北京大学学报(自然科学版), 第 卷, 第 期, 2012 年 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. , No. (. 2012)

# 中文同义词自动抽取研究*

孙玉霞 狄颖 曹冉 孙玉杰 周俊生 曲维光

南京师范大学计算机科学与技术学院，实验室，南京 210046

**摘要** 本文对中文同义词自动抽取的多种方法进行了研究，充分利用现有字典资源和网络百科知识。使用同义词词林和中文概念词典进行字典方法研究；基于百度百科，使用了模式匹配和并列结构的方法。基于模式匹配的方法在少量手动获取模式的基础上，从百科资源中自动获取上下文模式，从而实现同义词获取和自扩展。同时，提出了基于并列结构的同义词自动抽取方法，采用基于词素的过滤方法和基于知网的过滤方法进行过滤，大大提高同义词抽取性能。实验结果表明，多种方法的综合使用，使得本文的中文同义词自动抽取性能大幅度提高，并且适应于多种词类的同义词获取。

**关键词** 同义词词林; CCD; 百度百科; 模式提取; 并列结构
**中图分类号 TP391**

# The Research of Chinese Synonyms Extraction

**SUN Yuxia , DI Ying , CAO Ran , SUN Yujie , QU Weiguang , Zhou Junsheng**
School of Computer Science, Nanjing Normal University, Nanjing,210046, China

**Abstract** This paper studies some methods on Chinese synonyms extraction, making full use of dictionary and encyclopedia resources. Cilin and Chinese Concept Dictionary are used as dictionary resources. Based on baidu encyclopedia , we take pattern matching and coordinate structure based methods. Pattern matching is based on a small amount of manual patterns and obtaines comprehensive context pattern automatically from baidu encyclopedia to the access and expansion of synonyms. Meanwhile, the coordinate structure based method is combined with the morpheme and HowNet filter, which have greatly improved the synonym extraction performance. The experiment shows that the combination of a variety of methods achieved good performance and is adapted for various types of words.

**Key words** Cilin ; CCD; Baidu Encyclopedia; Coordination

在语言学中，同义词是指表达意义相同或者相近，但表达形式不同的词汇。同义词是世界上各种语言都存在的一种普遍现象。在信息表示和信息检索领域中，同义词的概念主要是指一个或多个能够相互替换、表达相同概念的词或词组，并不考虑词汇的感情色彩和语气，与语言学上严格定义的同义词相比，它的含义要宽泛一些[1,2]。面向信息检索的中文同义词主要分为学名与俗名、全称与简称、新称与旧称、型号或代号、译词（专有名词）、字母缩略词、音译词、同一外文词的不同音译词等[3,4]。同义词发现是自然语言处理领域的一个基础研究课题，在机器翻译、词义消歧、信息检索领域中具有十分重要的意义。例如，在搜索引擎中，使用同义词对搜索关键词进行扩展，能够获取到更多更全面的相关信息，有效提高搜索效率。

*基金项目：国家自然科学基金资助项目（61372221，61073119）；江苏省自然科学基金资助项目（BK2010547）

传统的同义词字典存在着一些缺陷，如收录的词数有限，复合词收录很少，专有词汇，如机构名、电影名、人名等涉及较少，词典更新滞后。同时，随着互联网的不断发展，出现了大量的网络用语，词汇的用法呈现爆炸式增长。如“粉丝”，原是指“中国常见的食物之一”，而现在成为英语“Fans”的音译，有“狂热、热爱”之意，后引申为“影迷、追星”。因此，传统的同义词字典远远不能适应日新月异的语言实际的需求。与此同时，互联网的出现也带来了更多的百科资源，如百度百科、维基百科等，这些资源包含了大量专有词汇的描述和解释，同时更新速度较快。但是，由于网络百科资源是由业余编纂者整理的，因此词条描述规范性不强。

本文将基于传统字典的方法和基于网络百科资源的方法进行整合来获取同义词，以便提高同义词抽取系统的性能。

# 1 同义词抽取方法介绍

本文将现有的同义词字典资源和网络资源结合起来抽取同义词。字典资源中同义词信息均是由人工识别，正确率高，但是覆盖率低[10]。网络资源广泛，相关及无关信息同时存在，需要进行甄别。将两者结合使用，则可以兼顾各自的优势，在达到高覆盖率的同时，又能保证更高的正确率。

