# 基于知识图谱的未登录词语义研究

朱 峰 顾 敏 郑 好 顾彦慧 周俊生 曲维光
(南京师范大学计算机科学与技术学院 南京 210023)

**摘 要** 传统的应用于未登录词语义研究的语料库包含许多限制,例如更新慢、语言相关等。为了解决此问题,提出了基于知识图谱的中文未登录词语义研究方法。知识图谱是一种包含实体、概念及语义关系的语义网络。它具有丰富的实体,并且实体及其关系的添加极为方便,使得弥补传统语料库更新慢的缺憾成为可能。在充分熟悉知识图谱的结构、数据获取方法及相关数据处理方法后,进行基于知识图谱的未登录词语义研究的探索工作,最后以百度百科(目前最大的中文知识图谱)为语料资源,在同一语义分析模型下分别进行基于知识图谱与传统语料的实验,对实验结果进行分析并提出改进方法。

**关键词** 汉语未登录词语义预测,语义标注,知识图谱
**中图法分类号** TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.01.018

## Research on Sense Guessing of Chinese Unknown Words Based on Knowledge Graph

ZHU Feng GU Min ZHENG Hao GU Yan-hui ZHOU Jun-sheng QU Wei-guang
(School of Computer Science and Technology, Nanjing Normal University, Nanjing 210023, China)

**Abstract** Semantic study based on traditional corpus has lots of limits, such as updating infrequently and being language-related. To tackle such issues, sense guessing of Chinese unknown words based on knowledge graph(KG) was proposed in this paper. KG is a semantic network containing entities, concepts and semantic relations. It has a huge number of entities and relations and it is very convenient to add them into the KG, which makes it possible to fix the infrequent updating problem. After the introduction of the structure of knowledge graph, how to get data and ways to process them, some exploration about KG-based sense guessing of Chinese unknown words were excuted. At last, BaiduBaike, which has the most abundant chinese-related data, is used as the corpus with traditional ones to do experiments that are particularly designed to use one specific sense guessing model. This paper also compared the results of experiments based on different knowledge bases and proposed some improvement work.

**Keywords** Sense guessing of Chinese unknown words, Semantic annotation, Knowledge graph

## 1 引言

随着互联网的普及与发展,越来越多的新词出现在网络中,其中许多词没有在语料库及词典中出现过,这部分词被称为未登录词。未登录词包括人、地名等。未登录词在自然语言处理的分词中占有重要地位,如果未登录词未能被正确识别,便不能进行有效的句法分析,因此未登录词的研究十分重要。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进行理解的基础之上。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,导致很难获取所有词语的语义,这就对许多自然语言处理(Natural Language Processing, NLP)技术和其他以语义为基础的研究提出了挑战。汉语未登录词的语义预测可以为未登录词预测语义,从而为研究者提供语义参考。这对一些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[1,2]。

经过前人的努力,未登录词的语义预测已经初成体系。目前的汉语未登录词语义分析方法包括基于知识的方法、基于语料的方法以及基于知识和语料的混合方法3大类。其中每一大类亦包含若干小类,如基于知识的方法大类下包含基于重叠字模型及字-类别关联模型等子算法。本次研究中着重使用上述两种子方法中的重叠子模型作为衡量语料优劣的主要算法。此外,用于未登录词研究的语料库数量有限,较为成熟的有《同义词词林(扩展版)》与《现代汉语语义词典》。《同义词词林(扩展版)》与《现代汉语语义词典》都是由人工标注的知识库,而它们有更新速度慢、词语数目有限等缺点。

为了改善未登录词的语义分析的效果,提出使用一种新的语料库即知识图谱来提高性能。知识图谱能弥补传统语料

到稿日期:2015-08-21 返修日期:2015-10-23 本文受国家自然科学基金(61272221,61073119),江苏省社科基金(12YYA002),江苏省高校自然科学基金项目(14KJB520022),山东省语言资源开发与应用重点实验室开放课题资助。

朱 峰(1993-),男,硕士生,主要研究方向为自然语言处理,E-mail:zfeng_beyond@163.com;顾 敏(1993-),女,硕士生,主要研究方向为自然语言处理;郑 好(1991-),男,硕士生,主要研究方向为自然语言处理;顾彦慧(1978-),男,副教授,主要研究方向为信息检索、人工智能、自然语言处理;周俊生(1972-),男,教授,主要研究方向为自然语言处理;曲维光(1964-),男,教授,主要研究方向为自然语言处理。

