文章编号：1003-0077(2018)08-0009-10

引入词性标记的基于语境相似度的词义消歧

孟禹光，周俏丽，张桂平，蔡东风

（沈阳航空航天大学 人机智能中心，沈阳 辽宁 110136）

摘　要：目前的语境向量模型在对语义空间建模的时候，没有考虑到同一个词的不同词性具有不同的含义，将它们看作同一个点进行建模，导致得到的语境向量质量不高，使用这种语境向量计算语境相似度效果不好。针对该类问题，提出了一种加入词性特征的语境向量模型，加入词性后，可以将原本用语义空间中一个点表示的几个语义区分出来，得到质量更好的语境向量和语境相似度，进而得到更好的消歧效果。实验结果表明，这种建模方式可以有效区分不同词性的语义，在 2004 年的 Senseval-3 测试集上进行测试，准确率达到了 75.3%，并在 SemEval-13 和 SemEval-15 公开测试集上进行了测试，消歧效果相比未引入词性特征的模型均得到了提升。

关键词：语境向量；语境相似度；词义消歧；词性特征

中图分类号：TP391　　**文献标识码**：A

Word Sense Disambiguation Based on Context Similarity with POS Tagging

MENG Yuguang, ZHOU Qiaoli, ZHANG Guiping, CAI Dongfeng

(Human Machine Intelligence Research Center, Shenyang Aerospace University, Shenyang, Liaoning 110136, China)

Abstract: While learning embeddings, context2vec represent words with different parts-of-speech as one point without considering they often have different meanings. To avoid this low-quality context vectors and context similarity, we present a context2vec model with POS features to differentiate different meanings represented by one point in the vector space. Experiments show that the accuracy of word sense disambiguation reaches 75.3% on 2004 Senseval-3, outperforming baselines on SemEval-13 and SemEval-15.

Key words: context vector; context similarity; word sense disambiguation; part of speech features

0　引言

确定文本中某单词的实际含义，即词义消歧，简称 WSD，是自然语言处理领域中历史久远的问题，有着广泛的应用。目前可分为有监督方法、无监督方法和基于知识的三类方法。虽然已发表的有监督词义消歧系统在提供特定语义的大规模训练语料时有很好的表现，但缺乏大规模标注语料是其存在的主要问题。使用预训练的词向量可以在一定程度上解决这个问题。因为使用预先在大规模语料上训练的词向量，包含了较多的语义语法信息，用它来训练有监督系统，会使性能得到提升。而想要对句中的词义做推断，目标词和目标词的语境都需要清楚地表示出来。我们将语境定义为从一个句子中去掉目标词之后剩余的部分。为了更好地计算语境相似度，语境也需要以向量的形式进行表示。

在此前的消歧任务中[1]，语境只是简单地对目标词在一定窗口内的词向量求和或者加权平均来表示。但是由于目标词和其整体语境之间具有的内在联系，使用这种方法预训练的词向量所包含的信息十分有限。而想要对句中的词义做推断，目标词和目标词的语境向量都需要包含整个句子的信息。目前很多消歧系统共同的缺点是不包含语序信息。而长短期记忆网络（LSTM），尤其是双向长短期记忆网络（BLSTM），克服了以上缺点，可以对目标词周围的所有词进行建模，并将语序考虑进去。然而目前 BLSTM 模型都将一个词的不同词性，看做一个词来进行建模，显然不是十分正确的，因为同一个词

收稿日期：2017-10-19　**定稿日期**：2017-11-30

基金项目：教育部人文社会科学研究规划基金（18YJA870020）

如果词性不同会有不同的含义。

本文提出的方法是在训练之前,把词性加入到语料中,在训练的过程中把不同词性区分出来。把同一个词的不同词性,映射到语义空间的不同位置,可以更好地提取语境中所包含的信息。本文对三种不同的词性特征加入方法进行实验,即细分类词性特征、粗分类词性特征和只用实词的词性特征。通过实验找到了一种最为合适的词性特征引入方法,得到的消歧准确率在2004年的SE-3(Senseval-3 lexical sample dataset)测试集上达到了75.3%,并在SemEval-13和SemEval-15测试集上对比了加入词性前后系统的消歧效果,加入词性之后,消歧性能均有所提升。

本文结构安排如下:第一部分介绍相关工作;第二部分介绍模型;第三部分介绍实验设置;第四部分讨论实验结果;第五部分总结并介绍未来工作。

1 相关工作

一般来说,有监督的词义消歧方法比其他词义消歧算法效果更好,但是需要更大的训练集以达到这种效果,而获得大的训练集代价很大。本文证明,不需要使用大量的训练语料也可以达到很好的消歧效果。

使用大规模语料训练的词向量可以在一定程度上弥补有监督词义消歧对训练集要求过高的缺点。词向量能够在紧凑低维空间表示中包含词的语义语法信息,在很多NLP任务中都得到了很好的应用。在词向量研究中最具有代表性的就是Word2vec[2]和GloVe[3]。而这些技术的主要目的是在低维空间中对词的语义、语法信息进行表达,还没有对句子及语境中包含的信息进行低维表示。