## 1.1 基于语义词典的方法

基于语义词典的方法主要是根据现有的《同义词词林（扩展版）》[12]和中文概念词典（Chinese Concept Dictionary，CCD）[7]进行同义词自动获取，找寻目标词的同义词。

《同义词词林（扩展版）》是哈工大实验室在梅家驹编纂的《同义词词林》的基础上扩展而成，最终包含 77,458 条词语。扩展版中，同义词编码共 8 位编码位，分为有五级：大类、中类、小类、词群、原子词群类别，编码末尾位是对第五级类别的说明，“=”代表相等、同义，“#”代表不等、同类，“@”代表自我封闭，独立，它在字典中既没有同义词也没有相关词。因此，目标词的同义词即为包含目标词的编码末尾为“=”的同义词集合。

“中文概念词典”是一部 WordNet 类型的汉英双语语义词典。一方面，CCD 在规格上要求与 WordNet 兼容，即在尽可能不破坏 WordNet 框架（以同义词集定义概念并描述这些概念之间的关系）的前提下细化汉语的语义描述，以便直接复用现有的 WordNet 理论、方法和技术；另一方面，考虑语种不同必然导致的描述结构不同，CCD 对汉语概念的内容和概念之间的关系必须进行一定的调整和发展，着力突出汉语的特点、反映汉语的语言事实。

CCD 使用 Synset 来描述概念，要求该字段中的每个词语，其词义都能相当准确地表达当前概念；能基本准确地表达当前概念的每个词语都应该出现在当前概念的 Synset 字段中。因此，一个词语可能出现在多个概念的 Synset 中，如词语“爱好”所属的部分概念如表 1：

表 1 “爱好”对应的概念集合

| 概念编号 | 定义 | Synset |
|---|---|---|
| 01745360 | 吸引普通大众 | 喜好 流行 爱好 盛行 风行 |
| 00273902 | 一种附属的活动 | 嗜好 爱好 业余爱好 |
| 04463325 | 特别的宠爱或爱好 | 亲信 嗜好 宠信 宠儿 心肝 心腹 爱好 心肝宝贝 |
| 04700175 | 对关照的对象持先入为主的观念 | 偏好 偏爱 偏袒 爱好 |
| 05565069 | 对某东西渴望的感情 | 倾向 本能 欲望 欲求 渴望 爱 |

| | | |
|---|---|---|
| | | 好 胃口 食欲 |
| 05608483 | 一种积极的喜爱的 | 情意 感情 柔情 深情 温情 爱好 真情 风情 |
| 05573285 | 一种愉快和享乐的感情 | 喜好 喜欢 嗜好 爱好 |

表 1 中，Synset 是使用同义词集合来对概念进行描述。但是，在一些概念描述中并不完全是同义词，而只是某种意义上的相关，如 { 友爱，情意，感情，柔情，深情，温情，爱好，爱情，真情，风情 } 。因此，本文提出使用典型同义词过滤的方法：基于包含目标词的概念集合统计出典型同义词，计算典型同义词在每个概念对应的同义词集合中所占的比例，按照一定的规则进行同义词的过滤。

给定目标词 W，包含 W 的概念集合为 Set={synset$_1$, synset $_2$,…, synset $_m$}，统计概念集合中所有词语出现的频率 FreList={$w_1$:$f_1$,$w_2$:$f_2$,…,$w_n$:$f_n$}，其中 $f_1$>$f_2$>…>$f_n$。对于词语 $w_i$（1<=i<=n），若其 $f_i$ 大于 1 且排名位于前 1/3 或者 $w_i$ 跟 W 字面相似度大于等于 1/3，则称 w 为典型同义词。字面相似度使用 Jaccard similarity coefficient 计算，如公式 1 所示：

$$JC(W, w_i) = \frac{|W \cap w_i|}{|W \cup w_i|} \quad (1)$$

其中，$|W \cap w_i|$ 指目标词和候选同义词包含的相同的汉字个数，$|W \cup w_i|$ 指目标词和候选同义词共包含的汉字个数。

给定目标词 W，包含 W 的概念集合为 Set={synset$_1$, synset $_2$,…, synset $_m$}，每个概念使用同义词集合 synset$_i$（1<=i<=m）来表征。若 m=1 或者 W 的概念集合包含的同义词个数小于 10 个，则无需进行过滤。否则，进行同义词过滤。首先基于概念集合统计典型同义词。然后，遍历概念集合中的每一个同义词集合 synset$_i$（1<=i<=m）：若典型同义词在该同义词集合中所占比例超过 0.8，则该同义词集合中所有词语为典型，将其加入到典型同义词表中；若其典型同义词所占比例超过 0.5，则该同义词集合中的词语是有效的同义词；若其典型同义词所占比例小于 0.5，则该同义词集合中的词语只是概念上的相关而不是同义词。

