跨语言特征分布表示学习。

此外，不同语言之间由于词序不同，导致很多依存结构迥异。例如，在某些语言中（如法语、西班牙语），形容词通常置于名词之后，从而产生了很多左指向的*amod*依存弧。假如我们采用英语作为唯一的源语言，那么目标语言中的这种依存结构很难被解析出来。针对这个问题，我们进一步提出基于多源语言迁移的方法，希望通过使用多种源语言来更多地覆盖在目标语言中可能出现的语言现象。多源语言也为现有的跨语言分布表示学习带来了新的挑战。因此，我们也针对多于两种语言的情形，提出了相应的多语言分布表示学习方法。

在多语言通用依存树库（Universal Dependency Treebank, UDT）[15]上的实验结果表明，在单源语言及多源语言情形下，我们的模型在“去词汇化”的基础之上都能够取得显著的性能提升。特别的，在多源语言情况下，我们取得了目前跨语言依存句法分析最好的模型迁移性能。最后，我们研究了如何利用少量目标语言标注数据（如50句）来进一步改进跨语言迁移模型。

3.2 背景与相关工作

3.2.1 依存句法分析

形式化地，对于一个输入词序列：$\mathbf{x} = [w_1, w_2, ..., w_n]$，依存句法分析的目标是构建一棵依存树：$\mathbf{d} = \{(h, m, l) : 0 \le h \le n; 0 < m \le n, l \in \mathcal{L}\}$，如图3-1所示。$(h, m, l)$表示由$w_h$指向$w_m$的一条有向依存弧，其中$w_h$表示核心结点，$w_m$表示修饰结点，$l$为依存关系的类别。

从解码算法的角度，目前主流的依存句法分析方法可以分为**基于图的方法**（Graph-based models，以下简称图方法）以及**基于转移的方法**（Transition-based models，以下简称转移方法）。图方法的主要思想是在由句子中所有词构成的

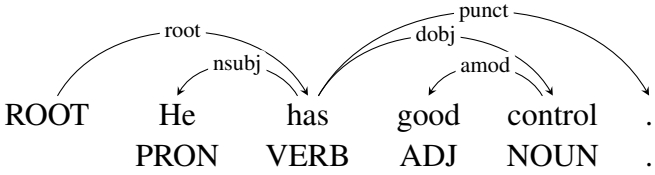


图 3-1 依存句法树示例。

Figure 3-1 An example labeled dependency tree.

有向完全图中寻找一棵最大生成树。为了计算依存树的分值，通常需要对其子结构引入特定的独立性假设，并利用动态规划算法进行精确解码。比如在一阶图模型（first-order graph-based model）[128]中，我们假设每条依存弧之间都是相互独立的，而高阶模型则考虑更复杂的子树结构[129–131]，比如二阶的兄弟（sibling）结构以及三阶的子孙（grandchild）结构。高阶模型对依存树刻画得更为准确，性能优于低阶模型，同时其解码空间也更大，时间复杂度更高。转移方法则是通过预测一个转移序列来增量式地产生依存弧，直到获得完整的依存树结构[132]。可能看出，转移方法的核心是学习一个转移动作的预测模型或分类器。转移模型通常采用非精确解码，如贪心或者柱搜索，因此会有一定的精度损失。然而，转移模型时间复杂度较低（线性），并且可以充分利用丰富的全局特征[133]，因此在近几年受到越来越多的关注。

从特征表示的角度，也可以将现有的依存句法分析模型分为基于符号表示的模型以及基于分布表示的模型两类。后者是近三年依存句法分析研究的重点，通常伴随着神经网络的应用。Chen和Manning首次成功地将两层神经网络应用于转移模型的学习，非常显著地提升了基于“贪心解码”以及“局部学习”的转移模型性能[71]。受该工作启发，研究者随后提出了各种不同的改进策略，典型的有基于长短时记忆网络（LSTM）来对转移状态进行建模[134]；在输出层使用感知机进行结构化学习[135]；以及全局柱搜索策略[136, 137]等。不失一般性，我们在本研究中将采用Chen和Manning的模型（简称C&M模型）。

3.2.2 跨语言依存句法分析

跨语言的依存句法分析主要有两种方法，一种是标注映射，也称数据迁移（data transfer）；另一种是模型迁移（model transfer）。

数据迁移方法的主要思想是通过双语平行数据，将源语言中自动标注的依存树结构映射至目标语言数据中，从而构建一个自动标注的、含噪声的目

标语言树库。进而，我们可以利用该树库训练一个目标语言的依存句法分析器。数据迁移方法最早由Yarowsky应用于词性标注、组块分析等词法分析任务[4]。Hwa等人[5]将其扩展至句法分析任务，并设计了一套句法结构映射的规则。Tiedemann进一步对映射规则进行了改进[94]。这种方法的主要缺点有两方面：1. 依赖双语平行数据；2. 跨语言标注映射的规则不容易制定，且受到词对齐错误的影响。当然，数据迁移方法的优点也较为明显，由于直接在目标语言树库上进行训练的，因此不受词序问题的影响。

模型迁移方法原则上不依赖双语平行数据，同时也不需要精心设计的句法映射规则。但是，由于模型是在源语言端进行训练的，因此在一定程度上受到词序不一致问题的影响。另一方面，在模型迁移方法中，词汇化特征难以有效地利用。因此，如何缓解词序不一致问题的影响? 如何更有效地利用词汇化特征? 这两个问题成为本研究的主要出发点。

3.3 基于神经网络的依存句法分析

根据第3.2.1的介绍，在基于转移的依存句法分析方法中，我们从转移系统的初始状态出发，通过一系列的转移动作达到一个终止状态。转移系统的每个状态（也称configuration）通常可以表示为一个三元组$\langle S, B, A\rangle$，其中S为一个栈，用以存储当前已经处理过的词以及部分生成的子树结构，B为一个缓存，按顺序保存着尚未处理过的词序列；A为当前状态下已经生成的依存弧集合。对于一个输入句子$\mathbf{x} = [x_1, x_2, ..., x_n]$，初始转移状态可以表示为$c_0 = \langle [w_0]_S, [w_1, w_2, ..., w_n]_B, \emptyset\rangle$，终止状态为$c_t = \langle [w_0]_S, []_B, A\rangle$。其中$w_0$是我们人为增加的一个伪词，表示整棵依存树的根结点。我们记S_i $(i = 0, 1, ...)$为栈中第i个元素（$i = 0$表示栈顶），B_i $(i = 0, 1, ...)$为缓存中第i个元素。

常用的转移算法有面向投射树分析（projective parsing）的*arc-standard*与*arc-eager*算法[132, 138]以及面向非投射树分析的*list-based*[139]与*swap-based*算法[140]等。不同的转移算法定义的转移动作也有所不同。以*arc-standard*算法为例，该算法定义了三类转移动作，分别是：左弧归约（Left-Arc(r)），右弧归约（Right-Arc(r)）以及移进（Shfit）。其中r为依存弧上关系。

- Left-Arc(r)：产生一条由S_0指向S_1的依存弧（$S_1 \overset{r}{\leftarrow} S_0$）（$S_0$为核心结点，$S_1$为修饰结点），从栈中删除$S_1$。
- Right-Arc(r)：产生一条由S_1指向S_0的依存弧（$S_1 \overset{r}{\rightarrow} S_0$）（$S_1$为核心结点，$S_0$为修饰结点），从栈中删除$S_0$。

- SHIFT：将缓存顶部元素B_0移入栈顶，前提条件是B不为空。

对基于贪心搜索的*arc-standard*算法而言，通常的做法是构建一个多元分类器（如支持向量机），根据当前时刻转移状态下抽取出来的特征向量来预测最有可能的转移动作。然而，传统的“特征工程”方法容易受三方面因素的影响：1. 数据稀疏；2. 特征覆盖不完全；3. 特征抽取的时间代价较高[71]。针对这三个问题，基于分布表示及神经网络的模型提供了一个有效的解决方案。

我们使用的基于神经网络的依存分析模型结构如图3-2所示。

与基于符号表示的句法分析模型不同，我们使用特征的分布表示。具体的，在Chen与Manning的模型中，主要抽取三类基本特征，分别是词特征、词性特征以及依存弧特征。本研究中，我们对其进行了扩展，并增加了两类非局部特征：距离特征（Distance）以及配价特征（Valency）[133]。距离特征表示两个元素之间的距离，配价特征则表示特定元素儿子结点（或修饰结点）的数目。接下来，我们根据特征值所在的区间对这两类特征进行离散化，然后将所有的特征都映射为相应的分布表示向量。完整的特征模板如表3-1所示：

所有特征的分布表示向量在隐含层通过Cube激活函数进行组合：

$$\boldsymbol{h} = g(\boldsymbol{x}) = (\boldsymbol{W}_1 \cdot [\boldsymbol{x}^w \oplus \boldsymbol{x}^t \oplus \boldsymbol{x}^r \oplus \boldsymbol{x}^d \oplus \boldsymbol{x}^v] + \boldsymbol{b}_1)^3 \tag{3-1}$$

其中，$\boldsymbol{W}_1$为隐含层权值矩阵，$\boldsymbol{b}_1$为偏置向量。

特征组合对于大部分自然语言处理任务而言都至关重要。可以看出，Cube激活函数$g(x) = x^3$实际上是低秩张量分解（low-rank tensor）的一种特殊情况。为了验证两者之间的联系，我们将$g(x)$扩展为：

$$\begin{aligned} &g(w_1x_1 + ... + w_mx_m + b) = \\ &\sum_{i,j,k}(w_iw_jw_k)x_ix_jx_k + \sum_{i,j} b(w_iw_j)x_ix_j + ... \end{aligned} \tag{3-2}$$

假如将偏置项b视为$b \times x_0$，其中$x_0 = 1$。那么对于每种特征组合$x_ix_jx_k$，其权重则为$w_iw_jw_k$。容易看出，$g(x)$等价于CP张量分解结果中的一个秩一张量（rank-1 tensor）。我们知道，低秩张量分解隐式地捕捉了各个维度特征的全组合（full combination）[141]，而Cube激活函数的主要作用也体现在这里。当然，多项式激活函数在误差反向传播的过程中容易产生梯度爆炸的问题，因此在实际应用中，需要对梯度进行限幅处理（clipping）。

接下来，经由特征组合的隐含层分布表示传递至输出层，通过softmax函数计算转移动作集合下的概率分布：$\boldsymbol{y} = \mathtt{softmax}(\boldsymbol{W}_2 \cdot \boldsymbol{h})$。我们使用交叉熵损失函数来训练模型：

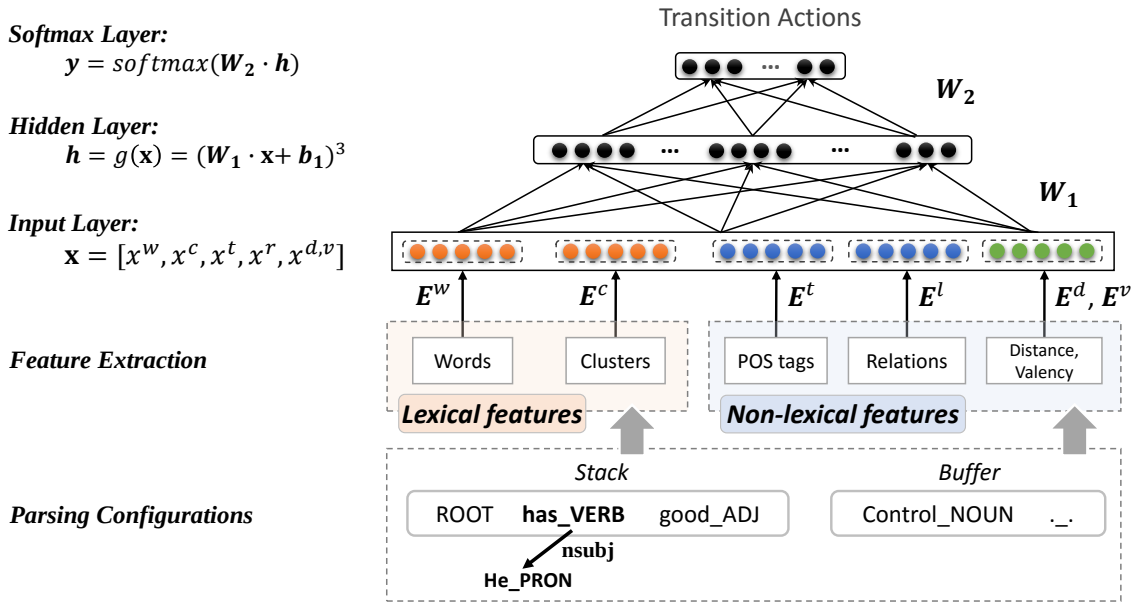


图 3-2 基于神经网络的依存句法分析模型结构。

Figure 3-2 Neural network model for dependency parsing.

$$\mathcal{J}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=0}^{N} \texttt{CrossEntropy}(\boldsymbol{d}_i, \boldsymbol{y}_i) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \tag{3-3}$$

其中$\texttt{CrossEntropy}(\boldsymbol{p}, \boldsymbol{q})$为概率分布$\boldsymbol{p}$与$\boldsymbol{q}$之间的交叉熵：

$$\texttt{CrossEntropy}(\boldsymbol{p}, \boldsymbol{q}) = \sum_{k} \boldsymbol{p}_k \ln \boldsymbol{q}_k \tag{3-4}$$

在我们的模型中，$\boldsymbol{\theta}$将包含所有的特征分布表示（词、词性等）以及网络权值矩阵。需要注意的是，在某些情况下（如第3.5.4节中所描述的弱监督实验），我们不更新词的分布表示向量（E^w），因而不需要计算E^w的梯度。

3.4 跨语言词汇表示学习

在本研究中，我们首先探讨较为简单的双语词汇分布表示学习方法，接着针对多源语言的情形提出相应的多语词汇分布表示学习模型。最后，我们介绍多语言词聚类表示的学习方法。

3.4.1 双语词汇分布表示学习

我们提出两种方法来学习双语词汇分布表示，分别是基于词对齐的跨语言映射以及典型关联分析。这两种方法都可以归类为线下处理方法（见

表 3-1 模型中所使用的特征模板。其中$E_p^{\{w,c,t,r,d,lv,rv\}}$表示位置$p$元素的不同特征分布表示向量，$lc1/rc1$表示左侧/右侧的第一个子结点，$lc2/rc2$表示左侧/右侧的第二个子结点。†词汇化特征，‡非词汇化特征。

Table 3-1 Feature templates of the neural network model for transition-based dependency parsing. $E_p^{\{w,c,t,r,d,lv,rv\}}$ indicates various feature embeddings of the element at position p. $lc1$ ($rc1$) is the first child to the left (right) and $lc2$ ($rc2$) is the second child to the left (right). † indicates the lexical features, ‡ indicates the non-lexical features.

Type	Feature Templates
Word†	$E^w_{S_i}, E^w_{B_i}, i = 0, 1, 2$
	$E^w_{lc1(S_i)}, E^w_{rc1(S_i)}, E^w_{lc2(S_i)}, E^w_{rc2(S_i)}, i = 0, 1$
	$E^w_{lc1(lc1(S_i))}, E^w_{rc1(rc1(S_i))}, i = 0, 1$
POS‡	$E^t_{S_i}, E^t_{B_i}, i = 0, 1, 2$
	$E^t_{lc1(S_i)}, E^t_{rc1(S_i)}, E^t_{lc2(S_i)}, E^t_{rc2(S_i)}, i = 0, 1$
	$E^t_{lc1(lc1(S_i))}, E^t_{rc1(rc1(S_i))}, i = 0, 1$
Relation‡	$E^r_{lc1(S_i)}, E^r_{rc1(S_i)}, E^r_{lc2(S_i)}, E^r_{rc2(S_i)}, i = 0, 1$
	$E^r_{lc1(lc1(S_i))}, E^r_{rc1(rc1(S_i))}, i = 0, 1$
Distance‡	$E^d_{\langle S_0,S_1\rangle}, E^d_{\langle S_0,B_0\rangle}$
Valency‡	$E^{lv}_{S_0}, E^{lv}_{S_1}, E^{rv}_{S_1}$

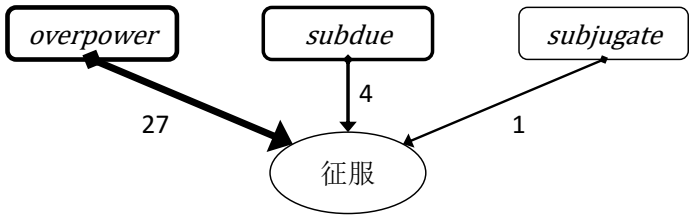
第1.2.2节）。

（1）基于词对齐的跨语言鲁棒映射

直觉上，如果两个词互为翻译，那么它们的分布表示向量应该在欧氏空间内接近。在实际情况中，一个词往往对应着多个翻译词，比如中文的“征服”在特定语料中会以一定概率分布翻译为“overpower”，“subdue”，“subjugate”等，如图3-3所示，图中的数字表示在双语数据中统计得到的对齐次数。因此，我们对源语言中译词的表示向量根据对齐次数（频率）进行加权求和，并以此作为目标词的分布表示向量。

形式化地，给定一个双语平行语料$\mathbb{D}$。首先我们对其进行无监督的词对齐，并根据对齐概率进行过滤，以选择可信度较高的对齐结果（阈值$\delta = 0.95$）。由此可以统计得到一个双语对齐词典，记为：$\mathcal{A}^{T|S} = \{(w_i^T, w_j^S, c_{i,j}), i = 1, 2, ..., N^T; j = 1, 2, ..., N^S\}$。其中$c_{i,j}$表示目标语言词$w_i^T$与源语言词$w_j^S$在$\mathbb{D}$中的对齐次数；$N_T$与$N_S$为词表大小。我们使用简略记号$(i, j) \in \mathcal{A}^{T|S}$表示$\mathcal{A}^{T|S}$中的一个双语词对，那么，跨语言分布表示映射方法可以写成以下形式：

$$\boldsymbol{e}(w_i^T) = \sum_{(i,j)\in\mathcal{A}^{T|S}} \frac{c_{i,j}}{c_{i,\cdot}} \cdot \boldsymbol{e}(w_j^S) \tag{3-5}$$



$$\boldsymbol{e}(\text{征服}) = 0.84 \cdot \boldsymbol{e}(overpower) + 0.13 \cdot \boldsymbol{e}(subdue) + 0.03 \cdot \boldsymbol{e}(subjugate)$$

图 3-3 跨语言分布表示映射示例。

Figure 3-3 Example of cross-lingual projection of word embeddings.

其中$c_{i,\cdot} = \Sigma_j c_{i,j}$。

跨语言映射的方法可以为双语词典$\mathcal{A}^{T|S}$中的目标词赋予分布表示，然而，注意到双语平行数据（$\mathbb{D}$）规模通常较为有限，可覆盖的词表也相对较小。为了增加跨语言映射方法的健壮性，我们采用基于词形学（morphology）的机制，引入了一个额外的单语分布表示传播的过程。在很多语言中，词形相近的词意义上也会有一定程度的重叠，比如英文中的时态变化。因此，我们可以利用词形上的相似性，获得词典$\mathcal{A}^{T|S}$中未出现过的目标语言词（也称未登录词，OOV）的分布表示。具体的，对于目标语言中的每个未登录词w_{oov}^T，我们根据编辑距离（Edit distance，也称Levenshtein distance）在双语词典中抽取出一系列与之相似的词，记为C，那么w_{oov}^T的分布表示向量则为C中所有词分布表示向量的平均：

$$\begin{gathered}\boldsymbol{e}(w_{oov}^T) = \underset{w' \in C}{\texttt{Avg}}(\boldsymbol{e}(w')) \\ \text{其中：}\ C = \{w | \texttt{EditDist}(w_{oov}^T, w) \le \tau\}\end{gathered} \tag{3-6}$$

为了减少单语分布表示传播的噪声，我们将编辑距离的阈值τ设为1。

上述过程可以概括为跨语言映射（cross-lingual projection）以及单语传播（monolingual propagation）两个主要过程。我们采用**矩阵运算**进一步对其形式化。

1. **跨语言映射**。记$\boldsymbol{A}^{T|S} \in \mathbb{R}_{|N^T| \times |N^S|}$为双语词典$\mathcal{A}^{T|S}$所对应的词对齐矩阵。$\boldsymbol{A}^{T|S}$中的元素为对齐频次的归一化值：

$$\boldsymbol{A}^{T|S}(i, j) = \frac{c_{i,j}}{\Sigma_j c_{i,j}} \tag{3-7}$$

假设源语言词汇分布表示矩阵为$\boldsymbol{E}^S$，则目标语言的分布表示矩阵可由下式进行计算：

$$\boldsymbol{E}_{in}^T = \boldsymbol{A}^{T|S} \cdot \boldsymbol{E}^S \tag{3-8}$$

2. **单语传播**。接下来，我们执行单语传播过程，即根据$\boldsymbol{E}_{in}^{T}$来计算未登录词分布表示矩阵$\boldsymbol{E}_{oov}^{T}$。记$\boldsymbol{M}^{T} \in \mathbb{R}_{|N_{oov}^{T}|\times|N_{in}^{T}|}$为单语相似度矩阵，这里我们使用编辑距离作为相似度的度量，因此：

$$\boldsymbol{M}^{T}(i,j)=\begin{cases}1, & if\ \texttt{EditDist}(w_{oov(i)}^{T}, w_{in(j)}^{T}) \le \tau \\ 0, & if\ \texttt{EditDist}(w_{oov(i)}^{T}, w_{in(j)}^{T}) > \tau\end{cases} \tag{3-9}$$

其中$\tau = 1$。在对$\boldsymbol{M}^{T}$按行进行归一化处理之后，未登录词的分布表示矩阵可由下式进行计算：

$$\boldsymbol{E}_{oov}^{T} = \boldsymbol{M}^{T} \cdot \boldsymbol{E}_{in}^{T} \tag{3-10}$$

（2）典型关联分析

根据第1.2.2节中的介绍，我们知道，CCA是一种度量两个多维变量之间相关性的统计分析方法。对于两个多维变量，CCA的结果是两个映射矩阵，将原始变量分别映射至新的子空间，使得两者相关性最大。这里，我们介绍如何将CCA应用于双语词汇分布表示学习。首先，我们从双语对齐词典$\mathcal{A}^{T|S}$中抽取出一个“一对一”的词典$\mathcal{D}: \Sigma \leftrightarrow \Omega$。其中$\Sigma \subseteq \boldsymbol{E}^{T}$中的每个词在$\Omega \subseteq \boldsymbol{E}^{S}$中只有一个译词，反之亦然。

基于CCA的双语词汇分布表示学习过程如图3-4所示。假设映射之后

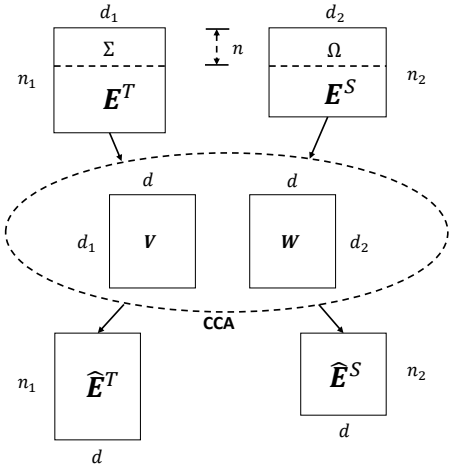


图 3-4 基于CCA的双语词汇分布表示学习过程。

Figure 3-4 Bilingual word representation learning using CCA.