最近几年,用神经网络训练词向量并建立语言模型,在情感分析、机器翻译和其他的自然语言应用中都取得了很大发展。对于词义消歧而言,Rothe和Schütze[4]将词向量扩展到了语义向量,并用语义向量作为特征训练了一个支持向量机分类器[5]。就像词向量一样,语境也可以以向量的形式表示,研究发现,在计算句子相似度[6]、词义消歧[1]、词义归纳[7]、词汇替换[8]、句子补全[9]等任务中,语境的向量表示取得了较好的效果,但他们的语境只是简单地对目标词一定窗口内的词向量求和或者加权平均来表示,起到的作用有限。

2016年Google[10]使用LSTM进行消歧,使用了1 000亿词新闻语料训练模型,在SemEval-2015测试集上达到了最好的消歧效果,它的消歧方法是使用一个词的语境预测目标词来进行消歧。LSTM模型是一种特殊的循环神经网络(RNN)模型,在1997年由Hochreiter和Schmidhuber[11]提出,它可以使RNN在对序列建模时更好地提取长距离信息。但是单向LSTM只能提取句子中位于目标词之前的信息。而语境向量模型(Context2vec)[12]使用了BLSTM训练,仅用了20亿词的语料作训练就在SemEval-2015测试集上达到了比Google的单向LSTM模型消歧高0.1%[13]的结果,它的目的是训练词向量和语境向量,利用得到的语境向量和词向量进行消歧,而且可以在其他的语言任务中得到较好的应用,从某种方面来说,BLSTM的性能更好。BLSTM[14]每个时刻的状态都包含了两个LSTM的状态,一个从左到右,另一个反之。这就意味着一个状态既可以包含前面的词的信息,也可以包含后面词的信息,在很多情况下,这对选择一个词的语义是非常必要的。本文中,通过向Context2vec中加入词性特征,进而训练得到语境向量,这种语境向量可以很好地提取语境中的语义语法信息。在得到的语境向量基础上,使用简单的k近邻算法进行消歧,实验证明本文的词义消歧系统不需要大量的训练语料,也可以达到很好的效果。

2 Context2vec

2.1 模型综述

Context2vec可以对目标词的语境进行学习,得到一个独立于特定任务的向量表示。这种模型基于Word2vec的连续词袋模型(CBOW)(图1),图中左侧计算目标词submitted的语境向量,仅考虑目标词周围一定大小窗口内的词,paper被忽略了,将John和a的词向量进行简单的加权平均得到submitted的语境向量,将右侧的目标词向量与得到的目标词的语境向量输入目标函数,训练模型。Context2vec用更有效的BLSTM替代了原来模型中固定窗口内词向量取平均的建模方式,如图2所示。

