北京大学学报(自然科学版) 第 52 卷 第 1 期 2016 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 52, No. 1 (Jan. 2016)
doi: 10.13209/j.0479-8023.2016.009

基于《现代汉语语义词典》的未登录词语义预测研究

尚芬芬[1,2] 顾彦慧[1,2,†] 戴茹冰[3] 李斌[3] 周俊生[1,2] 曲维光[1,2]

1. 南京师范大学计算机科学与技术学院，南京 210023; 2. 江苏省信息安全保密技术工程研究中心，南京 210023;
3. 南京师范大学文学院，南京 210097; † 通信作者, E-mail: gu@njnu.edu.cn

摘要 基于《现代汉语语义词典》，首先建立不同语义层次的词典，根据词典分别构建模型并进行语义预测，然后将各个模型进行集成，通过集成模型再对未登录词进行语义预测，得到较好的预测性能。利用预测模型对2000年《人民日报》语料进行未登录词语义预测和标注，最终得到带有未登录词语义义项标注的语料资源。

关键词 汉语未登录词；语义预测；语义标注；集成学习

中图分类号 TP391

Research on the Sense Guessing of Chinese Unknown Words Based on “Semantic Knowledge-base of Modern Chinese”

SHANG Fenfen[1,2], GU Yanhui [1,2,†], DAI Rubing[3], LI Bin[3], ZHOU Junsheng[1,2], QU Weiguang[1,2]

1. School of Computer Science and Technology, Nanjing Normal University, Nanjing 210023; 2. Jiangsu Research Center of Information Security & Privacy Technology, Nanjing 210023; 3. School of Chinese Language and Culture, Nanjing 210097;
† Corresponding author, E-mail: gu@njnu.edu.cn

Abstract Based on the research issue of sense guessing of Chinese unknown words, different levels of semantic dictionary were introduced by applying “Semantic Knowledge-base of Modern Chinese”. Models have constructed for sense guessing by using these dictionary. Each model was intergrated to predict the unknown words and obtained better performance. Based on each model, semantic prediction and annotation of the unknown words in People’s Daily which published in 2000 were evaluated. Finally, corpus resources with the sense annotation of unknown words were obtained.

Key words Chinese unknown words; sense guessing; semantic annotation; ensemble learning

语义问题一直是自然语言处理领域的研究热点。文本内容的理解必须建立在对文本中每一个词语的语义理解基础之上。然而，由于大量未登录词的存在，其语义未知，文本中没有标注未登录词的句法和语义类别标记，因此很难做到获取所有词语的语义，这对很多自然语言处理(natural language processing, NLP)技术和其他以语义为基础的研究是一个挑战。汉语未登录词的语义预测研究可以为未登录词提供语义预测，从而为研究者提供语义参考，对许多NLP 应用，如机器翻译、信息检索、语义分析、词典编纂等有重要意义。

汉语未登录词语义预测的研究难度较大，因此相关研究工作较少，除使用基于知识的模型和基于语料的模型及其混合模型外，很少有新的模型提出。在已有的研究中使用的词典资源也比较有限，使用较多的是《同义词词林》(Cilin)。本文通过构

国家自然科学基金(61272221, 61472191)、国家社会科学基金(11CYY030, 10CYY021)、江苏省社会科学基金(12YYA002)和江苏省高校自然科学基金(14KJB520022)资助

收稿日期: 2015–06–19; 修回日期: 2015–09–03; 网络出版日期: 2015–09–30

建多种语义预测模型，利用《现代汉语语义词典》进行未登录词语义预测，并对 2000 年《人民日报》语料中的未登录词进行语义预测和标注。

1　相关研究

在对汉语未登录词的语义预测研究中，学者们先后提出不同的模型方法，Chen 等[1–3]、Lu[4–5]、Tseng 等[6–7]以及 Qiu 等[8–9]等都为汉语未登录词语义预测的研究做出了贡献。有研究指出，对于一个 8 万词的词典而言，大约有 3.51%的未登录词存在[1]。这些未登录词中包含复合名词 51%，复合动词 34%，专业名词只占 15%[3]。目前对专业名词已有大量的研究来确定其语义类别。与只占 15% 的专业名词相比，占 85% 的复合词语的语义类别预测研究显得更为重要[10–13]。因此，近期的研究更多倾向于未登录词中复合词语的语义猜测，比如 Chen 等[3]和 Lua[14]的研究。

