文章编号：1003-0077（2016）00-0000-00

基于远监督的语义知识资源扩展研究*

卢达威[1]，王星友[2]，袁毓林[1]

（1.北京大学中文系，北京 100871；2. 北京语言大学信息科学学院，北京 100083）

摘要：语义知识资源蕴含了深刻的语言学理论，是语言学知识和语言工程的重要接口。本文以形容词句法语义词典为研究对象，探索对语义知识资源自动扩展的方法。本文的目标是利用大规模语料库，扩展原有词典的词表及其对应的句法格式。具体方法是根据词的句法格式将词典的词分类，将待扩展的新词通过分类器映射到原有词典的词中，以此把词典扩展问题转化为多类分类问题。依据的原理是词典词和待扩展新词在大规模语料中句法结构的相似性。本文通过远监督的方法构造训练数据，避免大量的人工标注。训练过程结合了浅层机器学习方法和深度神经网络，取得了有意义的成果。实验结果显示，深度神经网络能够习得句法结构信息，有效提升匹配的准确率。

关键字：资源扩展 远监督 语义知识资源

中图分类号：TP391　　**文献标识码：**A

Research on the Extension of Semantic Knowledge Resources Based on Distant Supervision

LU Dawei[1], WANG Xingyou[2], YUAN Yulin[1]

(1.Peking University, Beijing, 100871, China; 2. Beijing Language and Culture University, Beijing, 100083, China)

Abstract: The semantic knowledge resources containing extensive linguistic information are one of the important interfaces of linguistics and language engineering. In this paper, we study the automatic extension of semantic knowledge resources based on the *Adjective Syntactic-Semantics Dictionary*. We aim to extend the vocabulary of the dictionary and their syntactic patterns via the large corpus. More specifically, our method is to classify the words in dictionary into 97 categories by their syntactic patterns, and match the new words which are not existing in the dictionary to each category by the classifier, thereby the dictionary extension problem is transformed into a multi-class classification problem. The method is based on the fact that the new words and the dictionary words have the similar syntactic patterns in large corpus. We construct the training data by distance supervision, so as to reduce the effort of manual annotation. Training process combines the shallow learning and the deep neural network, which achieves the significant results. The experimental results show that the deep neural network is able to learn the syntactic information, and effectively improve the accuracy in the mapping task.

Keywords: resource extension; Distant Supervision; semantic knowledge resources

1. 引言

语义知识资源是在特定的语言学理论基础上，以词型（Type）为标注对象进行语言描写的语言工程实践的成果，是语言学理论和自然语言处理技术结合的重要手段。相比以词例（Token）为标注对象的语料库标注，语义知识资源库建设更为便捷，且对自然语料的覆盖面更大，对系统的可移植性更高。比较有名的语义知识资源有 WordNet[1][2]、VerbNet[3]、PropBank[4]、FrameNet[5][6]、ConceptNet[7]和国内的 HowNet[8]，《现代汉语语法信息词典》[9]，《同义词词林》[10]等，这些资源都有各自的语言学、心理学或哲学理论基础，成为了语言

* 本课题的研究得到国家社科基金重大招标项目《汉语国际教育背景下的汉语意合特征研究与大型知识库和语料库建设》（12&ZD175）和国家重点基础研究计划（973 计划）项目课题《语言认知的神经机制》（2014CB340502）的资助，谨此致以诚挚的谢意。

学理论和自然语言处理的重要接口。

语义知识资源作为一种专家型的资源，具有以下特征：

(1) 高质量。这些资源都凝聚了语言学家们的智慧和多年的积累，蕴含了深刻的理论和实践价值。

(2) 规模有限。由于资源的编纂需要大量的人力物力，语义资源的规模增长缓慢。

(3) 高频。由于规模所限，出于典型性考虑，词典所选词一般为高频词，这使得词典在语料覆盖度上有一定的保证。

在语言工程实践中，面对真实的大数据文本，语义资源常常因其规模有限，难以在计算中充当核心角色，通常仅作为一种特征参与计算，以辅助提高准确率和召回率。这既没有充分发挥其高质量的优势，也不能很好地将语言学知识融入计算中。同时，由于更新缓慢，语义知识资源的规模难以追赶日新月异的语言变化和网络生态。规模的有限性成为了语义知识资源在工程实践中的最大瓶颈。

本文的目标就是以语义知识资源为种子，根据特定语义资源的格式，从大数据中自动学习并扩充语义知识资源的词表及词所对应的内容，本文把这一任务称为语义知识资源扩展。

在多年的语言知识资源建设和语言处理工程实践中，我们认识到语义知识资源扩展对语言资源的使用有下列重要的意义。

(1) 在应用领域，有利于充分发挥语义知识资源的基础作用。若扩展了语义知识资源的规模，实现语义知识资源对语料的高覆盖，则语义知识资源将能够作为 NLP 应用的基础，如直接用于语义分析、语义理解等，使语义知识资源更好地应用于自然语言处理的各个领域。