虽然两个模型的本质都是同时在低维空间学习语境和目标词的表示,但基于BLSTM的Context2vec可以更好地提取句子语境中的本源信息。

图2左侧说明Context2vec是如何表示句子语境的。使用BLSTM循环神经网络,将句中的词从

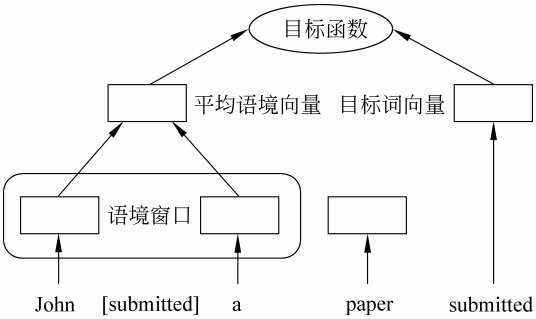

图 1　Word2vec 模型

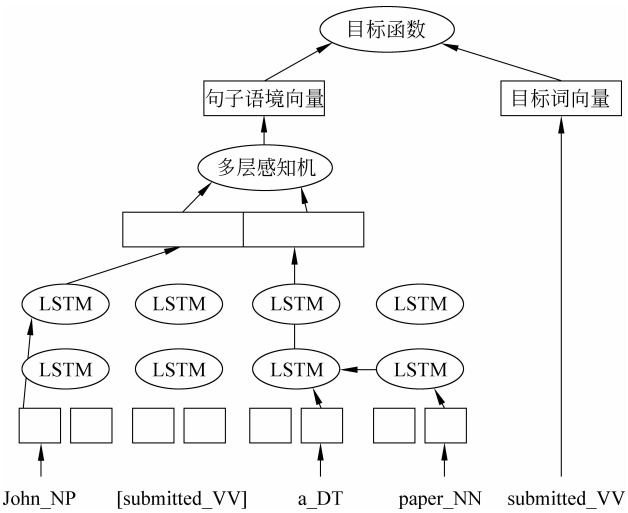

图 2　加入词性的 Context2vec 模型

左到右输入一个 LSTM，另一个从右到左进行输入。这两个网络的参数是完全独立的，包括一个从左至右和一个从右至左的语境向量。为表示句子中目标词的语境（John_NP［submitted_VV］a_DT paper_NN），首先将 LSTM 输出的左—右语境 John_NP 的向量表示和右—左语境 a_DT paper_NN 的向量表示拼接，这样做的目的是获得目标词语境中的相关信息。下一步将拼接后的向量输入到多层感知机中，就可以将两个方向的语境向量有机融合。将此层的输出作为目标词的句子语境表示。同时，目标词用它自己的向量表示（图 2 右侧），它的维度和句子语境向量的维度相同。我们注意到 Context2vec 和 Word2vec 模型之间的唯一的差别是 Word2vec 模型是将目标词的语境表示为周围一定窗口内语境词的简单平均，而使用了 BLSTM 的 Context2vec 则将整句话有机结合起来表示语境，不仅考虑了距离远一些的词，还考虑到了语序信息。

Word2vec 模型在内部使用语境建模，并将目标词的向量表示作为主要输出，Context2vec 更关注语境的表示。Context2vec 通过给目标词和目标词的语境指定相似的向量来进行建模。这就间接地导致了给具有相似句子语境的目标词指定了相似的词向量，相反地，给相似目标词的语境也指定了相似的向量表示。

2.2　语境向量计算公式

我们使用 Context2vec[12] 来得到一个句子级别的语境表示。lLS 表示从左到右读取句子的 LSTM，rLS 表示从右到左读取句子的 LSTM。给定一个有 n 个词（w）的句子（$w_{1:n}$），那么目标词 w_i 的语境就可以表示为两个向量的连接，如式（1）所示。

$$\mathrm{biLS}(w_{1:n},i)=\mathrm{lLS}(l_{1:i-1})\oplus\mathrm{rLS}(r_{n:i+1}) \quad (1)$$

其中 $l_{1:i-1}$ 和 $r_{n:i+1}$ 分别表示句中从左向右和从右向左输入到 lLS 和 rLS 中的词向量。下一步，对左右两侧语境表示的拼接输入到以下非线性函数，如式（2）所示。

$$\mathrm{MLP}(x)=L_2(\mathrm{ReLU}(L_1(x))) \quad (2)$$

其中 MLP 代表多层感知机，ReLU 是修正线性单元激活函数，而 $L_j(x)=W_jx+b_j$ 是一个全连接线性操作。则在句中位置 i 处的单词的语境表示定义如式（3）所示。

$$c=\mathrm{MLP}(\mathrm{biLS}(w_{1:n},i)) \quad (3)$$

将目标词和语境向量用相同维度的向量表示。模型训练使用 Word2vec 中所用的负采样方法。

2.3　特征选择

词性是一种很重要的语义语法信息，由于本文训练的 Context2vec 所使用的 ukWaC[15] 语料提供词性标注版本，我们直接在词的后面用下划线将词和词性结合，将其看作一个词，输入到 Context2vec 模型中进行训练，这样所得到的词向量，就会将同一个词的不同词性区分开。

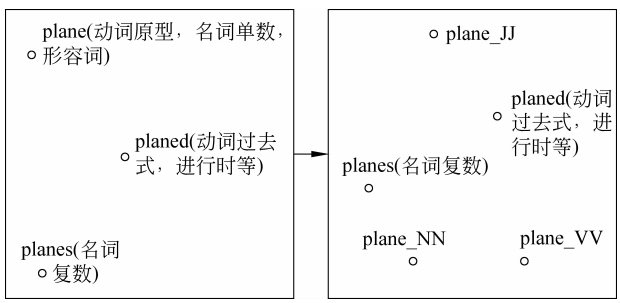

图 3　加入词性后，语义空间中的点的变化

从图 3 中可以发现，在加入词性之前，plane 的动词原型、名词单数和形容词三种用法，都通过

plane一个点表示,将三种不同的语义语法信息看作同一个点,显然是不合理的,这种情况在各种语言,尤其是英文中非常普遍。在加入词性之后,可以将不同的词性分别建模,plane的形容词(plane_JJ)、名词单数(plane_NN)、动词原型(plane_VV)等含义都在空间中重新分配了属于自己语义的点。通过这种方式,更好地捕捉语义语法信息,避免了语义上的混淆。

为了证明词性在Context2vec中的作用,我们还对三种不同的词性标注方式分别进行了实验:细分类词性标记、粗分类词性标记、单独用实词的词性标记。从中选择最好的词性特征加入方式。

2.4 模型说明

由于在模型的训练过程中要求语境和目标词更加接近,所以可以用一个语境向量求出和它余弦相似度最大的词向量,得到可以填补到空缺位置的词。得到的词越符合语义语法,说明模型训练得越好,和消歧准确率一样可以判断一个模型的好坏,用这种指标和准确率一起作为衡量模型质量的参考。

Context2vec在提出之时作者用k=1的近邻法消歧,得到了较好的效果,在实验过程中发现,使用k=1做开发集的实验,并不能完全说明模型的好坏。为了说明k值的选取过程,我们列出了加入词性前后在不同训练轮数e下得到的模型对给定语境的预测词,同时将模型在SE-3进行消歧达到的准确率进行对比(表1)。