关于汉语未登录词语义预测，现有研究大多采用基于词语结构信息和基于规则的方法，也有利用未登录词上下文信息，通过计算与已知词类词语上下文的相似度来进行预测。依据模型和算法的不同，归纳为以下 3 种方法。

1) 基于知识的方法。大部分学者对未登录词语义预测的研究是基于知识的模型，最早使用该方法的研究者之一是 Lua[14]，目的是把双音节中文词分类到同义词词林中的大类或者中类，使用三层反向传播神经网络，模拟双音节词的语义类别与其两个组成字的语义类别之间的依赖性。此后，又发展出基于实例的方法[3]以及基于相似度的方法[2]；文献[4–5]的研究涉及重叠字模型、字–类别关联模型以及基于规则的模型。此外，还有基于《知网》的模型[15–16]。

2) 基于语料的方法。Lu[4–5]提出的基于语料的模型是根据未登录词出现的上下文预测其语义类别，从语料中抽取出《同义词词林》中每个语义类别的广义上下文，再计算未登录词的上下文与每个候选语义类别的广义上下文之间的相似度，通过相似度的大小来确定未登录词的语义类别。

3) 基于知识和基于语料的混合方法。Lu[4–5]提出基于知识和基于语料的混合模型，使用基于知识的模型为每个未登录词提供候选语义类别，然后从语料中抽取《同义词词林》中每个语义类别的广义上下文，再计算出未登录词的上下文与每个候选语义类别的广义上下文之间的相似度。

早期的研究主要集中在基于知识的模型，随后出现加入上下文信息的模型研究，但效果不是很好，接着使用基于知识的模型与基于上下文信息松散结合的混合模型，效果也不理想。近期的研究将未登录词的知识与上下文信息更紧密地结合成混合模型，取得较好的预测效果。

2　语义资源及词典构建

汉语未登录词语义预测研究使用较多的语义资源是《同义词词林》，少部分研究使用《知网》(HowNet)，几乎没有相关研究使用《现代汉语语义词典》(The Semantic Knowledge-base of Contemporary Chinese, SKCC)[17–18]。《现代汉语语义词典》拥有丰富的语义义项分类，并且各个义项下有充分的成员词语，因此，本文利用该词典进行未登录词语义预测的研究。

2.1　语义资源介绍

本文未登录词语义预测研究使用的语义资源是《现代汉语语义词典》，这是一部面向 NLP 的语义知识库，收录 6.5 万余条汉语实词。作为综合型语言知识库(Comprehensive Language Knowledge Base, CLKB)的一部分，SKCC 广泛应用于计算词汇语义学的基础研究和应用研究中。SKCC 采用 Microsoft Access 数据库实现，其中包含全部词语的总库 1 个，每类词语(实词)各建一库，每个库文件中都包含词语与其语义的关系。由于名词库的分类较为详细，因此本文主要研究名词库的词语。

根据 SKCC 名词库的语义分类，可以分五级对语义词典 SKCC 名词库中不同语义层次的词语数目进行统计，如表 1 所示。

2.2　词典构建

利用 SKCC 进行未登录词语义预测，属于基于

表 1　语义词典 SKCC 名词库中不同语义层次的词语数目
Table 1　Word number of SKCC semantic dictionary under different semantic levels

语义层数	包含词语数目
1	3296
2	8220
3	6421
4	12211
5	9553

词典的方法，是根据词典中词语的信息构造预测模型，需要词典中词语位于词典树型结构的同一语义层次，便于统计每个语义类别中的词语信息。SKCC 的语义体系呈现树型结构，但是语义词典 SKCC 名词库中的词语并不是全都划分到树型结构的最底层，而是划分到不同的语义层次(如图 1 所示)，这样不便于语义分类。因此，先构造出语义类别的树型结构，再将所有词语都归为第一级来构造词典。由于划分到第一级语义类别粒度较大，因此再将词语尽量(当词语无法向下级语义划分时，则将该词去除)归为第二级和第三级，由第二级和第三级词语信息构造词典。本文分别构建 3 个 SKCC 词典。

