基于网络资源的词语语义关系自动获取

刘江鸣[1] 徐金安[1,†] 吴培昊[1] 张玉洁[1]

1. 北京交通大学计算机与信息技术学院，北京 100044

† 通讯作者，E-mail: jaxu@bjtu.edu.cn

摘要 针对中文词语语义关系自动获取的问题，提出了一种基于维基百科和百度百科等网络资源的同义及上下位语义关系的获取方案；实验结果显示，提出的语义关系获取方案，在同义关系和上下位关系自动识别中达到很好的效果。上下位关系自动识别宏平均达到 0.4185，微平均达到 0.5596。

关键词 词语语义关系；网络资源；同义关系；上下位关系

中图分类号 TP181

Automatic Acquisition of Lexical Semantic Relationship based on Web Resource

LIU jiangming[1], Xu jian[1,†], WU peihao[1], ZHANG yujie[1]

1. School of Computer and Information Techology, Beijing Jiaotong University, Beijing, 100044

† Corresponding Author, E-mail: jaxu@bjtu.edu.cn

Abstract The problem of automatic acquisition of Lexical Semantic Relationship is important and common. This paper tackles the problem by proposing a scheme on automatic acquisition of synonymy and hyponymy based on web resource. The experimental results show that the scheme performs well in the tasks of lexical semantic relationship extraction. Besides, macro average of automatic acquisition of hyponymy is 0.4185 and micro average of automatic acquisition of hyponymy is 0.5596.

Key words Lexical Semantic Relationship; Web Resource; Synonym; Hyponymy

词语语义关系研究是自然语言处理非常重要的关键任务之一。词语语义关系不仅是构造知识库的基本资源，而且在数据挖掘、情感计算、机器翻译等研究领域中占据着举足轻重的地位。

词语语义关系是指在语义范畴中建立起来的词汇间的逻辑关系，主要包括同义关系和上下位关系。存在同义关系的词汇，互称为同义词（同义异形词），其表达的意义相同或相近，但表达形式不同，主要包括别称、俗称、全称、简称、异形词、外来语译名差异和语义近似词汇等。存在上下位关系的词汇，称为上下位词，下位词指其语义包含在另一个词汇（称为上位词）之中的词汇。即下位词是上位词的一个特殊实例。

在语言处理初期，词语语义关系主要是由手工构建，这种模式费时费力。之后，由于网络信息膨胀和多变的语言现象，因此基于大规模数据的语义关系自动获取研究日渐成为了重要研究课题之一，备受国内外研究者的高度关注，同时，提出了很多语义关系自动识别策略。

词语语义关系来自于大量的文本信息。目前，语义关系自动获取方法主要是建立在大量文本信息的基础上，采用不同的研究取策略实现预期效果。传统方法主要包括：基于规则的方法[1-3]，其所使用的规则取决于语言的特点，包括词法和句法等，主要采用的是词级的词模式和句子级的依存模式，此方法能够准确获取语义关系，但是对于新词和非规范出现的词汇等不易于查全，再者，模式中是否使用句子结构问题，Sang ETK 等人[4-5]对此给出了分析和比较；基于多策略的方法[6]，主要寻求合适方法将多个策略融合，达到相互取长补短的效果，其关键思想体现在如何寻求一个较好的融合策略，改善抽取效果；基于图模型的方法[7-9]，通过使用词汇之间边权重得到语义关系，涉及大量的文本相关联信息，需要大规模计算数据。

随着互联网飞速发展，互联网成为最主要的文本信息来源，因此词语语义关系自动抽取离不开网络文本资源，网络文本资源成了语义关系最主要的载体[10]，如维基百科和百度百科等。本文借鉴前人的方法，利用维基百科和百度百科等网络资源，提出语义关系自动抽取的新策略，针对不同词语的特点，采取不同