的子空间维度为$d \le min(d_1, d_2)$，首先，我们使用CCA获得两个映射矩阵$\boldsymbol{V} \in \mathbb{R}^{d_1\times d}, \boldsymbol{W} \in \mathbb{R}^{d_2\times d}$：

$$\boldsymbol{V}, \boldsymbol{W} = \texttt{CCA}(\Sigma, \Omega) \tag{3-11}$$

接着，利用V与W对完整的词表进行映射：

$$\hat{\boldsymbol{E}}^{T} = \boldsymbol{E}^{T} \cdot \boldsymbol{V}; \quad \hat{\boldsymbol{E}}^{S} = \boldsymbol{E}^{S} \cdot \boldsymbol{W} \tag{3-12}$$

$\hat{\boldsymbol{E}}^{T}$与$\hat{\boldsymbol{E}}^{S}$则为最后得到的双语词汇分布表示矩阵。

3.4.2 多语词汇分布表示学习

为了处理多于两种语言的情形，我们进一步提出两种能够有效地用于多语分布表示学习的方法。

（1）多语言鲁棒映射

基于词对齐的跨语言鲁棒映射方法可以很自然地扩展至多语言的情况。如图3-5所示，我们首先一种语言（如英文）作为公共的源语言，并学习该语言上的单语词汇分布表示。接下来，我们利用“跨语言映射+单语传播”的方法将该语言的分布表示矩阵映射至其他的语言中。每种语言的映射过程相互独立。

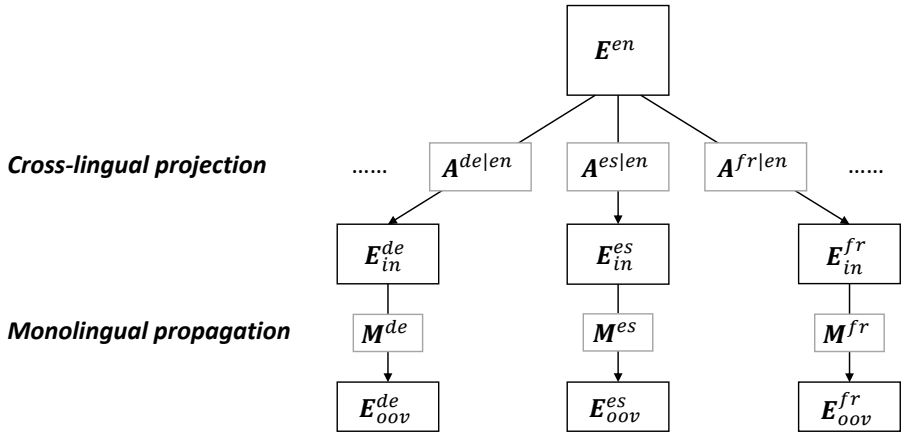


图 3-5 多语鲁棒映射方法示意图。

Figure 3-5 Illustration of the multilingual robust projection approach.

（2）多语言Skip-gram模型

根据第1.2.1节的介绍，我们知道，Skip-gram模型的主要基本假设是分布语义假设（distributional hypothesis），即：对于语义相似的词，与其共现的上下文分布也是相似的。因此，Skip-gram通过对上下文的建模来学习词语的分布表示。具体的，对于每一个由词和上下文构成的词对$\langle w, c\rangle$，Skip-gram模型采用对数线性函数来预测两者共现的概率：

$$P(w|c;\theta) = \frac{\exp(\boldsymbol{e}'(w)^{\top}\boldsymbol{e}(c))}{\sum_{w'\in\mathbb{V}}\exp(\boldsymbol{e}'(w')^{\top}\boldsymbol{e}(c))} \tag{3-13}$$

并通过优化似然函数：$J(\theta) = \sum_{(w,c)\in\mathbb{D}} \log p(c|w;\theta)$来学习模型参数。

在这里，我们将单语条件下的分布语义假设扩展至多语言的情况，并提出

跨语言分布语义假设（Cross-lingual Distributional Hypothesis）：对于语义相似的词，其跨语言上下文的分布也是相似的。从单语分布语义假设到多语分布语义假设，其中最为关键的桥梁是**词对齐**——在双语平行数据中，相互对齐的两个词通常互为翻译（语义一致）。当然，由于自动词对齐的错误所引入的噪声难以避免。根据这一假设，我们提出多语言Skip-gram模型，不仅仅对单语上下文进行预测，同时也在自动词对齐的基础上引入跨语言上下文的预测。

考虑到英文资源的广泛性，我们依然采用英文来联结其他的不同语言。以英语（EN）、法语（FR）、西班牙语（SP）为例进行说明，假设我们分别有英语到法语和西班牙语的双语平行数据。首先，我们进行自动词对齐。如图3-6所示，⟨**accepter**, **accept**⟩与⟨**accept**, **acceptan**⟩是两个相互对齐的词对。在多语言Skip-gram模型中，我们考虑多语上下文的预测，因此训练数据为：

$$\mathbb{D} = \underbrace{\mathbb{D}_{EN\leftrightarrow EN} \cup \mathbb{D}_{FR\leftrightarrow FR} \cup \mathbb{D}_{ES\leftrightarrow ES}}_{\text{单语}} \cup \underbrace{\mathbb{D}_{EN\leftrightarrow FR} \cup \mathbb{D}_{EN\leftrightarrow ES}}_{\text{跨语言}} \tag{3-14}$$

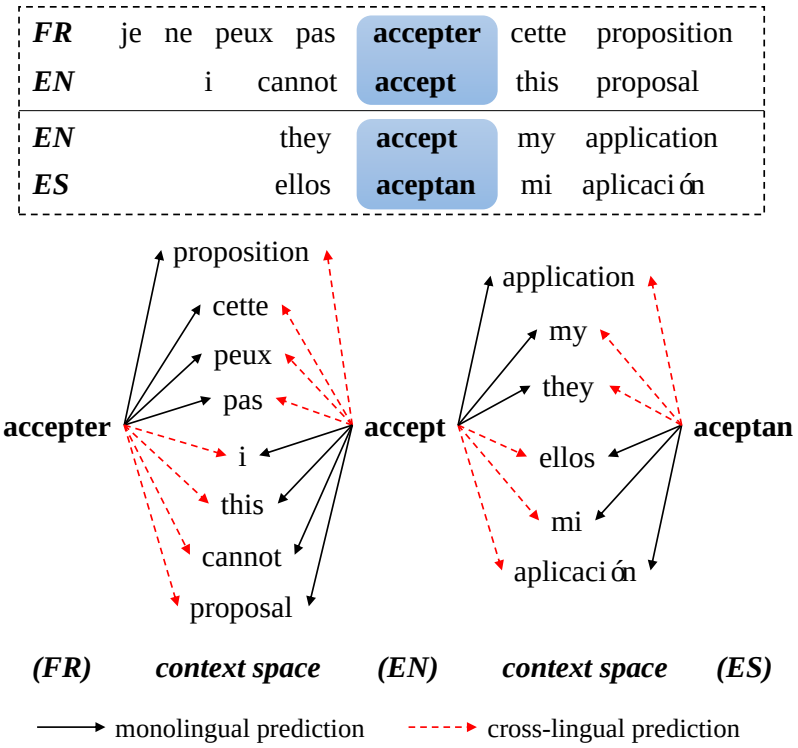


图 3-6 多语言Skip-gram模型。

Figure 3-6 Illustration of the multilingual skip-gram model.

考虑语言集合$\mathcal{L}$，多语言skip-gram模型优化如下目标函数：

$$J = \alpha \sum_{l\in\mathcal{L}} J_{mono_l} + \beta \sum_{l\in\mathcal{L}\setminus\{en\}} J_{bi_{l,en}} \tag{3-15}$$

其中：

$$J_{mono_l} = \sum_{(w,c)\in D_{l\leftrightarrow l}} \log p(c|w;\theta) \tag{3-16}$$

$$J_{bi_{l,en}} = \sum_{(w,c)\in D_{en\leftrightarrow l}} \log p(c|w;\theta) \tag{3-17}$$

在实际应用中，α与β可以根据具体任务进行调优。我们采用负采样（negative sampling）算法训练模型。需要注意的是，J_{mono_l}不仅可以从双语平行数据中进行计算，也可以利用额外的单语资源。这也是该模型的主要优势之一。

3.4.3 多语言词聚类表示学习

词聚类特征也是一种重要的半监督特征，在基于符号表示的模型中通常与其他的离散特征一起使用，能够很好地缓解数据稀疏问题。Brown聚类[61]是自然语言处理中常用的词聚类算法，它是一种层次聚类，将每个词表示成一个“01”序列，即层次化二叉树中的一条路径。

Täckström等提出双语词聚类特征并应用于跨语言依存句法分析，取得了显著的效果[127]。我们将该方法扩展至多语情形。与基于词对齐的跨语言鲁棒映射法类似，我们将英语作为种子语言，首先学习英语上的Brown词聚类，再通过每种语言对的词对齐信息映射至其他语言。具体地，对于其他语言的某个目标词，它的聚类可由下式确定：

$$c(w_i^T) = \arg\max_k \sum_{(i,j)\in\mathcal{A}^{T|S}} c_{i,j} \cdot \mathbb{1}[c(w_j^S) = k] \tag{3-18}$$

注意到，这种方法依然无法覆盖双语词典之外的未登录词。因此，我们同样采用单语传播的方法来获得未登录词的聚类结果：

$$\begin{aligned} c(w_{oov}^T) &= \arg\max_k \sum_{w'\in C} \mathbb{1}[c(w') = k] \\ \text{其中：}\ C &= \{w|\texttt{EditDist}(w_{oov}^T, w) \le \tau\} \end{aligned} \tag{3-19}$$

其中：$\tau = 1$。

3.5 实验与分析

3.5.1 实验设置

在本研究中，我们使用多语言通用依存树库（Universal Dependency Tree-

banks v2.0）[15]中印欧语系（Indo-European）语言的树库，并按照标准的“训练/开发/测试”划分进行实验。多语言通用依存分析是近几年依存句法分析研究的一大研究趋势，其中定义了通用的多语言词性（12种）以及依存句法标注规范（含40种依存关系），非常适合用于资源稀缺语言分析的研究。

我们考虑两种迁移句法分析场景，分别是单源语言迁移以及多源语言迁移。两种情况采用相同的神经网络依存句法分析设置，其中隐含层神经元数目为400，各类特征分布表示向量的维度如表3-2所示。我们使用小批量（mini-batch）自适应随机梯度下降算法（Adaptive Stochastic Gradient Descent, AdaGrad）对模型进行优化[142]。每个mini-batch大小为10,000。

表 3-2 不同特征表示向量维度。

Table 3-2 Dimensions of various types of feature embeddings.

	Word	POS	Label	Distance	Valency	Cluster
Dimension	50	50	50	5	5	8

接下来，我们分别对单源语言迁移以及多源语言迁移实验及结果进行详细的介绍。

3.5.2 单源语言迁移学习实验

在单源语言迁移实验中，我们将英语（EN）作为源语言，德语（DE）、法语（FR）、西班牙语（ES）作为目标语言。在单语词表示的预训练过程中，对于英语、德语与西班牙语，我们使用WMT-2011单语新闻语料[1]；对于法语，使用WMT-2011以及WMT-2012[2]的单语新闻语料。双语平行数据则包括WMT2006-10双语新闻评论语料以及Europarl语料[3]。我们使用cdec[143]中的fast-align工具对进行自动双语词对齐，并统计词对齐矩阵。

本实验中作为对比的基准系统包括：

- 非词汇化迁移系统（DELEX）。该系统同样基于前面所介绍的神经网络依存句法分析器，但是只使用非词汇化特征（见表 3-1），包括Distance以及Valency特征。
- McDonald等人提出的基于离散特征表示的非词汇化迁移系统[15]（McD13）。该系统使用的是基于感知机的依存句法分析，并使用柱

[1]http://www.statmt.org/wmt11/

[2]http://www.statmt.org/wmt12/

[3]http://www.statmt.org/europarl/

搜索进行训练/预测（柱搜索宽度为8）。同时，我们使用ZPAR[144]重现了该方法，重现系统记为McD13*。

- Täckström等人提出的基于跨语言词聚类特征的迁移系统[127]，该系统实际上是在McD13的基础之上增加了跨语言词聚类特征，因此我们记为McD13*+cls。

由于目标语言不存在标注数据，因此，在学习每个迁移系统时，我们都使用英语的开发集来控制训练进程（early-stopping）。实验结果如表3-3所示。同时，我们也在表 3-4中进一步概括了每种实验设置下取得的性能提升。

由于我们的DELEX系统使用的是贪心解码以及局部学习，因此性能相比基于柱搜索解码及全局学习的McD13系统略低。我们基于Zpar实现的McD13*与McD性能相当。首先，通过PROJ，CCA与DELEX系统的对比可以看出，两种双语词汇分布表示特征均能够显著提升迁移系统的性能。有趣的是，PROJ始终优于CCA，且与McD13*+cls性能相当。我们后续将对此展开进一步的分析。

另外，我们也发现，引入跨语言词聚类特征之后，PROJ与CCA迁移系统的性能均取得了进一步的显著提升。与非词汇化基准系统相比，UAS的相对错误率在最好情况下降低了13.1%，LAS为12.6%。

表 3-3 单源语言条件下迁移句法分析结果，采用UAS与LAS进行评价。由于模型在英语上的性能随着目标语言的不同而变化，因此使用†表示模型的平均性能。

Table 3-3 Cross-lingual transfer dependency parsing from English on the test dataset of 4 universal multilingual treebanks. Results measured by unlabeled attachment score (UAS) and labeled attachment score (LAS). † indicates the averaged UAS/LAS in EN since the model varies for different target languages in the CCA-based approach.

	Unlabeled Attachment Score (UAS)					Labeled Attachment Score (LAS)				
	EN	DE	ES	FR	AVG	EN	DE	ES	FR	AVG
DELEX	83.67	57.01	68.05	68.85	*64.64*	79.42	47.12	56.99	57.78	*53.96*
PROJ	91.96	60.07	71.42	71.36	*67.62*	90.48	49.94	61.76	61.55	*57.75*
PROJ+cls	92.33	60.35	**71.90**	**72.93**	***68.39***	90.91	**51.54**	**62.28**	**63.12**	***58.98***
CCA	90.62†	59.42	68.87	69.58	*65.96*	88.88†	49.32	59.65	59.50	*56.16*
CCA+cls	92.03†	**60.66**	71.33	70.87	*67.62*	90.49†	51.29	61.69	61.50	*58.16*
McD13	83.33	58.50	68.07	70.14	*65.57*	78.54	48.11	56.86	58.20	*54.39*
McD13*	84.44	57.30	68.15	69.91	*65.12*	80.30	47.34	57.12	58.80	*54.42*
McD13*+cls	90.21	60.55	70.43	72.01	*67.66*	88.28	50.20	60.96	61.96	*57.71*

表 3-4 不同系统之间的对比。所有的提升均是统计显著的（MaltEval）。

Table 3-4 Summary of each of the experimental gains detailed in Table 3-3, in both absolute LAS gain and relative error reduction. All gains are statistically significant using MaltEval [145] at $p < 0.01$.

Experimental Contribution			DE/ES/FR Avg. (Relative)
PROJ	*vs.*	DELEX	*+3.79 (8.2%)*
CCA	*vs.*	DELEX	*+2.19 (4.8%)*
PROJ	*vs.*	McD13*	*+3.33 (7.3%)*
CCA	*vs.*	McD13*	*+1.74 (3.8%)*
PROJ+cls	*vs.*	PROJ	*+1.23 (2.9%)*
CCA+cls	*vs.*	CCA	*+2.00 (4.6%)*
McD13*+cls	*vs.*	McD13*	*+3.29 (7.2%)*
PROJ+cls	*vs.*	DELEX	*+5.02 (10.9%)*
CCA+cls	*vs.*	DELEX	*+4.20 (9.1%)*
PROJ+cls	*vs.*	McD13*	*+4.46 (9.8%)*
CCA+cls	*vs.*	McD13*	*+3.74 (8.2%)*
PROJ+cls	*vs.*	McD13*+cls	*+1.27 (3.0%)*
CCA+cls	*vs.*	McD13*+cls	*+0.45 (1.1%)*

3.5.2.1 单语传播过程的影响

由于在PROJ以及跨语言词聚类特征中，我们都使用了基于编辑距离的单语传播来获取未登录词的表示。因此，我们以PROJ+cls系统为例，对该过程所带来的影响进行进一步的分析。

表 3-5显示了使用单语传播带来的性能提升，同时我们也观察了编辑距离阈值τ的变化（从1到3）对性能的影响。直觉上，跨语言分布表示对目标语言测试集中词语的覆盖率（coverage）对依存分析的性能有较大的影响。因此，我们在表 3-5中同时列出了覆盖率的变化。可以看出，在三种语言上，单语传播过程都能够带来较为明显的提升。尤其是在法语上提升最为显著（+1.76% UAS）。从覆盖率的变化也能够看出，单语传播在法语测试集上的覆盖率提升最大，这也验证了我们的假设。在德语以及西班牙语上，编辑距离为2时取得了最好的结果，但是继续增大编辑距离则会引入更多的噪声，从而导致性能反而有所下降。

3.5.2.2 词汇分布表示特征Fine-tuning过程的影响

PROJ优于CCA的另一个原因是可以在训练依存句法分析模型的过程中对源语言词表示特征进行调整（fine-tuning），在模型训练完成之后再执行跨语言映

表 3-5 单语传播过程的影响。

Table 3-5 Effect of robust projection. "Simple" indicates using cross-lingual projection only, whereas "Robust" includes the monolingual propagation procedure.

		Simple	Robust		
			τ=1	τ=2	τ=3
DE	coverage	91.37	94.70	96.50	97.47
	UAS	59.74	60.35	**60.53**	**60.53**
	LAS	50.84	51.54	**51.70**	51.69
ES	coverage	94.51	96.67	97.75	98.47
	UAS	70.97	71.90	**72.00**	71.93
	LAS	61.34	62.28	**62.34**	62.27
FR	coverage	90.83	97.60	98.33	98.58
	UAS	71.17	**72.93**	72.79	72.70
	LAS	61.72	**63.12**	63.02	62.94

射过程。因此，PROJ所得到的跨语言词汇分布表示可以认为是任务相关的，而在CCA中，由于线性映射的潜在假设，我们在模型训练过程中只能固定词表示特征，不能对其进行更新。为了观察fine-tuning过程对于性能的影响，我们设计了一组对比实验，即在PROJ系统中不更新词表示特征。实验结果如表

表 3-6 Fine-tuning过程的影响。

Table 3-6 Effect of fine-tuning word embeddings.

		Fixed	Fine-tuning	Δ
DE	UAS	59.74	**60.07**	+0.33
	LAS	49.44	**49.94**	+0.50
ES	UAS	70.10	**71.42**	+1.32
	LAS	61.31	**61.76**	+0.45
FR	UAS	70.65	**71.36**	+0.71
	LAS	60.69	**61.50**	+0.81

3.5.2.3 与其他双语词汇分布表示的对比

在本小节中，我们与现有的其他双语词汇分布表示学习方法进行对比。对比的方法有：

- 基于多任务学习的方法[48]（MTL）。
- 双语自编码器方法[50]（BIAE）。
- 双语组合语义模型[51]（BICVM）。

- 双语词袋模型[52]（BILBOWA）。

对于MTL以及BIAE，由于训练过程耗时较长，因此我们使用作者发布的双语分布表示数据。对于BICVM与BILBOWA，我们在本实验所用的数据上重新学习双语分布表示。实验结果如表 3-7所示。

表 3-7 与其他双语词汇分布表示的比较。‡MTL与BIAE使用的是作者发布的数据。

Table 3-7 Comparison with existing bilingual word embeddings. ‡For MTL and BIAE, we use their released bilingual word embeddings.

	DE		ES		FR	
	UAS	LAS	UAS	LAS	UAS	LAS
MTL[48]‡	56.93	46.22	67.71	58.43	67.51	57.27
BIAE[50]‡	53.74	43.68	58.81	46.66	60.10	49.47
BICVM[51]	56.30	46.99	67.78	58.08	69.13	58.13
BILBOWA[52]	54.51	44.95	67.23	56.16	64.82	52.73
CCA	59.42	49.32	68.87	59.65	69.58	59.50
PROJ	**60.07**	**49.94**	**71.42**	**61.76**	**71.36**	**61.55**

可以看出，PROJ与CCA在所有语言上均始终优于其他双语分布表示学习方法。MTL与BIAE方法之所以性能较低，一部分原因是其在目标语言测试数据上的词汇覆盖率较低。比如，在德语测试集上，MTL与BIAE只覆盖了31%的词，而CCA与PROJ均覆盖了70%以上。另外，BICVM与BILBOWA侧重句子级的语义分布表示对齐，而在词汇级别的语义对齐上表现并不理想。之前的分布表示学习研究通常是在文本分类任务上进行测试的，对词级别的对齐性要求不高；而在句法分析任务中，对句法结构的预测是以词为单位的，因此词级别的语义对齐就显得更为重要。

同时，我们也对不同的双语分布表示学习方法进行了定性分析。以英语/西班牙语为例，对于随机采样得到的西班牙语词汇，我们根据*cosine*相似度计算与其最相似的四个英语词汇（*k*近邻），如表 3-8所示。可以看出，PROJ与CCA所得到的*k*近邻更为准确，而BIAE，BICVM与BILBOWA有一定的语义/句法偏移。比如，根据BICVM计算出problematical（EN）是problemas（ES）的近邻，而实际上两者词性不一致，不太适用于句法分析任务。

表 3-8 在英语/西班牙语分布表示上的k近邻分析。

Table 3-8 Target words in Spanish and their 4 most similar words in English, as induced by various approaches.