库的典型缺点，例如更新速度慢。知识图谱基于互联网，以百科词条、豆瓣等垂直类网站作为数据源，并且每隔固定时间对数据源进行信息爬取，随后将爬取的信息处理成为知识集合，这能保证知识图谱及时获取到最新的知识。以百度百科为例，它每天会从互联网上爬取大量信息，且经过人工标注后再添加至百度百科，此外它也允许大众向百科中添加信息，因此百度百科知识图谱包含的信息内容每天都会增长和更新。此外，传统语料库也存在语言相关的缺点，即针对某种语言的语料库，例如《同义词词林》，它只能应用于中文语义分析。不同类型的语料库拥有各自不同的特性，因此针对不同类型语料库设计的算法很难用于其他类型的语料库，知识图谱能很好地解决此问题。自从 2012 年 Google 率先发布知识图谱后，学界越来越关注知识图谱的研究，包括不同语言的知识图谱的研究。在可预测的将来，知识图谱将会是一门很成熟的体系，各国学者将会研究出各个语种的知识图谱（例如中文的百度，英文的维基）。因此若能开发出基于知识图谱的通用分析算法，基于知识图谱的语义分析将有希望与语言无关。本文将借助重叠字模型进行基于知识图谱语料的相关实验，并且与基于《同义词词林》的结果相比较，以证明知识图谱应用于语义分析的可能性。

## 2 相关工作

关于汉语未登录词语义预测研究方法，现有研究大多采用基于词语结构信息和基于规则的方法，也有利用上下文并通过计算与已知词类的词语的相似度来进行预测的方法。依据研究者提出的多种模型和算法，归纳出以下方法。

### 2.1 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，如 Lua[3,4]，其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 3 层反向传播神经网络模拟双音节词的语义类别及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，如文献[5]；基于相似度的方法，如文献[6]；文献[7,8]涉及到重叠字模型、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，如文献[9]。

### 2.2 基于语料的方法

Lu[7,8]提出的基于语料的模型是根据未登录词出现的上下文来预测语义类别，首先从语料中抽取出《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和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。

### 2.3 基于知识和语料的混合方法

Lu[10,11]提出了基于知识和基于语料的混合模型。该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，然后从语料中抽取出《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和每个候选语义类别的广义上下文之间的相似度。

纵观前人研究成果，不难发现早前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，但是效果不是很好。接着使用基于知识的模型和基于上下文信息松散结合的混合模型，效果也不理想。本文使用基于知识的模型中基于重叠字的模型方法来测试知识图谱是否能进行语义分析，并且取得了不低于使用《同义词词林》的语义预测效果，证明了知识图谱作为语料库的可行性。

## 3 基于重叠字的语义分析模型

基于重叠字的模型是通过计算未登录词与每个语义类别的成员词的重叠字个数来预测未登录词的语义类别的模型。对于语义词典中的每个语义类别，统计其中的词和组成字的信息。提取其中所有的不重复字，并且统计每个字出现在各个位置（即词头、词中、词尾）的总字数。

根据以上信息，可提出 3 对变式。其中，在每一对变式的第一个变式中，通过计算未登录词与类别的重叠字的数目，计算出未登录词的一个类别的得分。它相应的第二个变式计算第一个分数的带权或归一化副本。第一对变量使用最直接的统计方法得到重叠字的基本信息。第二对变式根据每个组成字在未登录词和类别的成员词中出现的位置信息来构建。第三对变式只考虑根据每个类别的最后一个字和未登录词的最后一个字来构建。对每一个变量计算得分，将得分最高的类别推荐为未登录词的类别。重叠字模型使用以上统计信息计算未登录词和语义类别的关联强度。

基于重叠字的模型提出了 3 对变式[6]，每一对变式中的 $a$ 变式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。它相应的 $b$ 变式计算上述分数的一个带权的或者归一化的副本。在这些变式中，$S_{core}(Cat, w)$，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，$P_i$ 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 $Cat$ 中位于位置 $P_i$ 的 $C_i$ 的频率，$N$ 表示在 $Cat$ 中的字的总数，$N_{P_i}$ 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　在 $a$ 变式中，类别的得分是指在这个类别中未登录词的每个组成字出现次数的总和。在 $b$ 变式中，每个次数都由在类别中字的总数加权。