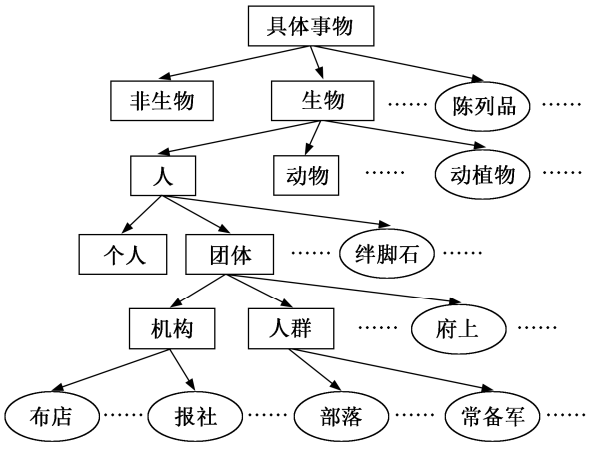
2.2.1 第一级语义类别 SKCC1

将 SKCC 中所有词语都归为第一级语义类别汇总，记为 SKCC1。第一级的各个语义类别所包含的词语数目如表 2 所示。

2.2.2 第二级语义类别 SKCC2

语义词典 SKCC 的词语划分到第二级语义类别中所构造的词典，记为 SKCC2。构造语义词典 SKCC2 时，语义词典 SKCC的词语语义类别向上划分，可以全部划分到父节点(也就是第一级语义类别中)，但是如果划分到第二级语义类别中时，所有归属于第一级语义类别的词语就无法向下划分到第二级的语义类别中。因此，基于 SKCC2 词典的研究只包含属于二级及以下类别的词语，并且将这些词语都向上划分到第二级父节点上的词语。

第二级语义类别分为16个。SKCC 名词库划到第二级的 16 个语义类别的词语数目如表 3 所示。



矩形表示语义类别，椭圆形表示词语

图 1 SKCC 中语义类别和词语结构

Fig. 1 Semantic category and word structure in SKCC

表 2 SKCC1 的各个语义类别词语数目

Table 2 Word number under different semantic levels of SKCC1

类别名	词数
过程	1908
时间	906
抽象事物	8643
空间	3195
具体事物	25149

2.2.3 第三级语义类别 SKCC3

语义词典 SKCC 的词语划分到第三级语义类别中所构造的词典，记为 SKCC3。将词语划归到第三级语义层次时，所有归属于第一级和第二级语义类别的词语由于所属语义节点层次高，难以向下划分到第三级的语义类别中。因此，基于 SKCC3 词典的研究只包含属于三级及以下类别的词语，并且把这些词语都向上划分到第三级父节点上。

第三级语义类别分为 17 个，具体语义类别和词语数目如表 4 所示。其中，语义类别“非生物构件”的词数为 0，原因是词语划分到其祖先类别中，

表 3 SKCC2 的各个语义类别及类别下词语数目

Table 3 Word number under different semantic levels of SKCC2

类别	词数
方位	210
心理特征	691
领域	725
相对时间	669
非生物	15241
构件	1474
动机	39
自然现象	172
生理	669
信息	757
绝对时间	109
属性	2944
法规	318
生物	8238
事件	1657
处所	2492

表 2　SKCC3 的各个语义类别及类别下词语数目
Table 4　Word number under different semantic levels of SKCC3

类别	词数
自然物	2003
颜色	88
外形	307
身体构件	1040
人	5923
情感	102
意识	574
植物	1155
模糊属性	2352
量化属性	448
微生物	76
可听现象	35
非生物构件	0
排泄物	100
可视现象	108
动物	1063
人工物	12811

