基于词典和 Web 的词汇关系抽取

范庆虎 昝红英 张坤丽 贾玉祥

郑州大学信息工程学院，河南 郑州 450001

Email:fanqinghude@163.com {iehyzan, ieklzhang, ieyxjia}@zzu.edu.cn

摘要 郑州大学参加了 NLP&&CC2012 中的词汇语义关系的任务，该任务包括两个子任务，子任务 1 是同义词抽取，采用中文概念词典、同义词词林等词典并且结合百度百科、有道翻译等 Web 方式进行抽取，子任务 2 是下位词抽取，采用中文概念词典并且结合百度百科标签、互动百科、维基百科等 Web 方式进行抽取。两个子任务均采用词典和基于 Web 的方法。其中子任务 1 提交的结果宏平均 F1 值取得了第二名，微平均 F1 值取得了第三名；任务 2 提交的结果宏平均 F1 值取得了第一名，微平均 F1 值取得了第二名，子任务 1 和子任务 2 的所有召回率均取得了第一名。

关键词 同义词；下位词；词汇关系；基于词典；基于 Web

Lexical Reiteration Extraction Based on Dictionary and Web

FAN Qinghu，ZAN Hongying，ZHANG Kunli，JIA Yuxiang

School of Information Engineering Zhengzhou University,Zhengzhou,Henan 450001

Email:fanqinghude@163.com {iehyzan,ieklzhang,ieyxjia}@zzu.edu.cn

Abstract Zhengzhou university had taken part in the task of NLP&&CC 2012. The task contains two sub-tasks. The sub-task one was synonym extraction. The method of the sub-task one that was based on dictionary that contains Chinese Concept Dictionary, Tongyici Cilin (Extended) and combined with the Web that contains Baidu Encyclopedia,Hudong Encyclopedia and Youdao translation. The method of the sub-task two that was based on Chinese Concept Dictionary and combined with Web that contains Baidu Encyclopedia Tag,Hudong Encyclopedia and Wikipedia. The macro-average of F1 value of the evaluation values of the sub-task one achieved the second, the micro-average of F1 value of sub-task one reached the third. The macro-average of F1 value of the evaluation values of the sub-task two reached the first, the micro-average of F1 value reached the second. Both the sub-task one and the sub-task two with Recall rate reached the first.

Key words synonym; hyponym; lexical reiteration; based on dictionary; based on Web

1 评测介绍

2012 年第一届自然语言处理与中文计算会议评测共设置 2 个任务。其中，任务 1 是面向中文微博的情感分析，任务 2 是中文词汇语义关系抽取。任务 2 中包括两个子任务（同义词发现和下位词发现）。下面对两个子任务的实现方法和实验过程做以详细介绍。

2 实现方法

2.1 同义词发现

同义词，指表达的意义相同或相近但表达形式不同的词汇。主要形式包括：别名/俗称，全称/简称，异形词，外来语译名，语义近似等类型。在对同义词进行抽取的过程中使用中文概念词典 (Chinese Concept Dictionary, CCD),以下称为CCD[1]，哈工大信息检索研究室同义词词林扩展版[2]等词典，并且结合百度百科[3]，有道翻译[4]等Web方式[5]对同义词进行抽

取。

2.1.1 使用 CCD 进行同义词抽取

CCD 是一个 WordNet 类型的汉英双语语义词典，从关系语义学的观点出发，以同义词集(Synset)定义概念(concept),在概念之间定义关系(relation)来描述语义,任务中主要通过对 CCD 中的同义词集进行同义词的抽取，一个词语的同义词集可能有多个，此时同义词集之间如果存在交集，要进行词语去重处理，然后将同义词集进行合并。