Word (ES)	Neighboring Words (EN)					
	PROJ	CCA	MTL	BIAE	BICVM	BILBOWA
china (china)	india	russia	china	korea	chinese	helsinki
	russia	indonesia	independent	india	chinois	bulgarians
	taiwan	beijing	sumitomo	chinese	sino	constituting
	chinese	chinese	malaysian	brazil	33.55	market
problemas (problems)	problem	problems	events	problem	problematic	deficiencies
	difficulties	woes	sanctions	greatly	problematical	situations
	troubles	troubles	conditions	highlighted	difficulties	omissions
	issues	dilemmas	laws	scale	troubles	attentively
septiembre (september)	october	december	december	month	11th	a.m
	august	july	february	april	11.00	p.m
	january	october	july	scheduled	11	twelve
	december	june	november	march	eleventh	1998-1999

3.5.3 多源语言迁移学习实验

针对不同语言之间由于词序等类型学（typology）特征差异而导致的依存句法结构差异，接下来，我们进行多源语言情形下的迁移学习实验。我们考虑UDT 2.0中所有印欧语系语言，其中英语（EN）作为公共的源语言、德语（DE）、西班牙语（SP）、法语（FR）、葡萄牙语（PT）、意大利语（IT）以及瑞典语（SV）为目标语言。对于每种目标语言，我们使用其他6种语言作为源语言进行训练。

记多语言鲁棒映射方法为MULTI-PROJ，多语言Skip-gram模型为MULTI-SG，对于英语、德语、西班牙语以及法语，我们采用与单源语言实验中相同的实验设置。对于葡萄牙语、意大利语以及瑞典语，我们使用Europarl双语平行数据。另外，在训练MULTI-SG模型时，我们使用WMT-2011英语新闻语料作为额外的单语资源。在单源语言迁移实验中，我们已经证明了跨语言词聚类的作用，因此，在接下来的实验中，词聚类将与分布词表示特征一起使用。

本实验中作为对比的基准系统包括：

- 单源语言迁移（Bi-Proj）。我们将对比在单源语言迁移实验中表现最好的Proj+cls系统。
- 多源语言非词汇化迁移（Multi-Delex）。在该系统中，我们不使用多语分布表示以及跨语言词聚类等词汇化特征。
- Zhang和Barzilay提出的层次化低秩张量模型[146]（ZB15）。该模型利用层次化的低秩张量分解来实现跨语言迁移过程中的选择性参数共享。

实验结果如表 3-9所示。首先，通过对比Multi-Delex与Bi-Proj可以看出，即使不使用词汇化特征，Multi-Delex在ES和FR上的性能也显著优于Bi-Proj。这也证明了**多源语言的有效性**。其次，对比Multi-SG，Multi-Proj与Multi-Delex可以知道，无论使用哪种多源语言词汇分布表示特征，都能够非常显著地提升非词汇化迁移系统的性能。具体的，Multi-SG相对于Multi-Delex平均提升了2.24%（UAS）与5.18%（LAS）；而Multi-Proj的提升更为明显，分别是3.90%（UAS）与6.53%（LAS）。可见，LAS的提升比UAS的提升更为显著，这也印证了我们在3.1中关于本文动机的描述，即：词汇化特征对于依存弧上关系的分类尤为重要。

整体而言，基于线下处理的Multi-Proj方法比基于联合学习的Multi-SG方法取得了更好的效果。这个观察与我们的直觉有一定的冲突。实际上，从单源语言迁移实验中可以看出，尽管基于线下处理的Proj方法较为简单，但是所得到的双语分布表示非常稳定（见表 3-8）。而在联合学习方法中，单语目标函数（J_{mono}）与双语目标函数（J_{bi}）在有些情况下容易产生冲突，尤其是在词对齐出现错误时。这时候，如果没有先验知识的引入，联合学习方法会在冲突中选择一个平衡点，而这并不是我们的目标任务所期望的。在未来工作中，我们可以通过调节α与β的值来进一步提升联合学习方法的效果。另外，我们的方法也显著优于当前最好的系统（ZB15），尤其是在LAS指标上。

3.5.4 弱监督条件下的目标语言自适应

本小节将探讨一种更加实际的情形。在之前的迁移学习实验中，我们都假设目标语言不存在任何标注数据。而在实际应用过程中，虽然为一种语言标注大规模的树库代价较高难以实现，但是标注少量的句子（如50句、100句）却是完全可行的。因此，我们借鉴Zhang与Barzilay的实验设置[146]，假设目标语言中标注了50句话，并尝试利用这少量的标注数据来进一步提升我们的多源语言迁移模型。

表 3-9 多源语言条件下的迁移依存句法分析结果。实验中采用标准词性。

Table 3-9 Transfer parsing accuracies on the test data using gold standard POS tags.

	Multi-Delex		Multi-SG		Multi-Proj		Bi-Proj		ZB15	
	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
DE	59.35	49.82	61.70	54.16	**65.01**	**55.91**	60.35	51.54	62.5	54.1
ES	75.54	64.68	78.42	71.56	**79.00**	**73.08**	71.90	62.28	78.0	68.3
FR	74.41	64.21	76.44	70.21	**77.69**	**71.00**	72.93	63.12	**78.9**	68.8
IT	76.60	65.49	77.48	70.04	**78.49**	**71.24**	–	–	**79.3**	69.4
PT	75.64	69.66	77.87	74.10	**81.86**	**78.60**	–	–	78.6	72.5
SV	73.38	62.90	76.45	67.74	**78.28**	**69.53**	–	–	75.0	62.5
AVG	*72.49*	*62.79*	*74.73*	*67.97*	***76.39***	***69.32***	–	–	*75.4*	*65.9*

一种最直接的方法是将目标语言的标注数据与源语言训练数据放在一起，重新训练模型。但是这种做法代价太高，而且目标语言数据量太少，起到的作用非常有限。因此，我们采用一种在线的方式，将之前实验中已经训练好的迁移模型在少量目标语言数据上进行精调（Fine-tuning），所有实验的参数设置均保持不变。由于目标语言上依然没有开发集，同时为了防止模型在目标语言数据上过拟合，我们限定Fine-tuning过程中的迭代次数为100。

实验结果如表 3-10所示。可以看出，对于Multi-Delex、Multi-SG以及Multi-Proj三个系统，增加少量目标语言数据之后均取得了显著的提升，其中使用词汇化特征之后的系统比非词汇化迁移系统（Multi-Delex）提升更为显著。

一个有趣的现象是，德语（DE）上的所有模型均取得了非常大的提升（>10% UAS/LAS）。回顾我们在第 3.1节中的介绍，迁移模型所面临的一个重要挑战是源语言与目标语言之间由于类型学特征不同（例如语序）而导致的句法结构差异。多源语言固然能够在一定程度上缓解这个问题，但是依然没有从本质上解决。而来自目标语言的弱监督信息才是解决这一问题的关键。因此，对于德语上所取得的提升，我们认为，一个关键的因素是德语与其源语言之间在句法结构上的差异较大。从之前的实验中可以看出，不管是单源语言迁移还是多源语言迁移，德语上的性能始终是低于其他目标语言的，这也在一定程度上说明了这一点。

从类型学的角度来分析，在德语中，动词常常出现在V2的位置，也就是在宾语之后，从而导致了大量左指向的*dobj*（动宾关系）依存弧。这种特性与大部分源语言不同，如图 3-7所示。

表 3-10 引入50句目标语言标注数据之后，不同的迁移模型所取得的提升。括号内的数表示相对于表 3-9中结果的提升。

Table 3-10 Parsing accuracies of different transfer models with 50 annotated sentences from target languages as minimal supervision. The numbers in parentheses are absolute improvements over the directly transferred models as shown in Table 3-9.

	Multi-Delex(50)		Multi-SG(50)		Multi-Proj(50)	
	UAS	LAS	UAS	LAS	UAS	LAS
DE	67.26$_{(+7.81)}$	57.40$_{(+7.58)}$	72.76$_{\mathbf{(+11.06)}}$	66.28$_{\mathbf{(+12.12)}}$	**73.61**$_{\mathbf{(+8.60)}}$	**66.79**$_{\mathbf{(+10.88)}}$
ES	73.46$_{(-2.08)}$	64.19$_{(-0.49)}$	79.07$_{(+0.65)}$	74.20$_{(+2.64)}$	**79.67**$_{(+0.67)}$	**74.27**$_{(+1.19)}$
FR	74.60$_{(+0.19)}$	64.72$_{(+0.51)}$	79.26$_{(+2.82)}$	73.10$_{(+2.89)}$	**79.99**$_{(+2.30)}$	**74.45**$_{(+3.45)}$
IT	75.68$_{(-0.92)}$	67.56$_{(+2.07)}$	**79.92**$_{(+1.43)}$	74.86$_{(+4.82)}$	79.85$_{(+0.46)}$	**74.94**$_{(+3.70)}$
PT	75.01$_{(-0.63)}$	68.37$_{(-1.29)}$	**81.44**$_{(+3.57)}$	**78.77**$_{(+4.67)}$	81.11$_{(-0.75)}$	78.22$_{(-0.38)}$
SV	74.93$_{(+1.55)}$	65.16$_{(+2.26)}$	**80.04**$_{(+3.59)}$	**74.10**$_{(+6.36)}$	80.03$_{(+1.75)}$	73.71$_{(+4.18)}$
AVG	*73.49*$_{(+1.00)}$	*64.57*$_{(+1.78)}$	*78.75*$_{(+4.02)}$	*73.55*$_{(+5.58)}$	***79.04***$_{(+2.65)}$	***73.73***$_{(+4.41)}$

因此，我们对德语及其源语言中*dobj*依存弧结构进行了统计。从表 3-11中可以看出，德语中的*dobj*依存弧与其他语言有着显著的差异。

表 3-11 不同语言中*dobj*依存弧指向的分布差异。

Table 3-11 Distribution divergence of left-directed and right-directed arcs with *dobj* relation across different languages.

		$dobj_{\curvearrowright}$	$dobj_{\curvearrowleft}$	ratio
Target	DE	4,277	3,457	1.2 : 1
Source	EN	38,395	764	50.3 : 1
	ES	10,551	1,175	9.0 : 1
	FR	10,015	2,667	3.8 : 1
	IT	4,714	695	6.8 : 1
	PT	8,052	773	10.4 : 1
	SV	2,724	163	16.7 : 1

鉴于此，我们进一步分析弱监督信息的引入对于*dobj*预测的精确率以及召回率的影响。结果如表 3-12所示，可以看到，弱监督信息极大地提升了*dobj*依存弧的召回率。

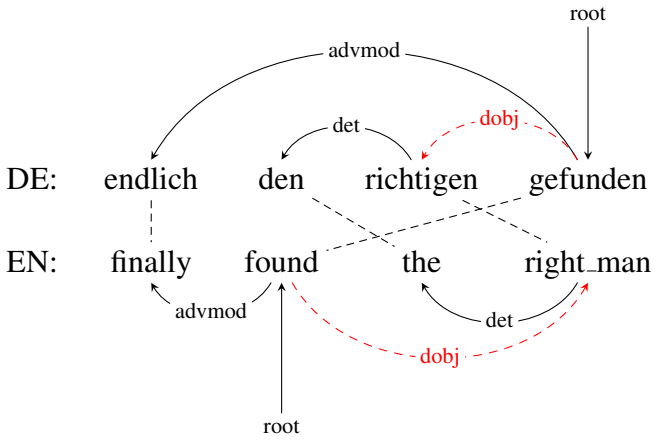


图 3-7 德语和英语中不同的*dobj*依存弧方向。

Figure 3-7 Reverse direction of the *dobj* relation in German and English.

表 3-12 弱监督信息对于*dobj*依存弧分类准确率的影响。

Table 3-12 Effect of minimal supervision on *dobj* of DE. Unsup indicates the (unsupervised) directly transfer models.

	Multi-Delex		Multi-SG		Multi-Proj	
	P	R	P	R	P	R
Unsup	36.84	35.69	36.10	38.65	50.47	35.69
+50	39.62	41.45	47.38	60.86	52.34	62.66
Δ	***2.78***	***5.76***	***11.28***	***22.21***	***1.87***	***26.97***

3.6 本章小结

本章针对依存句法分析模型在跨语言迁移过程中难以有效利用词汇化特征的问题，提出了基于跨语言分布表示学习的框架来弥补这一特征鸿沟。同时，对于不同语言之间由于类型学特征的差异而导致的句法结构差异，提出了多源语言迁移以及一种有效利用弱监督信息的目标语言自适应方法。

我们针对单源语言及多源语言的情况，分别提出了有效的跨语言分布表示学习方法。在多语言通用依存句法树库上的实验结果表明，我们所提出的跨语言分布表示能够显著地提升非词汇化迁移系统的性能。同时，多源语言能够带来极大的提升。另外，在实际应用中，我们还可以通过标注少量的目标语言数据（如50句），再利用本章所描述的框架来构建更精确的迁移系统。

在本章的工作中，为了实验方便，我们使用的是通用依存树库中的印欧语系语言。而实际上，这些语言并不是真正的资源稀缺语言。因此，在未来的工作中，我们希望将此技术应用于真正的资源稀缺语言中。

第 4 章 基于深度多任务学习的多类型树库融合

4.1 引言

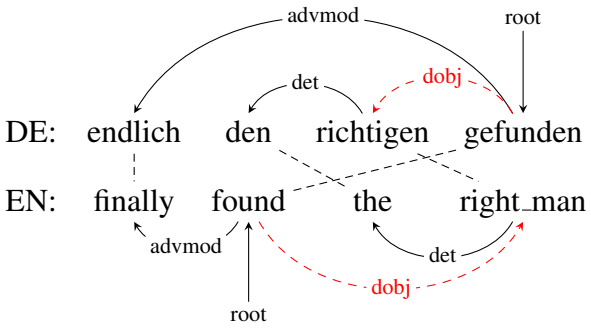
在前一章中，我们在资源稀缺语言的依存句法分析中证明了跨语言迁移的有效性。同时我们也通过实验发现，引入少量目标语言标注数据之后，迁移系统的性能能够取得显著的提升。由此而引出的一个问题是，当目标语言树库已经具有一定规模（如5,000句，1,0000句）时，跨语言资源是否仍然可以提供有效的帮助? 在本章中，我们将这个问题进一步展开，研究在目标树库具有一定规模的情况下，如何融合其他树库资源以提升依存句法分析的性能。

依存句法树库的标注不仅对专家知识的要求较高，而且对于很多语法较为自由的语言（如中文）而言，标注的一致性较低。这也在很大程度上限制了很多语言上句法树库的规模。另一方面，依存句法的标注规范近10年来也在不断地演变，从早期只反映句法关系的依存结构[13]到近几年逐渐成为主流的偏语义的依存结构[14, 15]，从而导致很多语言中存在不同标注规范的依存树库。这类树库通常称为异构树库（Heterogeneous Treebanks）。

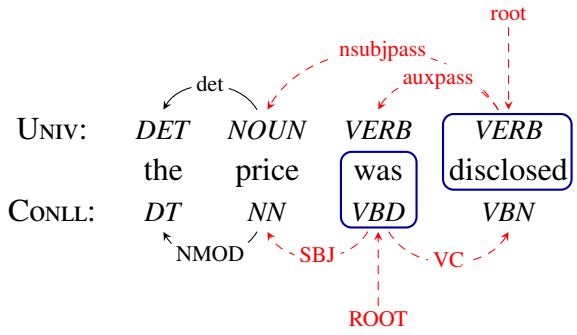
在本章的工作中，我们提出一个可融合多类型源树库的通用迁移句法分析框架。具体来讲，我们考虑两类源树库，分别为多语言通用树库，也就是前一章中所用到的UDT；以及单语异构树库。在此前的工作中，也有学者试图融合单语异构树库以提升目标树库上的依存句法分析。比如Niu等人所采用树库转换方法[147]，Li等人提出的基于准同步文法（Quasi-synchronous Grammar, QG）特征的方法[148]，以及Johansson所提出的基于特征共享的方法[149]。这些工作充分证明了单语异构树库之间确实存在信息互补，这也是本章研究的一个重要前提。

树库融合的挑战主要在于源树库与目标树库之间在句法结构上的差异，这种差异在不同类型的源树库上又有所不同。对于跨语言树库，其不一致性主要是由不同语言之间类型学（typology）特征的不同所导致的，如第 3.5.4节中所提到的德语、英语之间的动宾结构差异（图 4-1 a)）。对于单语异构树库，不一致性则主要体现在标注规范的差异。这种差异主要体现在两个层面，一是不同的文法，如短语结构文法以及依存结构文法；其次，对于同一种文法，也存在不同的标注规范。对于依存文法而言，则具体反映在核心结点的选择

以及依存关系的定义。比如在CoNLL依存体系中，往往是以功能词（functional words）作为核心结点，同时依存关系的定义也是以句法为主，语义粒度较粗；而在Stanford依存体系以及通用依存体系中，则主要以内容词（content word）作为核心结点，其依存关系集粒度较细，深入到了语义层面（见图 4-1 b)）。



a) Multilingual universal dependencies.



b) Monolingual heterogeneous dependencies.

图 4-1 不同依存句法标注的对比：多语通用依存结构（a）以及单语异构依存结构（b）。

Figure 4-1 Comparisons between multilingual universal dependencies (a) and monolingual heterogeneous dependencies (b).

尽管不同树库之间存在不一致性，但与此同时，也有相当一部分结构是一致的。我们期望从这些一致的结构中充分萃取知识，而尽量规避那些不一致性所带来的影响。在本研究中，我们提出一种简单且有效的深度多任务学习框架。该框架基于深度神经网络，将每种树库下的学习过程视作一个单独的任务，并通过**分布表示**层面上的参数共享来控制不同任务之间的信息交互。我们充分利用深度神经网络在表示学习上的层次性，针对不同源树库类型设计了不同的参数共享策略。

我们在不同语言、不同依存句法树库上进行了充分的实验。在UDT上的跨语言通用树库融合实验表明，我们的模型在六种语言上相比于目标树库上的有

监督模型均有显著的提升。在UDT与CoNLL-X[150]树库上的单语异构树库融合实验也证明了我们方法的有效性。同时，我们也在中文宾州树库（CTB5）（目标树库）以及中文依存树库（CDT）[1]（源树库）上进行实验，以进一步与前人工作进行对比。实验结果表明，相对于Li等人[148]基于准同步文法特征的方法，我们的模型能够更有效地利用单语异构树库。

4.2 背景与相关工作

4.2.1 面向依存句法分析的资源融合方法

4.2.1.1 单语资源融合

利用单语异构树库资源来帮助依存句法分析的思想由来已久。Niu等人[147]最早通过树库转化的方法，将依存结构的CDT转换为与CTB5一致的短语结构，从而直接扩充短语结构句法分析的训练数据规模。Li等人[148]则使用人工设计的转化模式（Transformation Patterns）对标注规范的不一致进行建模。然后基于转化模式构成准同步文法（QG）[151]特征，来增强依存句法分析模型。与本研究思想类似，Johansson[149]也采用参数共享的思想来融合异构树库资源，不同的是，在他们的工作中，参数是在离散特征层面（feature-level）上进行共享的。因此，他们的方法较难迁移至跨语言树库融合的情况，因为在跨语言树库中特征的表现形式可能完全不一致，无法确定需要共享的参数。与之相反，我们充分利用分布表示的通用性，使用表示层面（representation-level）上的参数共享，使得不同类型树库之间均能进行方便得融合。

4.2.1.2 跨语言资源融合

跨语言资源通常被用于资源稀缺语言分析的场景，或通过数据迁移[5, 94, 152]，或通过模型迁移[98, 127, 146]。与本文研究工作相似，Duong等人[153]与Ammar等人[154]也采用参数共享的方法来利用多语言树库来增强依存句法分析。在他们的工作中，模型的绝大部分参数都是共享的（除了词汇特征分布表示向量[2]），而忽略了不同语言树库之间句法结构上的不一致性。除此之外，Duong等人的工作主要集中在对于资源稀缺语言的处理（假设目标语言中只有~3K个词项），而在目标语言拥有一定规模数据的情况下，准确率会有所损失。Ammar等人的实验设置与本研究更为相似，我们将他们的方法记为浅

[1] https://catalog.ldc.upenn.edu/LDC2012T05

[2] Duong等人[153]根据双语词典在目标函数中引入了对于词汇特征的L2正则项。

层多任务学习（shallow multi-task learning, SMTL），并作为一个基准系统进行对比。需要注意的是，SMTL是本文方法的一种特例。除了多语言句法树库，未经标注的双语平行数据也能够为依存句法分析提供一定的帮助[155–157]。

4.2.2 基于神经网络的多任务迁移学习

多任务学习是一种归纳式（Inductive）迁移学习，旨在利用来自不同任务的信息来帮助目标任务上的学习。在多任务学习的一般定义中，多个任务是并行训练的，并通过一个共享的表示层来实现不同任务之间信息的迁移[16]。近年来，基于神经网络与分布表示的多任务学习在自然语言处理领域受到越来越多的关注，其主要原因在于神经网络对于特征表示的逐层抽象特性。我们知道，在深度神经网络中，对于特征的抽象程度随着网络层数的递增而越来越高，越接近底层的特征表示越具有泛化性，而越接近顶层的特征表示则与具体任务的相关性越强。因此，通过在泛化性较强的表示层进行参数共享，则可以实现不同任务之间的信息迁移。

与多任务学习非常相似的一种学习机制是联合学习（Joint learning）。联合学习在自然语言处理中有着广泛的应用，尤其是在一些相互之间存在一定依赖性的流水线任务上，如中文分词、词性标注、句法分析、语义角色标注等[158–162]。联合学习的主要优势是缓解任务之间的错误传播问题，如由词性标注错误而导致的句法分析错误。与多任务学习不同的是，联合学习通常要求数据中同时含有多个任务的标注信息，比如在词性与句法的联合学习[159]要求一个句子同时标注了词性序列以及句法结构；而多任务学习则没有这样的限制，因此在实际中可应用的范围更广。

近年来，在基于神经网络的多任务自然语言处理应用中，最值得一提的是Collobert与Weston所提出的基于深度卷积网络的多任务学习框架[9]。该框架使用统一的模型来解决多项自然语言处理任务，包括词性标注、命名实体识别、语义角色标注等。不同任务之间通过共享词汇特征分布表示来实现信息交互。Dong等人[163]在基于“编码——解码”框架的神经机器翻译系统中采用多个解码器（Decoder），从而实现将一种语言同时翻译成多种其他语言。该模型使用多任务学习进行训练，其中源语言到不同目标语言的翻译之间通过共享编码器（Encoder）参数来实现信息迁移。Luong等人[164]则将多任务“编码——解码”模型应用于更多的任务，包括句法分析、机器翻译、图片字幕生成等。最近，也有工作利用多任务神经网络在隐式篇章关系分类任务以及文本分类任务中融合

不同类型的训练数据[165, 166]，与本文出发点较为相似。

4.3 基于长短时记忆网络的依存句法分析模型

根据前一章的介绍，我们知道，在基于转移的依存句法分析中，最关键的因素是对于每个转移状态的表示。每个转移状态通常可以表示为由一个栈（S）、一个缓存（B）以及当前已经生成的依存弧集合（A）所构成的三元组。传统的依存句法分析模型通过人工定义的大量特征模板从该三元组中抽取特征[144]，一般为指定位置（如：栈顶，缓存顶等）的某种属性（如：词，词性等）或者组合。这种依赖“特征工程”的方式表达能力有限，对转移状态的表达并不充分。前一章所采用的基于前馈神经网络的模型[71]通过使用Cube激活函数在一定程度上解决了特征组合的完备性问题，但是，它所采用的基本特征仍然是由人工定义的局部特征，无法覆盖完整的栈、缓存以及当前依存弧集合的信息。

在本研究中，我们采用Dyer等所提出的基于长短时记忆（LSTM）网络的依存句法分析模型[134]，并使用双向LSTM对词表示进行建模[167]。采用该结构的原因如下：

- 与Chen和Manning的模型相比，该模型使用LSTM以及递归神经网络对当前转移状态的各个组成部分（栈、缓存、历史转移序列，以及当前依存子树集合）进行全局的编码，充分利用了历史转移信息。
- 通过对词、栈、缓存、转移序列等部分细粒度地建模，使得在多任务学习框架中能够更加灵活地控制参数共享。

图 4-2展示了模型的具体结构。

首先，我们使用基于字符的双向LSTM来获得句子中每个词的分布表示（图 4-3，记为Char-BiLSTM）。每个词项可以表示为：

$$\boldsymbol{x} = \mathrm{ReLU}(\boldsymbol{V} \cdot [\overrightarrow{\boldsymbol{w}} \oplus \overleftarrow{\boldsymbol{w}} \oplus \boldsymbol{t}] + \boldsymbol{b}) \tag{4-1}$$

其中$\boldsymbol{t}$为w的词性分布表示向量。接下来，我们分别使用LSTM对栈中的元素序列以及缓存中的反向词序列进行建模，并将最后时刻的隐层输出作为当前栈、缓存的表示。特别的，在某些转移算法中，栈及缓存中的每个元素都可能是当前状态下某颗依存子树的核心结点，因此，为了更全面地对栈中元素进行表示，我们采用递归神经网络（RecNN）对相应的子树自底向上地进行组合。另外，与传统的转移状态表示不同，该模型对历史转移动作序列也进行了建模。我们首先将历史上的转移动作表示为分布表示向量，再将其作为LSTM的输入。

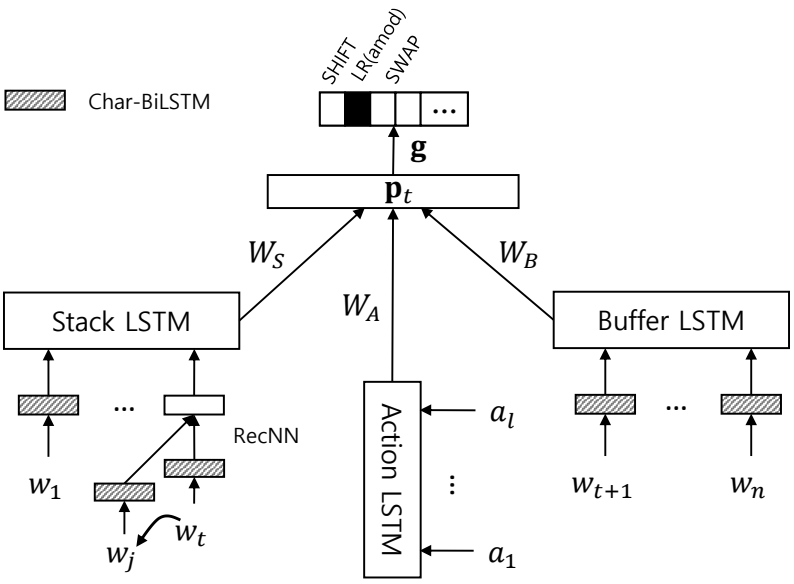


图 4-2 基于LSTM的依存句法分析模型结构。

Figure 4-2 LSTM-based model for dependency parsing.