(2) 在知识资源建设上，有助于提高语义资源的编纂效率。语义资源扩展虽然不能完全代替人工编纂，但可以通过人机结合的方法有效提高效率，减轻编者的负担。

不少学者尝试通过扩展知识资源的方法解决资源规模不足的问题，如 Kipper[11]利用 Levin[3]的动词分类扩展 WordNet 的动词词表，Strapparava[12]扩展了 WordNet 的情感词表等。Rothe[13]与本文的任务较为接近，他针对 WordNet 进行了扩展研究，巧妙地运用了 WordNet 中 Word、Synset 和 Lexeme 三种数据类型之间的关系，构造了无监督的自学习扩展方法，取得了较好结果。然而，不同的知识资源，由于其内容、结构、理论等方面的差异，其扩展方法并不能简单地移植。

本文以在建的北京大学形容词句法语义词典为研究对象，探索对语义知识资源自动扩展的方法。其主要思路是：从大规模的语料库中，得到其语料库的词表；再通过某种映射，匹配到语义资源的小词表中；并通过人机结合的方法，补充新的语义知识资源的细节。本文首先介绍形容词句法语义词典的结构及其特点；根据其结构特点，选择合适的特征，把资源扩展问题规约为分类问题；同时，使用远监督的方法构造训练语料，避免大量语料标注；依据的原理是词典的词和待扩展新词在大规模语料中句法结构的相似性；并结合深度神经网络和其他浅层机器学习方法进行语义资源的扩展。

2. 语义知识资源的结构特点分析

本文以北京大学《形容词句法语义词典》[14]为研究对象。句法语义词典中，每个词都带有句法信息和语义角色的信息，包括每个词所能接受的所有句法格式以及其对应语义角色。后文的讨论主要围绕该词典进行。

2.1 形容词句法语义词典及其结构特点

形容词句法语义词典收录形容词（含状态词）共 3000 多个，其理论基础是论元结构理论[15]。具体而言，词典的每个词条都包含三个部分：注音和释义，语义角色集，句法格式集。其中，语义角色集包括每个词在某个义项下各个论元的语义角色集合，共有主事（TH）、

感事（SE）、范围（RA）、与事（D）、量幅（EXT）、对象（TA）、系事（RE）、致事（CAU）、原因（RN）、目的（AI）、时间（T）、处所（L）、方向（DI）共 13 种语义角色，对每个词条的每种语义角色采用个例化的描述方法。句法格式指该谓词跟受其支配的这些论元角色在句子中的句法配置方式。如词条“美丽”的释义如下：

美丽 měilì 〈形容词〉好看；漂亮；看了使人产生美感的。多形容女性容貌或风光、景色、诗文、理想等。跟“丑陋”相对。

（1）语义角色：

主事 TH：具有好看、漂亮，看了使人产生美感这种属性的人或物；

与事 D：主事跟他在美丽这种属性上进行比较的参照者。

（2）句法格式：

S1：TH+（比 D+）__ [*注：括号中的部分是选择性成分，可以省略，下同。]

如：那位姑娘非常～。| 西湖的景色十分～。|湖边的天鹅雕塑在夕阳的照耀下显得非常～。| 眼前的首都比想象中的还要～得多。

S2：（比 D+）__+的+TH

如：～的姑娘 | ～的风景 | ～的地方 | ～的心灵 | 比七仙女还～的女孩儿 | 比未名湖更～的景点

词的句法格式在句法分析有重要的作用。当前句法分析的方法通常是使用依存语法或者上下文无关文法，自下而上进行递归分析。如果基于词的句法格式进行句法分析，则句法分析的过程就变成了模式匹配的过程，句子则看作多种模式的嵌套。这不仅大大减少了句子的层次，而且模式匹配更符合人对句子的理解和认知过程。例如上例“眼前的首都比想象中的还要美得多”,按照“美丽”的句法格式，该句可以分析为“眼前的首都$_{[TH]}$+比+想象中的$_{[D]}$+还要+美$_{[形容词]}$+得多”。更进一步，当我们使用句法格式完成句法分析后，语义角色分析也就同步完成了，这相当于一次性完成了浅层语义分析和句法分析，不仅高效，而且能避免在句法分析基础上做语义分析时，造成误差累积。这是使用句法语义词典进行句法分析的另一优势。当然，要使用句法格式进行句法分析，还需要动词和名词句法语义词典的配合，同时，也需要扩展词典对语料的覆盖度。这正是本文进行语义资源扩展研究的一大原因。

从词条“美丽”可知，词典涵盖了词语之间的聚合和组合关系，这两种关系是语言系统中的两种最根本的关系[16]。聚合关系指词语之间在意义上的关联，如词典标记了每个词条相对的同义词、反义词等，索绪尔称之为“联想关系”；组合关系指语篇中的共现关系，如每个词条的句法格式等，索绪尔称之为“句段关系”[17]。