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

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

变式 2　在 $a$ 变式中，类别的得分是指在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的次数的总和。在 $b$ 变式中，每个次数由在类别中字在未登录词相应位置出现的总数加权。

$$\text{Variant 2a: } Score(Cat, w) = \sum_{i=1}^{n} f(C_i, P_i) \tag{3}$$

$$\text{Variant 2b: } Score(Cat, w) = \sum_{i=1}^{n} \frac{f(C_i, P_i)}{N_{P_i}} \tag{4}$$

变式 3　在 $a$ 变式中，类别的得分是在这个类别中未登录词的尾字 $C_n$ 在未登录词的词尾 $P_n$ 出现的次数的总和。在 $b$ 变式中，得分由类别中所有词总数加权。

$$\text{Variant 3a: } Score(Cat, w) = f(C_n, P_n) \tag{5}$$

$$\text{Variant 3b: } Score(Cat, w) = \frac{f(C_n, P_n)}{N_w} \tag{6}$$

第一对变式使用了最直接的方法得到重叠字前提。第二对变式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。第三对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。

## 4　基于《同义词词林》的未登录词语义预测

### 4.1　《同义词词林(扩展版)》简介

《同义词词林》的语义体系为树形结构,如图 1 所示,其包含 12 个大类、94 个中类、1428 个小类,在此每个小类为一个段落,在此基础上继续划分至第五级,将段落中的每一行进行分类,并约定划分规则:大类对应第一级,中类对应第二级,小类对应第三级,新分的第四级与第五级分别对应词群与原子词群。特别地,对于第五级的分类,有的行是同义词,有的行是相关词,有的行只有一个词,具体情况如表 1 所列。

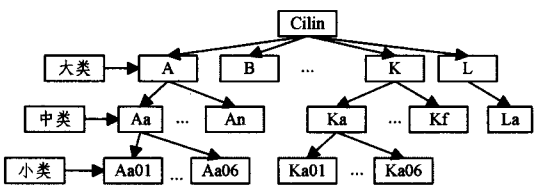



图 1　词林中语义类别和词语结构

表 1　本文划分规则

| 符号性质 | 大类 | 中类 | 小类 | 词群 | 原子词群 |
|---|---|---|---|---|---|
| 级别 | 第一级 | 第二级 | 第三级 | 第四级 | 第五级 |

### 4.2　基于《同义词词林》的未登录词语义预测

实验流程:得到《同义词词林(扩展版)》的统计信息后,从测试集/开发集读取词语及其在词林中的语义类别。再由公式计算每个语义类别的得分,将得分排序,取得分最高/前五的语义类别作为该词语的预测语义类别,把预测的语义类别与词语本身在词林中所属的类别作对比。实验流程图如图 2 所示。

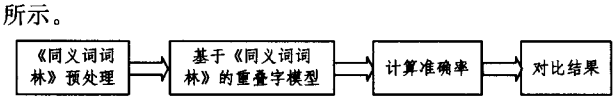



图 2　基于《同义词词林》语义的预测流程

## 5　基于知识图谱的未登录词语义预测

### 5.1　知识图谱简介

知识图谱(Knowledge Graph)是使用图来显示知识发展进程与结构关系的一种现代科学理论,它使用可视化技术描述知识资源及其载体,并且在此基础上挖掘、分析、构建、绘制和显示知识及它们之间的相互联系[12-15]。

知识图谱数据大都来自百科类站点和各种垂直类站点中的结构化数据,这些数据能够覆盖大部分常识性知识。虽然这些数据大都质量较高,但是更新速度极慢。除此之外,知识图谱可通过从各种半结构化数据中获取相关实体的属性-值来扩充对属性的描述,经典的半结构化数据包括 html 表格。另一方面,搜索日志(query log)也是发现新实体或新实体属性并不断扩展知识图谱覆盖率的好途径。相比高质量的常识性数据,通过数据挖掘抽取到的知识数据更大,更能反映当前用户的查询需求并能及时发现最新的实体及事实,但是这并不能保证知识的质量,它存在一定的错误。这些知识会暂时保留,在后续的挖掘中,通过相关聚合算法去除冗余信息,评估数据置信度并通过人工审核加入到知识图谱中。

### 5.2　基于知识图谱的重叠字模型

#### 5.2.1　实验所需数据处理

基于知识图谱的重叠字模型与基于传统语料库的重叠字模型使用方法相同。知识图谱是由大量事实组成的知识库,每条事实均是形如实体-属性-值和实体-关系-实体的键值对。此次研究需从知识图谱中提取如下信息:

1)在知识图谱每个类别中,词语各个字出现的全部频率;

2)在每个语义类别中,分别位于位置词首、词中、词尾的各个字的频率;

3)每个语义类别中,所有不重复的字的总数;

4)每个语义类别中,分别位于位置词首、词中、词尾的所有字的频率;

5)每个语义类别中,所有词语的总数。

#### 5.2.2　实验流程

基于知识图谱的语义预测实验流程:得到知识图谱的统计信息后,从测试集/开发集读取词语及其在词林中的语义类别;再由公式计算每个语义类别的得分,并将得分进行排序,取得分最高/前五的语义类别作为该词语的预测语义类别,把预测的语义类别与词语本身在词林中所属的类别作对比。实验流程图如 3 所示。

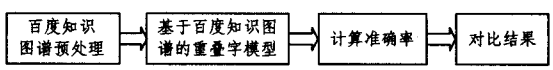



图 3　基于百度知识图谱的预测流程

#### 5.2.3　实验评价指标

本实验评价标准为语义预测的正确率。正确率=正确个数/返回结果个数。本次实验中,预先对照知识图谱标注未登录词语义,而后使用基于知识图谱的未登录词语义预测模型进行预测,并将结果与预先标注的语义进行比对,相同则正确,不同则不正确。

本实验所使用的语料资源是 1998 年 1 月《人民日报》语料。该语料是由北京大学计算语言学研究所(以下简称北大计算语言所)从 90 年代开始历时十几年针对现代汉语语料库的加工任务的研究。1999 年 4 月至 2002 年 4 月历时 3 年,依据 1999 年 3 月制订、2001 年 7 月修订的《现代汉语语料库加工规范—词语切分与词性标注》,完成了 1998 年《人民日报》整年语料的标注语料库,该语料库包含了 2600 多万汉字。将全部语料进行词语切分和词性标注等基本加工,其中 1 月份的语料近 200 万字,切分出来的汉语字或词数为 936065。

#### 5.2.4　实验结果

分别取基于《同义词词林(扩展版)》与知识图谱模型结果中的一组典型数据进行比较,结果如图 4 所示。

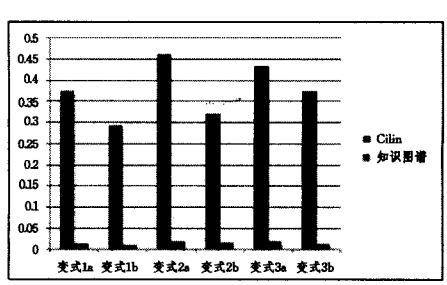



图 4　Cilin 与知识图谱语义预测正确率比较

将基于《同义词词林(扩展版)》所得结果与基于知识图谱语义预测所得结果进行比较可得,基于知识图谱的语义的预测正确率比《同义词词林(扩展版)》语义预测正确率低很多,并且知识图谱上 6 个变式所得正确率均很差。

实验结果分析:本次实验以百度百科作为知识图谱,而百度百科因为需要在短时间内迅速丰富自身实体,所以百度公司将百科编辑权开放给普通民众,这就造成百科中包含众多同义类别,例如"职业"与"工作"的意义相同,如果不对这对类别加以处理,判定算法预测结果"职业"与"工作"意义不同,将有可能出现计算机期望"职业",却因结果为"工作"而判定错误的情况,这也就导致了上述实验中使用知识图谱的预测正确率远低于同义词词林的预测准确率。因此需要将"职业"、"工作"等意义相同的类别进行合并,即聚类。

### 5.3 基于类别聚类后的知识图谱的重叠字模型

#### 5.3.1 基于 Agglomerative 算法对知识图谱的类别聚类

由以上实验分析可得,由于知识图谱中包含许多同义个体,使得最终的实验结果比基于《同义词词林》的方法的准确率低。因此,需要对意义相似的个体进行组合,使用聚类算法合并。

