

一种受限玻尔兹曼机的词义消歧方法

张春祥¹, 李海瑞², 高雪瑶²

(1. 哈尔滨理工大学 软件与微电子学院, 黑龙江 哈尔滨 150080;
2. 哈尔滨理工大学 计算机科学与技术学院, 黑龙江 哈尔滨 150080)

摘 要: 针对汉语一词多义现象, 根据上下文所蕴含的语言学知识, 采用受限玻尔兹曼机 (restricted boltzmann machine, RBM) 来确定歧义词汇的真实含义。选取歧义词汇左右邻接的四个词单元中的词形、词性和语义类作为消歧特征。同时, 使用 RBM 来构建词义消歧模型。结合 SemEval-2007: Task#5 的训练语料和哈尔滨工业大学的语义标注语料来优化 RBM 的参数。利用 SemEval-2007: Task#5 的测试语料对词义消歧模型进行测试。实验结果表明: 相对于贝叶斯词义消歧分类器而言, 受限玻尔兹曼机词义消歧方法的消歧准确率有所提高。

关键词: 受限玻尔兹曼机; 消歧特征; 词义消歧; 训练语料

DOI: 10.15938/j.jhust.2019.05.019

中图分类号: TP391.2 文献标志码: A 文章编号: 1007-2683(2019)05-0116-06

A Method of Word Sense Disambiguation with Restricted Boltzmann Machine

ZHANG Chun-xiang¹, LI Hai-rui², GAO Xue-yao²

(1. School of Software and Microelectronics, Harbin University of Science and Technology, Harbin 150080, China;
2. School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China)

Abstract: For polysemy phenomenon in Chinese, Restricted Boltzmann Machine (RBM) is adopted to determine the true meaning of ambiguous vocabulary where linguistic knowledge in context is used. Word form, part of speech and semantic categories in four left and right lexical units adjacent to an ambiguous word are selected as disambiguation features. At the same time, RBM is used to construct word sense disambiguation (WSD) model. Training corpus in SemEval-2007: Task#5 and semantic annotation corpus in Harbin Institute of Technology are used to optimize parameters of RBM. Test corpus in SemEval-2007: Task#5 is used to evaluate WSD model. Experimental results show that compared with Bayesian word sense disambiguation classifier, disambiguation accuracy of WSD method with RBM is improved.

Keywords: Restricted Boltzmann Machine; disambiguation features; word sense disambiguation; training corpus

收稿日期: 2017-11-08

基金项目: 国家自然科学基金(61502124, 60903082); 中国博士后科学基金(2014M560249); 黑龙江省自然科学基金资助项目(F2015041, F201420).

作者简介: 李海瑞(1992—), 男, 硕士研究生;
高雪瑶(1979—), 女, 博士, 教授.

通信作者: 张春祥(1974—), 男, 博士, 教授, E-mail: z6c6x666@163.com.