的抽取策略。实验结果在NLP&CC语义关系识别评测任务中取得较好效果。

本文组织如下：第一节介绍同义关系自动获取策略；第二节介绍上下位关系自动获取策略；第三节列出实验结果；第四节总结全文并展望未来工作。

1 同义关系

1.1 国内外资源

目前国内外存在相关的同义关系资源，例如哈工大信息检索研究室的《同义词林》扩展版、《中文概念词典 CCD》和 WordNet。这些词典所收入的是日常生活中的词语和一些专业术语。在网络飞速发展的今天，流行词汇等新词不断涌现，同义关系日益复杂。因此，有效的同义关系自动识别策略，能够更好地识别新词间的同义关系，进而动态扩展同义词典，为相关研究奠定坚实的基础。

1.2 同义关系自动识别

本文提出一种同义关系自动识别策略，并应用于语义关系自动获取系统的子系统——同义关系识别系统（图 1）。此子系统主要用于动态地查询和构造同义词词典。同义关系自动识别策略主要分为两个部分：基于模板匹配的方法和基于词集合处理的方法。

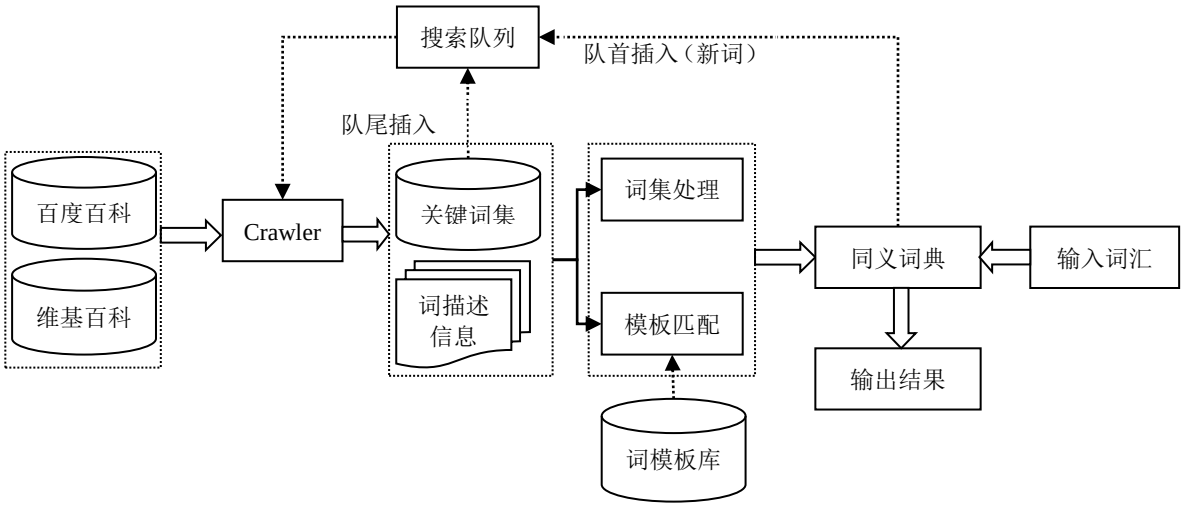

图 1 同义关系自动获取系统

Fig.1 the system of automatic acquisition of synonyms

1.2.1 模版匹配

基于模板匹配的主要任务是词模板库的构建，由于百度百科和维基百科的词条概念是以固定的规则描述，并且描述文本中蕴含着大量同义关系信息。因此，通过有效的词模板库进行模板匹配的方法，可以有效地提取同义关系。因此本文借鉴陆勇等提出的方法，通过筛选和增加模板信息，建立一个相对完善的词模板库，进而获得部分同义关系。模版选取规则如下（其中认为 A 与 B 存在同义关系）：