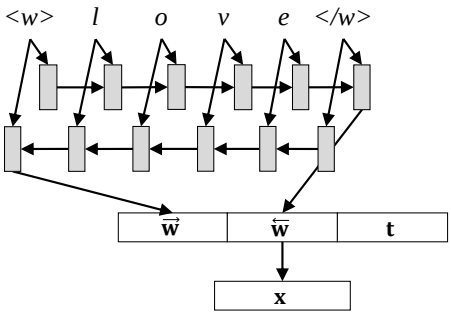


图 4-3 基于双向LSTM的词表示学习。

Figure 4-3 Char-BiLSTM for modeling words.

由此，我们可以得到栈、缓存、历史转移序列的向量表示，分别记为$\boldsymbol{s}_t, \boldsymbol{b}_t, \boldsymbol{a}_t$。形式化地，

$$\boldsymbol{s}_t = \texttt{LSTM}(r(S_0), r(S_1), ..., r(S_t)) \tag{4-2}$$

$$\boldsymbol{b}_t = \texttt{LSTM}(r(B_n), r(B_{n-1}), ..., r(B_{t+1})) \tag{4-3}$$

其中：$r(x) = \texttt{RecNN}(subtree(x))$，$subtree(S_i)$表示由$S_i$为根结点的依存子树。RecNN的计算过程如图 4-4所示。由于依存树中的一个根节点可能存在多个子结点，因此我们按照转移过程中依存弧的归约顺序递归地进行组合，组合过程考虑依存关系类型。以图 4-4中的依存树为例，首先组合“*overhasty*”与

“*decision*”，得到中间表示$\boldsymbol{c}_1$，再将$\boldsymbol{c}_1$与“*an*”进行组合，得到该子树的完整表示$\boldsymbol{c}_2$：

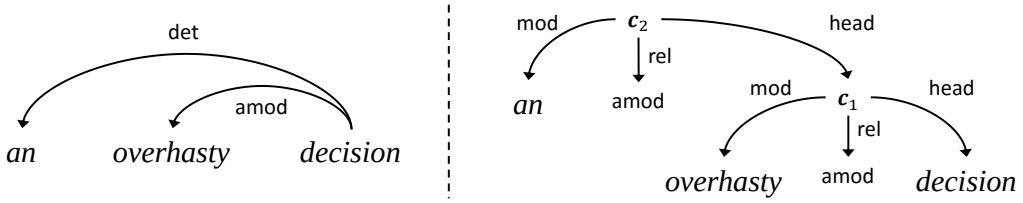


图 4-4 基于递归神经网络的依存子树表示。

Figure 4-4 Recursive neural network for computing the representation of a dependency tree.

$$\begin{aligned}\boldsymbol{c}_1 &= \tanh(\boldsymbol{U}\cdot[\boldsymbol{e}(decision)\ \oplus\ \boldsymbol{e}(overhasty)\ \oplus \boldsymbol{e}(amod)])\\ \boldsymbol{c}_2 &= \tanh(\boldsymbol{U}\cdot[\boldsymbol{c}_1\ \oplus \boldsymbol{e}(an)\ \oplus \boldsymbol{e}(det)])\end{aligned} \tag{4-4}$$

对于历史转移动作序列的表示则直接采用转移动作的分布表示向量作为输入：

$$\boldsymbol{a}_t = \texttt{LSTM}(\boldsymbol{e}(a_1), \boldsymbol{e}(a_2), ..., \boldsymbol{e}(a_l)) \tag{4-5}$$

接着，我们将$\boldsymbol{s}_t, \boldsymbol{b}_t, \boldsymbol{a}_t$拼接，再通过一个非线性隐含层。这里采用$\texttt{ReLU}(x) = \max(0, x)$作为非线性激活函数：

$$\boldsymbol{p}_t = \texttt{ReLU}(\boldsymbol{W}\cdot[\boldsymbol{s}_t \oplus \boldsymbol{b}_t \oplus \boldsymbol{a}_t] + \boldsymbol{d}) \tag{4-6}$$

$\boldsymbol{p}_t$即为当前转移状态的分布表示。最后，我们使用`softmax`函数计算下一步转移动作$z \in \mathcal{A}(S, B)$的概率分布：

$$p(z|\boldsymbol{p}_t) = \frac{\exp(\boldsymbol{g}_z \cdot \boldsymbol{p}_t + \boldsymbol{q}_z)}{\Sigma_{z' \in \mathcal{A}(S,B)} \exp(\boldsymbol{g}_{z'} \cdot \boldsymbol{p}_t + \boldsymbol{q}_{z'})} \tag{4-7}$$

其中$\mathcal{A}(S, B)$表示在当前转移状态下有效的转移动作集合[3]。

由于本研究中所考虑的大部分树库中都存在非投射的依存树，我们采用Nivre所提出的基于交换的转移算法[140]（Swap-based）。该算法定义了如下四类转移动作：

- Left-Arc(r)：产生一条由S_0指向S_1的依存弧（$S_1 \overset{r}{\leftarrow} S_0$）（$S_0$为核心结点，$S_1$为修饰结点），从栈中删除$S_1$。
- Right-Arc(r)：产生一条由S_1指向S_0的依存弧（$S_1 \overset{r}{\rightarrow} S_0$）（$S_1$为核心结点，$S_0$为修饰结点），从栈中删除$S_0$。
- Swap：将栈中元素S_1移入缓存顶部，从而使得S_0与S_1交换顺序。
- Shift：将缓存顶部元素B_0移入栈顶，前提条件是B不为空。

[3]某些转移状态对特定的转移动作有所约束，比如当缓存为空时，无法进行移进操作（Shfit）。

SWAP操作的引入，使得栈和缓存中都可能存在依存子树。因此，我们都使用递归神经网络对其中的元素进行表示。

4.4 基于深度多任务学习的树库融合框架

在本研究中，我们将每种树库的学习过程视为一个单独的任务，并提出一种深度多任务学习的框架来实现不同任务之间的信息交互，如图 4-5所示。我们将目标树库上的学习作为主任务（primary task），源树库上的学习为相关任务（related task）。受Ammar等人工作[154]的启发，我们引入了一个额外的任务表示e^t，将其与$\boldsymbol{s}_t, \boldsymbol{b}_t, \boldsymbol{a}_t$拼接之后再计算$\boldsymbol{p}_t$：

$$\boldsymbol{p}_t = \mathtt{ReLU}(\boldsymbol{W} \cdot [\boldsymbol{s}_t \oplus \boldsymbol{b}_t \oplus \boldsymbol{a}_t \oplus e^t] + \boldsymbol{d}) \tag{4-8}$$

同时，e^t也与$\boldsymbol{p}_t$一起用于转移动作概率分布的计算：

$$p(z|\boldsymbol{p}_t) = \mathtt{softmax}(\boldsymbol{g}_z \cdot [\boldsymbol{p}_t \oplus e^t] + \boldsymbol{q}_z) \tag{4-9}$$

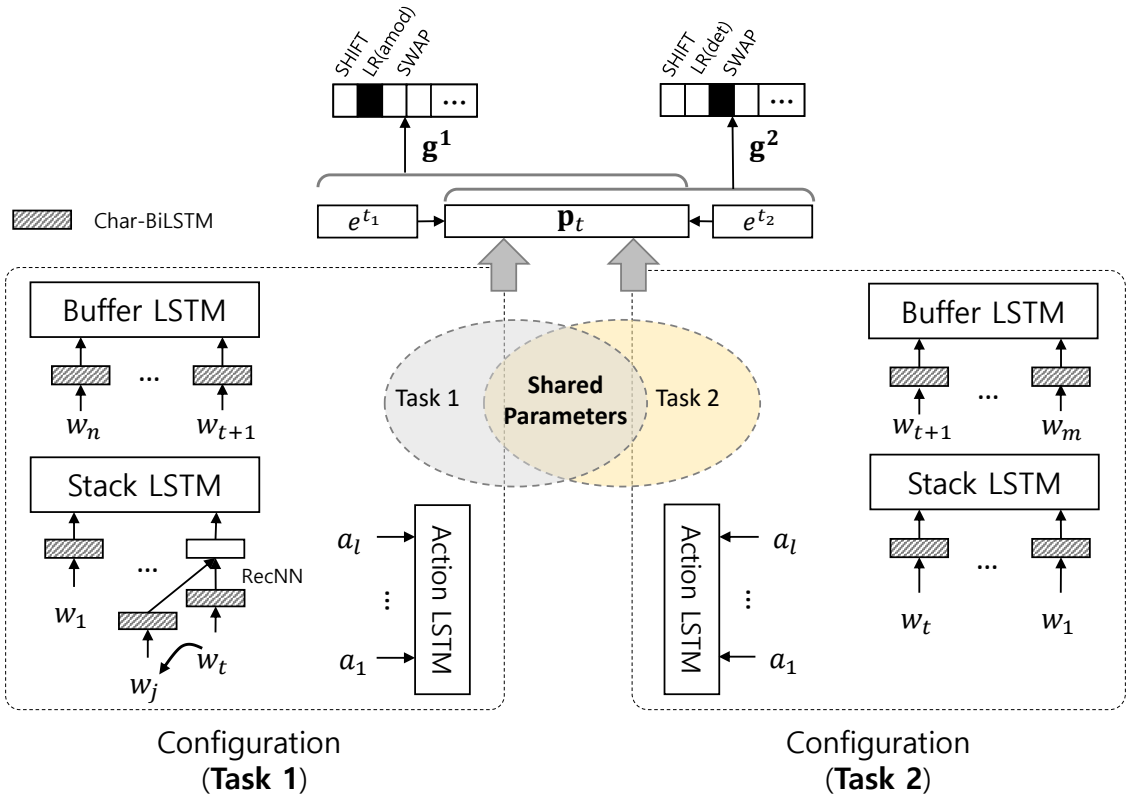


图 4-5 基于深度多任务学习的树库融合框架。

Figure 4-5 Illustration of the deep multi-task learning framework for treebank integration.

4.4.1 参数共享

多任务学习的关键在于参数共享。在本研究中，我们针对不同任务（不同源树库）设计了相应的参数共享策略。具体的，我们考虑两种类型的源树库，分别为跨语言通用树库（**MULTI-UNIV**）以及单语异构树库（**MONO-HETERO**）。

MULTI-UNIV 在通用依存体系中，不同语言的树库采用统一的通用词性（Universal POS）集合[168]以及依存关系集合。对于同一种转移算法而言，其转移动作集合也是一致的（三类转移动作与依存关系的组合）。然而，也应注意到，不同语言的词与字符集通常不同。因此，在选择共享参数时，我们可以共享词性、依存关系以及转移动作的分布表示（$\boldsymbol{E}_{pos}, \boldsymbol{E}_{rel}, \boldsymbol{E}_{act}$）；而不共享字符分布表示（$\boldsymbol{E}_{char}$）以及基于字符的双向LSTM参数（BiLSTM(chars)）。另一方面，由于不同语言之间存在类型学特征（如词序）的差异，因此，在生成依存树的归约过程中，转移动作的序列特性有所不同。因此，我们在多任务学习中也不共享用于历史转移动作序列建模的LSTM参数（LSTM(A)）。

MONO-HETERO 与跨语言通用树库不同，单语异构树库采用的是统一的字符集，因此$\boldsymbol{E}_{char}$以及BiLSTM(chars)可以在任务之间共享；但是，由于依存结构与依存关系之间存在差异，并由此其转移动作集合也有所不同，因此$\boldsymbol{E}_{rel}, \boldsymbol{E}_{act}, LSTM(A)$均为任务相关的。在本工作中，不失一般性，我们将单语异构树库的词性都映射为通用词性，因此$\boldsymbol{E}_{pos}$也是共享参数。

除了以上提到的参数，考虑到任务之间较强的相关性，我们共享大部分其他的参数，包括对于栈（LSTM(S)）、缓存（LSTM(B)）、用于子树建模的递归神经网络（RecNN）、以及用于计算转移状态表示$\boldsymbol{p}_t$的权值矩阵（$\boldsymbol{W}$）。而最顶层用于计算转移动作概率分布的权值矩阵（$\boldsymbol{g}$）由于与任务直接相关，因此不进行共享。

两种情况下的具体参数共享策略如表 4-1所示。

4.4.2 训练过程

我们采用一种随机的方式进行训练：

1. 随机选择一个任务；
2. 从该任务（所对应的树库）中随机选择一个句子，并构建用于学习分类器的训练实例；
3. 根据这些训练实例，利用反向传播计算相应参数的梯度，并更新参数；
4. 返回第1步。

表 4-1 跨语言通用树库以及单语异构树库条件下的参数共享策略。其中，LSTM(S)，LSTM(B), LSTM(A)分别表示建模栈、缓存、历史转移动作序列的LSTM参数；BiLSTM(chars)为基于字符的词表示双向LSTM参数；RecNN为建模子树的递归神经网络；$\boldsymbol{E}$为特征分布表示矩阵。

Table 4-1 Parameter sharing strategies for **MULTI-UNIV** and **MONO-HETERO**. LSTM(S) – *stack* LSTM; LSTM(B) – *buffer* LSTM; LSTM(A) – *action* LSTM; BiLSTM(chars) – Char-BiLSTM; RecNN – recursive NN modeling the subtrees; $\boldsymbol{W}$ – weights from A, S, B to the state ($\mathbf{p}_t$); $\boldsymbol{g}$ – weights from the state to output layer; $\boldsymbol{E}$ – embeddings.

	MULTI-UNIV	**MONO-HETERO**
Shared	$\boldsymbol{E}_{pos}, \boldsymbol{E}_{rel}, \boldsymbol{E}_{act}$ LSTM(S), LSTM(B), RecNN $\boldsymbol{W}$	$\boldsymbol{E}_{pos}, \boldsymbol{E}_{char}$ LSTM(S), LSTM(B), BiLSTM(chars), RecNN $\boldsymbol{W}$
Task-specific	$\boldsymbol{E}_{char}, e^t$ LSTM(A), BiLSTM(chars) $\mathbf{g}$	$\boldsymbol{E}_{rel}, \boldsymbol{E}_{act}, e^t$ LSTM(A) $\boldsymbol{g}$

我们使用目标任务中的开发集来控制训练进程（early-stopping）。

4.5 实验与分析

4.5.1 实验设置

在本研究中，我们使用通用依存树库（UDT v2.0）与CoNLL-X评测数据，并按照标准的划分进行实验。对于单语异构源树库，为了与Li等人的工作[148]进行对比，我们也进行了中文上的实验，其中使用CDT作为源树库，CTB5作为目标树库。所有数据的划分及规模如表 4-2所示。我们考虑以下实验设置：

- MULTILINGUAL (UNIV→UNIV)。该实验设置中，我们研究跨语言通用树库之间的融合。具体的，我们将德语、西班牙语、法语、葡萄牙语、意大利语以及瑞典语树库作为目标树库，而将资源最为丰富的英语树库作为公共的源树库。
- MONOLINGUAL (CONLL↔UNIV)。这里，我们研究单语异构树库CoNLL-X与UDT之间的融合。具体的，我们使用CoNLL-X语料与UDT语料中共有的语言（德语、西班牙语、葡萄牙语、瑞典语），对于每种语言，我们观察其UDT树库与CoNLL-X树库分别作为源树库时对目标树库的影响。
- MONOLINGUAL (CDT→CTB5)。该设置下，我们采用与Li等人[148]工作中相同

的实验设置，并考虑了使用自动词性以及正确词性两种场景。

本实验中作为对比的基准系统包括：

- 单语有监督学习系统（SUP）。该系统为在目标树库上的有监督学习系统。
- 级联学习系统（CAS）。该系统包含两个步骤，首先，我们在源树库上训练依存分析模型，然后将该模型中的参数用于目标树库学习模型的初始化。这种方式与我们在前一章第3.5.4节中研究的弱监督条件下的迁移学习方法相同。

此外，对于跨语言通用树库的融合实验，我们也对比了第4.2.1节中所提到的浅层多任务学习系统（SMTL），也就是Duong等人[153]以及Ammar等人[154]所使用的方法。在SMTL中，我们共享除了字符分布表示（$\boldsymbol{E}_{char}$）之外的其他所有参数，同时不使用任务表示向量e^t。与Duong等人和Ammar等人工作不同的是，我们没有使用诸如跨语言词聚类、跨语言分布表示以及词典等外部资源。

表4-2 本实验中所用到的依存句法树库资源统计信息。

Table 4-2 Statics of the treebanks used in our experiments.

	Train	Dev	Test	Train	Dev	Test
	UDT			CoNLL-X		
EN	39,832	1,700	2,416	–	–	–
DE	14,118	800	1,000	35,295	3,921	357
ES	14,138	1,569	300	2,976	330	206
FR	14,511	1,611	300	–	–	–
PT	9,600	1,200	1,198	8,164	907	288
IT	6,389	400	400	–	–	–
SV	4,447	493	1,219	9,938	1,104	389
	CDT			CTB5		
ZH	55,500	1,500	3,000	16,091	803	1,910

4.5.2 跨语言通用树库融合实验结果

跨语言通用树库融合的实验结果如表4-3所示。首先，我们可以看出，在大部分语言上，级联学习（CAS）相比于有监督学习的性能取得了较小的提升，且在瑞典语上提升最为显著（+1.52% UAS，+2.04% LAS）。这也意味着在英文源树库上的预训练确实能够为目标树库的学习提供一个较好的参数初始化。同

表 4-3 跨语言通用树库融合的实验结果。使用MaltEval进行显著性检验的结果显示MTL在所有语言上都以高于99%的置信度优于Sup。

Table 4-3 Parsing accuracies of Multilingual (Univ→Univ). Significance tests with MaltEval yield p-values < 0.01 for (MTL *vs.* Sup) on all languages.

	Multilingual (Univ → Univ)							
	Sup		Cas_{EN}		$SMTL_{EN}$		MTL_{EN}	
	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
DE	84.24	78.40	84.24	78.65	84.37	79.07	**84.93**	**79.34**
ES	85.31	81.23	85.42	81.42	85.78	81.54	**86.78**	**82.92**
FR	85.55	81.13	84.57	80.14	86.13	81.77	**86.44**	**82.01**
PT	88.40	86.54	88.88	87.07	89.08	87.24	**89.24**	**87.50**
IT	86.53	83.72	86.58	83.67	86.53	83.64	**87.26**	**84.27**
SV	84.91	79.88	86.43	81.92	**86.79**	**82.31**	85.98	81.35
Avg	*85.82*	*81.82*	*86.02*	*82.15*	*86.45*	*82.60*	***86.77***	***82.90***

时SMTL系统几乎在所有语言上都取得了比Cas更好的结果（在意大利语上性能接近）。因此，我们得到的初步结论是：即便使用同一个模型，在两个树库上进行联合学习也要优于级联学习的方式。

进一步的，通过合适的参数共享策略，我们的深度多任务学习方法在5种语言上取得了最好的性能。其中的一个例外是瑞典语。从表中可以看出，Cas与SMTL系统在瑞典语上的表现都显著优于MTL（注意MTL仍然显著优于Sup系统）。经分析，我们认为原因主要有以下两点：

1. 瑞典语中词语的形态学（mophology）特征与英语类似，因此我们期望与形态学相关的模型参数（如BiLSTM(chars)）也能够共享。
2. 瑞典语的训练树库规模较小，与英语源树库规模比例为1:9。我们猜想SMTL与Cas在目标树库较小（即：资源稀缺）的情况下比深度多任务学习（MTL）更加有效。直觉上，假如目标树库中蕴含的知识太少，那么在联合学习过程中，选择更强调源树库数据是一种更为合理的方式。而共享更多的参数，如SMTL则恰恰是通过共享更多的参数来实现这一目的。

为了验证第一个猜想，我们设计了一组简单的对比实验，将性能最好的SMTL系统中的BiLSTM(chars)参数设为任务相关（不共享），再观察其性能变化。实验结果发现，SMTL的性能显著下降，其中UAS与LAS分别降低了0.73%与0.81%。这个发现也启示我们，在深度多任务学习的框架中，其实可

以通过更细致的语言学分析来设计更加合理的参数共享策略。但是，在本研究中，我们希望用一个通用性较强的框架来解决多类型树库融合问题，因此我们不进一步将策略复杂化。

为了验证第二个猜想，我们参考Duong等人[153]在资源稀缺条件下的实验设置。该设置下，目标树库中只有3K个词项。我们采用与Duong等人一致的实验数据，结果如表 4-4所示。首先，我们可以看出，CAS，SMTL与MTL都显著优于SUP系统；同时，CAS与SMTL比MTL表现更好，其中CAS表现最好。这也证明，在目标树库规模较小的情况下，级联学习是最有效的方式。

另外，在本实验中，我们也发现对于SMTL与MTL而言，可以通过对任务的加权采样（weighted task sampling）进行优化。在多任务学习中，各个任务收敛速度的差异会在一定程度上影响系统的性能。理论上，我们期望各个任务在同一时刻收敛[16]。而在实际应用中，这一点很难保证。比如对于规模不均衡的两个树库的联合训练，大规模树库下的收敛速度肯定比小树库要慢。因此，这里我们提出一种基于加权任务采样的近似做法。具体的，假设两个任务单独运行时收敛速度比值为1:9，那么我们在多任务学习过程中，则以9:1的概率分布对两者进行采样，使得两者的收敛速率近似一致。如表 4-4所示，采用加权任务采样之后，SMTL与MTL的性能都取得了较为显著的提升。

表 4-4 目标树库规模较小情况下的实验结果，采用LAS评价。

Table 4-4 Low resource setup (3K tokens), evaluated with LAS.