基于表 1 的概念集合，根据频率统计，出现次数大于 1 且排名前 1/3 的词语为{嗜好：3，喜好：2}，拥有较高字面相似度的词语为{业余爱好：1/2，喜好：1/3，嗜好：1/3，偏好：1/3，偏爱：1/3，}，因此可获得典型同义词：{嗜好，喜好，业余爱好，偏好，偏爱 }。对于概念 00273902，典型同义词所占比例为 100%，则该同义词列表中所有的同义词均是典型；对于概念 04700175 和 05573285，典型同义词所占比例均为 75%，因此这两个同义词列表中包含的所有同义词均是有效的。最终获取到的同义词为：{ 嗜好，喜好，业余爱好，偏好，偏爱，偏袒，喜欢} 。

## 1.2 基于网络资源的方法

前人基于各种百科资源做了一系列的工作[6][8][9][11]，本文主要针对百度百科资源进行处理，主要采用两种方法：基于模式匹配的方法，基于并列结构的方法。

### 1.2.1 基于模式匹配的方法

模式的提取主要有两种。第一种是人工提取的模式，主要用在百度百科概述中，用以提取高质量的同义词。第二种是基于自动获取模式的同义词自扩展方法，从百度百科中自动学习模式，并对同义词进行自扩展。

在百度百科资源中，百科名片是最为规范的部分，百科名片位于百度百科词条页面的上

方，包括词条概述和基本信息栏两个部分。词条概述是通过图文方式，对词条内容进行简明阐述，其中文字内容较为规范，同时也包含了该词条的最基本相关的信息，包括同义词部分，并且同义词一般出现在概述开头，以“目标词+模式+候选同义词”的形式出现，如目标词“栀子花”的词条百科概述开头：

**例 1：** “栀子花又名栀子、黄栀子”。

而词条全文中也存在包含同义词共现模式的句子，并且同义词信息更加详尽，但是模式规范性较差，若使用“目标词+模式+候选同义词”框架进行提取，则容易遗漏不与目标词共现的包含同义模式的句子，如目标词“栀子花”的词条百科介绍：

**例 2：** “【别名】：栀子、黄鸡子、黄荼子、黄栀子、黄栀、山黄栀、玉荷花等。”。

在例 2 中，由于缺乏目标词“栀子花”，因此无法使用“目标词+模式+候选同义词”框架对其进行同义词提取。但是由于全文信息量较大，包含了大量非目标词语的同义词共现句子，若不要求目标词与同义词共现，仅根据模式，使用“模式+候选同义词”框架从目标词对应的词条百科全文中提取同义词，则会获取到大量无关词汇。如目标词“苏州”的词条百科中包含如下句子：

**例 3：** “碧螺春产自苏州市太湖洞庭山碧螺峰，俗名“吓煞人香”。”。

使用模式“俗名”能获取到词语“吓煞人香”，该词语是“碧螺春”的同义词而非目标词“苏州”。因此，对于全文的同义词发现必须对全文句子进行筛选，选取包含目标同义词对的句子作为限定语料进行同义词发现，假设限定语料中所包含的是目标词的同义模式。同时使用上下文环境模式进行同义词自扩展。本文首先基于词条概述和手动获取的模式进行高质量同义词发现，然后基于词条全文和高质量同义词获取限定语料和上下文环境模式，并进行同义词自扩展。

本文首先基于 770 个词条百科手动进行模式提取，然后使用“目标词+模式+候选同义词”框架对给定目标词的词条概述部分进行高质量同义词提取。然后对获取到的高质量同义词以及目标词两两组对，从百科全文中获取包含同义词对的限定语料。提取限定语料中同义词对后一个词语的语言环境作为模式，使用其左右相邻的共现词语作为模式框架，使用左三元组列表进行过滤，提取同义词，并将其加入到高质量同义词集合中，利用新获取到的同义词扩展限定语料，基于模式不断获取新同义词，直至获取不到新同义词。