如表1所示,对比2、4号相同训练轮数下加入词性前后的消歧准确率及预测词,虽然2号的消歧准确率高于4号,但是2号预测的目标词并不如4号预测的符合语义语法。同样对比1、3号结果,1号的消歧准确率虽然低于3号,但是预测的词比3号的更符合语义语法。这是由于表中使用k=1的k近邻法消歧得到的准确率,并不能完全说明模型的好坏,不能将其作为绝对的评价标准。我们对不同k值进行了测试,结果发现在给定模型的情况下,用k=5的k近邻法进行消歧,效果最好。因此本篇文章中,取k=5的k近邻算法进行消歧做对比实验。

表1 语境的预测词和SE-3测试集消歧结果

序号	有无词性特征	训练轮数e	消歧准确率/%	与John was [] last year语境最接近的词
1	无	1	71.2	born,interviewed,reelected
2	无	3	72.8	amazed,retiring,heartbroken
3	有	1	71.4	reelected,honored,opened
4	有	3	71.9	christened,born,hospitalised

3 实验

Context2vec模型可以训练得到高质量的语境向量和词向量,在多种自然语言处理任务中都有很好的应用。用于词义消歧也取得了较好的效果,为了测试加入词性特征后训练的模型的性能,进行了以下实验。

3.1 Context2vec训练语料

我们使用20亿词的ukWaC语料[15]的词性标注版本来训练模型,词性由TreeTagger自动标注,此工具的标注准确率约为95%,并将词性和词结合在一起,例如,apple我们将其写成了apple_NN构成一个新的词。这样可以将词性信息加入模型一同训练。用得到的模型求目标词的语境向量,就包含了这些词性信息。为加速训练过程,且方便和baseline实验作对比,并没有使用句子长度大于64的句子,这使得语料规模减小了10%。将所有的词都小写化并且把出现次数少于最小词频t的词看作未登录词,其中t取了不同的数值,做对比实验。实验证明不同的最小词频对实验结果有一定影响。

3.2 消歧方法

使用Python中的Chainer[16]工具包训练模型,用Adam[17]进行优化。为加速训练过程,使用了mini-batch训练,这样每次就只会将相同长度的句子加入同一个小的训练过程。

3.2.1 有监督词义消歧

Context2vec可以生成给定语境的语境向量,利用这些语境向量进行消歧,需要标注好语义的目标词的例句集合作为训练集。在消歧过程中,首先将训练集中包含目标词的多个例句的语境输入,得到每个语义若干例句的语境向量,然后计算待消歧句语境向量和这些语境向量的余弦相似度,将与待消歧句语境向量相似度最高的语境向量所对应的语义作为目标词的语义,得到消歧结果。

我们用的是2004年的Senseval-3 lexical sample dataset作为标注词义消歧数据集,其中包含了7 860个训练样本、57个目标词和3 944个测试样本。

在参数调优过程中，使用了留一法进行交叉验证，在训练出 Context2vec 模型之后，假如某个词在训练集中有 N 个样本，每次从训练集取出一个句子作为测试样本，其他 N-1 个样本作为训练样本，使用有监督词义消歧对测试样本进行消歧，对训练集中的 7 860 个句子进行以上操作，就得到了开发集的消歧结果。测试集是官方评测提供的，与开发集和训练集是完全独立的数据集。

还使用 SemEval-13 task 12 和 SemEval-15 task 13 两个公开评测集测试消歧准确率，目的是用最新的测试集验证这种方法的效果，这两个任务均为词义消歧任务，但是只提供了测试集，所以我们使用标记的 SemCor 和 SemCor＋OMSTI（One million sense-tagged instances）[18] 中的例句作为训练集，其中 SemEval-13 共有 621 个目标词、1 442 个测试样本。SemEval-15 共有 451 个目标词、924 个测试样本。

3.2.2　消歧流程

在消歧之前，使用非词性标注语料和词性标注语料分别训练两个语境向量模型，记为 M、MP。其中 M 用于未加词性特征的句子消歧，MP 用于加入词性特征的句子消歧。

对于待消歧句 Sen：Sodalities have an important role in【activating】laity for what are judged to be religious goals both personally and socially .

加入词性特征后句子为 SenP：Sodalities_NN have an important_JJ role_NN in【activating_VV】laity_NN for what are judged_VV to be religious_JJ goals_NN both personally_RB and socially_RB .

对应的语境向量分别为 V、VP，目标词在【】中，此目标词有五个不同的语义 S1、S2、S3、S4、S5，各语义对应解释如下：

S1：to initiate action in; make active.

S2：in chemistry, to make more reactive, as by heating.

S3：to assign (a military unit) to active status.

S4：in physics, to cause radioactive properties in (a substance).

S5：to cause decomposition in (sewage) by aerating.