	DE	ES	FR
SUP	58.93	61.99	60.45
CAS	64.08	**70.45**	**68.72**
SMTL	63.57	69.01	65.04
+ *weighted sampling*	*63.50*	*70.17*	*68.52*
MTL	62.43	66.67	64.23
+ *weighted sampling*	***64.22***	*68.42*	*66.67*
Duong et al.	61.2	69.1	65.3
Duong et al. + Dict	61.8	70.5	67.2

4.5.3 单语异构树库融合实验结果

单语异构树库融合的实验结果如表 4-5所示。总体而言，在两种设置下

（CONLL→UNIV与UNIV→CONLL），MTL都显著优于SUP系统[4]。同时，在大部分情况下，MTL优于CAS系统。

表 4-5 单语异构树库融合的实验结果。

Table 4-5 MONOLINGUAL (CONLL↔UNIV) performance.

	MONOLINGUAL (CONLL→UNIV)						MONOLINGUAL (UNIV→CONLL)					
	SUP		CAS		MTL		SUP		CAS		MTL	
	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
DE	84.24	78.40	85.02	80.05	**85.73**	**80.64**	89.06	86.48	89.64	86.66	**89.98**	**87.50**
ES	85.31	81.23	**85.90**	**81.73**	85.80	81.45	85.41	80.50	**86.46**	81.37	86.07	**81.41**
PT	88.40	86.54	89.12	87.32	**89.40**	**87.60**	**90.16**	**85.53**	89.50	85.03	89.98	85.23
SV	82.61	77.42	**85.39**	80.60	85.29	**81.22**	79.61	72.71	82.91	74.96	**84.86**	**77.36**
AVG	*85.14*	*80.90*	*86.35*	*82.43*	***86.56***	***82.73***	*86.06*	*81.31*	*87.13*	*82.01*	***87.72***	***82.88***

为了进一步与Li等人的单语异构树库融合工作[148]进行比较，我们在中文树库CTB5与CDT上进行了实验。参考Li等人的实验设置，我们考虑使用自动词性以及正确词性两种情况。实验结果如表 4-6所示。需要注意的是，与本文所使用的基于转移的依存句法分析模型不同，他们使用的是基于图的二阶模型[129]，因此我们的基准系统略低。尽管如此，相比之下我们的MTL系统在两种情况下都取得了更加显著的提升。这也证明了我们方法的优势。

表 4-6 中文上CTB5与CDT的融合实验结果。LI12-O2采用基于图的二阶模型，并使用兄弟结构以及祖孙结构特征；而LI12-O2SIB只使用兄弟结构特征。

Table 4-6 Parsing accuracy comparisons of MONOLINGUAL (CDT→CTB5). LI12-O2 use the O2 graph-based parser with both sibling and grandparent structures, while LI12-O2SIB only use the sibling parts.

		Auto-POS			Gold-POS		
		SUP	CAS	MTL	SUP	CAS	MTL
OURS	UAS	79.34	80.25 (+0.91)	**81.13 (+1.79)**	85.25	86.29 (+1.04)	**86.69 (+1.44)**
	LAS	76.23	77.26 (+1.03)	**78.24 (+2.01)**	83.59	84.72 (+1.13)	**85.18 (+1.59)**
		SUP	with QG		SUP	with QG	
LI12-O2	UAS	79.67	81.04 (+1.37)		86.13	86.44 (+0.31)	
LI12-O2SIB	UAS	79.25	80.45 (+1.20)		85.63	86.17 (+0.54)	

[4]葡萄牙语在UNIV→CONLL的情况下，CAS与MTL表现均不及SUP。与UDT作者Ryan McDonald讨论之后，我们认为，导致这种现象的原因是葡萄牙语UDT的自动转化过程存在缺陷，其中包含很多错误标注。

4.6 本章小结

本章可以认为是第3章工作的延续。我们探讨在目标树库具备一定规模的情况下，如何利用已有的树库来增强目标树库上的分析。我们考虑了多种类型的源树库，包括跨语言的通用树库以及单语异构树库，并提出一个通用的深度多任务学习框架，使得源树库与目标树库能够有效地融合。该框架主要基于多任务学习的思想，利用深度神经网络对于特征表示的学习能力与抽象特性，通过选择性的参数共享来实现不同树库学习时的信息交互，以提升目标树库上的分析性能。

我们在多种语言、多种设置下进行了充分的实验，并与传统的有监督学习系统以及级联学习系统进行了对比。实验结果表明，我们的框架在绝大多数情况下能够取得比基准系统更优的性能，证明了多任务学习在树库融合上的有效性。同时，我们也发现，在目标树库资源较为稀缺的情况下，采用级联学习是一种更有效的做法。

在本章的工作中，参数共享策略的制定仍然依赖少量的先验知识（比如跨语言的类型学特征、形态学特征等）。未来我们希望找到一种能够自动发现神经网络中特征表示之间相关性的机制，从而实现自适应的参数共享。

第 5 章 面向语义角色标注与关系分类的统一模型

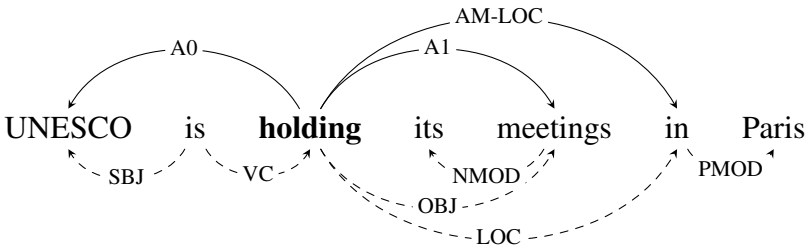
5.1 引言

在前几章的工作中，我们研究了不同语言、不同数据类型之间如何通过分布表示以及神经网络来实现知识迁移。在处理多类型树库融合问题时，我们采用的是一种基于深度多任务学习的框架，将每种树库的学习视作一个任务并通过参数共享来实现任务之间的信息交互。在本章中，我们将该思想延展至语义分析的层次，并考察在真实的多任务场景下，如何有效地进行迁移学习。

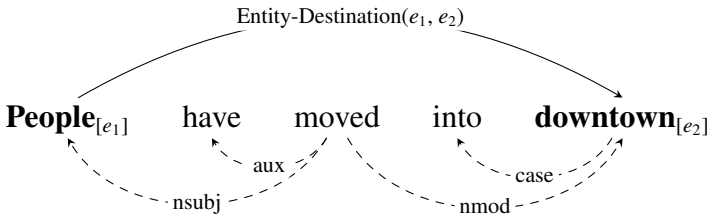
语义关系的识别与分类是自然语言理解中的重要问题。很多自然语言处理任务都涉及对句子中不同词项（或者成分）之间的语义关系进行分析，较为典型的有语义角色标注（SRL）、关系抽取（Relation Extraction, RE）与分类（Relation Classification, RC）等。不同任务对于语义关系的定义不尽相同。比如，语义角色标注任务中定义了谓词与论元之间的语义依存关系，进而反映该论元的语义角色（包括施事者、受事者、时间、地点、方式等）。而在关系抽取与分类任务中，则面向的是实体或者名词之间的语义关系，如“产品–生产者”（Product-Producer），“部分–整体”（Component-Whole）等。可以看出，后者所定义的关系类型所表达的语义层次更深，而语义角色标注任务主要表达的是浅层语义。图 5-1分别给出了两个任务下的语义关系标注示例。

在之前的研究中，语义角色标注与关系抽取/分类被认为是两个截然不同的任务，其研究群体也互不相同。由此而产生的一个问题是两者之间在数据资源、模型方法上的联系通常被忽略。在本研究中，我们发现语义角色标注与关系分类任务存在多方面的一致性，并且可以通过一个统一的模型来解决。为了说明这一点，我们对两个任务的主流模型中所用到的关键特征进行分析：

- **上下文特征**。对于大部分自然语言处理任务（如分词、词性标注、句法分析）而言，上下文特征都非常重要。在语义关系的分类中，上下文也同样重要。以图 5-1 b)中的句子为例，待标注的两个实体（名词）分别为**People**与**downtown**，关系为“实体–目的地”（Entity-Destination），显然，上下文中“moved”为该关系的判别提供了非常重要的证据，起到了“触发词”的作用。然而，目前在SRL与RC上的绝大部分研究使用特征工程的方式来提取上下文特征，通常只考虑局部窗口中的上下文。在语义任务中，这种方式极有可能覆盖不到真正的“触发词”。



a) Semantic Role Labeling.



b) Relation Classification.

图 5-1 语义角色标注（a）与关系分类（b）的标注示例。注意到两者可能采用不同的依存句法标注。

Figure 5-1 Examples of *semantic role labeling* (a) and *relation classification* (b). Note that they may use different syntactic parse annotations.

- **句法特征**。语义角色标注与关系分类任务的另一个重要的共同点是对于句法特征的依赖性。在图 5-1中的两个示例中，我们同时也标注了其句法结构。以语义角色标注为例（图 5-1 a)），论元候选词“meetings”与谓词“**holding**”之间的句法依存关系（“holdings” $\xrightarrow{OBJ}$ “meetings”，OBJ表示动宾关系）表明其语义角色为A1（受事者）的概率非常大。在之前的语义角色标注研究中，人们通常使用由词性或者依存关系所构成的句法路径特征（离散表示），而由词所构成的路径特征因为过于稀疏而鲜被使用。
- **词汇语义特征**。对于（深层）语义分析任务而言，词汇语义特征尤为重要。在很多情况下，待分类的名实体本身所蕴含的语义属性决定了其语义关系。因此，近年来，越来越多的语义分析研究开始使用词语分布表示、原型等特征，甚至从WordNet，同义词词林等知识库中抽取相应的语义类别[169, 170]。

在本研究中，我们首先提出一个用于语义角色标注与关系分类的统一模型。该模型使用双向长短时记忆网络来获取全局的上下文特征以及句法路径

特征，同时使用丰富的词汇语义分布表示特征作为模型的输入。与关系分类有所不同的是，语义角色标注实际上是一个结构预测问题（structure prediction），其输出结果需要满足一定的结构化约束。因此，对于语义角色标注任务，我们额外增加了一个基于整数线性规划（Integer Linear Programming, ILP）的后推断（Post-Inference, PI）过程。该模型为SRL与RC这两个任务建立了一座桥梁，使得我们能够进一步利用多任务学习来实现两者之间的迁移学习。

我们使用CoNLL-2009评测[171]所提供的语义角色标注数据，以及SemEval-2010任务8[172]所提供的关系分类数据进行实验。在语义角色标注上，我们的模型在所有语言上的表现都显著优于当前最好的系统；特别的，在中文上，我们的性能比当前最好的系统高了6个百分点（F1值）；在关系分类任务上，我们也取得了与当前最好的结果相近的水平。另外，多任务学习能够为语义角色标注的性能带来进一步的显著提升。

5.2 背景与相关工作

5.2.1 语义角色标注

语义角色标注是浅层语义分析（Shallow Semantic Parsing）的一种实现方式[173]。旨在识别句子中给定谓词的语义角色（Argument）。例如在图 5-1 a)的例子中，对于谓词“**holding**”，“UNESCO”、“meetings”、“in Paris”分别为其施事（A0）、受事（A1）以及动作发生的地点。语义角色标注综合考虑了分词、词性标注、句法分析等信息，因此早期的研究主要是通过从这些信息中挖掘丰富的特征，再结合传统的机器学习算法来训练语义角色标注模型。Gildea与Jurafsky最早采用这种方法并提出了一系列有效的特征集合[174]，并启发了之后的大量相关工作。比如，在CoNLL-2009评测任务中，表现最好的系统一共采用了50多个语言相关的特征模板[70]。这些特征主要考虑了谓词、论元候选词、上下文以及句法路径等[175–177]。同时，有的工作中也考虑了更高阶的特征，比如多个论元或者多谓词等[178–180]。

为了减少对特征工程的依赖，同时增加模型的泛化性，人们大致进行了三个方向的探索。首先是核方法。Moschitti等人在语义角色标注引入卷积树核（tree kernel），以捕捉句法特征之间的结构相似性[181, 182]；Che等人对其进行了改进，提出了混合卷积树核，并与二次多项式核进行结合，更显著地提升了语义角色标注的性能[183]。尽管核方法能够进行有效的自动特征学习，但是其计

算代价通常较高，扩展性不足。其次是基于张量的方法。Lei等人利用四阶低秩张量分解实现了（谓词，论元，上下文，句法）四类特征之间的组合[184]。这种方法的缺点在于其无法有效地对句法结构特征进行泛化。同时，他们的工作仍然依赖特征工程的方式来抽取局部上下文特征。近年来，越来越多的研究开始转向特征学习能力更强的神经网络模型[9, 10, 185, 186]。最近，Roth和Lapata采用长短时记忆网络模型来学习句法路径的分布表示[187]，显著地提升了语义角色标注的性能。该思想与本文的工作相似。同时，对于其他特征（如上下文），他们借鉴了Anders等人[188]所采用的离散特征。

5.2.2 关系分类

和语义角色标注类似，针对早期的关系抽取与分类研究过于依赖特征工程[169]的问题，人们同样探索了核方法[189]与张量方法[190, 191]。而近年来，基于神经网络的关系分类已经成为绝对的主流。Socher等人最早使用递归神经网络来学习待分类的两个实体之间句法路径的表示[89]。Zeng等人则使用卷积神经网络来学习句子的全局表示以用于关系分类[192]。他们的模型在不使用句法特征的条件下，取得了当年的最好结果。这也充分证明了神经网络在该任务上的有效性。受该工作启发，很多学者开始尝试使用更复杂的模型来更好地对句子中的长距离依赖进行建模，比如链式结构或者树状结构的长短时记忆网络[193–195]。

5.3 问题定义

在本小节中，我们给出本研究中对于语义角色标注与关系分类这两个任务的详细定义。

对于语义角色标注，我们采用CoNLL-2009评测任务中的定义。在该评测所提供的数据中，每个句子（**s**）的所有词项都被标注了自动词性以及原型，有些词被标记成了谓词。此外还提供了每个句子的依存句法自动分析结果（y_{syn}）。该任务的目标是，对于句中的每个谓词p_i，预测出其语义依存结构，即：标注出该谓词的所有论元及其语义角色类型[1]。

事实上，一个完整的语义角色标注任务还包含谓词的识别以及词义消歧。而在本研究中，我们假设谓词是给定的，而将任务的重心放在论元的识别与

[1]PropBank中包含20多种语义角色，其中核心的语义角色为A0~A5六种。A0通常表示动作的施事者，A1通常表示动作的影响（如受事者），A2~A5则根据谓词的不同会有不同的定义。

分类任务上。形式化地，对于**s**中的一个谓词p_i，我们遍历**s**中除p_i之外的所有词项：$\{w \in \mathbf{s} | w \neq p_i\}$，并判断它们的语义角色类别。可以将这个过程看作对于词对$\langle p_i, w \rangle$的一个分类任务。除了标准的语义角色类别集合，我们还增加了一个额外的*NULL*类别，表示该词不是p_i的一个论元。另外，为了保证对于一个句子的标注结果满足某些结构上的约束，我们进一步采用整数线性规划（ILP）对句子标注结果进行后推断（第 5.4.4节）。

对于关系分类，从图 5-1 b)中的示例可以看出，其中定义的语义关系类型和语义角色标注截然不同。我们采用SemEval-2010评测任务8[169]中的定义。在该任务中，每个句子都标注了一对名实体（nominal）e_1与e_2，而我们的目标则是对该实体对进行分类。该任务一共定义了九种关系类型，分别为：Cause-Effect，Entity-Origin，Message-Topic，Product-Producer，Entity-Destination，Member-Collection，Instrument-Agency，Component-Whole与Content-Container。此外，不在这9类之中的关系记为*Other*（无向）。由于在评价时我们既考虑关系类型也考虑关系的方向，因此一共有19种关系类别。

5.4 基于神经网络的统一模型

如前所述，语义角色标注与关系分类均可以形式化为对句子中词对进行分类的任务。我们提出一个统一的基于神经网络的模型来对两者进行建模，模型结构如图 5-2所示，其主要包含以下三个组成部分。

5.4.1 词汇语义特征表示

词汇语义特征对于语义关系的分类尤为重要。比如对于图 5-3中的例子，在不考虑两个实体“regulations”与“force”自身语义的情况下，根据其上下文特征（“entered into”）很容易将两者之间的语义关系错误标注为Entity-Destination。

因此，我们首先抽取句中每个词的词汇语义特征。基本的特征包括词（或者原型）以及词性。在关系分类任务中，通常人们还使用目标实体类别（NE）以及在WordNet中所定义的上位词特征。对于所有的特征，我们都采用其分布表示。对于词分布表示，可以利用*word2vec*在大规模外部资源上进行预训练。不同特征表示经过一层非线性变换进行组合，从而获得该词项的完整表示：

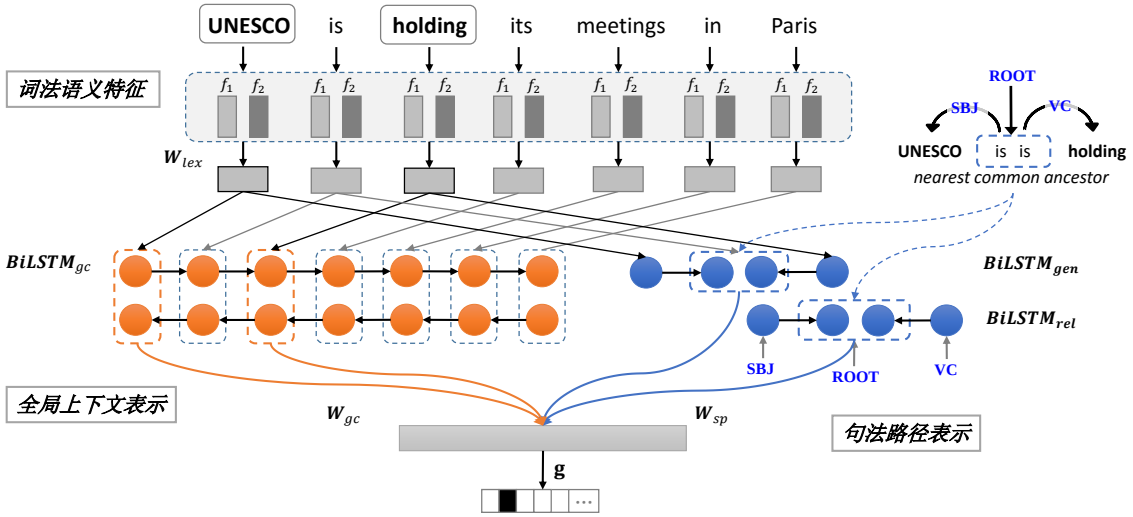


图 5-2 用于语义角色标注与关系分类的统一神经网络模型。

Figure 5-2 The unified architecture for SRL and RC.

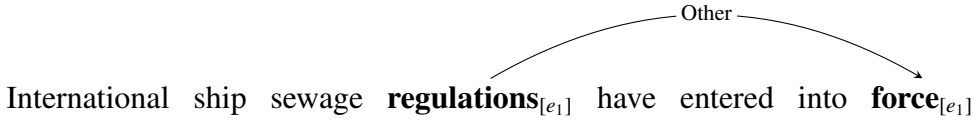


图 5-3 词汇语义特征重要性示例。

Figure 5-3 Illustration of the importance of lexical semantic features.

$$\boldsymbol{x}_i = \mathrm{ReLU}(\boldsymbol{W}_{lex} \cdot \boldsymbol{\Phi}_i + \boldsymbol{b}_{lex}),\ \text{其中:}$$

$$\text{对于 SRL:}\quad \boldsymbol{\Phi}_i = [\boldsymbol{e}(w_i) \oplus \boldsymbol{e}(t_i)] \tag{5-1}$$

$$\text{对于 RC:}\quad \boldsymbol{\Phi}_i = [\boldsymbol{e}(w_i) \oplus \boldsymbol{e}(t_i) \oplus \boldsymbol{e}(ne_i) \oplus \boldsymbol{e}(wn_i)]$$

其中w_i, t_i, ne_i, wn_i分别表示词（或原型）、词性、实体类型以及WordNet上位词。

5.4.2 全局上下文表示

接下来，我们使用双向长短时记忆网络来获取目标词对的全局上下文表示。具体的，将每个位置的词项表示作为双向LSTM的输入，并将目标词对所在位置的双向隐层表示进行拼接之后的向量作为其全局上下文表示：

$$\boldsymbol{R}_{e_1}^{gc} = [\overrightarrow{\boldsymbol{h}}_{e_1} \oplus \overleftarrow{\boldsymbol{h}}_{e_1}];\quad \boldsymbol{R}_{e_2}^{gc} = [\overrightarrow{\boldsymbol{h}}_{e_2} \oplus \overleftarrow{\boldsymbol{h}}_{e_2}] \tag{5-2}$$

需要注意的是，由于我们使用的是目标词对所在位置上的隐层表示，而非句子表示，因此我们无需使用相对位置等特征[10, 192, 196]。

5.4.3 句法路径表示

对于待分类的两个词项e_1, e_2，我们记它们的最近公共祖先结点为$nca(e_1, e_2)$，那么可以得到e_1, e_2分别到$nca(e_1, e_2)$的依存句法路径：$e_1 \rightarrow \ldots \rightarrow nca(e_1, e_2)$与$nca(e_1, e_2) \leftarrow \ldots \leftarrow e_2$。我们按照依存弧的指向采用双向LSTM对此路径进行建模，如图 5-2（右）所示。我们考虑两类句法路径，首先是由词法特征构成的路径，记为*generic path*；其次是由依存关系构成的路径，记为*relation path*。这两类路径分别由$\boldsymbol{BiLSTM}_{gen}$与$\boldsymbol{BiLSTM}_{rel}$进行建模。接着，我们同时将$nca(e_1, e_2)$处的双向词法特征路径表示以及双向依存关系路径表示进行拼接，从而获得完整的句法路径表示：

$$\boldsymbol{R}^{gen}_{(e_1,e_2)} = [\overrightarrow{\boldsymbol{h}}^{gen}_{nca(e_1,e_2)} \oplus \overleftarrow{\boldsymbol{h}}^{gen}_{nca(e_1,e_2)}]; \quad \boldsymbol{R}^{rel}_{(e_1,e_2)} = [\overrightarrow{\boldsymbol{h}}^{rel}_{nca(e_1,e_2)} \oplus \overleftarrow{\boldsymbol{h}}^{rel}_{nca(e_1,e_2)}] \tag{5-3}$$

接下来，全局上下文表示以及句法路径表示通过非线性层进行组合，从而获得最终的表示向量，并用于关系分类：

$$\boldsymbol{p} = \mathrm{ReLU}(\boldsymbol{W}_{gc} \cdot \underbrace{[\boldsymbol{R}^{gc}_{e_1} \oplus \boldsymbol{R}^{gc}_{e_2}]}_{\text{全局上下文表示}} + \boldsymbol{W}_{sp} \cdot \underbrace{[\boldsymbol{R}^{gen}_{(e_1,e_2)} \oplus \boldsymbol{R}^{rel}_{(e_1,e_2)}]}_{\text{句法路径表示}} + \boldsymbol{b}) \tag{5-4}$$

$$p(c|\boldsymbol{p}) = \mathrm{softmax}(\boldsymbol{g}_c \cdot \boldsymbol{p} + \boldsymbol{q}_c) \tag{5-5}$$

我们使用交叉熵损失函数对该模型进行训练：$\mathcal{L}(\theta) = -\sum_{i=0}^{N} \log p(c_i|\mathbf{p}_i)$，其中$N$为训练样本数目。

5.4.4 基于整数线性规划的后推断

语义角色标注实际上是一个结构预测问题，对于一个句子的预测结果需要满足一定的内在结构化约束。比如对于句中的一个谓词而言，大部分角色中通常只出现一次。而上述的基于分类的模型结构无法保证其语义角色标注结果满足这样的全局约束。受Punyakonok等人[197]与Che等人[198]工作的启发，我们在语义角色分类模型的输出（即句中每个词在语义角色类别集合上的概率分布）上进一步采用整数线性规划进行后推断，以获得句子上的全局最优标注结果。

令W为句子中词的集合，R为语义角色集合（包含*NULL*）。在给定谓词的条件下，对于每个词$w \in W$以及语义角色类别$r \in R$，我们使用一个二元变量$v_{wr} \in (0, 1)$来表示w的语义角色是否被标注为r。p_{wr}为w标记为r的概率，由公式 5-5计算得到。在后推断过程中，我们希望优化以下目标函数：

$$\mathcal{L}_{ilp} = \sum_{w,r} \log(p_{wr} \cdot v_{wr}) \tag{5-6}$$

通过求解v_{wr}的值，即可获得最终的语义角色标注结果。需要注意的是，虽然ILP是一个NP难的问题，但是对于问题规模（变量、约束数目）较小的情形，可以利用lp_solve[2]等工具进行高效地求解。

我们采用Che等人[198]的工作中所定义的三类约束：

- C1：标注结果中，每个词有且只能有一个类别（包含*NULL*）。即：

$$\sum_{r} v_{wr} = 1 \tag{5-7}$$

- C2：分类概率小于某个阈值（这里设为0.3）的语义角色不被标注（不包含*NULL*）。即：

$$v_{wr} = 0, \quad \text{if } p_{wr} < 0.3 \text{ and } r \neq NULL \tag{5-8}$$

- C3：一个谓词的大部分角色在句子中只出现一次，除了少量例外。该条件是语言相关的，因此，对于每种语言，我们都维护一个“单例语义角色”列表（R_{no_dup}，见表 5-1）。即：

$$\sum_{w} v_{wr} \leq 1, \quad \text{if } r \in R_{no\_dup} \tag{5-9}$$

表 5-1 不同语言所使用的“单例语义角色”列表。

Table 5-1 No-duplicated-roles for different languages.