我们认为，对聚合关系和组合关系，应该分别进行扩展。对于聚合关系的扩展，已有不少研究，如文献[18][19]；同时，汉语中也有一些反映聚合关系的资源，如 HowNet 和《同义词词林》等。而对于组合关系的扩展，则研究较少，每个词的句法格式也是词典的特色与核心。另外，对于词典的释义部分，由于该词典特色之一是以体验性认知的释义为原则，对计算机来说，这种自动释义要求过高。因此，我们把研究的重心放在语义角色和句法格式的扩展中。

2.2. 句法格式的统计和分析

形容词句法格式的类型包括主谓结构和偏正结构两类。主谓结构用于陈述事物的状态、性状等，如上例“美丽”中的 S1：“TH+（比 D+）__”（西湖的景色十分美丽）；偏正结构用于指称具有某种性状的事物，如上例中的 S2：“（比 D+）__+的+TH”（比仙女还漂亮的女孩）。句法格式的成分包括：语义角色、形容词和语义角色的相对位置，引介词语（引出语义角色的介词或动词，“比、对、让、使”等），助词或后缀（“的、地”等），谓词前的动词（“感到、显得”等），状语（“彼此、相互”等）。由于句法成分众多，位置多样，造成不同类型的句法格式共有 1000 多种（表 1 展示了使用数量较多的前 10 种句法格式）。

表 1 词典中使用数较多的句法格式（前 10）

	句法格式	具有该句法格式的词
1.	TH+(RA+)_	1177
2.	(RA+)_+的+TH	1155
3.	_+的+RA	1090
4.	TH+(比 D+)_	995
5.	TH+(RA+)比 D+_	943
6.	(比 D+)_+的+TH	923
7.	TH+比 D+RA+_	914
8.	比 D+RA+_+的+TH	893
9.	(RA+)比 D+_+的+TH	866
10.	_+的+TH	802
	……	……

在扩展词典时，我们将根据语义角色和句法格式的特点，简化及合并类似的句法格式，并把词典按照句法格式集归类，将词典的扩展问题规约为新词语的分类问题。

2.3 句法格式判别集的选取与简化

构造句法格式判别集的目的，是将形容词词典按照句法格式集分类。因此，在构造句法格式集时，我们牺牲一定的精确度，舍去个性较强的句法格式，而选择具有普遍性和典型性的句法格式，作为归类标准。同时，候选的句法格式还必须具有较强的完备性和一致性；即每个词所列的句法格式不是举例性的，而是排他性的；若某词不包含某种句法格式，则该句法格式一定不能用于该词。

在简化句法格式时，我们考虑了以下几个原则：（1）句法格式中，最重要的差别是区分句法结构类型，是主谓结构还是偏正结构？因为有些形容词或状态词只能充当主谓结构谓语，而不能充当偏正结构的定语，如“安好、尽然”等。对于主谓结构的句法格式，其论元角色在前，形容词在后；对于偏正结构的句法格式，其形容词在前，论元角色在后。因此，位于句法格式首和末的语义角色至关重要。（2）在句法格式的诸多成分中，最具完备性和一致性的成分是：语义角色，及其与形容词的相对位置。其他成分带有较多特定的词的特殊性，并可能带有编纂时的人为误差而造成不一致。而且，引介词语和语义角色之间有较大的论元的可预测性。例如，与事 D 前的引介词语是“比”，对象 TA 前的引介词语一般是“对”等。（3）在句法格式中，主事 TH 和感事 SE 是形容词的必有成分，是呈互补分布的对比特征，即二者不会出现在同一个句法格式中。形容词跟主事搭配还是感事搭配，反应了该词的性质，是形容词重要的区别特征。

根据以上的原则，我们对已有的句法格式进行简化：仅保留主事 TH、感事 SE 以及位于句法格式首末位置的语义角色，除句法格式括号内的可选性成分，并去除其他成分。经过此简化合并，整理得到句法格式的 45 个（见表 2）。根据每个词所含的句法格式，词典的形容词可以分为 97 类（见表 3）。

表 2 选取和简化后的句法格式及其在词典中的数量（前 15）

	句法格式	数量
1.	TH+_	3046
2.	_+TH	2993
3.	_+RA	1239
4.	TH+_+RA	544
5.	TH+_+RE	294
6.	SE+_	267

	句法格式	数量
7.	_+SE	250
8.	SE+_+RE	15
9.	TH+_+EXT	14
10.	_+RE	14
11.	SE+_+TA	12
12.	TH+_+TA	9