训练集包含了此目标词的 228 个例句，由于官方给出的语义粒度过小，具有某些重叠，每个例句对应 1 或 2 个语义，若对应两个语义，则每个语义的权重为 0.5，否则为 1，将例句语境输入到语境向量模型 M、MP 中得到每个例句的语境向量，通过语境向量之间的余弦相似度，并应用 k＝5 的 k 近邻法，分别得到与 Sen、SenP 语境最接近的五个例句，如表 2 所示。

表 2　与待消歧句语境最接近的例句和语义

待消歧句	与待消歧句余弦相似度最高的例句	例句对应语义及权重
Sen	You step on to【activate】it .	S1/1.0
	Which parts of the sensory system are【activated】.	S2/1.0
	The【activating】field producing resonance is the electric and magnetic field .	S2/1.0
	There are ions【activated】by thermal motion in blood .	S2/1.0
	This clause has never yet been【activated】.	S1/0.5 或 S3/0.5
SenP	You step_VV on to【activate_VV】it .	S1/1.0
	Which parts_NN of the sensory_JJ system_NN are【activated_VV】.	S2/1.0
	Different_JJ genes_NN are【activated_VV】in different_JJ cells_NN .	S3/1.0
	The dysfunctional_JJ assumptions_NN are【activated_VV】.	S1/1.0
	This clause_NN has never_RB yet_RB been【activated_VV】.	S1/0.5 或 S3/0.5

表 2 中，通过计算句子的语境向量之间的相似度，能够得到与待消歧句语境接近的例句，而这些例句是标记好的，通过这些例句对应的语义出现的次数及权重，加权求和得到每个语义的打分：

Sen：Score(S1)＝1.5、Score(S2)＝3.0、Score(S3)＝0.5、Score(S4)＝0、Score(S5)＝0

SenP：Score(S1)＝2.5、Score(S2)＝1.0、Score(S3)＝1.5、Score(S4)＝0、Score(S5)＝0

根据得到的打分，加入词性前选择的语义为 S2，加入词性后选择的语义为 S1，显然加入词性通

过计算语境相似度,我们得到了正确的语义。

4 实验结果

为找出最有效的词性特征引入方法,本文使用了三种不同的词性标注方式训练模型:细分类词性标记、粗分类词性标记和单独用实词的词性标记。

实验中所使用的模型参数如表3所示。我们每次的最小取样数是850,在ukWaC语料训练一轮的时间大约是24小时。

表3 模型参数

单向语境词维度	300
LSTM隐藏层单元/输出节点数	600
多层感知机输入节点数	1 200
多层感知机隐藏节点数	1 200
语境向量维度	600
目标词向量维度	600
负采样样本数	10

4.1 细分类词性标记

使用TreeTagger标注的ukWaC语料,进行训练。找到在开发集中消歧效果最好的模型用来测试,开发集在最小词频 $t=100$,训练轮数 e 不同时对应的结果如表4所示。

表4 细分类词性标记下的开发集结果

序号	参数	消歧准确率/%
1	$t=100,e=1$	73.7
2	$t=100,e=2$	73.9
3	$t=100,e=3$	74.4
4	$t=100,e=4$	74.2
5	$t=100,e=5$	74.3

序号3所训练的模型最好,与加入词性前效果最好的模型作对比,结果如表5所示。

表5 细分类词性标记下的语境预测词和消歧结果对比

序号	参数	有无词性特征	John was [ ] last year	消歧准确率/%
1	$t=100,e=3$	无	amazed,retiring,heartbroken	74.8
2	$t=100,e=3$	有	re-elected,honored,opened	74.4

由表5中的结果对比可以发现,虽然1号的消歧准确率高于2号,但是它语境预测的目标词并不十分符合语义,甚至是语法(加入词性后,可以将词性一同预测出来,为了方便对比,此处并不列出)。而加入词性之后,可以看到2号可以较好地预测出目标词。说明词性的加入,还是对模型起到了较好的作用。虽然加入词性后,语境预测词的效果较好,但是在词义消歧任务上的准确率并没有达到加入词性之前的效果,究其原因可能是因为我们所使用的TreeTagger的词性标记种类(表6)过多。

表6 细分类词性标记

序号	TreeTagger词性标记	简单词性分类	序号	TreeTagger词性标记	简单词性分类
1	NN, NNS, NP, NPS	名词	7	CD	数词
2	JJ,JJS,JJR	形容词	8	DT,PDT,WDT	冠词
3	PP,WP,PP$	代词	9	IN	介词
4	RB, RBR, RBS, WRB	副词	10	CC	连词
5	MD, VB, VBD, VBG,VBN,VBP, VBZ, VD, VDD, VDG,VDN,VDP, VDZ, VH, VHD, VHG,VHI,VHD, VHZ, VV, VVD, VVG,VVN,VVP, VVZ	动词	11	EX, FW, POS, RP,SYM	其他
6	UH	感叹词			

由表6的词性标记方式可以发现,其中一个词的同一词性的不同时态也会有不同的标记,而这显然是没有必要的,这导致词表过大,降低了训练模型的效率,也影响了词性标注的准确率,进一步也影响了消歧的准确率。