规则1. A(简称|简称为|中文简称|又称|又称为|亦称|亦叫|亦作|又叫|也称|也称为|俗称|又译|又译作|全称为|全称是){左引号|冒号}B{右引号}

在百度百科和维基百科中存在描述如下：

聚乙烯:简称 PE，是乙烯经聚合制得的一种热塑性树脂

杠杆原理：亦称“杠杆平衡条件”。

古琴，亦称瑶琴、玉琴、七弦琴，…

证券经营机构：也称证券商或证券经纪人

规则2. A{是|即}B(的简称|的全称|的对称|的缩写)

在百度百科和维基百科中存在描述如下：

马哲：是马克思主义哲学的简称

VOD 是 Video On Demand 的缩写

默示保证：明示保证的对称。

规则3. A{:|【}<中文别名|通用名称>{】|:}B

在百度百科和维基百科中存在描述如下：

何首乌:中文别名：首乌、夜交藤、赤首乌、铁秤砣

规则4. A(和)B(是同义词)

在百度百科和维基百科中存在描述如下：

万维网和 www 是同义词，已合并。

1.2.2 词集合

基于词集合处理的方法主要使用的是文本描述中的关键词集，并按照本文给出的定义 1 和定义 2 确定同义关系。

定义 1: 关键词集（KS）定义为相关词集和超链词集的并集。相关词集定义为两部分，其一为百度百科描述信息中，相关词条一栏中的所有词汇；其一为维基百科描述信息中相关条目中所有词汇。超链词集定义为概念描述信息中，能够链接到其它网页的词汇的集合。

定义 1: 词 A 的关键词集为 A_KS，词 B 的关键词集为 B_KS，当满足条件词 A 和词 B 互相出现于关键词集时，即满足 A∈B_Pset & B∈A_Pset 时，认为词 A 与词 B 存在同义关系。

2 上下位关系

本文提出的上下位关系自动识别策略，并应用于语义关系自动识别系统的子系统——上下位关系识别系统（图 2）。此子系统主要用于动态地查询和构造上下位关系表。系统采用基于多策略并行的方法，其中包括学科分类、词汇细化、开放分类和模板匹配。