	句法格式	数量
13.	SE+TH+_	6
14.	TH+_+L	5
15.	TH+SE+_	5
…	……	……

表 3 按照句法格式的词典分类（前 5）

类别	句法格式	所含词数	该类包含的词	句法格式例句
C1	_+TH TH+_	1173	安稳，博学，白蒙蒙，……	博学的历史学家[TH] 这条船[TH]很安稳。
C2	_+TH _+RA TH+_	903	矮小，卑贱，阳刚，……	矮小的植物[TH] 矮小的身材[RA] 他[TH]很阳刚。
C3	_+TH TH+_ TH+_+RA	383	饱满，晦暗，奢华，……	晦暗的灯光[TH] 这些谷粒[TH]十分饱满。 生活[TH]奢华的家庭[RA]
C4	_+TH TH+_ TH+_+RE	140	冷飕飕，匆促，哩哩啦啦，……	哩哩啦啦的小雨[TH] 屋子[TH]冷飕飕的。 敌军[TH]匆促应战[RE]。
C5	_+TH _+RA TH+_ TH+_+RE	93	恶狠狠，昏沉，雄赳赳，……	雄赳赳的哨兵[TH] 雄赳赳的气势[RA] 姐姐[TH]这几天有点昏沉 他[TH]恶狠狠地骂了一句[RE]
……	……	……	……	……

这些分类反应了形容词在句法结构方面的特点。例如，表 3 的 C1 类的形容词有两个句法格式“_+TH”和“TH+_”，C2 类比 C1 类多了一种句法格式“_+RA”。TH 是主事的标记，RA 是范围标记。主事（TH）表示性质、状态等事态的非感知性的主体，范围（RA）一般表示性状所涉及的主体的具体方面，如“身材、性格、气势、规模”等。C1 类和 C2 类句法格式的不同反映了词典中两类形容词的多方面差异。首先，这两类语义角色数量不同，类 1 的形容词不存在范围（RA）这一语义角色。因为若存在范围（RA）这一语义角色，则必然有“_+RA”这一句法格式。第二，C2 类的形容词能够同时受主事（TH）和范围（RA）两个维度的词语来修饰，如“他[TH]性格[RA]很阳刚。”而类 1 中的形容词则只能受一个维度的词语修饰，如“河面[TH]一片白蒙蒙”。

再如 C3 类和 C4 类，其差别在于 C3 类有句法格式“TH+_+RA”，而 4 类有“TH+_+RE”。RE 是系事标记，指主事呈现出形容词表示的某种性状时所处的状态或所进行的活动，一般是动词性成分。而如上所述，范围（RA）通常是名词性成分。也就是说，在类 4 的形容词可以状语，修饰动词行成分 RE，如“敌军[TH]匆促应战[RE]”，而 C3 类的形容词却不行，C1 类、C2 类的形容词也不行。

又如，有的形容词分类，只有“TH+_”一种句法格式，如 C13 类的“不赖、安好、不要紧、牢、枉然”等，表示该类形容词只能做谓语，不能做定语修饰名词。

在所有 97 类中，多于 1 个词的类有 45 类，共有词 3149 个词，占词典词数的 98.38%。只有一个词的类有 52 类，由于它们不便构造训练集和测试集，在下文的研究和实验中，我们以前 45 类词构造训练集和测试集，进行词典的扩展研究。

3. 实验

3.1 实验方案概述

本文语义资源的扩展思路是根据词典的特点构造分类器，将不在词典中的词映射到词典中。为了构造合适的分类器并检验其分类效果，我们首先利用现有词典构造训练集和测试集。

本文的实验方案如下：（1）从大规模语料库中训练出每个词的词向量。（2）将词典的每个词按照句法格式分类（如第 2 节所述），并把每一类的词随机分成训练集和测试集。（3）从语料库中，抽取含有训练集和测试集每个词的所有句子，并按训练集和测试集中词的类别来对这些句子贴上相应词及其类别的标签（如表 4 所示）。（4）这些句子通过深度神经网络进行训练。（5）通过若干种分类器对测试集进行分类，并比较这些方法。

我们的实验方案，基于以下两个假设：一是词向量表示一个词近距离上下文的语义表示；二是深度神经网络，特别是循环神经网络对句子训练的过程，看作是系统融入了句子的结构知识的过程。故认为该实验方案包含了近距离的上下文搭配知识和远距离的句法知识。我们

使用远监督方法，在词的训练集和测试集基础上，构造句子的训练集和测试集。所谓远监督是 Mintz[20]提出的，他们利用语义资源库 FreeBase 中的关系（relations）来构造训练数据，进而用这些训练数据解决关系抽取的问题，从而避免人工标注。我们实验方案参考 Mintz 的方法，利用词典句法格式的分类，构造用于深度神经网络的句子训练集和测试集，故称为远监督方法。本实验具体构造方法见 3.2 节。