2.1.2 使用哈工大信息检索研究室同义词词林扩展版进行同义词抽取

《同义词词林(扩展板)》保留《同义词词林》原有的三层分类体系，并在此基础上对词语继续细分类，增加两层，得到最终的五层分类体系。唯一的代表词典中出现词语的编码如：Ba01A02= 物质 质 素，Cb02A01=东南西北 四方 。例子中的编码的第八位标记有 3 种，分别是“=”、“#”、“@”，“=”代表“相等”、“同义”。末尾的“#”代表“不等”、“同类”，属于相关词语。末尾的“@”代表“自我封闭”、“独立”，它在词典中既没有同义词，也没有相关词。任务中主要通过对《同义词词林》扩展版中表示“同义”的词语集进行抽取。

2.1.3 使用百度百科进行同义词抽取

对百度百科进行同义词抽取，分为如下步骤：

1)从百科中抽取纯文本

首先用网络爬虫将词语所对应百科词条的源代码文件进行抽取，然后对源代码进行过滤，转换为文本文档。将文本文档中的噪音包括“词条统计”、“贡献光荣榜”、“最新动态”、“百科消息”、“合作编辑者”、广告、导航等信息中的内容进行过滤。在进行抽取百度百科词条中为了区分不同行的内容，在抽取的每一行后面加上“#”。

2）从百科纯文本中提取特征词

通过对百度百科文本文档的分析，提取其中的一些同义词抽取特征词语，例如“别称”、“简称”、“别名”、“谥号”等一共找了 34 个特征词。

3)根据特征词抽取同义词

在抽取其同义词时可先定位到“别名”，然后截取其后面的内容直到句号，中间遇到顿号时此时判断为并列的别名，最后将顿号的并列词语进行拆分，分别得到不同的同义词。词语“讨论”的百度词条中的内容“近义词：议论#”在抽取其同义词时先定位到“近义词:”，然后截取其后面的内容直到井号，就是其同义词。在进行同义词词的抽取过程中，先对以上总结的 34 个同义词特征进行定位，然后开始抽取其后面的内容，需要对后面内容的标点符号进行定位，比如逗号、井号、括号、句号，取在内容中出现的离同义词特征最近的符号，然后从特征词开始到所取符号就为所找的同义词。

4)根据百科中同义词明显特征进行抽取

在百度百科的词条内容中，有比较明显的同义词特征，比如词语“北宁市”的百度百科词条内容中“北宁市和北宁是同义词，已合并”，先定位到“是同义词，已合并”，然后抽取其前面的内容，抽取的内容中“和”字两边的词语即为同义词。

5)对抽取处理的同义词进行噪音过滤

对从百科中抽取的结果进行噪音过滤，比如标点符号(逗号、句号)、乱码、同义词的长度大于 4 倍的词语长度等情况要进行过滤。

2.1.4 使用有道在线翻译进行同义词抽取

从 Web 中获取所需语料的有效方法就是人工构造合适的查询串，提交搜索引擎，检索到符合查询意图的结果。在本次任务中利用百度搜索引擎，在百度搜索引擎中输入合适的查询串，获取查询词语的英语，例如在百度搜索框中输入“阿肯色州 英语”或“阿肯色州 翻译”得到的搜索结果如图 1 所示：

图 1 同义词在百度中的英语

通过对搜索结果源代码的分析，提取文本，可以得到阿肯色州的英语为“Arkansas”。然后在有道在线翻译中输入“Arkansas”，可以得到在线翻译的翻译结果，如图 2 所示：

图 2 同义词英语的汉语翻译

从图 2 可以看到，翻译结果中的“网络释义”下面的标有“-”和“+”后面的标题文字即为所要抽取的结果。通过对翻译结果的源代码进行分析，首先获得其中的纯文本，在纯文本中获得“网络释义”下面的翻译结果。

2.2 下位词发现

下位词指其语义内涵包含在另一个词汇(称为上位词)内涵之中的词汇。即下位词是上位词的一个特殊实例。例如“水果”的下位词包括“苹果”、“梨”、“菠萝”等。“国家”的下位词包括“中国”、“美国”、“日本”等。下位词的发现，本次任务主要使用CCD、百度百科标签、互动百科[6]、维基百科[7]等词典和Web方式进行下位词抽取。最后将合并后的结果进行噪音的过滤。