4.2 粗分类词性标记

由于细分类的词性标记准确率的影响,消歧效果并没有得到提升。改用表7中的词性映射方式,对同一类简单词性不再进行区分。

根据语言学知识,并没有将TreeTagger中的助动词进行映射,因为某些助动词对应的实意动词已经映射为了VV,而助动词没有映射,就可以将其区分开。

表 7　粗分类词性标记

序号	TreeTagger 词性标记	对应的词性标记	序号	TreeTagger 词性标记	对应的词性标记
1	CC	CC	7	NP,NPS	NP
2	DT,PDT,WDT	DT	8	PP,PP$,WP,WP$	PP
3	IN	IN	9	RB,RBR,RBS,WRB	RB
4	JJ,JJR,JJS	JJ	10	VV,VVD,VVG,VVN,VVP,VVZ	VV
5	MD	MD	11	EX, FW, LS, POS,RP,SYM, UH	O
6	NN,NNS	NN			

为了使词表和加入词性之前尽量一致，把原始 Context2vec 中没有用到的低频词都替换为＜unk＞，为了保证训练速率，在此基础上取最小词频 $t=50$ 和 $t=100$，把此方法在开发上得到最好结果的模型用于测试，并与其他词性标注方式训练的模型的消歧结果和语境预测词作对比 ，开发集结果如表 8 所示。

表 8　粗分类词性标记的开发集结果

序号	参数	消歧准确率/%	序号	参数	消歧准确率/%
1	$t=100,e=1$	75.1	6	$t=50,e=1$	74.8
2	$t=100,e=2$	75.2	7	$t=50,e=2$	75.0
3	$t=100,e=3$	75.2	8	$t=50,e=3$	75.4
4	$t=100,e=4$	75.2	9	$t=50,e=4$	74.9
5	$t=100,e=5$	75.3	10	$t=50,e=5$	75.0

选取开发集表现最好的模型作测试，结果如表 9 所示。

表 9　粗分类词性标记下的语境的预测词和消歧结果对比

序号	参数	词性特征加入方式	John was [ ] last year	消歧准确率/%
1	$t=100,e=3$	无	amazed, retiring, heartbroken	74.8
2	$t=100,e=3$	细分类词性标记	reelected,honored, opened	74.4
3	$t=50,e=3$	粗分类词性标记	reelected,baptised, due	74.3

对表 9 的结果进行对比，发现加入粗分类的词性标记后，3 号并没有达到 1 号的消歧效果，只是语境预测词更符合语义语法。而根据 2 号和 3 号的结果对比，消歧准确率和语境预测词都十分接近。这是因为两种词性标记方式都对虚词进行了标注。对训练语料进行检查发现，其中有些本该标注为 DT 的 that 标注成了 IN，本该标注成 IN 的 upon 标注成了 RP。这种虚词虽然数量有限，但是每一个虚词在训练语料中出现的频率都很高，一般用来构成句子框架，这些词如果标注错误，对语义空间所造成的影响更大。

4.3　实词的词性标记

基于 4.1 节和 4.2 节的实验结果，决定只使用实词对训练语料进行标注。实词，有实在意义，在句子中能独立承担句子成分。我们选择了几种对语义影响较大的实词，而代词和数词也属于实词，但是它们一般只具有单一词性，而且不同词性表达的语义也是相同的。其他的虚词不再进行标注，如表 10 所示。

表 10　实词的词性标记

TreeTagger 词性标记	实词标记 1	实词标记 2
RB,RBR,RBS	RB	RB
NNS,NN	NN	NN
NP,NPS	NP	NN
JJS,JJR,JJ	JJ	JJ
VV,VVD,VVG,VVN,VVP,VVZ	VV	VV

表 10 列出的两种实词标记方式，分别取名为实词标记 1 和实词标记 2。它们的区别仅在于是否将名词细分为普通名词(NN)和专有名词(NP)。这是因为，某些普通名词同时还有专有名词的含义，如 plane 既有普通名词飞机，又有专有名词作为人名的作用。同时在实验中我们发现，有些在命名实体中出现的普通名词，TreeTagger 工具将其标记为了 NP。到底是应该将所有的名词都标记为 NN 还是将 NN 和 NP 区分标记能达到更好的消歧效果，我们需要通过实验来进行判断。

为保证训练速度和词表大小，将实词标记 2 实验中的最小词频 t 取 10，将实词标记 1 中的最小词频 t 取 50，此时词表大小均为 22 万左右，训练效率接近。若实词标记 2 取 $t=50$，则词表大小为 19 万，会相比 $t=10$ 减少 3 万低频词，我们还对取相同的 t 值(100)训练的模型的消歧结果进行了对比，此时实词标记 1 词表大小为 19 万，实词标记 2 词表大小为 17 万。实词标记 1 下的开发集结果如表 11 所示，

实词标记2下的开发集结果如表12所示。

表11 实词标记1下的开发集结果

序号	参数	消歧准确率/%	序号	参数	消歧准确率/%
1	$t=100,e=1$	74.8	6	$t=50,e=1$	74.8
2	$t=100,e=2$	75.5	7	$t=50,e=2$	74.8
3	$t=100,e=3$	75.2	8	$t=50,e=3$	75.1
4	$t=100,e=4$	75.3	9	$t=50,e=4$	75.2
5	$t=100,e=5$	75.1	10	$t=50,e=5$	75.0