Hierarchical clustering 算法[16]是一种建立集群中个体间等级的聚类算法,它包含自顶向下与自底向上的两种模式,分别对应实现将一个单一集群切分为单一个体和将众多零散个体聚类至一个单独集群。本次研究中使用自底向上模式,即 Agglomerative 算法。首先找出所有个体中最相似的两个个体,接着将两个个体合并为一个新的个体,并计算新个体与其他所有个体间的相似度,即更新群体相似度;随后对新的个体群进行递归聚类,直到只剩下一个个体。聚类可通过零散值,即对算法执行过程中的个体数目进行控制,进而在达到合适的零散度时终止。个体数目越多,集合越零散;个体数目越少,集合越集中。图 5 为 Agglomerative 算法对某些特定个体聚类的过程的图形化。先找出最相似的两个个体"冰酒"和"威士忌",并合并为新个体。而对于"茅台"、"人头马"以及"红方",下一步需要确认这 3 个词中哪个与新类别最相似。由计算所得,"红方"与上一步所得新类别最相似,因此,在第二步将"红方"与上一步所得新类别进行合并,得到"冰酒"、"威士忌"以及"红方"的新合并类。经过几个步骤后,可以分别计算"人头马"以及"茅台"与新合并类的关系。同样地,对于"总经理"、"辅导员"、"教授"、"教师"和"老师",可以使用同样的方式进行聚类。

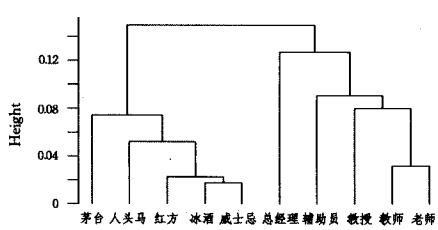



图 5 Agglomerative 聚类展示

为了能够使用 Agglomerative 算法,需要获得每个个体与其他个体的初始相似度的大小,随后进行迭代合并更新群体相似度。鉴于知识图谱的本质是图,本次使用 Random Walk With Restart(以下简称 RWR)算法[17,18]获取个体间的初始相似度。在获得初始相似度后,每步迭代需进行个体合并后的相似度更新。新合并个体与其他个体间的相似度定义为合并所需的两个个体与其他个体相似度和的平均值,计算可得 Agglomerative 算法的复杂度为 $O(n^2)$。

#### 5.3.2 使用 RWR 计算类别间初始相似度

RWR 算法的计算公式如下:

$$\vec{r_l}=c\widetilde{W}\vec{r_l}+(1-c)\vec{e_l} \tag{7}$$

其中,$W$ 为实体和概念间的关系矩阵,$c$ 为系数,本次实验中取值为 0.9,向量$\vec{e_l}$为初始点矩阵,通过设第 $m$ 个值为 1 来设置起始点位置。向量$\vec{r_l}$为相似度矩阵,初始值设置为起始点矩阵。一轮计算过后将结果设置为下一轮$\vec{r_l}$的值。经过多次循环,直到相似度矩阵趋于平稳时循环结束,得到相似度向量,向量中的第 $i$ 个值即为第 $m$ 个点与第 $i$ 个点之间的相似度大小。

由实践可知,每轮循环 6~8 次便可达到平稳。单次循环中计算集中于关系矩阵 $W$ 与相似度向量$\vec{r_l}$的乘积,且由于 $W$ 为稀疏矩阵,可得单次时间复杂度为 $O(n)$。为得到个体间所有相似度,需进行 $n$ 轮,故最终 RWR 执行的时间复杂度为 $O(n^2)$。

#### 5.3.3 实验结果

前文给出了语义预测所需要的相关策略说明,包括用于语义预测的重叠字模型介绍、新的语料库、知识图谱介绍以及用于后半部分实验的 RWR 算法和 Agg 算法的介绍。本节将进行语义预测实验方案设计、正确率比较、结果分析及方案优化。

本实验使用的知识图谱原始类别数目为 4907,分别将类别数目聚类至 50,100,300,500,并进行实验,分别记为 C50,C100,C300,C500。实验结果如表 2 及图 6 所示。

表 2 基于知识图谱重叠字模型的实验结果

(a) 第一对变式

| 测试集 | 变式 1a | 变式 1b |
|---|---|---|
| C50 | 0.039 | 0.014 |
| C100 | 0.029 | 0.018 |
| C300 | 0.085 | 0.067 |
| C500 | 0.094 | 0.072 |