如：使用“栀子花+又名+候选同义词”框架可从例 1 中获取到高质量同义词“栀子”、“黄栀子”。对其两两组对，{“栀子”，“黄栀子”}，{“栀子花”，“栀子”}，{“栀子花”，“黄栀子”}。根据同义词对百科全文中获取到如例 2 的限定语料。基于限定语料，获取上下文语言环境，如左词语“、”和右词语“、”，三元组“、「同义词」、”。然后，基于左词语“、”和右词语“、”，获取候选同义词：“黄鸡子”，“黄荼子”，“黄栀”，“山黄栀”，“玉荷花”，根据三元组“、「同义词」、”进行过滤，判定“黄栀”，“黄鸡子”为同义词，以此类推可以获取到该句子中的全部同义词。

由于网络资源存在不规范性，需要对获取到的同义词进行筛选。首先对获取到的同义词进行分词，然后使用如下规则进行筛选：

**规则 1** 若候选同义词是一个完整的分词结果，同义词有效。

**规则 2** 若候选同义词分词结果中包含{ 共和国，市，州，镇，县，郡，帝国 }中任一后缀，同义词有效。

**规则 3** 若候选同义词分词结果是两个字数词性一致的词语，同义词有效。

**规则 4** 若候选同义词分词结果包含目标词语，同义词无效。

上述四个规则，优先度依次递减。如目标词“白兰地”，候选同义词为“白兰地酒”，分词结果为“白兰地酒/n”，根据规则 1，为有效同义词；目标词“奥地利”，候选同义词为“奥地利共和国”，分词结果为“奥地利/nsf 共和国/n”，根据规则 2，为有效同义词；目标词“霸

道”候选同义词为“强横霸道”，分词结果为“强横/a 霸道/a”，根据规则 3，为有效同义词；目标词“大运河”候选同义词为“京杭大运河”，分词结果为“京/b 杭/b 大运河/n ”，根据规则 4，该候选同义词无效。

### 1.2.2 基于并列结构的方法

并列结构包含的两个词语一般有共同的形态，语义相似或相关[5]，中间使用 “和，或，逗号，顿号”等并列连词进行连接，这为同义词获取提供了一条思路。假设并列结构所包含的两个词语中的其中一个词是目标词，则另一个作为该目标词的候选同义词。

**例 4：** 浅谈责任感与责任心。

其中，“责任感”与“责任心”为并列结构，他们由并列连词“与”连接，所以“责任心”是目标词“责任感”的候选同义词。但是由于这种并列结构并不总是能提供出满意的同义词，如例 5：

**例 5：** 可贵的责任感与事业心。

其中，“责任感”与“事业心”虽然也是并列结构，但并不是同义词，而是相关词。所以本文采用基于词素和知网的方法进行严格过滤，确保同义词的正确性。

#### 1.2.2.1 基于词素的方法

假设目标词 X 和词语 Y，X 的汉字个数为 m，Y 的汉字个数为 n。

**规则 5** 如果满足 m-n=0，且 X 和 Y 中相同的中文字数=m-1，则 X 和 Y 为同义词，否则 Y 不是 X 的同义词。

例如“责任感”与“责任心”，字数同为 3，相同的字数为 2=3-1，所以是同义词，而“责任感”与“事业心”则不是。考虑到有些情况，如，“文档”和“档案”是目标词“文件”的同义词，根据规则 5，可以推出“文档”，但是无法推出“档案”，所以本文提出二次匹配规则。

**规则 6** 假设，根据规则 5，Y 是目标词 X 的同义词，将 Y 加入到待匹配集中去，如果词语 Z 和待匹配集中的任何一个词语进行匹配，若符合规则 5，则 Z 也是 X 的同义词。

例如，根据规则 5，“文档”是目标词“文件”的同义词，将“文档”也加入到待匹配集中，词语“档案”与待匹配集中的“文档”进行匹配符合规则 5，所以词语“档案”也是目标词“文件”的同义词。

这里最多只能进行二次匹配，否则会引起同义词扩展误差，例如“案例”与“档案”进行匹配也符合规则 5，但是“案例”却不是目标词“文件”的同义词。

#### 1.2.2.2 基于知网的方法

知网[1]（HowNet）是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。本文利用目标词 X 和词语 Y 在知网中的义项（也称“词语的概念”）进行候选同义词的过滤。