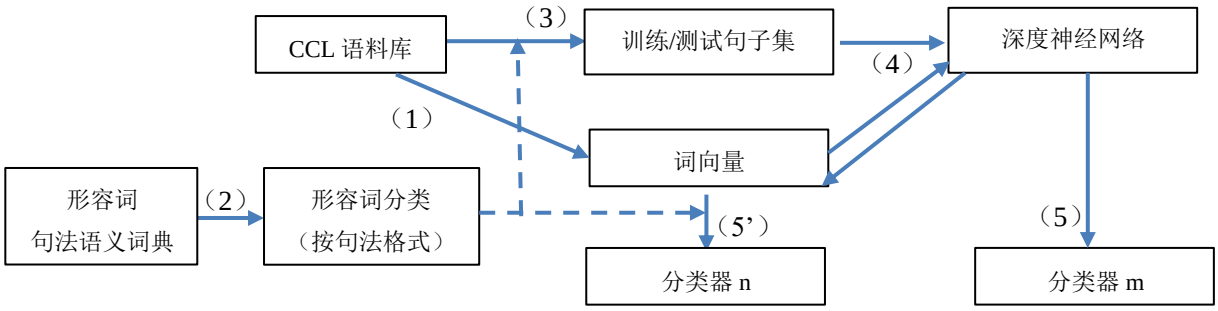

图 1 语义知识资源扩展实验模型

3.2 实验语料

本文的实验语料是北京大学中国语言学研究中心（CCL，Center for Chinese Linguistics PKU）的现代汉语语料库，简称 CCL 语料库。

词向量的训练：我们对语料库进行分词，并使用 word2vec[21][22]训练 300 维的词向量，使用 skip-gram 模型，迭代次数为 50 次。为了更好地训练词语上下文的语义信息，在分词时，我们基于规则对数字、时间、日期，以及人名、地名、机构名等专有名词进行了简单的识别与合并，规则如“姓氏+身份词/亲属称谓→人名”“省/市+字符串+地名后缀（镇/乡/村/街等）→地名”等，大大减少了语料中的词汇量。本文的目标是获得形容词的词向量，将数字、时间、日期、人名、地名、机构名等大量且低频的名词合并后，上下文更为简单一致，可以提高形容词的训练效果。

形容词分类训练集和测试集：句法语义词典根据句法格式分类（见第 2 节），并取词数大于 1 的类，共 45 类，其中有 3 类的测试集或训练集所有词都没有在 CCL 语料库中出现，故除去这 3 类，保留 42 类 3149 词。按 9:1 且至少测试集有 1 词的原则，构造训练集和测试集，得到训练集 2835 词，测试集 314 词。

句子训练集和测试集：在 CCL 语料库中抽取出含有训练集或测试集的形容词所对应的所有句子，分别构成句子训练集和测试集，并以形容词所在的分类作为句子的标签。得到训练集 81.9 万句，测试集 18.5 万句，平均句长 30.6 词。句子训练和测试样本见表 4。

表 4 句子训练及测试样本举例

1. 篱笆 不 <blank> ， 怎能 怪 别人 来 钻 ？ 【C3】【严实】
2. <m> 的 思想 包袱 卸掉 了 ， 他 <blank> 地 扛着 铁锹 走向 生产 地 。 【C4】【乐滋滋】
3. 在 <blank> 的 学术 殿堂 里 ， 印刷 符号 也 分 三六九等 。【C1】【幽深】
……

注：表 4 中每句是一个训练或测试样本，已分词。“【严实】”代表该句的目标词，“【C3】”代表目标词的分类。在样本句中抽走目标词，并用<blank>填充。<m>表示人名，在分词阶段使用规则对专有名词进行了基本识别，每类专有名词用一种符号表示，如<m>。

3.3 深度神经网络

使用深度神经网络（DNN，Deep Neural Networks）对句子训练集进行训练有两个作用：一是通过神经网络对句子进行分类，从而对句子所对应的标签词分类；二是在训练过程中，

会修改词向量，使得直接基于词向量的分类器能够取得更好的分类效果。

本文的神经网络模型采用了卷积神经网络（CNN，Convolutional Neural Networks）和循环神经网络（RNN，Recurrent Neural Networks）叠加的结构。卷积神经网络能够抽取句子的局部特征，而循环神经网络能够处理输入顺序和长距离依存的问题，这一模型组合在语音识别[23]、语言模型的建模[24]上都取得了较好的效果。本文神经网络模型如图 2 所示。