Language	No-duplicated-roles
English	A0, A1, A2, A3, A4, A5
Chinese	A0, A1, A2, A3, A4, A5
Catalan	arg0-agt, arg0-cau, arg1-pat, arg2-atr, arg2-loc
German	A0, A1, A2, A3, A4, A5
Spanish	arg0-agt, arg0-cau, arg1-pat, arg1-tem, arg2-atr, arg2-loc, arg2-null, arg4-des, argL-null, argM-cau, argM-ext, argM-fin

5.5 多任务学习

我们已经对语义角色标注与关系分类在问题定义、所用方法上的共同点进行了充分的描述，并提出了一个统一的模型结构。这使得我们不禁思考：这两个任务之间能否进行融合，从而达到相互促进的效果? 根据最短句法路径假

[2]http://lpsolve.sourceforge.net/5.5/

设[199]，如果句子中的两个实体e_1, e_2之间存在某种语义关系的话，它们通常也是该句中某个谓词或者一连串谓词的论元。为了更直观的理解，我们考察图5-4中的例子。这里“author”与“disassembler”之间存在Instrument-Agency的关系，这就为我们提供了一个强烈的信息：“author”与“disassembler”分别是句中某个谓语动词的施事者（A0）与受事者（A1）。根据该句子的依存结构，可以很容易地分析出，其所对应的谓词是“uses”。

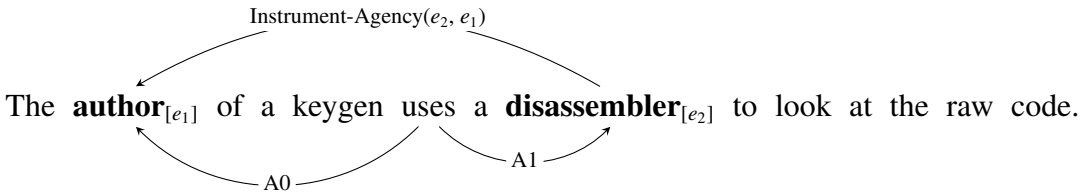


图 5-4 实体关系（上）和语义角色标注（下）的潜在联系。

Figure 5-4 The implicit connection between RC (top) and SRL (bottom).

可以看出，实体关系的识别有助于语义角色的判断。类似的，根据语义角色标注的结果（“author” $\overset{A0}{\longleftarrow}$ “uses” $\overset{A1}{\longrightarrow}$ “disassembler”），我们也能够推断出两者之间可能存在某种语义关系（关系抽取）。然而，其对于具体语义关系类别的判定（关系分类）却没有提供更多的证据。根据A0，A1语义角色，我们几乎难以判断两个词之间的语义关系是Instrument-Agency，Product-Producer还是Cause-Effect。

综上，我们认为，关系分类任务应有助于语义角色标注，而语义角色标注对于关系分类任务则没有帮助。其根本原因，在于关系分类中所定义的关系在语义粒度以及层次上比语义角色标注更细，更深。因此，本文将研究的重点放在语义角色标注任务的提升上。

基于第 5.4节中提出的统一模型，我们可以方便地通过参数共享来实现两个任务之间的知识迁移。与前一章的工作类似，我们考虑两种迁移学习机制：

- 级联学习（Cas）。在该方案中，我们首先训练一个关系分类模型，然后使用该模型中的参数（特征分布表示，网络权值）作为语义角色标注模型中参数的初始化。
- 多任务学习（MTL）。在该方案中，我们对语义角色标注与关系分类这两个任务进行联合学习。我们采用以下训练方式：
  1. 根据某种概率分布对任务进行采样；
  2. 从该任务中随机抽取一个批次的训练样本，执行前馈操作；
  3. 使用误差反向传播算法计算梯度，并更新相应任务下的参数；

4. 返回第1步。

在多任务学习中，有两个关键的因素需要考虑。首先，我们希望联合学习的两个任务接近于同时收敛[16]。为了达到这一目的，我们采用第4.5节中所介绍的加权任务采样方法。从单任务学习的实验中我们观察到，语义角色标注任务的收敛速率大约为关系分类任务的四分之一，因此我们以4:1（SRL:RC）的概率分布来进行任务采样。其次，多任务学习的关键在于参数共享。由于我们使用的是一个统一的模型，因此绝大部分参数都是可以在两个任务之间共享的。需要注意的是，不同任务中采用的依存句法标注可能不同，比如CoNLL-2009评测官方所提供的自动句法标注是CoNLL依存规范的，而大部分关系分类的研究则使用的是Stanford依存体系[200]。两者在核心结点的确定规则以及依存关系集上都有较大的差异，因此，与句法路径相关的参数都无法共享，包括$\boldsymbol{BiLSTM}_{gen}$, $\boldsymbol{BiLSTM}_{rel}$与$\boldsymbol{W}_{sp}$（见图5-2）。

5.6 实验与分析

5.6.1 实验设置

在语义角色标注任务上，我们使用CoNLL-2009评测所提供的英文数据进行实验。另外，我们也在中文、加泰罗尼亚语（Catalan），德语以及西班牙语上进行了多语实验。实验中采用标准的训练/开发/测试集划分。此外，评测中还提供了Brown语料的一个子集作为跨领域测试数据。为了观察我们模型在不同领域上的泛化性，我们也在英文的Brown测试数据上进行了评价。具体数据规模及划分如表5-2所示。本实验使用评测中所提供的自动词性、词的原型以及依存句法分析结果作为输入。在训练及测试过程中，我们假定所有的谓词均已被标注；同时，我们也不对谓词进行词义消歧，从而将研究的重心放在语义角色的识别与分类任务上。在对于大部分实验结果的评价中，我们采用不包含谓词词义的评价指标。对于某些情况，为了与之前研究工作中所公布的实验结果进行直接的对比，我们采用Lei等人[184]的方式，将Anders等人[188]的谓词词义分类结果与本实验中所得到的语义角色标注结果进行合并，从而获得包含谓词词义的评价结果。

我们与以下系统进行实验对比：

- CoNLL-2009评测中语义角色标注结果排名第一的系统；
- Roth与Lapata所提出的基于PathLSTM的模型[187]；

表 5-2 实验中使用的CoNLL-2009语义角色标注数据统计。

Table 5-2 Statistics of the SRL data from CoNLL-2009.

Language	Train	Dev	Test	OOD-Test
English	39,279	1,334	2,399	425
Chinese	22,277	1,762	2,556	–
Catalan	13,200	1,724	1,862	–
German	36,020	2,000	2,000	–
Spanish	14,329	1,655	1,725	–

- FitzGerald等人所提出的神经网络模型[186]；
- Lei等人所提出的张量分解模型[184]。
- Anders等人所采用的基于*liblinear*的分类模型[188]。

同时，我们也与上述模型的重排序改进结果以及多模型集成结果进行了对比。

对于关系分类任务，我们使用SemEval 2010任务8所提供的评测数据集。该数据集中一共标注了10,717个句子，其中训练集8,000句，测试集2,717句。我们采用5–折交叉验证的方式选择最佳迭代次数。我们对比了以下系统：

- SemEval-2010任务8中排名第一的系统（SVM）[169]；
- Socher等人提出的矩阵–向量递归神经网络模型（MVRNN）[89]；
- Zeng等人提出的卷积神经网络模型（CNN）[192]；
- Yu等人提出的基于张量的模型（FCM）[190]；
- dos Santos等人所采用的基于排序损失（ranking loss）的卷积神经网络模型（CR-CNN）[196]；
- 基于依存句法路径的神经网络模型（DepNN[193]，depLCNN[194]）。

我们在大规模未标注文本上使用*word2vec*对词汇分布表示进行预训练。对于英语、加泰罗尼亚语、德语以及西班牙语，我们采用最新的Wikipedia数据；对于中文，我们使用中文Gigaword第五版（LDC2011T13）中的新华社新闻语料（2000–2010），并使用语言技术平台（LTP）[201]进行分词处理。

在训练过程中，对于语义角色标注，我们使用动态批次（mini-batch）大小的随机梯度下降（SGD）算法来训练模型，其中一个批次为句中一个谓词所对应的实例集合；对于关系分类任务，由于每个句子只标注一对实体，因此批次大小为1。初始学习率设为：$\eta_0 = 0.1$，并随着迭代次数进行衰减：$\eta_t = \eta_0/(1 + 0.1t)$。表 5-3中为本实验中所使用的超参数设置。

表 5-3 实验中的超参数设置。

Table 5-3 Hyper-parameters settings.

Dimension of embeddings				Dimension of layers		
word	*POS*	*NE*	*WordNet*	*LSTM input*	*LSTM hidden*	*hidden*
200	25	25	25	100	100	200

5.6.2 语义角色标注实验结果

我们首先在英文数据上进行实验，结果如表 5-4所示。可以看出，我们的有监督学习系统（SUP）在同领域与跨领域测试数据上的性能优于6个基准系统（表的中间区域），同时，与两个使用重排序或者多模型集成的系统性能相当（表的下方区域）。

5.6.2.1 关系分类任务对于语义角色标注的影响

通过级联学习系统（CAS）、多任务学习系统（MTL）与SUP系统的对比，我们可以发现，关系分类任务的引入能够显著地提升语义角色标注模型的性能。另外，MTL始终优于CAS，这也进一步支撑了我们在前一章中关于联合学习与级联学习优劣对比的结论。我们的多任务学习系统（MTL）性能超越了之前所有的系统，取得了当前在CoNLL-2009评测数据上的最好实验结果。

我们进一步观察了SUP、CAS与MTL训练过程中在开发集上性能的变化曲线，如图 5-5所示。可以看出，在迭代初期，CAS与MTL性能非常接近且优于SUP，甚至在某些时刻优于MTL。这说明了关系分类任务确实能为语义角色标注模型提供一个更好的参数初始化。随着迭代次数的增加，MTL逐渐显现出联合学习的优势。

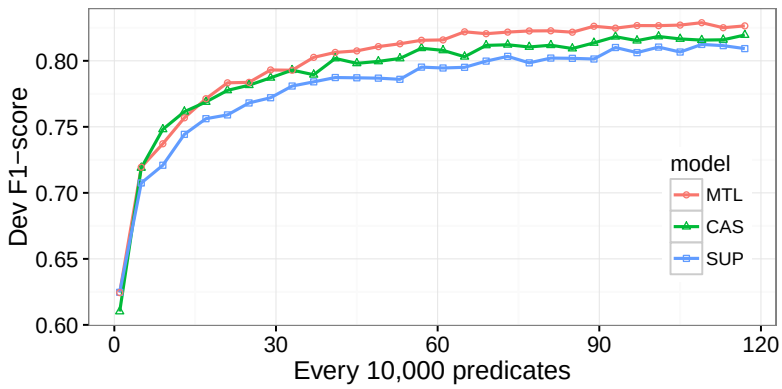


图 5-5 训练过程中模型的F1值在开发集上的变化曲线。

Figure 5-5 SRL F1-scores on the development data w.r.t. the number of predicates trained.

表 5-4 CoNLL-2009实验数据上的语义角色标注实验结果（F1），以及与其他基准系统的对比。显著性检验结果表明MTL以高于99%的置信度优于SUP。

Table 5-4 SRL labeled F1-score of our model variants, compared with the state-of-the-art systems on the CoNLL-2009 shared task. Statistical significance (MTL *vs.* SUP) with $p < 0.01$ is marked with *.

Model	**Excluding pred. senses**			**Including pred. senses**	
	WSJ-dev	WSJ-test	Brown-test	WSJ-test	Brown-test
SUP	82.32	84.06	72.12	87.67	76.56
CAS	83.33	84.73	73.00	88.14	77.15
MTL	**83.51***	**85.04***	**73.22***	**88.37***	**77.34***
CoNLL-2009 1st place	–	82.08	69.84	86.15	74.58
(Roth and Lapata, 2016)[187]	–	–	–	86.7	75.3
(FitzGerald et al., 2015)[186]	82.3	83.6	71.9	87.3	75.2
(Lei et al., 2015)[184]	81.03	82.51	70.77	86.58	75.57
(Roth and Woodsend, 2014)[202]	–	80.87	69.33	85.50	74.67
(Anders et al., 2010)[188]	78.85	81.35	68.34	85.80	73.92
Model + Reranker/Ensemble	WSJ-dev	WSJ-test	Brown-test	WSJ-test	Brown-test
(Roth and Lapata, 2016)+R,E	–	–	–	87.9	76.5
(FitzGerald et al., 2015)+E	83.0	84.3	72.4	87.8	75.5
(Roth and Woodsend, 2014)+R	–	82.10	71.12	86.34	75.88
(Anders et al., 2010)+R	80.50	82.87	70.91	86.86	75.71

5.6.2.2 后推断过程的影响

接下来，我们观察基于整数线性规划（ILP）的后推断过程对于语义角色标注性能的影响。如表 5-5所示，ILP对于SUP、CAS以及MTL均能够起到一定的提升作用。

5.6.2.3 多语言语义角色标注实验结果

此外，我们也在CoNLL-2009评测任务中所提供的其他几种语言上进行了实验。由于在其他语言上我们没有相应的关系分类数据来进行迁移学习，因此我们只考虑单个语义角色标注模型，也就是SUP模型。实验结果如表 5-6所示。可以看出，我们的模型在四种语言上均优于当前最好的系统。尤其在中文上，我们的F1值比最好的结果高6.3%。需要注意的是，我们没有针对不同语言进行特征或者超参数上的调整。

表 5-5 基于整数线性规划的后推断过程的影响。

Table 5-5 Effect of post-inference, evaluated excluding predicate senses.

Model	WSJ-dev	WSJ-test
SUP	82.32	84.06
w/o ILP	81.87	83.53
CAS	83.33	84.73
w/o ILP	82.90	84.40
MTL	83.51	85.04
w/o ILP	83.15	84.75

表 5-6 中文、加泰罗尼亚语、德语以及西班牙语上不含谓词词义的评价结果。

Table 5-6 SRL labeled F1-score excluding predicate senses on Chinese, Catalan, German and Spanish. All results are evaluated excluding predicate senses.

Language	Test set			
	Ours	(Lei et al., 2015)[184]	CoNLL 1st	CoNLL 2nd
Chinese	**75.46**	69.16	68.52	68.71
Catalan	**79.24**	74.67	76.78	74.02
German	**77.41**	76.94	74.65	76.27
Spanish	**79.17**	75.58	77.33	74.01

5.6.3 关系分类实验结果

在关系分类实验中，我们使用了两类额外的特征，分别是命名实体类型以及WordNet上位词。在SemEval 2010任务8上的实验结果如表 5-7所示。我们的模型取得了83.9%的F1值，与dos Santos等人[196]的结果（84.1%）相当。其中dos Santos等人的系统中针对*Other*关系所带来的混乱进行了特殊的处理。而类似的处理方式也可以用于我们的模型，以取得更高的性能。

根据第 5.5节中的分析，语义角色标注对于关系分类任务是没有促进作用的。为了验证该假设，我们设计了相应的级联学习与多任务学习实验。结果表明，CAS与MTL相比于SUP系统，性能均有一定的下降，其中CAS降低了0.9%，MTL降低了0.7%。

5.7 本章小结

本章将基于分布表示以及神经网络的多任务学习思想延展至语义分析层

表 5-7 关系分类实验结果，以及与之前系统的对比。

Table 5-7 Comparison with previously published results for SemEval 2010 Task 8.

Model	**Features**	**F1**
SVM[169] (Best in SemEval 2010)	POS, prefixes, morphological, WordNet, Levin classes, PropBank, FrameNet, dependency parse, NomLex-Plus, Google n-gram, paraphrases, TextRunner	82.2
MVRNN[89]	syntactic parse	79.1
MVRNN[89]	syntactic parse, POS, NER, WordNet	82.4
CNN[192]	position, WordNet	82.7
FCM[190]	dependency path, NER	83.0
DepNN[193]	dependency parse, NER	83.6
CR-CNN[196]	position	**84.1**
depLCNN[203]	WordNet, words around nominals	83.7
Ours	dependency path, POS, NER, WordNet	**83.9**
Model + Ensemble/Additional data		
ER-CNN+R-RNN[204]	position	84.9
depLCNN+NS[203]	WordNet, words around nominals	85.6

次。首先，我们针对语义角色标注任务与关系抽取任务在问题定义、特征以及模型上的共同点提出了一个统一的神经网络模型。该模型能够有效地表达目标词对的全局上下文特征以及词汇化的句法路径特征。接着，我们对两个任务之间可能存在的相互促进作用进行了分析，并提出基于级联学习以及多任务学习的方式，利用关系抽取任务中所蕴含的知识来提升语义角色标注的性能。

我们在CoNLL-2009语义角色标注评测数据以及SemEval-2010任务8所提供的关系分类数据上进行了实验。实验结果表明，我们的模型在多种语言上均取得了当前最高的语义角色标注性能，尤其在中文上，相比于之前发表的最好结果提升了6.3%。另外，迁移学习的实验结果也表明，与关系分类任务的联合学习能够显著地提升语义角色标注系统的性能。

在本章的工作中，我们主要面向的是语义关系的分类。从实验中可以看出，两个任务的联合学习收益是单向的，其中语义角色标注对于关系分类任务并没有提升。究其根源，主要在于两个任务之间所定义的关系类型的语义层次不同。在未来的工作中，我们希望能够有效地联合更深层次的语义分析技术（如语义依存分析）以及信息抽取任务。

结　论

特征表示是统计自然语言处理中的基础工作，好的特征表示对于模型的可扩展性以及泛化能力至关重要。目前主流的特征表示方法可以概括为高维离散的符号表示以及低维连续的分布表示，其中分布表示在近年来受到越来越广泛的关注，也为深度学习在自然语言处理领域取得的成功发挥了重要的作用。

现有的大多数研究主要集中在分布表示学习本身，以及结合深度神经网络在具体任务上的应用。本文则重点关注分布表示的另一个重要性质，即其在语义表示上的通用性。利用这种通用性，我们可以在不同语言、不同数据类型，甚至不同任务之间进行知识的迁移。迁移学习对于统计自然语言处理而言具有重要的意义。一方面，通过融合不同来源的数据，可以在一定程度上缓解统计模型所面临的数据稀疏问题。另一方面，不同数据或者任务之间蕴含一定的归纳偏置，这种偏置能够有效地防止有监督学习模型在特定数据或者特定任务上的过拟合，从而增强其泛化能力。

本文从自然语言处理的三个主要研究层面：词法，句法以及语义分析出发，系统深入研究了基于分布表示的跨语言、跨任务、跨数据类型知识迁移在不同任务中的关键技术。具体地讲，本文的贡献主要表现在以下几个方面：

1. 提出了基于跨语言资源的词义表示学习方法。传统的词分布表示无法有效地表达自然语言中“一词多义”的现象。我们利用跨语言平行文本中所蕴含的词义映射关系来对单语文本进行词义归纳，进而使用循环神经网络语言模型来学习词义分布表示。同时，我们也提出了一套基于循环神经网络的词义消歧算法，使得词义分布表示能够以特征形式被用于命名实体识别任务。实验表明，词义分布表示不仅能够更好地表达多义词之间的语义相似性，而且也能够显著地提升命名实体识别的性能。

2. 提出了基于分布表示学习的跨语言依存句法分析。针对跨语言依存句法分析中的“词汇化特征鸿沟”问题，我们设计了多种有效的跨语言分布表示学习算法，使得不同语言的词语能够映射至统一的语义向量空间之内，从而为不同语言之间的词汇化知识迁移构建了一座桥梁。接着，在基于特征分布表示与神经网络模型的依存句法分析框架之下，我们可以利用资源丰富的源语言数据来实现资源稀缺语言的自动依存句法分析。同时，针对不同语言由于类型学特征差异而导致的句法结构不一致的情况，我们提出了多源语言下的分布表示学

习技术。实验表明，跨语言词汇分布表示能够非常显著地提升对于资源稀缺语言的句法分析性能。

3. 提出了基于深度多任务学习模型的多类型依存句法树库融合方法。不同类型的树库之间存在一定的相似性，也存在不一致之处。如何充分利用不同树库之间的相似性，而规避其中的不一致性，进而提升目标树库上的分析性能，是我们所关注的主要问题。我们将每种树库下的学习过程视为一个独立的任务，并利用深度多任务学习来实现不同表示层次之间的信息传递或迁移。在实验中，我们具体考虑了多语言同构树库以及单语异构树库之间的融合，结果表明，该框架能够有效地实现不同树库之间的迁移学习，从而显著提升目标树库上的句法分析效果。

4. 提出了面向语义角色标注与关系分类的统一模型。我们进一步将深度多任务学习的思想扩展至语义分析的层次，并研究两种语义关系分类任务（语义角色标注、关系分类）之间的联合学习收益。针对这两个任务之间的共同点，我们首先提出了一个通用的神经网络模型。在此基础之上，我们能够方便地通过参数共享来实现不同任务之间的知识迁移。实验结果表明，该模型在语义角色标注任务上取得了当前最好的水平；同时，通过多任务学习，关系分类任务能够进一步显著提升语义角色标注任务的性能。

总体而言，本文的研究内容在不同层次上证明了基于分布表示的模型对于不同语言、不同任务、不同类型数据之间知识迁移的有效性。同时，迁移学习作为一种相对于有监督、无监督学习更接近人类学习方式的机制，还有非常广阔的研究空间。从本文研究内容出发，我们认为对于以下几个问题，还需要做进一步的研究：

1. 任务相关的跨语言表示学习。从目前的研究现状来看，不同的跨语言分布表示学习方法在不同自然语言处理任务上的表现各有优劣。比如双语自编码器模型在跨语言文本分类上表现较好，而对于句法分析则几乎没有帮助。对于情感分析等任务而言，或许情感极性的对齐比词义对齐更为重要。因此，未来的一个研究方向是学习任务相关的跨语言分布表示。

2. 有限资源条件下的跨语言表示学习。目前表现较好的跨语言表示仍然严重依赖双语平行资源，而对于大部分真正的资源稀缺语言而言，大规模双语数据难以获取。因此，如何充分利用单语数据以及代价相对较低的小规模词典资源来学习高质量的跨语言分布表示，也是一个非常值得探索的方向。

3. 任务表示相关性的自动建模。在本文的多任务学习研究中，我们所采用的参数共享策略是根据对于任务本身的先验知识来确定的。在未来的研究中，

我们希望找到一种有效的机制，能够自动地评估多任务深度神经网络中不同表示层次之间的任务相关性，从而自适应地进行参数共享。

参考文献

[1] Klein D, Manning C. Corpus-Based Induction of Syntactic Structure: Models of Dependency and Constituency[C]//Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume. Barcelona, Spain: [s.n.] , 2004:478–485.