**规则 7** 若 X，Y 都在知网中，且有相同的义项，则是同义词。例如，词语“依赖”和“依靠”在知网中有相同的义项“DEF=depend|依靠”，因此“依靠”是目标词“依赖”的同义词。

**规则 8** 若 X，Y 都在知网中，且没有相同的义项，则不是同义词。例如，词语“信任”的义项“DEF=believe|相信”，与目标词“依赖”的义项“DEF=depend|依靠”不同，所以词语“信任”不是目标词“依赖”的同义词。

**规则 9** 若 X 与 Y 之有一个不在知网中，则不是同义词。由于知网不可能收录所有的词语，所以一些在知网中没有出现的词语如果作为目标词或者候选同义词就会被直接过滤掉，

[1] http://www.keenage.com/html/c_index.html

例如词语“城管”是一个新词语，没有在知网中出现，进行过滤时直接过滤即可。

有的词语在知网中有多个义项，例如“打”有义项“DEF=compile|编辑”（打草稿），也有义项“DEF=buy|买,commercial|商”（打饭）等。这种情况下，只要该词有一个义项与候选同义词的义项相同，就认为候选同义词是该词的同义词。例如词语“购”在知网中也有义项“DEF=buy|买,commercial|商”，所以“购”是“打”的同义词。

#### 1.2.3 传递性扩充

对于获取到的同义词，可能有交叉现象，例如：目标词“北京”的同义词是“北京市”和“京”，目标词“北京市”的同义词是“京”，目标词“北京市”是目标词“北京”的同义词，所以这两个目标词应该共享同义词。然而并不是所有含有相同同义词的两个目标词都应该集成，例如：目标词“巴黎”的同义词是“灯城”和“花都”；目标词“花都”的同义词是“花都区”和“花县”，这里的“花都”表示广州的一个花都区，和目标词“巴黎”没有任何关系，因此，本文提出了传递性验证规则 10。

**规则 10** 假设目标词 X1 有 m 个同义词，目标词 X2 有 n 个同义词，X1 和其 m 个同义词组成 set1，X2 和其 n 个同义词组成 set2，且 set1 和 set2 中相同词语的个数为 equ，若 equ 大于等于 set1 或 set2 长度的一半，则认为目标词 X1 和 X2 是同义词，并将 X1 和 X2 的同义词进行扩充至相同。

例如，目标词“北京”和“北京市”所对应的 set1{北京，北京市，京}和 set2{北京市，京}长度分别为 3 和 2，相同词的个数 equ 为 2，大于 set1 和 set2 长度的一半 3/2 和 2/2，符合规则 10，所以目标词“北京”和“北京市”是同义词，并将其同义词分别进行扩充；目标词“巴黎”和“花都”对应的 set1 和 set2 长度均为 3，相同词的个数 equ 为 1，小于 set1 和 set2 长度的一半 3/2，不符合规则 10，所以目标词“巴黎”和“花都”不是同义词，不用进行同义词扩充。

## 2 实验与结果分析

### 2.1 实验数据

本文使用 NLP&CC2012 提供的同义词语料，共包含 9455 个目标词，主要词语类型为普通名词、动词、形容词，地名，人名，机构名，专业术语等。

### 2.2 基于语义词典的同义词获取实验

基于同义词词林获取到 5277 个目标词的同义词，平均每个词语对应 13 个同义词；基于 CCD 是 5727 个，平均每个对应 6 个同义词。基于字典的方法获取到的同义词的目标词大多是普通名词、动词、形容词。

### 2.3 基于模式匹配的同义词获取实验

#### 2.3.1 人工模式提取

基于 770 个词条百科，手动进行模式提取，这里的模式是指同义词提示词，如“栀子花又名栀子”，其中“又名”就是所要提取的模式。获取到的模式规模如表 2 所示：

表 2 获取到的模式

| 模式类型 | 模式数量 | 实例 |
|---|---|---|
| 前置名词模式 | 39 | 模式：昵称<br>例：周笔畅的昵称是笔笔。 |
| 前置动词模式 | 64 | 模式：又名<br>例：栀子花又名栀子。 |
| 后置模式 | 8 | 模式：美称<br>例：杭州素有人间天堂的美称。 |