该类别没有词语和子类别。

3　模型构建

根据词典词语信息，分别构建基于重叠字的模型、基于字–类别关联的模型(Character-Category Association Model)和基于规则的模型。

3.1　基于重叠字的模型

根据现代汉语的构词规则，大多数新词的语义都与其组成词素相关，两者之间有着相同或者相近的语义，不同词语共享相同的组成词素极为常见，因此利用词语组成词素相重叠的知识可以更好地预测新构成词语的语义义项。重叠字模型通过计算未登录词与每个语义类别成员词的重叠字个数来预测未登录词的语义类别。

对于Cilin中的每个语义类别，抽取其成员词的所有不重复的字，并且统计每个字现在词头、词中、词尾的总频数。根据这些信息，提出3对变式。在每一对变式中，变式a通过计算类别和未登录词的重叠字的数目，计算出未登录词的一个类别的得分。相应地，变式b计算上述分数的一个带权值的或归一化的副本。这些变式中，Score(Cat, w)表示分配类别 Cat 为未登录词类别的得分；n 代表未登录词 w 的长度；c_i 代表未登录词 w 的第 i 个字；P_i 表示第 i 个字 c_i 在词 w 中的位置，包括{词头，词中，词尾}；$f(c_i)$表示类别 Cat 中第 i 个字的全部频率；$f(c_i, p_i)$表示在 Cat 中位于 p_i 的 c_i 的频率；N 表示在 Cat 中的字的总数；N_{pi} 表示在类别 Cat 中，位于位置 p_i 的字的总数；N_w 表示在类别 Cat 中词的总数。

变式 1：变式 1a 中，类别的得分是这个类别中未登录词的每个组成字出现次数的总和；变式 1b 中，每个次数都由类别中字的总数加权得到。

$$\text{变式 1a：}\ \text{Score (Cat, } w) = \sum_{i=1}^{n} f(c_i)\ ; \tag{1}$$

$$\text{变式 1b：}\ \text{Score (Cat, } w) = \sum_{i=1}^{n} \frac{f(c_i)}{N}\ 。 \tag{2}$$

变式 2：变式 2a 中，类别的得分是这个类别中未登录词的每个组成字在未登录词的相应位置出现次数的总和；变式 2b 中，每个次数由类别中字在未登录词相应位置出现的总数加权得到。

$$\text{变式 2a：}\ \text{Score (Cat, } w) = \sum_{i=1}^{n} f(c_i, p_i)\ ; \tag{3}$$

$$\text{变式 2b：}\ \text{Score (Cat, } w) = \sum_{i=1}^{n} \frac{f(c_i, p_i)}{N_{p_i}}\ 。 \tag{4}$$

变式 3：变式 3a 中，类别的得分是这个类别中未登录词的尾字 c_n 在未登录词的词尾 p_n 出现的数的总和；变式 3b 中，得分是由类别中所有词总数加权得到。

$$\text{变式 3a：}\ \text{Score (Cat, } w) = f(c_n, p_n)\ ; \tag{5}$$

$$\text{变式 3b：}\ \text{Score (Cat, } w) = \frac{f(c_n, p_n)}{N_w}\ 。 \tag{6}$$

变式1用最直接的方法得到重叠字语义的预测，变式 2 与每个组成字在未登录词和类别的成员词中出现的位置相关，变式 3 只考虑未登录词的最后一个字和每个类别成员词的最后一个字。每一个变式，得分最高的类别被推荐为未登录词的类别。

3.2　基于字–类别关联的模型

字–类别关联模型采用多种复杂的信息理论模型来估算词语组成字与语义类别之间的关联，再估算词语与语义类别之间的关联，为未登录词预测合适的语义。字–类别关联模型计算字与语义类别之间的关联值，使用的统计量包括互信息和 χ^2，如式(7)~(9)所示：

$$\mathrm{Asso}_{\mathrm{MI}}(\mathrm{Char}, \mathrm{Cat}_j) = \log \frac{P(\mathrm{Char}, \mathrm{Cat}_j)}{P(\mathrm{Char})P(\mathrm{Cat}_j)}, \quad (7)$$