0 引言

词义消歧是计算语言学领域的关键性研究课题。近年来,随着文本数量的激增,自动判别词语的含义有着越来越广泛的需求。针对这一问题,国内外许多学者开展了大量的研究。

高雪霞^[1]针对词义错误配对问题,提出了基于 Jaccard 系数的词义消歧算法,利用 WordNet 知识库中的知识源来表示歧义词的词义信息并生成词义资源库。同时,结合提出的基于 Jaccard 系数词义消歧算法来完成信息检索。杨陟卓^[2]提出了一种基于上下文翻译的有监督词义消歧方法,将由译文所组成的上下文当作伪训练语料,并利用真实训练语料和伪训练语料共同确定歧义词的词义。陈浩^[3]提出了一种基于语言模型的无指导词义消歧方法,在基于术语抽取的基础上,使用基于统计的语言模型来提高消歧性能。史兆鹏^[4]提出了一种多特征词义消歧方案,通过依存句法分析提取上下文中多义词及义项的词性、依存结构和依存词特征。同时,细化特征粒度,根据多特征构造权值函数,选择权值最大的义项作为多义词的义项。李冬晨^[5]将句法分析与词义消歧相结合。根据层次化语义知识的句法分析框架,并利用句法结构信息对文法模型进行调整。同时,给出了一种句法分析和词义消歧一体化方法。闫蓉^[6]提出了一种上下文边界可变的中文词义消歧模型。通过分词来调整消歧上下文边界,构建多义词义项搭配库,来计算词语之间的语义相关度。钱涛^[7]提出了一种基于超图的词义归纳模型。首先根据词汇链来发现目标单词的上下文实例之间的高阶语义关系;然后使用上下文实例表示结点,利用词汇链发现超边以构建超图;最后使用基于最大密度超图谱聚类算法来发现词义。王少楠^[8]将出现在歧义词上下文语境中有明确含义的实词作为模型的输入。在上下文中,获取可以表示歧义词词义的其它特征。利用贝叶斯模型来整合这两种信息,共同实现歧义词的词义表示和归纳。张仰森^[9]针对中文文本语义错误,提出了一种基于语义搭配知识库和证据理论的语义错误侦测模型。利用知网来提取词语搭配的语义信息,使用词语搭配聚合度来进行辅助过滤。同时,利用证据理论来判定语义搭配错误。Ivan^[10]提出了一种特定领域的词义消歧方法。该方法使用了特定领域的测试语料库和特定领域的辅助语料库。通过抽取相关词语来获得特

定领域的辅助语料库。Koppula^[11]提出了一种基于图的词义消歧方法。使用词汇知识库来构建无向图模型,利用页面排序算法和随机游走算法来进行词义消歧。Iacobacci^[12]研究了如何利用词语嵌入来进行词义消歧。将词语嵌入应用于有监督词义消歧之中,并深入分析了不同参数对性能的影响。Ben-net^[13]分析了现有的语义分布学习方法,统计语义频率,生成了一个大规模的语义数据集。Hender-son^[14]以分布式语义为基础提出了一个向量空间模型,使用逻辑向量代替消歧特征,利用半监督方法进行词义消歧。唐共波^[15]将知网(HowNet)中表示词语语义的义原信息融入到语言模型的训练中,利用义原向量对多义词进行向量化表示,并将其应用于词义消歧。通过义原向量对词语进行向量化表示,实现了词语语义特征的自动学习,提高了特征学习效率。

本文将歧义词左右邻接的4个词单元的词形、词性和语义类作为消歧特征,将受限玻尔兹曼机作为消歧模型,提出一种基于RBM的词义消歧方法来判别歧义词的语义类别。

1 消歧特征的选择

大多数的文本中存在着一词多义^[16]的现象,这种现象给机器翻译带来了很大的困扰。只有先判别歧义词的真实语义,才能对文本进行有效快速的分类和翻译。在消歧过程中,结合相关的语言学知识,根据歧义词所在的上下文语境中的消歧特征可以有效地判别歧义词的真实含义。因此,消歧特征对于语义分类而言将是至关重要的。

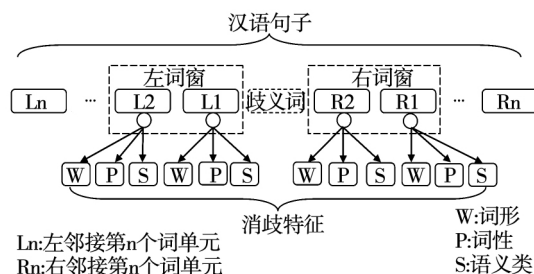


图1 消歧特征的提取过程

Fig.1 The process of extracting disambiguation features

本文以歧义词为中心,使用左右词窗来提取上下文中的语言学信息作为消歧特征。消歧特征的提取过程如图1所示。在左右词窗中,分别包含了两个邻接的词单元。选取左右词窗中的词形(W)、词

性(p)和语义类(s)作为消歧特征,来判别歧义词的语义类别。

歧义词“菜”具有两种语义类别,分别是“dish”和“vegetable”。以包含歧义词“菜”的汉语句子为例,来说明消歧特征的选取过程。

汉语句子“在厨师大赛上,他完成的这道菜令所有评委都赞口不绝。”。

对该汉语句子进行分词处理,其结果如下所示:

分词结果“在 厨师 大赛 上, 他 完成 的 这 道 菜 令 所有 评委 都 赞口不绝。”。

以该汉语句子的分词结果为基础,标注出每个单词的词性,其结果如下所示:

词性标注结果“在/p 厨师/n 大赛/n 上/nd, /wp 他/r 完成/v 的/u 这/r 道/q 菜/n 令/v 所有/b 评委/n 都/d 赞口不绝/i。/wp”。

根据《同义词词林》,标注出每个单词的语义类别^[17],其结果如下所示:

语义类标注结果“在/p/K 厨师/n/A 大赛/n/H 上/nd/C, /wp/-1 他/r/A 完成/v/I 的/u/K 这/r/E 道/q/D 菜/n/B 令/v/J 所有/b/E 评委/n/D 都/d/K 赞口不绝/i/K。/wp/-1”。

以歧义词“菜”为中心,开设长度为2的左右词窗,获得了歧义词左右邻接的4个词单元。它们分别是“这/r/E”、“道/q/D”、“令/v/J”、“所有/b/E”。从每个词单元中,抽出词形、词性和语义类作为消歧特征。词形特征分别为“这”、“道”、“令”和“所有”;词性特征分别为“r”、“q”、“v”和“b”;语义类特征分别为“E”、“D”、“J”和“E”。此处,共得到了12个消歧特征,分别为:这、r、E、道、q、D、令、v、J、所有、b、E。

从该实例中所提取的词义消歧特征如图2所示。12个词义消歧特征分别为: w_{l2} 、 p_{l2} 、 s_{l2} 、 w_{l1} 、 p_{l1} 、 s_{l1} 、 w_{r1} 、 p_{r1} 、 s_{r1} 、 w_{r2} 、 p_{r2} 、 s_{r2} 。

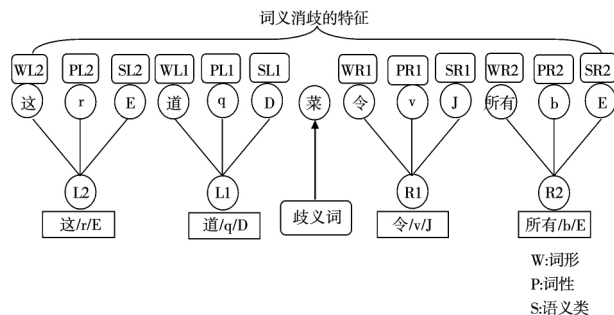


图2 词义消歧特征的示意图

Fig. 2 Features of word sense disambiguation

2 基于受限玻尔兹曼机的消歧过程

受限玻尔兹曼机是一种随机递归的神经网络^[18],具有接收输入数据集并学习其概率分布的功能。RBM有两层,一层是可视层,另一层是隐藏层^[19]。每层都由若干个神经元组成,每个神经元取1或0两种状态。其中,状态1表示激活状态;状态0表示关闭状态。在同一层中,神经元之间是相互独立的。在不同层中,神经元相互之间处于连接状态。因此,每个神经元的激活条件是相互独立的。

RBM可以作为多种深度学习网络的基本组成单元^[20]。在降维、分类和特征学习中,RBM有着广泛的应用。通常,使用监督学习方法来训练RBM的相关参数。

在哈尔滨工业大学人工语义标注语料中,每个汉语句子都进行了分词、词性标注和语义标注处理。从语义标注语料中,分别提取词形、词性和语义类。分别构造词形表、词性表和语义类表。在词形表中,每个单词都有唯一的序号;在词性表中,每个词性都有唯一的序号;在语义类表中,每个语义类都有唯一的序号。选取歧义词的左右邻接四个词单元的词形、词性和语义类作为消歧特征。从词形表、词性表和语义类表中,可以得到消歧特征所对应的序号(N)。这些序号构成了消歧特征向量 Feature。