表12 实词标记2下的开发集结果

序号	参数	消歧准确率/%	序号	参数	消歧准确率/%
1	$t=100,e=1$	75.2	6	$t=10,e=1$	75.2
2	$t=100,e=2$	75.4	7	$t=10,e=2$	75.1
3	$t=100,e=3$	75.8	8	$t=10,e=3$	75.3
4	$t=100,e=4$	75.2	9	$t=10,e=4$	67.4
5	$t=100,e=5$	75.1	10	$t=10,e=5$	67.4

从表12可以看出，用实词标记2标注的训练语料所训练的模型效果普遍高于实词标记1。对于表12中的9、10号实验结果准确率突然降低的现象，在Context2vec达到最优点之后均会出现这种过拟合现象，由于此时参数t取10，相比t取50而言，词表中包含了三万个出现频率高于10低于50的低频词，导致这种现象过早地出现。因为命名实体中的普通名词数量远远高于有专有名词含义的普通名词。如：在训练语料中，如果将名词细分为NP和NN，会出现Valley_NP Park_NP，Web_NP Author_NP Dr._NP Walton_NP等词语，其中的人名虽然被区分出来，但是也将本应属于普通名词范畴的词标记为了专有名词，把原本语义空间中的一个点强行划分为了两个点，导致了消歧效果的降低。

把此方法得到的最好模型与其他词性标注方式得到的最好模型的消歧结果和语境预测词作对比，如表13所示。

表13 实词标记2下的语境的预测词和消歧结果对比

序号	参数	词性特征加入方式	John was [ ] last year	消歧准确率/%
1	$t=100,e=3$	无	amazed, retiring, heartbroken	74.8

续表

序号	参数	词性特征加入方式	John was [ ] last year	消歧准确率/%
2	$t=100,e=1$	细分类词性标记	reelected, honored, opened	74.4
3	$t=50,e=2$	粗分类词性标记	reelected, baptised, due	74.3
4	$t=100,e=3$	实词标记2	reelected, elected, reappointed	75.3

由表13的结果可以看出，序号4不管在预测词还是消歧效果上，都要好于之前三种方法训练的模型。通过两种指标的对比，我们可以得出结论：使用实词标记2标注的训练语料训练Context2vec模型，模型性能得到了提升。

以上的结果都是用单一的k值得到的结果进行对比，不一定具有代表性。为了进一步说明结果所具有的代表性，我们将序号4得到的模型和序号1模型使用不同的k值(1～10)进行消歧得到的结果进行对比，通过对比发现，在不同的k值下，消歧效果基本都有所提升，符合我们之前得到的结论。

我们还和在SE-3做过消歧的其他系统进行了比较，结果如表14所示。

表14 不同系统在SE-3测试集的结果/%

2006 Ando	2015 Rothe	2016 Melamud	2004 Grozea	2004 Strapp-arava	Ours-1	Ours-2
74.1	73.6	72.8	72.9	72.6	74.8	75.3

从表14可以看出，最好的结果是由Ando[19]达到的，他使用了一种交替结构优化的半监督方法，过程十分繁琐。而我们比它提升了1.2%，且方法简单很多，仅仅需要在模型训练好之后，应用一次k近邻算法。而第二好的结果由Rothe和Schutze[4]达到，他们通过向已有的系统中加入词向量特征达到此结果，他们的主要目的是训练词义向量，没有对词义消歧做过于深入的研究，因此没有达到超过Ando的消歧效果。SE-3评测的前两名结果分别为2004 Grozea、2004 Strapparava。2006年达到了74.1%准确率之后，除了2015年达到了较好的效果之外，一直没有超过Ando的消歧系统。2016 Melamud提出的Context2vec模型使用k=1的k近邻法得到的结果为72.8%，这是因为Melamud提出的Context2vec在多种自然语言处理任务中都能得到较好的应用，没有单独对词义消歧进行深入研究，

我们将 Context2vec 在消歧上的应用进行了扩展，Ours-1 是我们用未加词性的 Context2vec 模型通过开发集选取参数 k，取参数 k＝5 时消歧效果最好，可以达到 74.8%，Ours-2 在 Ours-1 的基础上加入了词性特征，选取同样的 k 值，可以看出，加入词性特征之后，消歧准确率提高了 0.5%。

在这一部分，我们使用了三种不同的词性标记方式来标注训练语料并训练模型。其中前两种无法达到未加词性之前的效果，分析原因是虚词在训练语料中出现次数很多，它们是组成句子框架的主要成分，它们的词性标记错误会给建模带来更大的影响，而且这些词的不同词性表示的语义往往是相同的。而实词标记 2 得到了比不加词性更好的效果。