### 2.3.2 同义词自扩展

对于给定的 9455 个目标词，其中有 8268 个词语能够获取到对应的词条百科。基于词条概述能够获取到同义词的有 815 个目标词。

基于百科全文获取包含同义词词对的句子。在获得的句子集合上，提取左右词语列表、左三元组列表，并根据获得的列表进行同义词穷尽自动抽取。如表 3 所示：

表 3 获取到的左右列表

| | 左词语列表 | 右词语列表 | 左三元组列表 |
|---|---|---|---|
| 排名前 5 | 、<br>，<br>称<br>和<br>又 | 、<br>，<br>。<br>；<br>等 | 、「同义词」、<br>，「同义词」，<br>，又称<br>称「同义词」、<br>”、“ |
| 总计 | 74 | 24 | 131 |

最终，使用过滤规则进行过滤，部分过滤实例如表 4：

表 4 原结果和过滤后结果对比

| 目标词 | 过滤前（候选同义） | 过滤后（候选同义） |
|---|---|---|
| 巴西 | 巴西文化 巴西联邦共和国 巴西郡 | 巴西联邦共和国 巴西郡 |
| 霸道 | 强横霸道 横行霸道 | 强横霸道 横行霸道 |
| 白金 | 人造白金 | 无 |
| 白兰地 | 白兰地酒 | 白兰地酒 |

## 2.4 基于并列结构的同义词获取

在基于并列结构的同义词获取中，首先要确定并列连词，包括“和，与，或，逗号，顿号”。然后在百度网页中搜索“目标词+并列连词”以及“并列连词+目标词”，因为目标词有可能在前，也可能在后，取前 500 条搜索结果中的有效内容（即包含目标词和并列词且不能和之前的结果重复）。在此基础上，再进行编码的转换，分词和词性标注（调用 ICTCLAS2012[2]），因为分词有可能将目标词拆分，所以需扫描句子将目标词合并处理。在分词和词性标记的基础上，抽取目标词的并列词作为候选同义词（名词，动词和形容词）。这里不取专有名词，

[2] ICTCLAS（Institute of Computing Technology,Chinese Lexical Analysis System）是由中国科学院计算技术研究所研制的汉语词法分析系统，主要功能包括中文分词，词性标注等。ICTCLAS2012 是 2012 年发布的新版本。

因为利用并列结构抽取的专有名词同义词候选大多为相关词，例如：句子“刘备与孔明实质关系”中，“刘备”与“孔明”为相关词。部分目标词的同义词候选如表 5 所示，有些目标词候选太多，只展示部分同义词候选。

表 5　基于并列结构获取的同义词候选

| 专辑 | 写真 | 乐队 | 名字 | 专刊 | 音乐 | 专集 | 增刊 | 乐团 | 电影 |
|---|---|---|---|---|---|---|---|---|---|
| 书摊 | 书店 | 旧书 | 书亭 | 书市 | 城管 | 地铁 | 出版社 | 书城 | 教授 |
| 劳累 | 疲乏 | 舒服 | 辛苦 | 轻松 | 忙碌 | 疲倦 | 疲劳 | 紊乱 | 紧张 |
| 丰盈 | 丰沛 | 饱满 | 丰满 | 强大 | 充沛 | 美好 | 纤细 | 简约 | 伟大 |
| 修订 | 完善 | 修正 | 改造 | 构建 | 修改 | 新增 | 调整 | 建立 | 改进 |
| 依赖 | 恋爱 | 依恋 | 驾驭 | 依靠 | 喜欢 | 高估 | 依附 | 滥用 | 信赖 |
| 伪装 | 化妆 | 掩埋 | 欺骗 | 隐匿 | 化装 | 佯装 | 隐瞒 | 掩饰 | 假装 |
| 俊俏 | 婀娜 | 英俊 | 鲜艳 | 俊秀 | 美丽 | 帅气 | 绚丽 | 挺拔 | 秀美 |

由上表可以看出，目标词的候选同义词大多是相关词或者无关词。所以需要进行过滤处理，只保留同义词，从而提高准确率。

### 2.4.1 基于词素的过滤方法

基于词素的过滤规则（规则 5，规则 6）在上文中已详细阐述，该方法对于词语的字面形式要求极其严格，表 6 即为经过词素过滤的同义词候选展示。

表 6 经过词素过滤的同义词候选

| 专辑 | 专集 | 专刊 | 增刊 | | | |
|---|---|---|---|---|---|---|
| 书摊 | 书店 | 旧书 | 书亭 | 书市 | 书城 | 城管 |
| 劳累 | 疲劳 | 疲乏 | 疲倦 | | | |
| 丰盈 | 丰沛 | 丰满 | 饱满 | 充沛 | | |
| 修订 | 修正 | 修改 | 改造 | 改进 | | |
| 依赖 | 依恋 | 依靠 | 依附 | 信赖 | 信任 | 恋爱 |
| 伪装 | 化妆 | 化装 | 佯装 | 假装 | | |
| 俊俏 | 英俊 | 俊秀 | 秀美 | | | |