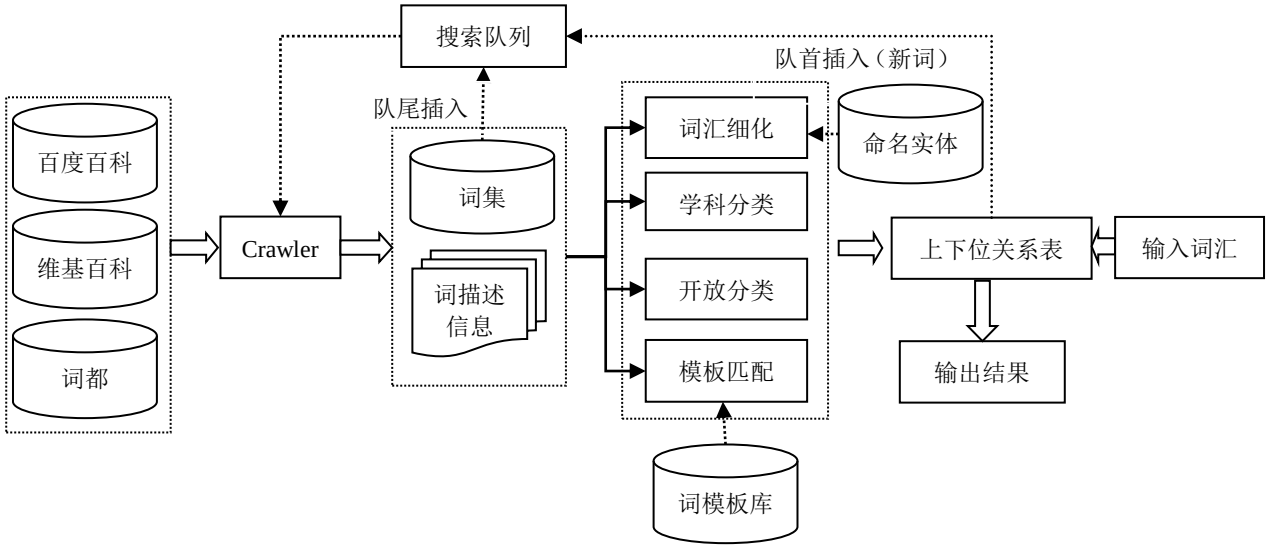

图 2 上下位关系自动获取系统

Fig.2 the system of automatic acquisition of hyponymy

2.1 学科分类策略

中文大量词汇集合中，词汇是按照一定知识分为多个大类，并且每个大类中的词汇具有相同的共性，例如常用领域。文本提出学科分类策略，在某一专业学科中，词汇的上下位关系有明显的定义。目前系统仅对生物学领域做初步的处理，如图 3 所示，生物学中存在的界、门、纲、目等有明显的上下位关系。图 3 中有明显的上下位关系，因此对此类词汇单独处理能够取得好的效果，例如鹱鹱目的下位词包括鹱鹱科、鼹科、沟齿鹱科、岛鼩科、麝鹱亚科、鹱鹱亚科、非洲白齿鹱亚科、美洲鼹亚科、鼹亚科和鹱鼹亚科等。

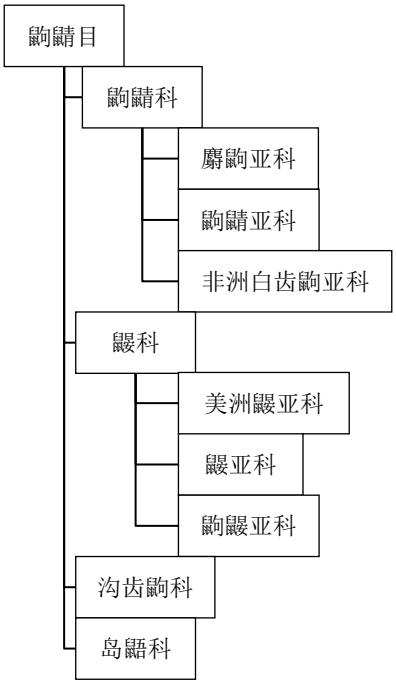

图 3　生物学上下位关系
Fig.3　hyponymy of biology

2.2 开放分类策略

词条在网络资源的描述中存在标签标注。在百度百科中，苹果的开放分类标签有植物，水果，蔷薇科，落叶乔木和苹果属。标签所表示的是此词条所在的类别，反之，词条正是对该类别的细化，这符合上下位关系的定义。因此类别标签为词条的上位关系词，即苹果为水果、蔷薇科等的下位词。

2.3 模版匹配策略

在模版匹配策略中，最主要是模版库的构建，再者就是后处理的问题。借鉴英文中上下位关系的 is-a 结构，库中的模板主要遵循的是类似的中文 is-a 结构[11]。

规则：ASeg{10}<是一|为一><种|个|名|篇|片|块|堆|群|批|章|节|环|部|次|步|颗|套|本|条|张|幅|款|代|缕|位|卷|册|只|双|件|台|门|棵|株|朵|根|头|尾|…>BSeg{10}

其中 ASeg{10}表示片段 ASeg 的长度为 10 个字，BSeg{10}表示片段 BSeg 的长度为 10 个字。认为片段 A{10}与片段 B{10}存在上下位关系。后处理使用词典匹配的方法的，获取片段中的词信息，即词 A 和词 B，从而确定词 A 为词 B 的下位词。模版匹配策略，在大量的词描述信息的基础上，结合模板库中的模板，通过模板匹配的方法自动识别上下位关系。

2.4 词汇细化策略

许多专有词汇都是有相关词汇构成的，例如勳章（勋章）包括大紫荊勳章、铜十字英勇勋章、维多利亚十字勋章和乔治勋章等，称为复合专有名词。复合专有名词构成规则可定义如下：