[2] Liang P. Semi-supervised learning for natural language[M]. Cambridge, MA, USA: Massachusetts Institute of Technology, 2005.

[3] Zaidan O F, Callison-Burch C. Crowdsourcing Translation: Professional Quality from Non-Professionals[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics, 2011:1220–1229.

[4] Yarowsky D, Ngai G, Wicentowski R. Inducing multilingual text analysis tools via robust projection across aligned corpora[C]//Proceedings of the first international conference on Human language technology research. San Diego, CA, USA: Association for Computational Linguistics, 2001:1–8.

[5] Hwa R, Resnik P, Weinberg A, et al. Bootstrapping parsers via syntactic projection across parallel texts[J]. Natural language engineering, 2005, 11(03):311–325.

[6] Blitzer J, McDonald R, Pereira F. Domain Adaptation with Structural Correspondence Learning[C]//Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Sydney, Australia: Association for Computational Linguistics, 2006:120–128.

[7] Daume III H. Frustratingly Easy Domain Adaptation[C]//Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic: Association for Computational Linguistics, 2007:256–263.

[8] 张博, 史忠植, 赵晓非, et al. 一种基于跨领域典型相关性分析的迁移学习方法[J]. 计算机学报, 2015, 38(7):1326–1336.

[9] Collobert R, Weston J. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning[C]//Proceedings of the 25th International Conference on Machine Learning. 2008. Helsinki, Finland: ACM, ICML '08.

[10] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12:2493–2537.

[11] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on knowledge and data engineering, 2010, 22(10):1345–1359.

[12] 庄福振, 罗平, 何清, et al. 迁移学习研究进展[J]. 软件学报, 2015, 26(1):26–39.

[13] Yamada H, Matsumoto Y. Statistical dependency analysis with support vector machines[C]//Proceedings of the 8th International Workshop on Parsing Technologies (IWPT). Nancy, France: Association for Computational Linguistics, 2003:195–206.

[14] De Marneffe M C, Manning C D. The Stanford typed dependencies representation[C]//COLING 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation. Manchester, UK: Association for Computational Linguistics, 2008:1–8.

[15] McDonald R, Nivre J, Quirmbach-Brundage Y, et al. Universal Dependency Annotation for Multilingual Parsing[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Sofia, Bulgaria: Association for Computational Linguistics, 2013:92–97.

[16] Caruana R. Multitask learning[J]. Machine learning, 1997, 28(1):41–75.

[17] Berger A L, Pietra V J D, Pietra S A D. A maximum entropy approach to natural language processing[J]. Computational linguistics, 1996, 22(1):39–71.

[18] Collins M. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms[C]//Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. Philadelphia, PA, USA: Association for Computational Linguistics, 2002:1–8.

[19] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the eighteenth international conference on machine learning, ICML. Williamstown, MA, USA: Morgan Kaufmann Publishers Inc., 2001，1:282–289.

[20] Ernest Bellman R. Dynamic programming[M]. USA: Princeton University Press, 1957.

[21] Dumais S T. Latent semantic analysis[J]. Annual review of information science and technology, 2004, 38(1):188–230.

[22] Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research, 2003, 3:1137–1155.

[23] Firth J R. THE TECHNIQUE OF SEMANTICS.[J]. Transactions of the philological society, 1935, 34(1):36–73.

[24] Firth J R. A synopsis of linguistic theory 1930–1955[J]. Studies in linguistic analysis, 1957:1–32.

[25] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8):1798–1828.

[26] Cho K, van Merrienboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014:1724–1734.

[27] Vinyals O, Le Q. A neural conversational model[C]//Proceedings of ICML Deep Learning Workshop. Lille, France: JMLR, 2015.

[28] Hinton G, McClelland J, Rumelhart D. Distributed representations[J]. Parallel distributed processing: explorations in the microstructure of cognition, vol. 1, 1986:77–109.

[29] Hinton G E. Learning distributed representations of concepts[C]//Proceedings of the eighth annual conference of the cognitive science society. Amherst, Mass: Taylor & Francis Group, 1986.

[30] Landauer T K, Foltz P W, Laham D. An introduction to latent semantic analysis[J]. Discourse processes, 1998, 25(2-3):259–284.

[31] Schutze H. Dimensions of meaning[C]//Supercomputing'92., Proceedings. Minneapolis, MN, USA: IEEE, 1992:787–796.

[32] Padó S, Lapata M. Dependency-based construction of semantic space models[J]. Computational Linguistics, 2007, 33(2):161–199.

[33] Lebret R, Collobert R. Word Embeddings through Hellinger PCA[J]. EACL 2014, 2014:482.

[34] Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014:1532–1543.

[35] Xu W, Rudnicky A I. Can artificial neural networks learn language models?[C]//International Conference on Spoken Language Processing. Beijing, China: CMU, 2000:202–205.

[36] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]//Proceedings of 11th Annual Conference of the International Speech Communication Association (Interspeech). Makuhari, Chiba, Japan: ISCA, 2010:1045–1048.

[37] Mnih A, Hinton G E. A Scalable Hierarchical Distributed Language Model[M]//Advances in Neural Information Processing Systems 21. Vancouver, B.C., Canada: Curran Associates, Inc., 2009:1081–1088.

[38] Mnih A, Hinton G. Three New Graphical Models for Statistical Language Modelling[C]//Proceedings of the 24th International Conference on Machine Learning. 2007. Corvalis, Oregon, USA: ACM, ICML '07.

[39] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. International Conference on Learning Representations (ICLR) Workshop, 2013.

[40] Levy O, Goldberg Y. Dependency-Based Word Embeddings[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Baltimore, Maryland: Association for Computational Linguistics, 2014:302–308.

[41] Levy O, Goldberg Y. Neural word embedding as implicit matrix factorization[C]//Advances in neural information processing systems. Montreal, Canada: Curran Associates, Inc., 2014:2177–2185.

[42] Baroni M, Dinu G, Kruszewski G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Baltimore, Maryland: Association for Computational Linguistics, 2014:238–247.

[43] Levy O, Goldberg Y, Dagan I. Improving distributional similarity with lessons learned from word embeddings[J]. Transactions of the Association for Computational Linguistics, 2015, 3:211–225.

[44] Mikolov T, Le Q V, Sutskever I. Exploiting similarities among languages for machine translation[J]. arXiv preprint arXiv:1309.4168, 2013.

[45] Hardoon D R, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: An overview with application to learning methods[J]. Neural computation, 2004, 16(12):2639–2664.

[46] Faruqui M, Dyer C. Improving Vector Space Word Representations Using Multilingual Correlation[C]//Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden: Association for Computational Linguistics, 2014:462–471.

[47] Lu A, Wang W, Bansal M, et al. Deep Multilingual Correlation for Improved Word Embeddings[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado: Association for Computational Linguistics, 2015:250–256.

[48] Klementiev A, Titov I, Bhattarai B. Inducing Crosslingual Distributed Representations of Words[C]//Proceedings of COLING 2012. Mumbai, India: The COLING 2012 Organizing Committee, 2012:1459–1474.

[49] Zou W Y, Socher R, Cer D, et al. Bilingual Word Embeddings for Phrase-Based Machine Translation[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics, 2013:1393–1398.

[50] Chandar A P S, Lauly S, Larochelle H, et al. An Autoencoder Approach to Learning Bilingual Word Representations[C]//Advances in Neural Information Processing Systems. Montreal, Canada: Curran Associates, Inc., 2014:1853–1861.

[51] Hermann K M, Blunsom P. Multilingual Models for Compositional Distributed Semantics[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Baltimore, Maryland: Association for Computational Linguistics, 2014:58–68.

[52] Gouws S, Bengio Y, Corrado G. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments[C]//Proceedings of the 32nd International Conference on Machine Learning (ICML). Lille, France: JMLR W&CP, 2015:748–756.

[53] Dagan I, Lee L, Pereira F C. Similarity-based models of word cooccurrence probabilities[J]. Machine Learning, 1999, 34(1-3):43–69.

[54] Finkelstein L, Gabrilovich E, Matias Y, et al. Placing search in context: The concept revisited[C]//Proceedings of the 10th international conference on World Wide Web. Hong Kong, China: ACM, 2001:406–414.

[55] Jin P, Wu Y. SemEval-2012 Task 4: Evaluating Chinese Word Similarity[C]//*SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). Montréal, Canada: Association for Computational Linguistics, 2012:374–377.

[56] Socher R, Huang E H, Pennin J, et al. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection[M]//Advances in Neural Information Processing Systems 24. Granada, Spain: Curran Associates, Inc., 2011:801–809.

[57] Mikolov T, Yih W t, Zweig G. Linguistic Regularities in Continuous Space Word Representations[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia: Association for Computational Linguistics, 2013:746–751.

[58] 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 (Volume 1: Long Papers). Baltimore, Maryland: Association for Computational Linguistics, 2014:1199–1209.

[59] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C]//Advances in Neural Information Processing Systems. Stateline, NV, USA: Curran Associates, Inc., 2013:2787–2795.

[60] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion[C]//Advances in Neural Information Processing Systems. Stateline, NV, USA: Curran Associates, Inc., 2013:926–934.

[61] Brown P F, Desouza P V, Mercer R L, et al. Class-based n-gram models of natural language[J]. Computational linguistics, 1992, 18(4):467–479.

[62] Miller S, Guinness J, Zamanian A. Name Tagging with Word Clusters and Discriminative Training[C]//HLT-NAACL 2004: Main Proceedings. Boston, Massachusetts, USA: Association for Computational Linguistics, 2004:337–342.

[63] Huang F, Yates A. Distributional Representations for Handling Sparsity in Supervised Sequence-Labeling[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec, Singapore: Association for Computational Linguistics, 2009:495–503.

[64] Owoputi O, O'Connor B, Dyer C, et al. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia: Association for Computational Linguistics, 2013:380–390.

[65] Koo T, Carreras X, Collins M. Simple Semi-supervised Dependency Parsing[C]//Proceedings of ACL-08: HLT. Columbus, Ohio: Association for Computational Linguistics, 2008:595–603.

[66] Turian J, Ratinov L A, Bengio Y. Word Representations: A Simple and General Method for Semi-Supervised Learning[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: Association for Computational Linguistics, 2010:384–394.

[67] Wang M, Manning C D. Effect of Non-linear Deep Architecture in Sequence Labeling[C]//Proceedings of the Sixth International Joint Conference on Natural Language Processing. Nagoya, Japan: Asian Federation of Natural Language Processing, 2013:1285–1291.

[68] Guo J, Che W, Wang H, et al. Revisiting Embedding Features for Simple Semi-supervised Learning[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014:110–120.

[69] Zhang M, Zhang Y. Combining Discrete and Continuous Features for Deterministic Transition-based Dependency Parsing[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015:1316–1321.

[70] Che W, Li Z, Li Y, et al. Multilingual Dependency-based Syntactic and Semantic Parsing[C]//Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task. Boulder, Colorado: Association for Computational Linguistics, 2009:49–54.

[71] Chen D, Manning C. A Fast and Accurate Dependency Parser using Neural Networks[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014:740–750.

[72] Jordan M I. Serial Order: A Parallel Distributed Processing Approach[J]. ICS Technical Report 8604, 1986.

[73] Elman J L. Finding structure in time[J]. Cognitive science, 1990, 14(2):179–211.

[74] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[75] 刘伟权, 钟义信. 基于SRNN神经网络的汉语文本词类标注方法[J]. 计算机研究与发展, 1997, 34(6):421–426.

[76] 王海峰, 高文, 李生. 基于神经网络的汉语口语多义选择[J]. 软件学报, 1999, 10(12):1279–1283.

[77] 王海峰, 高文, 李生. 基于神经网络的汉语口语言语行为分析[J]. 计算机学报, 1999, 22(10):1014–1018.

[78] Mikolov T. Statistical Language Models Based on Neural Networks[D]. Czech Republic: Brno University of Technology, 2012.

[79] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8):1735–1780.

[80] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE transactions on neural networks, 1994, 5(2):157–166.

[81] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks.[J]. Proceedings of The 30th International Conference on Machine Learning (ICML), 2013, 28:1310–1318.

[82] Hubel D H, Wiesel T N. Receptive fields and functional architecture of monkey striate cortex[J]. The Journal of physiology, 1968, 195(1):215–243.

[83] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. Stateline, NV, USA: Curran Associates, Inc., 2012:1097–1105.

[84] Kim Y. Convolutional Neural Networks for Sentence Classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014:1746–1751.

[85] Johnson R, Zhang T. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado: Association for Computational Linguistics, 2015:103–112.

[86] Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification[C]//Advances in Neural Information Processing Systems. Montreal, Canada: Curran Associates, Inc., 2015:649–657.

[87] Socher R, Manning C D, Ng A Y. Learning continuous phrase representations and syntactic parsing with recursive neural networks[C]//Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop. Vancouver, Canada: Curran Associates, Inc., 2010:1–9.

[88] Socher R, Lin C C, Manning C, et al. Parsing natural scenes and natural language with recursive neural networks[C]//Proceedings of the 28th international conference on machine learning (ICML-11). Bellevue, WA, USA: JMLR W&CP, 2011:129–136.

[89] Socher R, Huval B, Manning C D, et al. Semantic Compositionality through Recursive Matrix-Vector Spaces[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea: Association for Computational Linguistics, 2012:1201–1211.

[90] Socher R, Pennington J, Huang E H, et al. Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions[C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK.: Association for Computational Linguistics, 2011:151–161.

[91] Socher R, Bauer J, Manning C D, et al. Parsing with Compositional Vector Grammars[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Sofia, Bulgaria: Association for Computational Linguistics, 2013:455–465.

[92] Tai K S, Socher R, Manning C D. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long

Papers). Beijing, China: Association for Computational Linguistics, 2015:1556–1566.

[93] Zhu X, Sobhani P, Guo H. Long short-term memory over recursive structures[C]//Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR W&CP, 2015:1604–1612.

[94] Tiedemann J. Rediscovering Annotation Projection for Cross-Lingual Parser Induction[C]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014:1854–1864.

[95] Tiedemann J. Cross-Lingual Dependency Parsing with Universal Dependencies and Predicted PoS Labels[J]. 2015:340–349.

[96] Tiedemann J. Improving the Cross-Lingual Projection of Syntactic Dependencies[C]//Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015). Vilnius, Lithuania: Linköping University Electronic Press, Sweden, 2015:191–199.

[97] Padó S, Lapata M. Cross-lingual annotation projection for semantic roles[J]. Journal of Artificial Intelligence Research, 2009, 36(1):307–340.

[98] McDonald R, Petrov S, Hall K. Multi-Source Transfer of Delexicalized Dependency Parsers[C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK.: Association for Computational Linguistics, 2011:62–72.

[99] Zhou H, Chen L, Shi F, et al. Learning Bilingual Sentiment Word Embeddings for Cross-language Sentiment Classification[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, 2015:430–440.

[100] Socher R, Ganjoo M, Manning C D, et al. Zero-shot learning through cross-modal transfer[C]//Advances in neural information processing systems. Stateline, NV, USA: Curran Associates, Inc., 2013:935–943.

[101] Socher R, Karpathy A, Le Q V, et al. Grounded compositional semantics for finding and describing images with sentences[J]. Transactions of the Association for Computational Linguistics, 2014, 2:207–218.

[102] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: Computer Vision Foundation, 2015:3156–3164.

[103] Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks?[C]//Advances in neural information processing systems. Montreal, Canada: Curran Associates, Inc., 2014:3320–3328.

[104] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

[105] Arora S, Li Y, Liang Y, et al. Linear algebraic structure of word senses, with applications to polysemy[J]. arXiv preprint arXiv:1601.03764, 2016.

[106] Reisinger J, Mooney R J. Multi-Prototype Vector-Space Models of Word Meaning[C]//Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, California: Association for Computational Linguistics, 2010:109–117.

[107] Huang E, Socher R, Manning C, et al. Improving Word Representations via Global Context and Multiple Word Prototypes[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Jeju Island, Korea: Association for Computational Linguistics, 2012:873–882.

[108] Mikolov T, Kombrink S, Deoras A, et al. RNNLM-Recurrent neural network language modeling toolkit[C]//Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding. Hawaii, USA: IEEE Signal Processing Society, 2011:1–4.

[109] Werbos P J. Generalization of backpropagation with application to a recurrent gas market model[J]. Neural Networks, 1988, 1(4):339–356.

[110] Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11):39–41.

[111] Iacobacci I, Pilehvar M T, Navigli R. SensEmbed: Learning Sense Embeddings for Word and Relational Similarity[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, 2015:95–105.

[112] Navigli R, Ponzetto S P. BabelNet: Building a Very Large Multilingual Semantic Network[C]//Proceedings of the 48th Annual Meeting of the Association for

Computational Linguistics. Uppsala, Sweden: Association for Computational Linguistics, 2010:216–225.

[113] Moro A, Raganato A, Navigli R. Entity linking meets word sense disambiguation: a unified approach[J]. Transactions of the Association for Computational Linguistics, 2014, 2:231–244.

[114] Chen X, Liu Z, Sun M. A Unified Model for Word Sense Representation and Disambiguation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014:1025–1035.

[115] Rothe S, Schütze H. AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, 2015:1793–1803.

[116] Tian F, Dai H, Bian J, et al. A Probabilistic Model for Learning Multi-Prototype Word Embeddings[C]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014:151–160.

[117] Gale W A, Church K W, Yarowsky D. Using bilingual materials to develop word sense disambiguation methods[C]//Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation. Montreal, Canada: TMI, 1992:101–112.

[118] Chan Y S, Ng H T. Scaling up word sense disambiguation via parallel texts[C]//Proceedings of the Twentieth AAAI Conference on Artificial Intelligence. Pittsburgh, PA, USA: AAAI Press, 2005，5:1037–1042.

[119] Liang P, Taskar B, Klein D. Alignment by Agreement[C]//Proceedings of the Human Language Technology Conference of the NAACL, Main Conference. New York City, USA: Association for Computational Linguistics, 2006:104–111.

[120] Apidianaki M. Translation-oriented Word Sense Induction Based on Parallel Corpora[C]//Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). Marrakech, Morocco: European Language Resources Association (ELRA), 2008, 28-30. http://www.lrec-conf.org/proceedings/lrec2008/.

[121] Frey B J, Dueck D. Clustering by passing messages between data points[J]. science, 2007, 315(5814):972–976.

[122] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python[J]. Journal of Machine Learning Research, 2011, 12:2825–2830.

[123] Finkelstein L, Gabrilovich E, Matias Y, et al. Placing Search in Context: The Concept Revisited[J]. ACM Transactions on Information Systems, 2002, 20(1):116–131.

[124] 董振东, 董强. 知网和汉语研究[J]. 当代语言学, 2001, 3(1):33–44.

[125] Dong Z, Dong Q. HowNet and the Computation of Meaning[M]. Singapore: World Scientific, 2006.

[126] 李正华. 汉语依存句法分析关键技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2013.

[127] Täckström O, McDonald R, Uszkoreit J. Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure[C]//Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Montréal, Canada: Association for Computational Linguistics, 2012:477–487.

[128] McDonald R, Crammer K, Pereira F. Online Large-Margin Training of Dependency Parsers[C]//Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05). Ann Arbor, Michigan: Association for Computational Linguistics, 2005:91–98.

[129] McDonald R T, Pereira F C. Online Learning of Approximate Dependency Parsing Algorithms[C]//Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics. Trento, Italy: The Association for Computer Linguistics, 2006:81–88.

[130] Carreras X. Experiments with a Higher-Order Projective Dependency Parser[C]//Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007. Prague, Czech Republic: Association for Computational Linguistics, 2007:957–961.

[131] Koo T, Collins M. Efficient Third-Order Dependency Parsers[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: Association for Computational Linguistics, 2010:1–11.

[132] Nivre J. An efficient algorithm for projective dependency parsing[C]//Proceedings of the 8th International Workshop on Parsing Technologies (IWPT). Nancy, France: Association for Computational Linguistics, 2003:149–160.

[133] Zhang Y, Nivre J. Transition-based Dependency Parsing with Rich Non-local Features[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics, 2011:188–193.

[134] Dyer C, Ballesteros M, Ling W, et al. Transition-Based Dependency Parsing with Stack Long Short-Term Memory[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, 2015:334–343.

[135] Weiss D, Alberti C, Collins M, et al. Structured Training for Neural Network Transition-Based Parsing[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, 2015:323–333.

[136] Zhou H, Zhang Y, Huang S, et al. A Neural Probabilistic Structured-Prediction Model for Transition-Based Dependency Parsing[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, 2015:1213–1222.

[137] Andor D, Alberti C, Weiss D, et al. Globally normalized transition-based neural networks[J]. arXiv preprint arXiv:1603.06042, 2016.

[138] Nivre J. Incrementality in deterministic dependency parsing[C]//Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together. Barcelona, Spain: Association for Computational Linguistics, 2004:50–57.

[139] Nivre J. Algorithms for deterministic incremental dependency parsing[J]. Computational Linguistics, 2008, 34(4):513–553.

[140] Nivre J. Non-Projective Dependency Parsing in Expected Linear Time[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing

of the AFNLP. Suntec, Singapore: Association for Computational Linguistics, 2009:351–359.