以目标词“俊俏”为例，经过规则 5 过滤，找到同义词“英俊”和“俊秀”，根据二次匹配的规则 6，找到同义词“秀美”。以目标词“书摊”为例，根据规则 5，找到符合条件的同义词“书店”、“旧书”、“书亭”、“书市”、“书城”，根据规则 6 再找到同义词候选“城管”。不过，有些词语虽然符合了词素过滤规则，字面相似，但是语义并不相似，例如“城管”。因此，本文继续使用基于知网的语义知识进行二次过滤。

### 2.4.2 基于知网的过滤方法

利用知网的义项来进行过滤，首先将知网转换为一个“词语-义项”字典，词语作键，义项作键值（例如键“依赖”的键值是义项“依靠”），利用字典可以提高查找效率。有的词语又多个义项，则在词语前添加一个 id 号来区分。将目标词与候选同义词分别在字典中查找，比较义项，根据义项判定是否是同义词。表 7 为经过知网过滤的同义词候选。

表 7 经过知网过滤的同义词候选

| | |
|---|---|
| 专辑 | 专集　专刊 |
| 书摊 | 书店　书市 |
| 劳累 | 疲劳　疲乏　疲倦 |
| 丰盈 | 丰沛　丰满　充沛 |
| 修订 | 修正　修改 |
| 依赖 | 依靠　依附 |
| 伪装 | 化装　佯装　假装 |
| 俊俏 | 英俊　俊秀　秀美 |

利用知网中的义项信息，可以把与目标词的义项不相同的候选同义词过滤掉。如上表中的目标词“依赖”在知网中的义项“依靠”，而候选同义词“信任”的义项“相信”，两者义项不相同，根据规则 8 所以过滤。目标词“书摊”的候选同义词“城管”由于不在知网中，根据规则 9 所以过滤掉。

## 2.5 汇总集成

由各种方法获得的同义词候选最后需要进行汇总集成。根据本文提出的同义词传递方法，对方法内结果进行同义词传递性验证。

根据传递性验证规则 10，部分符合规则并成功合并的目标词及其同义词如下表 6 所示。

表 6 原同义词结果与传递性扩充后的结果对比

| | 目标词 | 原同义词结果 | 扩充后结果（粗体为增加的） |
|---|---|---|---|
| 根据规则 11 需要扩充的词 | 整洁 | 清洁 | 清洁　**洁净**　**清新** |
| | 洁净 | 清洁　清新 | 清洁　清新　**整洁** |
| | 山芋 | 红薯 | 红薯　**甘薯**　**红芋**　**番薯**　**白薯**　**白芋**　**地瓜**　**红苕** |
| | 甘薯 | 山芋　红芋　番薯　红薯　白薯　白芋　地瓜　红苕 | 山芋　红芋　番薯　红薯　白薯　白芋　地瓜　红苕 |
| 根据规则 11 不需要扩充的词 | 巴黎 | 灯城　花都 | 灯城　花都 |
| | 花都 | 花都区　花县 | 花都区　花县 |
| | 花椒 | 香椒　大花椒　青椒　山椒 | 香椒　大花椒　青椒　山椒 |
| | 青椒 | 大椒　灯笼椒　柿子椒　甜椒　菜椒 | 大椒　灯笼椒　柿子椒　甜椒　菜椒 |

通过上表可以看出，传递性扩充可以将互为同义词的目标词在满足一定条件的情况下（见规则 10）进行合并，在保证正确率的前提下，提高目标词的同义词抽取召回率。必须要在原同义词结果正确率高的前提下进行，否则进行传递性扩充，只会降低同义词抽取的正确率。

本研究参加了中国计算机学会中文信息技术专业委员会举办的“中文微博情感分析&词汇语义关系抽取评测”中的同义关系抽取评测并取得了在微平均 F 值和宏平均 F 值上均得第一，验证了该方法的有效性，具体结果如表 7。