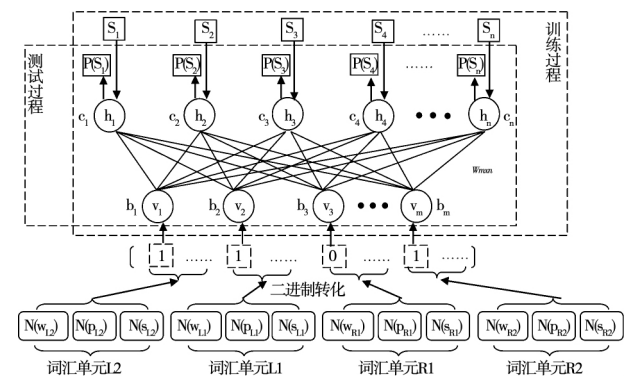


图3 基于RBM的词义消歧过程

Fig. 3 The process of WSD based on RBM

本文使用RBM来对歧义词进行语义分类。基于RBM的词义消歧框架如图3所示。在图3中, $v = \{v_1, v_2, \dots, v_m\}$ 表示RBM可视层中的 m 个神经元, $h = \{h_1, h_2, \dots, h_n\}$ 表示RBM隐藏层中的 n 个神经元。 $W_{m \times n}$ 表示连接可视层与隐藏层之间的权值矩阵。参数 $b = \{b_1, b_2, \dots, b_m\}$ 和 $c = \{c_1, c_2, \dots, c_n\}$ 为RBM的偏移量。 $Feature = \{N(w_{l2})\}$,

$N(p_{l2}), N(s_{l2}), N(w_{l1}), N(p_{l1}), N(s_{l1}), N(w_{r1}), N(p_{r1}), N(s_{r1}), N(w_{r2}), N(s_{r2}), N(s_{r2})$ 作为受限玻尔兹曼机的消歧特征向量。

在训练过程中,输入消歧特征向量 $Feature = \{N(w_{l2}), N(p_{l2}), N(s_{l2}), N(w_{l1}), N(p_{l1}), N(s_{l1}), N(w_{r1}), N(p_{r1}), N(s_{r1}), N(w_{r2}), N(s_{r2}), N(s_{r2})\}$ 和所对应的语义类 $S = \{S_1, S_2, \dots, S_n\}$ 。经过 k 次训练后,可以得到 RBM 的优化参数,即权重矩阵 $W_{m \times n}$ 、偏移量 $b = \{b_1, b_2, \dots, b_m\}$ 和 $c = \{c_1, c_2, \dots, c_n\}$ 。在测试过程中,输入消歧特征向量 $Feature = \{N(w_{l2}), N(p_{l2}), N(s_{l2}), N(w_{l1}), N(p_{l1}), N(s_{l1}), N(w_{r1}), N(p_{r1}), N(s_{r1}), N(w_{r2}), N(s_{r2}), N(s_{r2})\}$ 。优化后的 RBM 将输出歧义词在每个语义类下的概率分布。对于歧义词汇而言,它有 n 种语义类别 $S_1, S_2, \dots, S_n$ 。基于 RBM 的词义消歧模型的输出结果为概率分布 $P(S_1), P(S_2), \dots, P(S_n)$ 。

3 基于 RBM 的词义消歧模型训练

在训练过程中,RBM 的神经元只能接收 0 或 1 形式的二进制数。因此,需要将训练数据转换成二进制数。训练数据包括消歧特征向量 ($Feature$) 和语义类 (S) 两个部分。

$Feature$ 和 S 转化为二进制数的过程如图 4 所示。在图 4 中, $N^i(w_{l2})$ 表示 $Feature$ 的第 i 个特征向量中左侧邻接的第 2 个词单元的词形特征序号; $N^i(p_{l2})$ 表示 $Feature$ 的第 i 个特征向量中左侧邻接的第 2 个词单元的词性特征序号; $N^i(s_{l2})$ 表示 $Feature$ 的第 i 个特征向量中左侧邻接的第 2 个词单元的语义类特征序号。 S_j^i 表示第 i 个特征向量所对应的语义类。