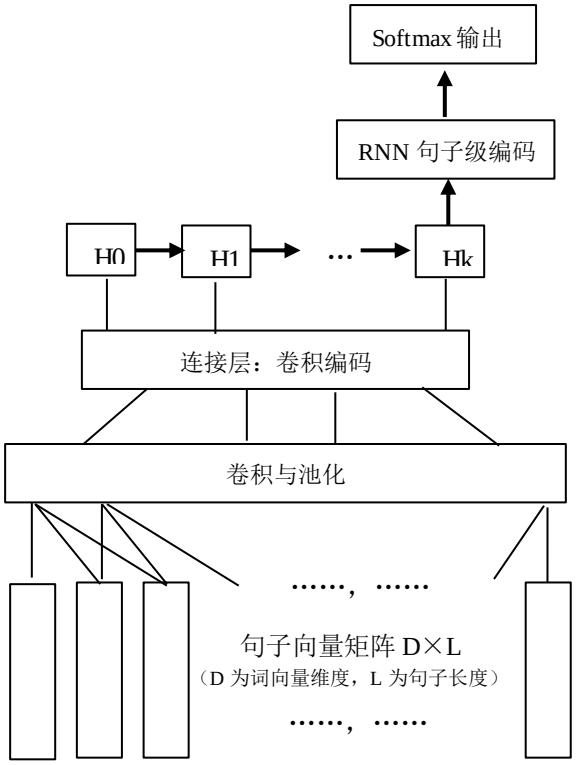

图 2 CNN+RNN 神经网络模型结构

实验首先使用 word2vec 得到预训练的词向量，词向量维度为 300，然后通过对词向量的查表，将长度为 L 的句子表示为 300*L 的向量表示形式；将这个句子向量通过卷积神经网络，学习句子的局部特征。实验中，我们使用 4 和 5 的卷积窗口来学习不同长度的特征，对于同一种窗口，我们分别采用 200 个不同的卷积核，对于一个窗口为 W 的卷积核，应用卷积核函数，将 300*L 的输入变化成 L-W+1 的输出。然后将卷积的输入通过最大池化技术，进行特征筛选，并且降低输出维度。我们使用缩放为 2 的池化比例，将卷积输出变为 floor((L-W+1)/ 2)的池化输出。对于不同的卷积核和不同的卷积窗口的输出，得到不同的池化结果。通过连接层，将这些结果连接起来，并且保持位置关系。卷积与池化后的局部特征表示作为序列输入，输入到循环神经网络，为了解决长时依赖问题，实验中采用 GRU 循环神经网络单元。通过循环神经网络，我们可以获得处理局部特征的位置关系并学习长距离特征。循环神经网络的输出为 100 维，作为句子级别的特征表示，再连接隐藏层 400 维的全连接的网络，并通过 softmax 输出，实现多类分类功能。

3.4 分类器及实验结果

对于测试集的分类预测，我们构造了若干分类器，以比较最好的方法：

（1）基于深度神经网络的分类器。3.3 小节的深度神经网络是以句子为单位作分类预测的。对于测试词或扩展新词的预测，由包含该词的所有句子的 DNN 分类投票结果决定，该分类器记为：DNN 分类器。

除了基于深度神经网络的分类器，我们还构造了基于词向量的若干分类器。

（2）K 近邻分类器。K 近邻（KNN，k-Nearest Neighbor）算法由 Altman[25]提出，本实

验中，测试词的分类由与之距离最近的若干个训练集词的分类决定，距离用词向量之间的余弦距离计算，记为 KNN 分类器。由于词典类数较多，且非常不均衡，存在许多元素个数较少甚至为 1 的类，故 K 的取值不宜太大。实验中从我们取 K=3 和 K=1（即最近邻）进行比较。同时，我们还比较基于原始的词向量的 KNN 结果和基于深度神经网络调整后的词向量的 KNN 结果。以上分类器分别记为：KNN（K=3，训练前），KNN（K=1，训练前），KNN（K=3，训练后），KNN（K=1，训练后）。

（3）SVM 分类器。支持向量机（SVM，Support Vector Machine）由 Cortes 和 Vapnik[26] 首先提出的。它在解决小样本、非线性及高维模式识别中表现出许多特有的优势。SVM 中，我们以词向量的各维度作为特征，共 300 维，对词典中训练集和测试集所涉及的 21 个句法格式分别构造分类器，使用径向基函数[27]作为核函数。测试时，对每个词对这 21 个句法格式独立判断是否存在该句法格式，全部判断正确才认为该词分类正确，否则算错。同时，我们基于原始词向量和 DNN 训练后词向量分别构造分类器进行测试。

各分类器准确率情况如表 5 所示。

表 5 各分类器的准确率对比（测试样本词总数：314 词）

	DNN 分类器		SVM 分类器		KNN_K=1		KNN_K=3	
	正确率 (%)	正确数 (词)	正确率 (%)	正确数 (词)	正确率 (%)	正确数 (词)	正确率 (%)	正确数 (词)
DNN 训练前	/	/	42.04	132	38.54	121	41.40	130
DNN 训练后	34.08	107	48.41	152	47.77	150	49.68	156