<复合专有名词>　::= <名词> | <复合词>

<复合词>　　::= <形容词><复合专有名词> | <名词><复合专有名词>

在大量的复合专有名词（人名、地名和机构名等）的基础上，使用反向最长匹配的方式，获取词汇的细化词汇（下位词）。例如名词勳章（勋章），按照词汇细化策略，获得所有勋章的复合专有名词（下位词），包括大紫荊勳章和铜十字英勇勋章等。

3　实验

实验数据采用的是 NLP&CC2012 语义关系识别任务中的评测标准数据。评测方法使用 NLP&CC2012 语义关系识别任务中的评测方法。

3.1 实验数据

同义关系数据集包含 10000 个词汇。数据来源包括普通词典、百科词条、叙词表等多种资源。词汇的词性包括普通名词、专有名词、动词和形容词。上下位关系评测数据集包括 10000 个词汇。数据来源包括普通词典、百科词条、叙词表等多种资源。词汇的词性包括普通名词和专有名词。

3.2 评测方法

评测采用三个指标：正确率（Precision），召回率（Recall）和 F 值（F-measure），并分别计算其微平均和宏平均值。

3.2.1 微平均

微平均以每个语义关系为一个计算单元，具体计算公式如下：

1) 正确率

表示发现的语义关系（同义或下位）中出现在标准结果中的比例，计算公式如下：

$$\text{正确率} = \frac{\text{发现的语义关系中出现在标准结果中的数量}}{\text{发现的语义关系总数}} \times 100\%$$

其中，词表中的每个词汇与发现的每个同义词（或下位词）为一条语义关系。发现的同义词之间的关系不计算在内。

2) 召回率

表示标准结果中被正确发现的语义关系比例，计算公式如下：

$$\text{召回率} = \frac{\text{发现的语义关系中出现在标准结果中的数量}}{\text{标准结果中的语义关系总数}} \times 100\%$$

3) F 值

是正确率和召回率的调和平均数，计算公式如下：

$$\text{F 值} = \frac{2 \times \text{正确率} \times \text{召回率}}{\text{正确率} + \text{召回率}} \times 100\%$$

3.2.2 宏平均

宏平均以每个词为一个计算单元，每个词的评价指标计算公式如下：

$$\text{词}i\text{的正确率} = \frac{\text{发现的词}i\text{的语义关系在标准结果中的数量}}{\text{发现的词}i\text{的语义关系数量}} \times 100\%$$

$$\text{词}i\text{的召回率} = \frac{\text{发现的词}i\text{的语义关系在标准结果中的数量}}{\text{标准结果中词}i\text{的语义关系总数}} \times 100\%$$

$$\text{词}i\text{的 F 值} = \frac{2 \times \text{词}i\text{的正确率} \times \text{词}i\text{的召回率}}{\text{词}i\text{的正确率} + \text{词}i\text{的召回率}} \times 100\%$$

宏平均值计算公式如下：

$$\text{正确率} = \frac{1}{N}\sum_{i} \text{词}i\text{的正确率} \times 100\%$$

$$\text{召回率} = \frac{1}{N}\sum_{i} \text{词}i\text{的召回率} \times 100\%$$

$$\text{F 值} = \frac{1}{N}\sum_{i} \text{词}i\text{的 F 值} \times 100\%$$

其中，N 为评测词汇总数。

3.3 实验结果

同义词识别实验结果如表 1 所示，上下位词识别实验结果如表 2 所示。表 1 中 SF1 表示原始词典，SF2

表示规则库，SF3 表示词集合。表 2 中 HF1 表示开放分类，HF2 表示词汇细化，HF3 表示学科分类，HF4 表示规则库。

同义关系识别实验结果（表 1）中，由于网络资源的噪声，本文以加入规则库和词集合的方法，从网络上大规模地获取同义关系，微平均的结果有所下降，但是无论是在准确率还是召回率宏平均都有所增加。这说明文本提出在识别同义关系中，加入词集合和规则的方法在宏观上是有效的。