$$\mathrm{Asso}_{\chi^2}(\mathrm{Char}, \mathrm{Cat}_j) = \frac{\alpha(\mathrm{Char}, \mathrm{Cat}_j)}{\mathrm{Max}_k \alpha(\mathrm{Char}, \mathrm{Cat}_j)}, \quad (8)$$

$$\alpha(\mathrm{Char}, \mathrm{Cat}_j) = \sqrt{\frac{\left[f(\mathrm{Char}, \mathrm{Cat}_j)\right]^2}{f(\mathrm{Char}) + f(\mathrm{Cat}_j)}}。 \quad (9)$$

其中, Asso(Char, Cat_j)表示字符Char与语义类别 Cat_j 的关联, $P(x)$和 $f(x)$分别表示 x 的概率和频率。

计算出字–类关联后, 词–类关联就可以通过对类别和词的每个组成字的关联加权求和计算出来, 如式(10)所示:

$$\mathrm{Asso}(W, \mathrm{Cat}_j) = \sum_{i=1}^{|W|} \lambda_i \mathrm{Asso}(\mathrm{Char}_i, \mathrm{Cat}_j), \quad (10)$$

其中, Char_i 表示词 W 的第 i 个字符, $|W|$表示词 W 的长度, λ_i 表示 Char_i 与 Cat_j 之间关联的权重, λ_i 的和为 1。

3.3 基于规则的模型

基于规则模型的原理是观察未登录词的组成结构信息, 对之进行归纳总结, 获得可以匹配到更多未登录词词语结构的规则。通过设定的规则模式进行未登录词语义的预测, 实际上是依据未登录词组成字的句法和语义类别来预测未登录词子集的语义类别。基于规则的方法是对不同长度的未登录词分别设计不同的规则集。例如: 对于三字长的未登录词 ABC, 如果 BC 与“学家”相同, 猜测 ABC 为 SKCC1 的类别“具体事物”, 如表 5 所示。

表 5 三字词 ABC 规则 A+“学家”举例

Table 5 Examples of 3-gram words ABC under A+ “expert”

词语	规则	语义类别
文学家	A+BC: “文”+“学家”	具体事物
神学家	A+BC: “神”+“学家”	具体事物
农学家	A+BC: “农”+“学家”	具体事物
史学家	A+BC: “史”+“学家”	具体事物
医学家	A+BC: “医”+“学家”	具体事物

4 模型实验

4.1 实验语料与预处理

实验中使用 1998 年 1 月的《人民日报》语料, 该语料主要用于抽选测试词。测试词抽取条件是: 分别从构造的语义词典中随机抽取; 存在于 1998 年 1 月的《人民日报》语料中; 词语长度为 2~4 个字; 词语词性为名词。对 1998 年 1 月的《人民日报》语料做如下的预处理: 1) 处理为包含词语、词性标记和词频信息的格式; 2) 过滤掉停用词和命名实体; 3) 抽取出词性标记为 n 的词语。

4.2 实验与分析

从 SKCC1 中随机抽取 3000 个测试词, 这些是已知语义类别的词语, 再从 SKCC1 中去除这 3000 个词语。然后, 基于去除测试词的 SKCC1, 利用构建的模型进行语义预测, 并对比其正确的语义类别, 计算语义分类的正确率。

实验 1 基于重叠字模型的 6 个变式的未登录词语义预测正确数和正确率。抽取未登录词的总数为 3000, 实验结果如表 6 所示。结果显示, 这些模型的正确率都较高, 其中最高值是变式 2a 得到的 77.0%。

实验 2 基于字–类别关联模型不同统计量的未登录词语义预测正确数和正确率。抽取未登录词的总数为 3000, 实验结果如表 7 所示。结果显示, 统计量 MI 与 χ^2 相比, χ^2 得到更高的语义预测正确率, 为 74.3%。