我们还在其他的测试集中对此方法进行了验证，如 SemEval-13 task 12 和 SemEval-15 task 13，由于此任务没有提供每个消歧词的例句，我们使用标记的 SemCor 和 SemCor＋OMSTI(One million sense-tagged instances)[18] 中的例句作为训练语料进行消歧。而 Raganato[20] 使用 Context2vec 和其他系统使用相同的训练语料进行了比较，其中 Context2vec 在多个公开评测集中均取得了非常好的消歧效果。我们使用了与之相同的训练语料，但是由于参数不同的原因，无法达到和 Raganato 相同的准确率。我们使用自己的参数，进行了对比实验。只对加入实词标记方法 2 的词性特征前后的消歧结果进行对比。结果如表 15 所示。

表 15　加入词性前后消歧结果对比

训练语料	测试集	加入词性前/%	加入词性后/%
Semcor	SemEval-13	63.4	63.8
Semcor	SemEval-15	69.6	70.1
Semcor＋OMSTI	SemEval-13	65.0	65.4
Semcor＋OMSTI	SemEval-13	69.7	70.3

由表 15 中的结果我们可以发现，相比加入词性之前，使用我们的方法加入词性特征的消歧效果在两个公开的测试集上也都得到了提升。其中 SemEval-13 和 SemEval-15 测试结果分别提升了 0.4%和 0.3%。提升效果不明显的原因主要有两个，首先每个目标词对应的例句数量较少，其次是词性标注的准确率限制了这种方法的作用。

5　结束语

我们通过实验，提出了一种将词性特征加入到 Context2vec 中建模的方法，更好地对语义进行建模，在消歧任务中达到了更好的效果，其中最好的结果已经比未加词性之前提高了 0.5%，并在多个国际公开评测集中有所提升。虚词和代词在句子中出现频率很高，它们的词性标注准确率对 Context2vec 是否能在语义空间上正确建模起着至关重要的作用。而且虚词的不同词性语义是基本相同的，不需要在语义空间中分配多个点进行建模。由于词性标注的准确率无法达到 100%，而且同一词性之间的细分准确率更低，限制了这种方法的作用。这种语境相似度消歧的方法依赖例句的质量和数量，可以考虑使用标签扩展(LP)算法对例句进行扩展。

参考文献

[1] Chen X, Liu Z, Sun M. A Unified Model for Word Sense Representation and Disambiguation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014:1025-1035.

[2] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv: 1301.3781V3,2013,9.

[3] Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014:1532-1543.

[4] Rothe S, Schütze H. Autoextend: Extending word embeddings to embeddings for synsets and lexemes [J]. arXiv preprint arXiv:1507.01127, 2015.

[5] Zhi Z, Ng H T. It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text [C]//Proceedings of the Meeting of the Association for Computational Linguistics, DBLP, 2010:78-83.

[6] Huang E H, Socher R, Manning C D, et al. Improving word representations via global context and multiple word prototypes[C]//Proceedings of the Meeting of the Association for Computational Linguistics: Long Papers. 2012:873-882.

[7] Kågebäck M, Johansson F, Johansson R, et al. Neural context embeddings for automatic discovery of word senses[C]//Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, 2015: 25-32.

[8] Melamud O, Levy O, Dagan I. A Simple Word Embedding Model for Lexical Substitution[C]//Proceedings of the The Workshop on Vector Space Modeling for Natural Language Processing. 2015:1-7.

[9] Liu Q, Jiang H, Wei S, et al. Learning semantic word

embeddings based on ordinal knowledge constraints [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015, 1: 1501-1511.

[10] Yuan D, Doherty R, Richardson J, et al. Word sense disambiguation with neural language models[J]. arXiv preprint arXiv:1603.07012, 2016.

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

[12] Melamud O, Goldberger J, Dagan I. context2vec: Learning Generic Context Embedding with Bidirectional LSTM[C]//Proceedings of the Signll Conference on Computational Natural Language Learning, 2016:51-61.

[13] Raganato A, Camacho-Collados J, Navigli R. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison[C]//Proceedings of the EACL, 2017: 99-110.

[14] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5-6):602.

[15] Baroni M, Bernardini S, Ferraresi A, et al. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora[J]. Language resources and evaluation, 2009, 43(3): 209-226.

[16] Tokui S, Oono K, Hido S, et al. Chainer: a next-generation open source framework for deep learning [C]//Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS), 2015, 5: 1-6.

[17] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv: 1412.6980, 2014.

[18] Taghipour K, Ng H T. One million sense-tagged instances for word sense disambiguation and induction [C]//Proceedings of the Nineteenth Conference on Computational Natural Language Learning, 2015: 338-344.

[19] Ando R K. Applying alternating structure optimization to word sense disambiguation[C]//Proceedings of the Tenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2006: 77-84.

[20] Raganato A, Camacho-Collados J, Navigli R. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison[C]//Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, 2017.

孟禹光(1991—),硕士研究生,主要研究领域为自然语言处理。
E-mail:465611370@qq.com

张桂平(1962—),硕士生导师,主要研究领域为自然语言处理、知识工程。
E-mail: zgp@gesoft.com

周俏丽(1977—),硕士,副教授,主要研究领域为自然语言处理。
E-mail:27401082@qq.com