表 7 同义关系评测结果

| | 宏平均准确率 | 宏平均召回率 | 宏平均F1值 | 微平均准确率 | 微平均召回率 | 微平均F1值 |
|---|---|---|---|---|---|---|
| 中科院声学所 | 0.1328 | 0.1034 | 0.1033 | 0.4737 | 0.0687 | 0.1199 |
| 北京理工大学 | 0.1999 | 0.2441 | 0.1874 | 0.2115 | 0.2299 | 0.2203 |
| 北京交通大学 | 0.2878 | 0.3394 | 0.2733 | 0.3088 | 0.3737 | 0.3382 |
| 华为 1 | 0.3641 | 0.5176 | 0.3664 | 0.2754 | 0.5829 | 0.3740 |
| 华为 2 | 0.3305 | 0.5506 | 0.3635 | 0.2615 | 0.6102 | 0.3662 |
| 华侨大学 | 0.0382 | 0.0111 | 0.0151 | 0.2996 | 0.0115 | 0.0221 |
| 南京师范大学 | 0.3588 | 0.6041 | 0.3968 | 0.3025 | 0.6358 | 0.4100 |
| 哈尔滨工业大学 | 0.3225 | 0.3885 | 0.2842 | 0.2303 | 0.3676 | 0.2832 |
| 郑州大学 1 | 0.2975 | 0.6395 | 0.3588 | 0.2530 | 0.6762 | 0.3682 |
| 郑州大学 2 | 0.3256 | 0.6930 | 0.3919 | 0.2540 | 0.7040 | 0.3734 |

## 3 结语

中文同义词抽取在自然语言处理和信息检索领域中具有重要意义，本文主要对中文同义词自动抽取的多种方法进行了研究，充分利用词典和网络百科资源，基于字典的抽取方法主要获取普通名词、动词、形容词的同义词。基于模式匹配方法和基于并列结构方法主要获取专有名词等的同义词。基于模式匹配的方法在少量手动获取模式的基础上，从百科资源中自动获取全面的上下文模式，从而实现同义词获取和自扩展。同时，提出了基于并列结构的同义词自动抽取方法，并且由于基于词素的过滤方法和基于知网的过滤方法的应用，大大提高了基于简单并列结构的同义词自动抽取方法的性能。多种方法的综合使用，使得本文的中文同义词自动抽取性能有较大地提高。

本文也存在一些问题，为了确保正确率，在使用基于词素的过滤方法时要求字面相似度高，因此会漏掉那些字面相似度稍低的同义词，在未来的工作中，将考虑针对这个问题进行改进。

## 参考文献


[1] 陆勇. 面向信息检索的汉语同义词自动识别[D]. 南京: 南京农业大学,2009.

[2] 陆勇,侯汉清. 基于 PageRank 算法的汉语同义词自动识别. 西华大学学报(自然科学版) ,2008, 27(2):13-16.

[3] 吴志强.经济信息后控制词表的研究[D].南京:南京农业大学,1999.

[4] 陆勇,章成志,侯汉清. 基于百科资源的多策略中文同义词自动抽取研究. 中文图书馆学报,2010(1):56-62.

[5] 吴云芳,李素建,李芸,等. 双向考察和验证:并列成分中心语的语义关系和 CCD 的名词语义

分类体系. Computational Linguistics and Chinese Language Processing , 2005, 10(4):543-552.
[6] 陆勇,侯汉清. 基于模式匹配的汉语同义词自动识别. 情报学报,2006,25(6):720-724.
[7] 于江生,俞士汶. 中文概念词典的结构. 中文信息学报,2002,16(4):12-20.
[8] Andrey Simanovsky, Alexander Ulanov. Mining Text Patterns for Synonyms Extraction . The 22nd International Workshop on Database and Expert Systems Applications, 2011 ,473-477.
[9] Christian Bohn, Kjetil Norvag. Extracting Named Entities and Synonyms from Wikipedia . Proceedings of the 2010 24th IEEE International Conference on Advanced Information Networking and Applications, AINA '10, 2010.
[10] T. Wang, G. Hirst. Extracting synonyms from dictionary definitions. Proceedings of RANLP 2009. RANLP, 2009.
[11] David Milne, Olena Medelyan, Ian H. Witten. Mining Domain-Specific Thesauri from Wikipedia: A Case Study.International Conference on Web Intelligence (WI'06), 2006.
[12] 《同义词词林》扩展版. http://www.ir-lab.org/