(b) 第二对变式

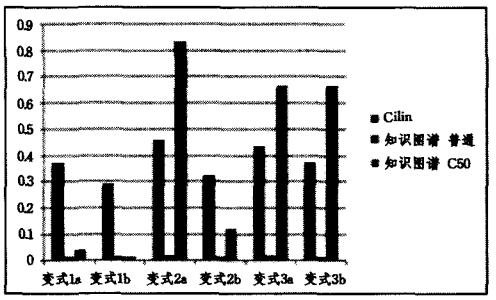
| 测试集 | 变式 2a | 变式 2b |
|---|---|---|
| C50 | 0.833 | 0.122 |
| C100 | 0.818 | 0.104 |
| C300 | 0.462 | 0.097 |
| C500 | 0.473 | 0.099 |

(c) 第三对变式

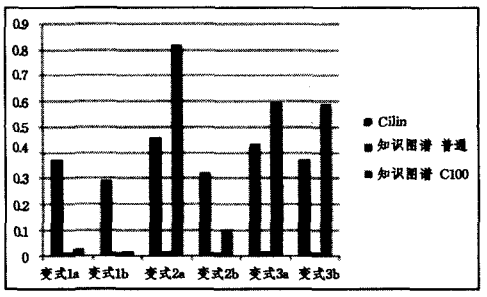
| 测试集 | 变式 3a | 变式 3b |
|---|---|---|
| C50 | 0.669 | 0.669 |
| C100 | 0.596 | 0.593 |
| C300 | 0.174 | 0.174 |
| C500 | 0.267 | 0.264 |

表 2 列出了 6 种变式在 C50,C100,C300,C500 下的结果,从中可以看出,测试集 C50 在变式 2a 中取得了最好的效果,准确率达到了 0.833,这是由于本文实验中使用的数据在类别较少时,变式 2a 表现了较好的性能。而同样地,C50 在其他变式中表现效果一般,在变式 1a 以及变式 1b 中只取得了 0.0386 以及 0.0137 的准确率。

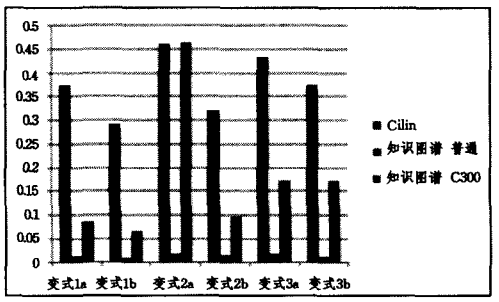
为了更好地展现实验结果，使用图 6 来表示实验结果。



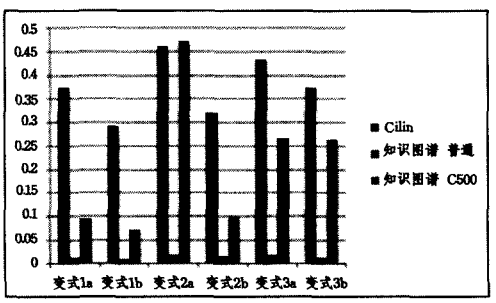

(a)C50 的对比结果




(b)C100 的对比结果




(c)C300 的对比结果




(d)C500 的对比结果

图 6　不同类大小下的实验结果

5.3.4　实验结果分析

实验结果并未达到最优，这是由于知识图谱中含有部分类似“楼主”、“东东”、“恐龙”等网络流行词汇，表示与传统意义相异的含义，或者并不存在于传统词汇中，导致正确率下降。

变式 1 的思想是词的所有组成字对词的含义具有同等影响，比如“教授”便是一个符合变式 1 思想的例子，所有组成字与词意义相近。而诸如“水笔”等词，仅单一或部分组成字决定了词语含义，传统语料库由专家编写，此类词语数量远少于自由开放的百度百科中的此类词语数量，极大影响了变式 1 结果的准确率。

从上述 4 组实验对比结果可以看出，本文所提方法在变式 2a、变式 2b、变式 3a 以及变式 3b 中都取得了较好的结果。另外，由于所提聚类方法将意义相近的个体进行组合，使得结果更加精确，具体表现在变式 2a 中。而聚类的方法在变式 3a 以及变式 3b 中表现出了较好的效果，即聚类可以使得相近的个体组合在一起，从而减少了同义个体的干扰，使得结果更加精确。