[141] Lei T, Xin Y, Zhang Y, et al. Low-Rank Tensors for Scoring Dependency Structures[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Baltimore, Maryland: Association for Computational Linguistics, 2014:1381–1391.

[142] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization[J]. Journal of Machine Learning Research, 2011, 12:2121–2159.

[143] Dyer C, Lopez A, Ganitkevitch J, et al. cdec: A Decoder, Alignment, and Learning Framework for Finite-State and Context-Free Translation Models[C]//Proceedings of the ACL 2010 System Demonstrations. Uppsala, Sweden: Association for Computational Linguistics, 2010:7–12.

[144] Zhang Y, Clark S. Syntactic processing using the generalized perceptron and beam search[J]. Computational Linguistics, 2011, 37(1):105–151.

[145] Nilsson J, Nivre J. MaltEval: an Evaluation and Visualization Tool for Dependency Parsing.[C]//Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). Marrakech, Morocco: European Language Resources Association (ELRA), 2008:161–166.

[146] Zhang Y, Barzilay R. Hierarchical Low-Rank Tensors for Multilingual Transfer Parsing[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015:1857–1867.

[147] Niu Z Y, Wang H, Wu H. Exploiting Heterogeneous Treebanks for Parsing[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec, Singapore: Association for Computational Linguistics, 2009:46–54.

[148] Li Z, Liu T, Che W. Exploiting Multiple Treebanks for Parsing with Quasi-synchronous Grammars[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Jeju Island, Korea: Association for Computational Linguistics, 2012:675–684.

[149] Johansson R. Training Parsers on Incompatible Treebanks[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia: Association for Computational Linguistics, 2013:127–137.

[150] Buchholz S, Marsi E. CoNLL-X Shared Task on Multilingual Dependency Parsing[C]//Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X). New York City: Association for Computational Linguistics, 2006:149–164.

[151] Smith D A, Eisner J. Parser Adaptation and Projection with Quasi-Synchronous Grammar Features[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2009:822–831.

[152] Rasooli M S, Collins M. Density-Driven Cross-Lingual Transfer of Dependency Parsers[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015:328–338.

[153] Duong L, Cohn T, Bird S, et al. A Neural Network Model for Low-Resource Universal Dependency Parsing[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015:339–348.

[154] Ammar W, Mulcaire G, Ballesteros M, et al. One Parser, Many Languages[J]. CoRR, 2016, abs/1602.01595.

[155] Chen W, Kazama J, Torisawa K. Bitext Dependency Parsing with Bilingual Subtree Constraints[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: Association for Computational Linguistics, 2010:21–29.

[156] Huang L, Jiang W, Liu Q. Bilingually-Constrained (Monolingual) Shift-Reduce Parsing[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2009:1222–1231.

[157] Burkett D, Klein D. Two Languages are Better than One (for Syntactic Parsing)[C]//Proceedings of the 2008 Conference on Empirical Methods in Natural

Language Processing. Honolulu, Hawaii: Association for Computational Linguistics, 2008:877–886.

[158] Hatori J, Matsuzaki T, Miyao Y, et al. Incremental Joint Approach to Word Segmentation, POS Tagging, and Dependency Parsing in Chinese[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Jeju Island, Korea: Association for Computational Linguistics, 2012:1045–1053.

[159] Li Z, Zhang M, Che W, et al. Joint Models for Chinese POS Tagging and Dependency Parsing[C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK.: Association for Computational Linguistics, 2011:1180–1191.

[160] Bohnet B, Nivre J. A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea: Association for Computational Linguistics, 2012:1455–1465.

[161] Henderson J, Merlo P, Titov I, et al. Multilingual joint parsing of syntactic and semantic dependencies with a latent variable model[J]. Computational Linguistics, 2013, 39(4):949–998.

[162] Lluís X, Carreras X, Màrquez L. Joint arc-factored parsing of syntactic and semantic dependencies[J]. Transactions of the Association for Computational Linguistics, 2013, 1:219–230.

[163] Dong D, Wu H, He W, et al. Multi-Task Learning for Multiple Language Translation[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, 2015:1723–1732.

[164] Luong M T, Le Q V, Sutskever I, et al. Multi-task sequence to sequence learning[J]. arXiv preprint arXiv:1511.06114, 2015.

[165] Liu Y, Li S, Zhang X, et al. Implicit Discourse Relation Classification via Multi-Task Neural Networks[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI). Phoenix, Arizona: AAAI Press, 2016:2750–2756.

[166] Liu P, Qiu X, Huang X. Recurrent Neural Network for Text Classification with Multi-Task Learning[C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York, USA: AAAI Press, 2016:2873–2879.

[167] Ballesteros M, Dyer C, Smith N A. Improved Transition-based Parsing by Modeling Characters instead of Words with LSTMs[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015:349–359.

[168] Petrov S, Das D, McDonald R. A Universal Part-of-Speech Tagset[C]//Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012). Istanbul, Turkey: European Language Resources Association (ELRA), 2012:2089–2096.

[169] Rink B, Harabagiu S. UTD: Classifying Semantic Relations by Combining Lexical and Semantic Resources[C]//Proceedings of the 5th International Workshop on Semantic Evaluation. Uppsala, Sweden: Association for Computational Linguistics, 2010:256–259.

[170] 李国臣, 吕雷, 王瑞波, et al. 基于同义词词林信息特征的语义角色自动标注[J]. 中文信息学报, 2016, 30(1):101–114.

[171] Hajič J, Ciaramita M, Johansson R, et al. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages[C]//Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task. Boulder, Colorado: Association for Computational Linguistics, 2009:1–18.

[172] Hendrickx I, Kim S N, Kozareva Z, et al. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals[C]//Proceedings of the 5th International Workshop on Semantic Evaluation. Uppsala, Sweden: Association for Computational Linguistics, 2010:33–38.

[173] 车万翔. 基于核方法的语义角色标注研究[D]. 哈尔滨: 哈尔滨工业大学, 2008.

[174] Gildea D, Jurafsky D. Automatic Labeling of Semantic Roles[J]. Computational Linguistics, 2002, 28(3):245–288.

[175] Surdeanu M, Harabagiu S, Williams J, et al. Using Predicate-Argument Structures for Information Extraction[C]//Proceedings of the 41st Annual Meeting of

the Association for Computational Linguistics. Sapporo, Japan: Association for Computational Linguistics, 2003:8–15.

[176] Xue N, Palmer M. Calibrating Features for Semantic Role Labeling[C]//. Lin D, Wu D. Proceedings of EMNLP 2004. Barcelona, Spain: Association for Computational Linguistics, 2004:88–94.

[177] Pradhan S, Hacioglu K, Krugler V, et al. Support Vector Learning for Semantic Argument Classification[J]. Machine Learning, 2005, 60(1-3):11–39.

[178] Toutanova K, Haghighi A, Manning C D. A Global Joint Model for Semantic Role Labeling[J]. Computational Linguistics, 2008, 34(2):161–191.

[179] Martins A F T, Almeida M S C. Priberam: A Turbo Semantic Parser with Second Order Features[C]//Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: Association for Computational Linguistics and Dublin City University, 2014:471–476.

[180] Yang H, Zong C. Multi-Predicate Semantic Role Labeling[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014:363–373.

[181] Moschitti A. A Study on Convolution Kernels for Shallow Statistic Parsing[C]//Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume. Barcelona, Spain: Association for Computational Linguistics, 2004:335–342.

[182] Moschitti A, Pighin D, Basili R. Tree kernels for semantic role labeling[J]. Computational Linguistics, 2008, 34(2):193–224.

[183] Che W, Zhang M, Liu T, et al. A Hybrid Convolution Tree Kernel for Semantic Role Labeling[C]//Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. Sydney, Australia: Association for Computational Linguistics, 2006:73–80.

[184] Lei T, Zhang Y, Màrquez L, et al. High-Order Low-Rank Tensors for Semantic Role Labeling[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado: Association for Computational Linguistics, 2015:1150–1160.

[185] Zhou J, Xu W. End-to-end learning of semantic role labeling using recurrent neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, 2015:1127–1137.

[186] FitzGerald N, Täckström O, Ganchev K, et al. Semantic Role Labeling with Neural Network Factors[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015:960–970.

[187] Roth M, Lapata M. Neural Semantic Role Labeling with Dependency Path Embeddings[J]. arXiv preprint arXiv:1605.07515, 2016.

[188] Anders B, Bernd B, Hafdell L, et al. A High-Performance Syntactic and Semantic Dependency Parser[C]//Coling 2010: Demonstrations. Beijing, China: Coling 2010 Organizing Committee, 2010:33–36.

[189] Zelenko D, Aone C, Richardella A. Kernel methods for relation extraction[J]. Journal of machine learning research, 2003, 3(Feb):1083–1106.

[190] Yu M, Gormley M, Dredze M. Factor-based compositional embedding models[C]//NIPS Workshop on Learning Semantics. Montreal, Canada: Curran Associates, Inc., 2014.

[191] Yu M, Gormley M R, Dredze M. Combining Word Embeddings and Feature Embeddings for Fine-grained Relation Extraction[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado: Association for Computational Linguistics, 2015:1374–1379.

[192] Zeng D, Liu K, Lai S, et al. Relation Classification via Convolutional Deep Neural Network[C]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014:2335–2344.

[193] Liu Y, Wei F, Li S, et al. A Dependency-Based Neural Network for Relation Classification[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Beijing, China: Association for Computational Linguistics, 2015:285–290.

[194] Xu Y, Mou L, Li G, et al. Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Paths[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015:1785–1794.

[195] Miwa M, Bansal M. End-to-end Relation Extraction using LSTMs on Sequences and Tree Structures[J]. arXiv preprint arXiv:1601.00770, 2016.

[196] dos Santos C, Xiang B, Zhou B. Classifying Relations by Ranking with Convolutional Neural Networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, 2015:626–634.

[197] Punyakanok V, Roth D, Yih W t, et al. Semantic Role Labeling Via Integer Linear Programming Inference[C]//Proceedings of Coling 2004. Geneva, Switzerland: COLING, 2004:1346–1352.

[198] Che W, Li Z, Hu Y, et al. A Cascaded Syntactic and Semantic Dependency Parsing System[C]//CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning. Manchester, England: Coling 2008 Organizing Committee, 2008:238–242.

[199] Bunescu R, Mooney R. A Shortest Path Dependency Kernel for Relation Extraction[C]//Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada: Association for Computational Linguistics, 2005:724–731.

[200] Manning C, Surdeanu M, Bauer J, et al. The Stanford CoreNLP Natural Language Processing Toolkit[C]//Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore, Maryland: Association for Computational Linguistics, 2014:55–60.

[201] Che W, Li Z, Liu T. LTP: A Chinese Language Technology Platform[C]//Coling 2010: Demonstrations. Beijing, China: Coling 2010 Organizing Committee, 2010:13–16.

[202] Roth M, Woodsend K. Composition of Word Representations Improves Semantic Role Labelling[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014:407–413.

[203] Xu K, Feng Y, Huang S, et al. Semantic Relation Classification via Convolutional Neural Networks with Simple Negative Sampling[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015:536–540.

[204] Vu N T, Adel H, Gupta P, et al. Combining Recurrent and Convolutional Neural Networks for Relation Classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics, 2016:534–539.

攻读博士学位期间发表的论文及其他成果

（一）会议论文

[1] **Jiang Guo**, Wanxiang Che, David Yarowsky, Haifeng Wang, Ting Liu. Cross-lingual Dependency Parsing Based on Distributed Representations. *In Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics* (**ACL 2015** long paper). Beijing, China. 2015.07. (CCF A类，他引27次)

[2] **Jiang Guo**, Wanxiang Che, David Yarowsky, Haifeng Wang, Ting Liu. A Representation Learning Framework for Multi-Source Transfer Parsing. *In Proceedings of the 30th AAAI Conference on Artificial Intelligence* (**AAAI 2016**). Phoenix, AZ, USA. 2016.02. (CCF A类，他引8次)

[3] **Jiang Guo**, Wanxiang Che, Haifeng Wang, Ting Liu. Revisiting Embedding Features for Simple Semi-supervised Learning. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (**EMNLP 2014** long paper). Doha, Qatar. 2014.10. (CCF B类，他引44次)

[4] **Jiang Guo**, Wanxiang Che, Haifeng Wang, Ting Liu. Learning Sense-specific Word Embeddings By Exploiting Bilingual Resources. *In Proceedings of the 25th International Conference on Computational Linguistics* (**COLING 2014**). Dublin, Ireland. 2014.08. (CCF B类，他引30次)

[5] **Jiang Guo**, Wanxiang Che, Haifeng Wang, Ting Liu, Jun Xu. A Unified Architecture for Semantic Role Labeling and Relation Classification. *In Proceedings of the 26th International Conference on Computational Linguistics* (**COLING 2016**). Osaka, Japan. 2016.12. (CCF B类)

[6] **Jiang Guo**, Wanxiang Che, Haifeng Wang, Ting Liu. A Universal Framework for Inductive Transfer Parsing across Multi-typed Treebanks. *In Proceedings of the 25th International Conference on Computational Linguistics* (**COLING 2016**). Osaka, Japan. 2016.12. (CCF B类)

[7] Ruiji Fu, **Jiang Guo**, Bing Qin, Wanxiang Che, Haifeng Wang, Ting Liu. Learning Semantic Hierarchies via Word Embeddings. *In Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics* (**ACL 2014** long paper). Baltimore, MD, USA. 2014.06. (CCF A类，他引67次)

[8] Yijia Liu, Wanxiang Che, **Jiang Guo**, Bing Qin, Ting Liu. Exploring Segment Representations for Neural Segmentation Models. *In Proceedings of the 25th International Joint Conference on Artificial Intelligence* (**IJCAI 2016**). New York City, NY, USA. 2016.07. (CCF A类)

[9] Yuxuan Wang, **Jiang Guo**, Wanxiang Che, Ting Liu. Transition-based Chinese Semantic Dependency Graph Parsing. *In Proceedings of the 15th China National Conference on Computational Linguistics* (**CCL 2016**). Yantai, China. 2016.10. (**best paper award**)

（二）期刊论文

[1] **Jiang Guo**, Wanxiang Che, David Yarowsky, Haifeng Wang, Ting Liu. A Distributed Representation-based Framework for Cross-lingual Transfer Parsing. *Journal of Artificial Intelligence Research* (**JAIR 2016**). 2016, 55:995–1023. (SCI, IF=2.363)

[2] Ruiji Fu*, **Jiang Guo***, Bin Qing, Wanxiang Che, Haifeng Wang, Ting Liu. Learning Semantic Hierarchies: A Continuous Vector Space Approach. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (**TASLP 2015**). 2015, 23(3):461–471. (* Co-First Author) (SCI, IF=1.225)

[3] Wanxiang Che, **Jiang Guo** and Ting Liu. ReliAble Dependency Arc Recognition. *Expert Systems with Applications*. 2014, 41(4):1716–1722. (SCI, IF=2.240)

哈尔滨工业大学学位论文原创性声明及使用授权说明

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《基于分布表示的跨语言跨任务自然语言分析》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果。据本人所知，论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。本声明的法律结果将完全由本人承担。

作者签名：　　日期：　　年　　月　　日

学位论文使用授权说明

本人完全了解哈尔滨工业大学关于保存、使用学位论文的规定，即：

（1）已获学位的研究生必须按学校规定提交学位论文；（2）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（3）为教学和科研目的，学校可以将学位论文作为资料在图书馆及校园网上提供目录检索与阅览服务；（4）根据相关要求，向国家图书馆报送学位论文。

保密论文在解密后遵守此规定。

本人保证遵守上述规定。

作者签名：　　日期：　　年　　月　　日

导师签名：　　日期：　　年　　月　　日

致　谢

写本文第一章的时候，我常想，每个人何尝不是社会网络中由许多维度进行分布表示的一个节点? 在每一个维度上我们最终写下的是什么，都和与我们相连的那些“上下文”息息相关。而我是如此幸运，在这四年多的时光里，能够遇到那么多优秀，真诚，志趣相投的师长、好友和知己。是他们让我不断地成长，成熟和完善；让我看到了自己，也看到了天地。在博士论文的最后，我发自内心地感谢他们。

首先，我要感谢我的导师王海峰教授。我第一次遇见王老师是在七年前的百度大厦。作为一个小实习生，我并没有很多机会接触王老师，然而很少的几个细节却令我印象深刻：他坐在电脑前的时候腰背永远是挺直的，仿佛坐多久都不会感到疲惫；他记忆力惊人，似乎能够存储下身边人说过的每一句话；他讲话时声音不大，却令身边的所有人安静聆听。两年后，我有幸成为王老师的第一个博士生。在这几年的科研历程中，我渐渐体会到王老师对于科研深刻的洞察力与前瞻性，尤其是他无比严谨认真的工作态度。从博士课题的选择，到我第一篇论文的发表，到我成功申请到约翰·霍普金斯大学的访学机会，到博士论文的最终完成，这条漫漫读博路上的每一步脚印、每一次坚持、每一个收获都与王老师的指导与帮助分不开。我想，也许我毕其一生都无法达到王老师的高度，但我将一直以他为人生的灯塔。

我更要感谢我的副导师车万翔副教授。车老师是我在自然语言处理领域的启蒙老师，我与他有着超过八年的缘份。无论我在实验室，还是在国外，车老师对我的科研工作以及日常生活都关怀备至。我博士期间所取得的任何成绩都有着车老师的心血，可以说，车老师完整地见证了我人生中最重要的蜕变过程。我常常惊叹于车老师渊博的学识，以及在研究中独特的思考角度；也常为车老师充沛的精力，豁达的胸怀，宽广的视野与格局所折服。尽管车老师一头白发尽显威严，但是从来都和我们打成一片，一起看电影，一起滑雪，亦师亦友。尤其感谢车老师为我创造的普林斯顿大学访学机会，让我有了更开阔的研究视野以及更真诚的求知之心。

我要特别感谢刘挺教授。赛尔实验室是我成长的土壤，而这片沃土离不开刘老师的耕耘。我的言谈举止和为人处事无不受到刘老师以及实验室文化的影响。感谢刘老师在我曾经一度处于低谷时帮我重塑对于科研的信心，以及在我每次遇到重要选择时悉心为我解惑，并为我提供宝贵的帮助与支持。同时也非

常感谢在赛尔朝夕相处的秦兵教授和张宇教授。他们是我遇到过的最像亲人的老师，让我在赛尔感到家一般的温暖。

感谢我在约翰·霍普金斯大学的导师David Yarowsky教授。在JHU的那一年是我博士期间最专注和充实的一年，在JHU的研究工作奠定了我博士论文的方向。我与David不仅讨论学术，也讨论政治，旅游，人生。他对待科学的严谨，对科研工作的品味，探索世界的热情都对我产生了巨大的影响。同时也感谢我在普林斯顿大学访问时的导师Han Liu助理教授。他对工作无比的投入深深感染了我，也在我初入科研之门时给了我许多宝贵的经验。

感谢本论文责任专家李生教授对论文一丝不苟的审阅，也感谢各位答辩委员会的专家以及外审专家对本论文提出的宝贵意见和改进建议。

特别感谢我的武术老师董平教授。学习武术的一年半彻底改变了我的身体与精神状态，让我在科研之路上更加自信与从容。董老师为人豪迈、坦荡，不拘泥于师生之别与世俗礼法。我平素最爱读武侠小说，董老师是我心中的侠。

感谢赛尔实验室和我一起并肩战斗过的所有兄弟姐妹，以前的，现在的。他们是我科研路上最重要的伙伴与最坚强的后盾。特别感谢妍妍师姐，伟男师兄事无巨细的帮助。感谢一起学习武术的森栋、王栋。感谢133寝室的室友严浩、王鑫、车凯。感谢在CLSP一起奋斗过的小伙伴们。感谢我的所有Co-author，尤其感谢瑞吉师兄，与我合作了第一篇ACL。感谢所有与我讨论、为我论文提出宝贵意见的朋友们。

感谢海浪群的晓修，瑾晨，然然，来明，马彪，传钊，王星。我常常怀念我们在美国上山下海、浪迹天涯的时光。那是我博士生活中最阳光灿烂的日子。

最后，我要将本文献给我挚爱的父母。感谢他们对我数十年如一日的爱。

个人简历

郭江，男，1989年7月生，出生于江西省莲花县。

教育经历

- 2012.09至今：于哈尔滨工业大学计算机学院社会计算与信息检索研究中心攻读工学博士学位，导师为王海峰教授，副导师为车万翔副教授。
- 2010.09–2012.07：于哈尔滨工业大学计算机学院社会计算与信息检索研究中心获得工学硕士学位，导师为车万翔副教授。
- 2006.09–2010.07：于哈尔滨工业大学计算机学院获得工学学士学位。

实习经历

- 2014.10–2015.10：于美国约翰霍普金斯大学（Johns Hopkins University）访问，导师为David Yarowsky教授。访问期间的主要工作是跨语言表示学习以及面向资源稀缺语言的依存句法分析研究。
- 2012.09–2012.12：于美国普林斯顿大学（Princeton University）访问，导师为Han Liu助理教授。访问期间的主要工作为高斯图模型的结构预测并行算法以及在线平台开发。
- 2010.04–2010.06：于北京百度公司自然语言处理部门实习。实习期间的主要工作为面向搜索引擎查询的依存句法分析。

获奖情况

- 2016.10：获第二十六届国际计算语言学会议（COLING）学生资助。
- 2016.10：获全国第十五届计算语言学会议（CCL）最佳论文奖。
- 2016.07：获国家开发银行奖学金。
- 2015.11：获百度奖学金（共奖励全球10名华人学生）。
- 2014.12：获博士生国家奖学金。
- 2013.10：获哈工大武术比赛太极拳项目第二名。
- 2012.06：获黑龙江省优秀硕士毕业生。
- 2010.02：大学生创新性实验计划杰出项目奖。
- 2009.08：国家信息安全竞赛三等奖。
- 2009.04：全美大学生数学建模竞赛二等奖。
- 2006–2010：获校三好学生、人民奖学金多次。

参与项目

- 2015.09–2016.09：面向资源稀缺语言的高精度依存句法分析。Google专注研究项目。
- 2014.01–2018.06：面向三元空间的互联网中文信息处理理论与方法。国家973计划项目（编号2014CB340503）。
- 2014.01–2017.12：依存句法分析子结构可信度计算研究。国家自然科学基金（编号61370164）。
- 2013：语义角色标注。与索尼公司合作项目。
- 2012：基于图以及基于转移的依存句法分析工具。与腾讯公司合作项目。
- 2011至今：语言技术平台。实验室自主项目。
- 2010：雷同报告检测系统。全国大学生创新性实验计划SUN项目。

系统开发和数据标注

- 2011–2015：开发和维护“语言技术平台（Language Technology Platform, LTP）”。期间独立完成了基于神经网络的依存句法分析模块的原型代码；独立开发了准确率更高、模型更小、速度更快的语义角色标注模块。
- 2010：参与标注了6万句依存句法树库。目前，这个大规模中文依存句法树库已经在Linguistic Data Consortium（LDC）上发布。
- 2013：参与组织与标注了5万句微博依存句法树库，该数据目前在数据堂（datatang.com）上发布。

学术论文

- 在自然语言处理与人工智能领域的国际顶级和重要会议及期刊：ACL, EMNLP, COLING, AAAI, JAIR, TASLP上以第一作者发表论文。