上下位关系识别实验结果（表 2）中，考虑词的学科类别，专业名称的构成等。在微平均评测方法中，由于网络资源噪音的干扰，虽然准确率有所下降，但是召回率和 F 值都有显著的提高。学科分类和词汇细化都是以词的特点做不同的处理，通过实验结果证明是明显有效的。因此，在宏平均评测方法中，如上文所述，以词特点为依据的特殊处理，能够查全查准上下位关系，达到较好效果，从而在宏观上有显著提升。

表 1　同义关系实验结果

Table 1　The results of experiments on synonymy

	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
SF1	0.2476	0.3233	0.2525	0.3194	0.3685	0.3422
SF1+SF2	0.2773	0.3370	0.2687	0.3109	0.3740	0.3396
SF1+SF2+SF3	0.2876	0.3406	0.2737	0.3088	0.3753	0.3389

表 2　上下位关系实验结果

Table 2　The results of experiments on hyponymy

	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
HF1	0.4463	0.1724	0.2173	0.7082	0.2397	0.3581
HF1+HF2+HF3	0.6192	0.3573	0.3958	0.7057	0.4533	0.5520
HF1+HF2+HF3+HF4	0.6611	0.3776	0.4185	0.7043	0.4642	0.5596

3　总结和展望

本文提出的词语同义关系和上下位关系自动获取策略，主要利用词语的自身特点及其特殊的概念描述等特性，对不同特性的词汇采取不同策略处理。本文将其应用于语义关系自动识别系统中，并参与 NLP&CC2012 语义关系识别任务，取得较好效果。特别在上下位关系识别中取得突出效果。

网络噪音是不可避免的，因此我们将在今后的工作中，寻求更加有效的方法，减小网络噪音，在尽量不影响准确率的基础上提高 F 值。下一步，会在三个方面继续努力提高识别准确率。第一，丰富同义关系和上下位关系识别所用到的模板库。第二，考虑词的分类问题，对生物医学等专业领域的专业词汇做特殊处理。第三，在较好的上下位关系基础上，以一种搜索策略提高同义关系的识别。

参考文献

[1] Yildiz T, Yildirim S. Association Rule Based Acquisition of Hyponym and Hypernym Relation from a Turkish Corpus. Innovations in Intelligent Systems and Applications (INISTA), 2012: page 1-5

[2] Tian F, Yuan C, Ren F. Hyponym Extraction from the Web by Bootstrapping. IEEJ Transaction on Electronic Engineering. 2012, 7: 62-68

[3] Ritter A, Soderland S, Etzioni O. What Is This, Anyway: Automatic Hypernym Discovery. In Proceedings of AAAI-09 Spring Symposium on Learning by Reading and Learning to Read, 2009: 88–93

[4] Sang ETK, Hofmann K. Lexical Patterns or Dependency Patterns:Which Is Better for Hypernym Extraction?. In Proceedings

of CoNLL, 2009: 174-182

[5] Sang ETK. To Use a Treebank or Not –Which Is Better for Hypernym Extraction?. In Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories, 2009.

[6] 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. 中国图书馆学报, 2010, 1.

[7] Weale T, Brew C, Fosler E. Using the Wiktionary Graph Structure for Synonym Detection. Proceedings of the ACL-IJCNLP Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, 2009: 28-31

[8] Minkov E, Cohen WW. Graph Based Similarity Measures for Synonym Extraction from Parsed Text. Proceedings of the TextGraphs-7 Workshop at ACL, 2012: 20-24

[9] 吴云芳,石静,金澎. 基于图的同义词集自动获取方法. 计算机研究与发展, 2011, 4: 610-616.

[10] Kliegr T, Chandramouli K, Nemrava J, et al. Wikipedia as the Premiere Source for Targeted Hypernym Discovery. WBBT ECML08, 2008

[11] Liu L, Cao C, Wang H. Extracting hyponymic relations from Chinese free corpus. ACOS'06 Proceedings of the 5th WSEAS international conference on Applied computer science. 2006: 962-968