**结束语**　本文重点探索了基于知识图谱的语义预测模型。在介绍了语义预测的基本方法与基本语料之后，提出了新的语料库，即知识图谱，并且重新设计了基于知识图谱的语义预测实验，将其与基于传统语料库的语义预测结果进行比较。在此基础上进一步分析了知识图谱的特点以及基于知识图谱的语义预测模型的不足，同时提出了改进方案。实验表明，改进后的知识图谱能作为语义预测合适的语料库，有进一步研究的价值。此外，各国学者均热衷于知识图谱的研究，目前已有百度、Google 等投入使用的知识图谱，在可预见的未来，各语种将会有属于自己的成熟的知识图谱，因此基于知识图谱的通用的未登录词语义预测算法或理念将有希望突破语义分析中语言相关的障碍。

汉语未登录词语义预测已成为不容忽视的重要研究问题之一，迫切需要研究学者们的进一步关注。在未来的工作中，主要从以下几个方面继续研究：探索改进语义预测方法；将未登录词语义预测应用于实际应用中；寻找更加准确的方法使得预测结果更加精确；研究知识图谱的结构，寻找规避与处理“风险词”的方法。

## 参 考 文 献

[1] SUN Mao-song, ZOU Mao-song. Several problems in Automatic Chinese Word Segmentation[J]. Applied Linguistics, 1995, 16(4): 40-46. (in Chinese)
孙茂松，邹嘉彦. 汉语自动分词研究中的若干理论问题[J]. 语言文字应用，1995，16(4)：40-46.

[2] CHEN Xiao-he. A package scheme for identifying unlisted words in Chinese segmentation[J]. Applied Linguistics, 1993, 13(3): 103-109. (in Chinese)
陈小荷. 自动分词中未登录词问题的一揽子解决方案[J]. 语言文字应用，1999，13(3)：103-109.

[3] LUA K T. Prediction of Meaning of Bi-syllabic Chinese Compound Words Using Back Propagation Neural Network[J]. Computational Processing of Oriental Languages, 1997, 11(2): 133-144.

[4] SHANG Feng-feng, GU Yan-hui, DAI Ru-bing, et al. Research on the Sense Guessing of Chinese Unknown Words Based on Semantic Knowledge-base of Modern Chinese [J]. Acta Scientiarum naturalium Universitatis Pekinensis, 2016, 52(1): 10-16. (in Chinese)
尚芬芬，顾彦慧，戴茹冰，等. 基于《现代汉语语义词典》的未登录词语义预测研究[J]. 北京大学学报：自然科学版，2016，52(1)：10-16.

[5] CHEN K, CHEN C. Automatic Semantic Classification for Chinese Unknown Compound Nouns[C]// Proceedings of the 18th International Conference on Computational Linguistics (COLING), 2000. USA, 2000: 173-179.

(下转第 127 页)

USA：ACM，2013：13-18.

[6] GREENBERG A，HJALMIYSSON G，MALTZ D A，et al. A clean slate 4D approach to network control and management [J]. ACM SIGCOMM Computer Communication Review，2005，35(5)：41-54.

[7] YU M，REXFORD，FREEDMAN M J，et al. Scalable Flow-based Networking with DIFANE[J]. ACM SIGCOMM Computer Communication Review，2010，40(4)：351-362.

[8] TOOTOONCHINA A，GANJALI Y. HyperFlow：A distributed control plane for OpenFlow[C]// Proceedings of the 2010 Internet Network Management Conference on Enterprise Networking. USENIX Association，2010：3.

[9] LIN Ping-ping，BI Jun，HU Hong-yu，et al. A Mechanism for Scalabe Intra-domain Control Plane in SDN[J]. Journal of Chinese Computer Systems，2013，34(9)：1790-1794. (in Chinese)
林萍萍，毕军，胡虹雨，等. 一种面向 SDN 域内控制平面可扩展性的机制[J]. 小型微机计算机系统，2013，34(9)：1970-1974.

[10] LIN P，BI J，HU H. Asic：an architecture for scalable intra-domain control in openflow[C]// Proceedings of the 7th International Conference on Future Internet Technologies. ACM，2012：21-26.

[11] KOPONEN T，CASADO M，GUDE N，et al. Onix：a distributed control platform for large-scale production networks[C]// OSDI. 2010：101-106.

[12] VOLKAN Y，ALI O. Controlling a Software-Defined Network via Distributed Controllers[C]// Proceedings of the 2012 NEM Summit，2014. Istanbul：arXiv preprint arXiv，2014：19-27.

[13] HASSAS Y S，GANJALI Y. Kandoo：a framework for efficient and scalable offloading of control applications[C]// Proceedings of the First Workshop on Hot Topics in Software Defined Networks. ACM，2012：19-24.