3.5 讨论

从分类结果看，基于深度神经网络的分类结果并不理想，仅为 34.08%，甚至不如简单的最近邻方法。但是经过深度神经网络有目标地训练后调整的词向量，对基于词向量的浅层分类器的分类效果有了显著提升。SVM 分类器基于原始词向量的分类准率为 42.04%，基于深度神经网络训练后准确率为 48.41%，提升了 6.37%；KNN（K=1）分类器从基于原始词向量的 38.54，提升到 44.77%，提升了了 9.23%；KNN（K=3）分类器准确率从 41.40%提升到 49.68%，提升了 8.28%。

在各分类器中，KNN（K=3，训练后）分类器效果最好，达到 49.68%，接近一半的准确率。而且，由于算法简单，过拟合现象少，一些比例上不占优势的类也能够预测准确。而 SVM 分类器和 DNN 分类器中预测准确的词都集中在训练集数量最多的几个类中，这也是多数分类器对非均衡分类容易造成的问题。

对于经过 DNN 调整后的词向量有效提升分类器准确率的问题，我们认为：由于词典扩展的目标，需要根据词典确定。而原始自动学习的词向量，是从一定窗口的上下文学习出来的，更偏重于近上下文的语义。形容词句法语义词典的扩展则更侧重于句法格式。语义相似的词语，句法格式未必相同。深度神经网络训练针对句子训练，通过 CNN 和 RNN 的层叠，融入了句法结构的因素，故在调整词向量时，有了具体调整目标。因此，对基于词向量的分类器由较大幅度的提升。

例如，原始的词向量中，与测试词“累”最接近的是“辛苦”，从语义上看似乎很合理；然而从句法上看，由于他们所能进入的句法格式不一样，属于不同类（见表 5）。“辛苦“是一种生存状态，主事 TH 表示具有身心劳累、艰辛困苦这种属性的人，范围 RA 表示辛苦的具体方面，如句法格式①“TH+_：孩子们学习十分辛苦。（TH=孩子们）”；而“累”不仅可以表示主体的身体或生存状态，还强调主体的主观感受，所以不仅能带主事 TH，还带语义角色感事 SE，表示感到疲惫乏力的人或动物，比如句法格式④ “SE+_：我感到浑身都十分累。（SE=我）”，这里的“累”就不能替换成“辛苦”。可见，句法格式与分布的不同，深刻地反映了词汇更深层、细致的语义差别。

表 5 “累”和“辛苦”的句法格式

累（类 9）//测试集数据	辛苦（类 2）//训练集数据
① TH+_ ② _+TH ③ _+RA ④ SE+_ ⑤ _+SE	① TH+_ ② _+TH ③ _+RA

经过 DNN 训练调整后，最接近“累”的词变成了“伤感”和“苦”。它们虽然在语义上不相同，但在句法格式上，却是比较一致的，属于同一类（类 9）；即既能够表达具有“伤感”或“苦”状态的人或事，如“这首诗很伤感。（TH=这首诗）”“他的命很苦。（TH=他的命）”，又能表示感到“伤感”或“苦”的人，如“方鸿渐正因情场失意而感到伤感。（SE=方鸿渐）”“他感到很苦。（SE=他）”。可见，经过 DNN 训练调整后，词向量更好地反映了词汇的句法结构能力。

4. 结论和展望

本文的目标是扩展现有的语义知识资源，以期使语义知识资源更好地应用于 NLP 的各个领域，乃至作为 NLP 应用的基础。本文以北京大学《形容词句法语义词典》为研究对象，其资源扩展的主要思路是：从大规模的语料库中，得到其语料库的词表。再通过某种映射，匹配到语义资源的小词表中；并通过人机结合的方法，补充新的语义知识资源的细节，达到语言资源扩展的目的。而这种映射，体现在本文中就是：根据词典的特点，利用每个词句法格式的不同，把词典扩展问题转化为分类问题。

解决分类问题需要的大量的训练语料，为避免人工语料标注消耗巨大的人力物力，我们利用远监督的方法进行机器学习。首先从大规模语料库中训练词向量，并以词向量的维度为词的特征进行机器学习。由于词向量的学习过程决定了词向量仅能表达有限窗口的上下文信息，因而我们通过训练集和测试集抽取相应的句子，组成句子的训练集和测试集；再利用由卷积神经网络和循环神经网络叠加的神经网络模型进行训练，并在训练中调整词向量，以此将句法结构的信息融入词向量中。

实验结果显示，利用经过调整的词向量，使用较简单的 K 近邻算法下，在 45 类的多分类问题中，能达到接近 50%的准确率；与使用 SVM 分类器相当，优于使用深度神经网络进行分类预测。

另一方面，本文的工作也可以辅助人工扩充词典。在人工扩充词典的词汇时，可先用词典所有的词在语料库中抽取句子进行训练；然后对候选的形容词表进行 KNN（K=3）的分类，以此确定词的基本句法格式集；在此基础上，再进行人工校对。人工校对的过程，不仅是编纂新词的过程，而且还能够发现原来词典的错误，比如原词的类属不当等。通过人机结合的方法，不仅能够提高词典编纂的效率，还能提高词典的准确率和一致性。