实验 3 基于规则模型的未登录词语义预测正确数和正确率。所抽取的未登录词总数为 861, 即

表 6 基于 SKCC1 词典的重叠字模型预测结果

Table 6 Prediction results of overlapping words based on SKCC1 dictionary

变式名	正确数	正确率/%
1a	2179	72.6
1b	2134	71.1
2a	2309	77.0
2b	1694	56.5
3a	2293	76.4
3b	1963	65.4

表 7 基于 SKCC1 词典的字–类别关联模型预测结果

Table 7 Prediction results based on SKCC1 dictionary under word-type models

统计量	正确数	正确率/%
MI	1702	56.7
χ^2	2230	74.3

表 8　基于 SKCC1 词典的规则模型的未登录词语义预测结果

Table 8　Results of unknown words based on SKCC1 dictionary

有语义返回词数	正确数	正确率/%	召回率/%
861	770	89.4	28.7

在 3000 个测试词语中，模型预测出语义的词语共有 861 个，实验结果如表 8 所示。可以看出，正确率很高，但是召回率很低。

实验 4　多模型的集成。由于基于规则的模型得到的预测正确率较高，但召回率较低，因此本实验设计基于规则的模型与其他模型的集成。集成模型的预测语义由以下两条确定。

1) 如果能够由基于规则的模型预测出语义，则将这个语义作为混合模型预测语义。

2) 如果基于规则的模型不能够给出预测语义，那么对基于重叠字模型和基于字–类别关联模型的语义预测进行投票，并对投票结果进行排序，取票数最高的语义类别作为该未登录词的混合模型预测语义。

集成模型对所抽取出的 3000 个未登录词进行语义预测，得到的正确数和正确率如表 9 所示。可见集成模型得到较高的正确率(77.9%)，同时也克服了基于规则模型召回率较低的问题，获得较好的未登录词语义预测性能。

表 9　基于 SKCC1 词典的集成模型语义预测结果

Table 9　Results of intergration models based on SKCC1

有语义返回词数	正确数	正确率/%
3000	2337	77.9

5　汉语未登录词语义预测应用

在基于不同词典和不同模型对 2000 年《人民日报》语料的未登录词标注中，基于规则的模型得到的预测正确率较高，但是其覆盖率较低。比如词语“股东会”，在基于规则模型中，基于 SKCC3 得到预测语义“人”；该词在字–类别关联模型的预测结果为 SKCC3 “人，人”；在重叠字模型中，该词语的预测结果为 SKCC3 “人”。预测语义都与对应人工标注相同。集成模型结合了基于规则的模型与其他模型，得到较高的正确率，可见集成模型对基于 SKCC3 的语义预测性能较好。本文根据基于 SKCC3 的集成模型所获得的未登录词预测语义标注到 2000 年《人民日报》语料中，所得到的语料示例如表 10 所示。

表 10 的语料示例中共有 3 个未登录名词，分别是“主景”、“凹版”、“凹凸感”。其中，词语“凹凸感”语义预测有误，正确语义应该为“意识”而不是“情感”，其余两个词语语义预测正确。

在基于 SKCC 对 2000 年《人民日报》语料的研究中，未登录词的语义可以划分到 SKCC 第二级和第三级。但是，由于尚无对 2000 年《人民日报》语料未登录词语义标注的标准语料，无法确定未登录词预测语义的正确性。针对这个问题，本文取基于 SKCC 已标注的第二级语义和第三级语义进行分析。

假设未登录词 W 预测出的第三级语义为 GuessThirdCat(W)，GuessThirdCat(W)在 SKCC 树型语义结构的上一级语义为 SecondCat(W)，预测出的在 SKCC 中第二级的语义为 GuessSecondCat(W)，如果 GuessSecondCat(W) = SecondCat(W)，那么认为该未登录词 W 所预测的二级语义为正确的。使用该评估方法可以判定 2000 年《人民日报》语料中 12162 个未登录词的预测语义正确，正确率为 72.2%。

6　总结与展望

本文首次使用《现代汉语语义词典》进行汉语未登录词语义预测的研究，通过构建的模型对 2000 年《人民日报》语料的未登录词进行语义预测和标注，得到具有未登录词语义标注的语料。在未来的