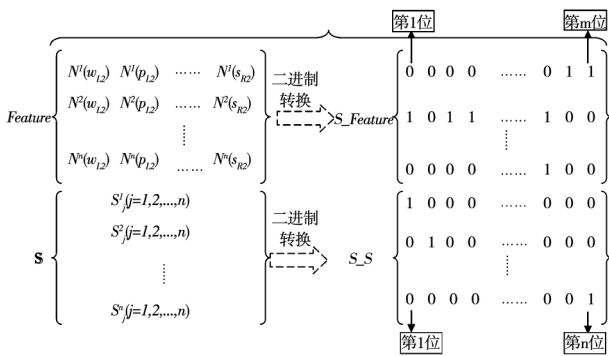


图4 消歧特征向量和语义类的二进制转换

Fig. 4 Binary transformation between disambiguation feature vector and semantic class

在图 4 中, $Feature$ 和 S 经过二进制转换,得到了图中右侧的标准消歧特征向量 ($S_Feature$) 和标准语义类 (S_S)。在标准消歧特征向量中,共有 n 个特征向量,每个特征向量用 m 位二进制数来表示。在标准语义类中,共有 n 个特征向量,每个特征向量用 n 位二进制数来表示。因此,标准消歧特征向量为 $n \times m$ 阶的二值矩阵,标准语义类为 $n \times n$ 阶的二值矩阵。

本文改进了对比散度 (contrastive divergence, CD) 算法^[21],并对基于 RBM 的词义消歧模型进行训练。

训练算法:

输入: $S_Feature, S_S, k, \theta(W, b, c)$

输出: 更新后的参数 $\theta(W, b, c)$

1. 初始化: $W=0, b=0, c=0$;

2. 将所有的 $v = \{v_1, v_2, \dots, v_m\} \in S_Feature$, $h = \{h_1, h_2, \dots, h_n\} \in S_S$ 的二进制特征向量,输入到 RBM 中;

3. for $t=1, 2, \dots, k$ do

1) 利用式 (1) 计算隐藏层的每个隐藏结点 h_j 值为 1 的概率 $P(h_j=1|v)$, 记 $P_{(j|v)} = P(h_j=1|v)$, 其中, $\sigma(x)$ 为激活函数,其表达形式如式 (2) 所示。

$$P(h_j^{(t)}=1|v^{(t)}) = \sigma(b_j^{(t)} + \sum_{i=1}^m w_{ij}^{(t)} v_i^{(t)}) \quad (1)$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

2) 产生 0 到 1 之间的随机数记为 r_j , 利用式 (3) 得到隐藏层结点 h_j 的值, 即

$$h_j = \begin{cases} 1, & \text{if } r_j < P_{(j|v)} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

3) 利用式 (4) 计算可视层的每个可视结点 v_i 值为 1 的概率 $P(v_i=1|h)$, 记 $P_{(i|h)} = P(v_i=1|h)$, 即

$$P(v_i^{(t+1)}=1|h^{(t)}) = \sigma(c_j^{(t)} + \sum_{j=1}^n w_{ij}^{(t)} h_j^{(t)}) \quad (4)$$

4) 产生 0 到 1 之间的随机数记为 r_i , 利用式 (5) 得到可视层结点 v_i 的值, 即

$$v_i = \begin{cases} 1, & \text{if } r_i < P_{(i|h)} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

5) 利用公式 (6)、(7) 和 (8) 计算参数 $\theta(W, b, c)$ 的更新值。

$$w_{ij} = w_{ij} + [P(h_i=1|v^{(1)}) v_j^{(1)} - P(h_i=1|v^{(k)}) v_j^{(k)}] \quad (6)$$

$$b_j = b_j + [v_j^{(1)} - v_j^{(k)}] \quad (7)$$

$$c_i = c_i + [P(h_i=1|v^{(1)}) - P(h_i=1|v^{(k)})] \quad (8)$$

6) $i++$, $j++$ 。

4 实验

在实验中,所使用的训练数据和测试数据选自于 SemEval - 2007: Task#5 的训练语料和测试语料。SemEval - 2007: Task#5 是 ACL2007 的一个组成部分,即 SemEval - 2007 国际语义评测的中英文词汇任务。该任务共包含 40 个歧义词,从中选取 18 个常见的歧义词。所选取歧义词的训练语料和测试语料分布如图 5 所示。