由于句法语义词典中每个词都带有句法信息和语义角色的信息，句法格式同时体现了论元的位置和论元的语义角色。因而，若利用句法语义词典进行句法结构分析，则可以同步解决句法分析和语义分析问题，而这正是自然语言处理的基础环节。

句法语义词典凝聚了语言学家们多年知识积累，本文的工作也是将语言学知识融入机器学习的一种探索，从实验结果看，取得了初步的成效。

参考文献

[1] Miller G, Fellbaum C. Wordnet: An electronic lexical database[J]. 1998.

[2] Miller G A, Fellbaum C. WordNet then and now[J]. Language Resources and Evaluation, 2007, 41(2): 209-214.

[3] Levin B. English verb classes and alternations: A preliminary investigation[M]. University of Chicago press, 1993.

[4] Palmer M, Gildea D, Kingsbury P. The proposition bank: An annotated corpus of semantic roles[J]. Computational linguistics, 2005, 31(1): 71-106.

[5] Fillmore C J. Frame semantics[J]. Linguistics in the morning calm, 1982: 111-137.

[6] Fillmore C J, Johnson C R, Petruck M R L. Background to framenet[J]. International journal of lexicography, 2003, 16(3): 235-250.

[7] Liu H, Singh P. ConceptNet—a practical commonsense reasoning tool-kit[J]. BT technology journal, 2004, 22(4): 211-226.

[8] 董振东, 董强. 2000, HowNet, http://www.keenage.com.

[9] 俞士汶. 现代汉语语法信息词典详解[M]. 清华大学出版社, 1998.

[10]梅家驹. 同义词词林[M]. 上海辞书出版社, 1983.

[11]Kipper K, Dang H T, Palmer M. Class-Based Construction Of A Verb Lexicon[C]// Seventeenth National Conference on Artificial Intelligence & Twelfth Conference on Innovative Applications of Artificial Intelligence. 2000:691--696.

[12]Strapparava C, Valitutti A. WordNet Affect: an Affective Extension of WordNet[C]//LREC. 2004, 4: 1083-1086.

[13]Rothe S, Schütze H. AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes[J]. Computer Science, 2015.

[14]袁毓林. 基于生成词库论和论元结构理论的语义知识体系研究[J]. 中文信息学报, 2013, 27(6):23-30.

[15]袁毓林. 汉语配价语法研究[M]. 商务印书馆, 2010.

[16]袁毓林, 李强. 怎样用物性结构知识解决“网球问题”?[J]. 中文信息学报, 2014, 28(5):1-12.

[17]索绪尔. 普通语言学教程[J]. 北京: 商务印书馆, 1980.

[18]宋文杰, 顾彦慧, 周俊生,曲维光. 多策略同义词获取方法研究[J]. 北京大学学报:自然科学版, 2015, 51(2):301-306.

[19]孙霞, 董乐红. 基于监督学习的同义关系自动抽取方法[J]. 西北大学学报: 自然科学版, 2008, 38(1): 35-39.

[20]Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data[C]// ACL 2009, Proceedings of the, Meeting of the Association for Computational Linguistics and the, International Joint Conference on Natural Language Processing of the Afnlp, 2-7 August 2009, Singapore. 2009:1003-1011.

[21]Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013a.

[22]Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013b: 3111-3119.

[23]Sainath T N, Vinyals O, Senior A, et al. Convolutional, long short-term memory, fully connected deep neural networks[C]//Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015: 4580-4584.

[24]Kim Y, Jernite Y, Sontag D, et al. Character-aware neural language models[J]. arXiv preprint arXiv:1508.06615, 2015.

[25]Altman N S. An introduction to kernel and nearest-neighbor nonparametric regression[J]. The American Statistician, 1992, 46(3): 175-185.

[26]Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.

[27]Lowe D, Broomhead D. Multivariable functional interpolation and adaptive networks[J]. Complex syst, 1988, 2: 321-355.

作者简介：

卢达威（1983——）男，博士，主要研究领域为中文信息处理、汉语语言学等；

E-mail：wedalu@163.com

北京市海淀区颐和园路 5 号北京大学中文系 100871 18046530552

王星友（1991——），男，硕士研究生，主要研究领域为自然语言处理；

E-mail：ultimate010@gmail.com

北京市海淀区学院路 15 号北京语言大学信息科学学院 100083

袁毓林（1962——），男，教授、博士生导师，主要研究领域为语言学和汉语语言学、句法学、语义学、语用学，计算语言学和中文信息处理等；

E-mail：yuanyl@pku.edu.cn

北京市海淀区颐和园路 5 号北京大学中文系 100871