表 10　基于 SKCC3 标注未登录词语义语料示例

Table 10　Examples of unknown words based on SKCC3

原始语料	标注语料
20000130-03-009-007/m 正面/b 主景/n 毛/nrf 泽东/nrg 头像/n , /wd 采用/v 手工/d 雕刻/v 凹版/n 印刷/vn 工艺/n , /wd 形象/n 逼真/a 、/wu 传神/a 、/wu 凹凸感/n 强/a ; /wf	20000130-03-009-007/m 正面/b 主景/n/人工物 毛/nrf 泽东/nrg 头像/n, /wd 采用/v 手工/d 雕刻/v 凹版/n/外形 印刷/vn 工艺/n , /wd 形象/n 逼真/a 、/wu 传神/a 、/wu 凹凸感/n/情感 强/a ; /wf

工作中，我们将探索改进语义预测方法，并尝试将未登录词语义预测拓展到实际应用中。

参考文献

[1] Chen H, Lin C. Sense-tagging Chinese corpus // Proceedings of ACL-2000 Workshop on Chinese Language. Hong Kong, 2000: 7−14

[2] Chen C. Character-sense association and compounding template similarity: automatic semantic classification of Chinese compounds // Proceedings of the 3rd SIGHAN Workshop on Chinese Language Processing. Barcelona, 2004: 33−40

[3] Chen K, Chen C. Automatic semantic classification for Chinese unknown compound nouns // Proceedings of the 18th International Conference on Computational Linguistics (COLING). Saarbrücken, 2000: 173−179

[4] Lu Xiaofei. Hybrid model for Chinese unknown word resolution [D]. Ohio: The Ohio State University, 2006

[5] Lu Xiaofei. Hybrid model for semantic classification of Chinese unknown words // Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Rochester, 2007: 188−195

[6] Tseng H. Semantic classification of Chinese unknown words // Proceedings of the Student Research Workshop at the 41st Annual Meeting of the Association for Computational Linguistics (ACL). Sapporo, 2003: 72−79

[7] Tseng H, Chen K J. Design of Chinese morphological analyzer // Proceedings of the First SIGHAN Workshop on Chinese Language Processing. Stroudsburg, 2002: 1−7

[8] Qiu Likun, Wu Yunfang, Shao Yanqiu. Combining contextual and structural information for supersense tagging of Chinese unknown words // Proceedings of CICLing, Part Ⅰ, LNCS 6608. Tokyo, 2011: 15−28

[9] Qiu Likun, Zhao Kai, Hu Changjian. A hybrid model for sense guessing of Chinese unknown words // Proceedings of 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC). Hong Kong, 2009: 464−473

[10] Cucerzan S. Large-scale named entity disambiguation based on wikipedia data // Procedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, 2007: 708−716

[11] 周俊生，戴新宇，尹存燕，等. 基于层叠条件随机场模型的中文机构名自动识别. 电子学报，2006, 34(5): 804−809

[12] 陈钰枫，宗成庆，苏克毅. 汉英双语命名实体识别与对齐的交互式方法. 计算机学报，2011, 34(9): 1688−1696

[13] 冯元勇，孙乐，张大鲲，等. 基于小规模尾字特征的中文命名实体识别研究. 电子学报，2008, 36(9): 1833−1837

[14] Lua K T. Prediction of meaning of bi-syllabic Chinese compound words using back propagation neural network. Computational Processing of Oriental Languages, 1997, 11(2): 133−144

[15] 张瑞霞，肖汉. 基于《知网》的词图构造. 华北水利水电学院学报, 2008, 29(3): 53−56

[16] 张瑞霞，杨国增，闫新庆. 基于《知网》的汉语普通未登录词语义分析模型. 计算机应用与软件，2012, 29(8): 126−130

[17] 王惠，詹卫东，俞士汶. 现代汉语语义词典规格说明书. 汉语语言与计算学报, 2003, 13(2): 159−176

[18] Bai M H, Hsieh Y M, Chen K J, et al. Translating Chinese unknown words by automatically acquired templates // Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCNLP). Nagoya, 2013: 839−843