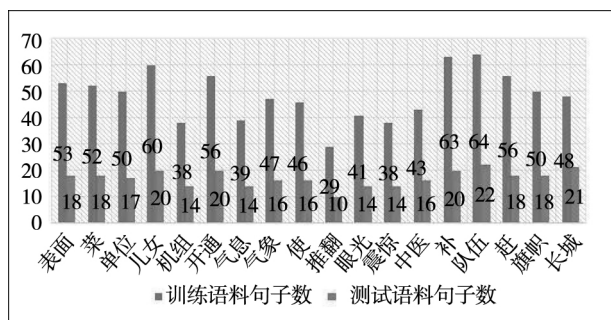


图 5 歧义词训练语料和测试语料的分布

Fig. 5 The distribution of training corpus and test corpus of ambiguous word

为了度量本文所提出方法的性能,共进行了两组实验。在第 1 组实验中,选取歧义词左右邻接的两个词单元的词形作为消歧特征,使用贝叶斯消歧模型来判别歧义词的语义。使用 SemEval - 2007: Task#5 的训练语料对贝叶斯消歧模型进行训练。利用优化后的贝叶斯消歧模型对 SemEval - 2007: Task#5 的测试语料进行词义消歧。

在第 2 组实验中,选取歧义词左右邻接的四个词单元的词形、词性和语义类作为消歧特征,利用 RBM 消歧模型来判别歧义词的语义。使用 SemEval - 2007: Task#5 的训练语料,结合词形表、词性表和语义类表来获得消歧特征向量 *Feature*。将消歧特征向量 *Feature* 转换成标准消歧特征向量。同时,将该特征向量所对应的语义类转换为标准语义类。使用标准消歧特征向量和标准语义类对 RBM 消歧模型进行训练。使用优化后的 RBM 对 SemEval - 2007: Task#5 的测试语料进行语义分类。两组实验的消歧准确率如表 1 所示。

为了更清楚地比较两组实验的消歧性能,根据表 1 画出了两组实验的折线图来对比它们的消歧准确率,如图 6 所示。

从图 6 可以看出,RBM 词义消歧分类器的准确率要高于贝叶斯词义消歧分类器。其原因是:在第

1 组实验中,只选取了词形一种消歧特征。在第 2 组实验中,选取了词形、词性和语义类 3 种消歧特征,能够覆盖更多的语言学现象。此外,RBM 的分类性能要强于贝叶斯模型。

表 1 测试语料的消歧准确率

Tab. 1 Disambiguation accuracy of test corpus

	实验 1: 词形 + 贝叶斯消歧模型	实验 2: 词形、词性、语义类 + RBM
表面	50.00%	61.11%
菜	33.33%	55.56%
单位	47.10%	58.82%
儿女	50.00%	50.00%
机组	71.43%	71.43%
开通	70.00%	70.00%
气息	64.30%	71.43%
气象	37.50%	62.50%
使	62.50%	62.50%
推翻	50.00%	60.00%
眼光	71.43%	71.43%
震惊	71.43%	71.43%
中医	37.50%	62.50%
补	40.00%	50.00%
队伍	36.40%	45.45%
赶	27.80%	50.00%
旗帜	55.56%	55.56%
长城	14.30%	38.10%

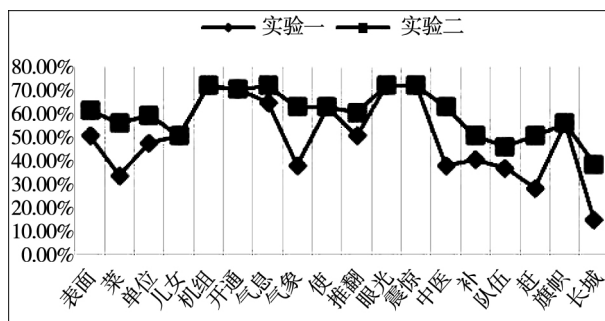


图 6 测试语料消歧准确率的对比

Fig. 6 The comparison of disambiguation accuracy of test corpus

5 结论

本文提出了一种受限玻尔兹曼机的词义消歧方法。以歧义词左右相邻的四个词单元的词形、词性

和语义类为消歧特征,使用 RBM 分类器来判别歧义词的语义类。使用 SemEval – 2007: Task#5 的训练语料结合哈尔滨工业大学的语义标注语料来优化受限玻尔兹曼机的参数,以提高词义消歧精度。使用优化后的 RBM 分类器对 SemEval – 2007: Task#5 的测试语料进行词义消歧。实验结果表明:所提出方法的词义消歧性能要优于贝叶斯分类器。