[14] TAM A S W，XI K，CHAO H J. Use of devolved controllers in data center networks[M]// Computer Communications Workshops(INFOCOM WKSHPS). IEEE，2011：596-601.

[15] KIM T，LEE K，LEE J，et al. A Dynamic timeout control algorithm in software defined networks[J]. International Journal of Future Computer and Communication，2014，3(5)：331-336.

[16] AKAIKE H. Fitting auto regressive models for prediction[J]. Annals of the institute of Statistical Mathematics，1969，21(1)：243-247.

[17] TSU T C，MUGELE R A，MCCLINTOCK F A. A statistical distribution function of wide applicability[J]. Journal of Applied Mechanics-Transactions of the Asme，1952，19(2)：233-234.

[18] ZHOU B，GAO W，WU C，et al. AdaFlow：Adaptive control to improve availability of OpenFlow forwarding for burst quantity of flows[M]// Testbeds and Research Infrastructure：Development of Networks and Communities. Springer International Publishing，2014：406-415.

[19] XIE L，ZHAO Z，ZHOU Y，et al. An adaptive scheme for data forwarding in software defined network[C]// 2014 Sixth International Conference on Wireless Communications and Signal Processing (WCSP). IEEE，2014：1-5.

[20] ZHU H，FAN H，LUO X，et al. Intelligent timeout master：Dynamic timeout for SDN-based data centers[C]// IFIP/IEEE International Symposium on Integrated Network Management (IM)，2015. IEEE，2015：734-737.

[21] BENSON T，AKELLA A，MALTZ D A. Network traffic characteristics of data centers in the wild[C]// Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. ACM，2010：267-280.

(上接第 99 页)

[6] CHEN C. Character-sense Association and Compounding Template Similarity：Automatic Semantic Classification of Chinese Compounds[C]// Proceedings of the 3rd SIGHAN Workshop on Chinese Language Processing. Barcelona. 2004：33-40.

[7] LU Xiao-fei. Hybrid Model for Chinese Unknown Word Resolution[D]. The Ohio State University，2006.

[8] LU Xiao-fei. Hybrid Model for Semantic Classification of Chinese Unknown Words[C]// Proceedings of North American Chapter of the Association for Computational Linguistics-Human Language Technologies 07，2007. New York，2007：188-195.

[9] ZHANG Rui-xia，XIAO Han. The construction of Lattice based on HowNet [J]. Journal of North China Institute of Water Conservancy and Hydro Electric Power，2008，29(3)：53-56. (in Chinese)
张瑞霞，肖汉. 基于《知网》的词图构造[J]. 华北水利水电学院学报，2008，29(3)：53-56.

[10] LU Xiao-fei. Hybrid Model for Chinese Unknown Word Resolution[D]. The Ohio State University，2006.

[11] LU Xiao-fei. Hybrid Models for Semantic Classification of Chinese Unknown Words[C]// Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics，2007. USA，2007：188-195.

[12] BORDES A，GABRILOVICH E. Constructing and Mining Web-Scale Knowledge Graphs：WWW 2015 Tutorial[C]// Proceedings of International Conference on World Wide Web，2015. Italy，2015：1523.

[13] MASS Y，SAGIV Y. Knowledge Management for Keyword Search over Data Graphs[C]// Proceedings of the 23rd ACM International Conference on Information and Knowledge Management，2014. China，2014：2051-2053.

[14] WANG Z，ZHANG J，FENG J L，et al. Knowledge Graph and Text Jointly Embedding[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing，2014. Qatar，2014：1591-1601.

[15] ROKACH L，MAIMON O. Data mining and knowledge discovery handbook(2nd ed)[M]. US：Springer，2005：321-352.

[16] ALFRED R，FUN T S，TAHIR A，et al. Concepts Labeling of Document Clusters Using a Hierarchical Agglomerative Clustering (HAC) Technique[C]// The 8th International Conference on Knowledge Management in Organizations. Springer Netherlands，2013：263-272.

[17] TONG H，FALOUTSOS C，PAN J Y. Fast Random Walk with Restart and Its Applications[C]// Proceedings of IEEE International Conference on Data Mining，2006. China，IEEE Computer Society，2006：613-622.

[18] XIA J，CARAGEA D，HSU W H. Bi-relational Network Analysis Using a Fast Random Walk with Restart[C]// Proceedings of IEEE International Conference on Data Mining，2009. USA，IEEE Computer Society，2009：1052-1057.