参考文献:

- [1] 高雪霞, 炎士涛. 基于 WordNet 词义消歧的语义检索研究 [J]. 湘潭大学自然科学学报, 2017, 39(2): 118.
- [2] 杨陟卓. 基于上下文翻译的有监督词义消歧研究 [J]. 计算机科学, 2017, 44(4): 252.
- [3] 陈浩. 基于统计语言模型的无导词义消歧 [J]. 电脑知识与技术, 2015(1): 178.
- [4] 史兆鹏, 邹徐熹, 向润昭. 基于依存句法分析的多特征词义消歧 [J]. 计算机工程, 2017, 43(9): 210.
- [5] 李冬晨, 张献涛, 樊扬, 等. 融合词义消歧的汉语句法分析方法研究 [J]. 北京大学学报(自然科学版), 2015, 51(4): 577.
- [6] 闫蓉, 高光来. 上下文边界可变的词义消歧 [J]. 计算机工程与设计, 2015(10): 2843.
- [7] 钱涛, 姬东鸿, 戴文华. 一个基于超图的词义归纳模型 [J]. 四川大学学报(工程科学版), 2016, 48(1): 152.
- [8] 王少楠, 宗成庆. 一种基于双通道 LDA 模型的汉语词义表示与归纳方法 [J]. 计算机学报, 2016, 39(8): 1652.
- [9] 张仰森, 郑佳. 中文文本语义错误侦测方法研究 [J]. 计算机学报, 2017, 40(4): 911.
- [10] IVAN L A, VICTOR J S S, FRANCO R L, et al. Improving Selection of Synsets from WordNet for Domain-specific Word Sense Disambiguation [J]. Computer Speech & Language, 2017, 41(1): 128.
- [11] KOPPULA N, RANI B P, RAO K S. Graph Based Word Sense Disambiguation [J]. Advances in Intelligent Systems and Computing, 2017, 507(1): 665.
- [12] IACOBACCI I, PILEHVAR M T, NAVIGLI R. Embeddings for Word Sense Disambiguation: An Evaluation Study [C] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 897.
- [13] BENNETT A, BALDWIN T, LAU J H, et al. LexSemTm: A Semantic Dataset Based on All-words Unsupervised Sense Distribution Learning [C] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 1513.
- [14] HENDERSON J, POPA D N. A Vector Space for Distributional Semantics for Entailment [C] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 2052.
- [15] 唐共波, 于东, 荀恩东. 基于知网义原词向量表示的无监督词义消歧方法 [J]. 中文信息学报, 2015, 29(6): 23.
- [16] 张春祥, 邓龙, 高雪瑶, 等. 结合语义知识的汉语词义消歧 [J]. 计算机工程与应用, 2016, 52(3): 119.
- [17] 李国臣, 吕雷, 王瑞波, 等. 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.
- [18] LU N, LI T, REN X, et al. A Deep Learning Scheme for Motor Imagery Classification based on Restricted Boltzmann Machines [J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2017, 25(6): 566.
- [19] 刘明珠, 郑云非, 樊金斐, 等. 基于深度学习法的视频文本区域定位与识别 [J]. 哈尔滨理工大学学报, 2016, 21(6): 61.
- [20] 吕淑宝, 王明月, 翟祥, 等. 一种深度学习的信息文本分类算法 [J]. 哈尔滨理工大学学报, 2017, 22(2): 105.
- [21] MA X, WANG X. Average Contrastive Divergence for Training Restricted Boltzmann Machines [J]. Entropy, 2016, 18(2): 35.

(编辑: 温泽宇)