2.2.1 使用 CCD 进行下位词的抽取

CCD 主要的语义关系有同义关系、反义关系、下位关系、整体部分关系等，在进行同义词抽取过程中，主要使用 CCD 的同义关系进行同义词的抽取，在进行下位词的抽取中主要使用 CCD 的下位关系对下位词的抽取。如果词语的下位词对应多个下位词集时，要进行下位词去重，将没有交集的下位词集进行合并。

2.2.2 使用百度百科标签进行下位词的抽取

从百度百科词条中抽取出的内容中有隐含的上下位关系的信息，比如“360 安全卫士”词条的文本内容中的“开放分类”，如图 3 所示：

图 3 百科开放分类

图 3 中的“IT”、“品牌”、“安全软件”、“360 软件”等词语即为“360 安全卫士”的上位词，利用此信息可以从百度搜索框中输入“tag：安全软件”，得到的结果如图 4 所示：

图 4 百科标签

从图 4 可以看到“安全软件”的下位关系词有“360 安全卫士”、“小红伞”和“金山卫士”等词语，在此只抽取“tag:安全软件”得到的第一页搜索结果，通过网络爬虫对第一页搜索结果的源代码进行抽取，从源代码内容中抽取各个搜索结果的标题，即抽取如上图中的“360 安全卫士”、“小红伞”、“金山卫士”等标题，抽取得到的标题即为“安全软件”的下位词。

2.2.3 使用互动百科进行下位词的抽取

从互动百科词条中抽取的文本内容隐含有上下位关系，例如在互动百科中输入“安全软件”得到的分类如图 5 所示：

图 5 互动百科分类

从图 5 可以看到“安全软件”的分类词条有“360”、“Comodo”、“小红伞”、“金山卫士”

等，在此也只抽取搜索结果中的第一页，和从百度百科中获取下位词的方法类似，通过抽取源代码，从源代码中抽取搜索结果的每个标题，即“360”、“Comodo”、“小红伞”、“金山卫士”等标题，抽取的标题即为“安全软件”的下位词。

2.2.4 使用维基百科进行下位词的抽取

从维基百科词条中抽取的内容同义隐含有上下位关系，通过使用URL进行下位词的发现，例如要发现“安全软件”的下位词，首先将“安全软件”转换为对应的UTF-8 编码即为“%E5%AE%89%E5%85%A8%E8%BD%AF%E4%BB%B6”，构造的完整的URL为“http://zh.wikipedia.org/wiki/Category: %E5%AE%89%E5%85%A8%E8%BD%AF%E4%BB%B6”，使用JAVA的URL类进行搜索结果源代码的抽取，搜索的结果如图 6 所示：

图 6 维基百科分类

从图 6 可以看到“安全软件”的子分类有 4 个子分类，此时也只抽取搜索结果中的第一页内容，抽取搜索结果的源代码，从源代码中抽取带有链接的标题，即“加密软件”、“反间谍软件”、“杀毒软件”、“防火墙软件”等标题，抽取的标题即为“安全软件”的下位词。

2.2.5 使用百度相关搜索进行下位词的抽取

从百度相关搜索[8]中进行下位词的抽取，通过使用URL进行下位词的发现，例如在百度中搜索“安全软件”，相关搜索的结果如图 7 所示：

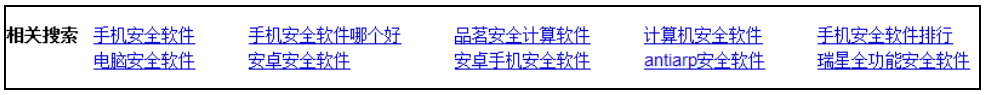

图 7 百度相关搜索

从图 7 中可以看到相关搜索中如果存在不符合下位词的词语要进行过滤，例如“手机安全软件哪个好”需要此类的词语(含有比如：哪，怎样)等进行过滤。

2.2.6 结果合并和过滤噪音

将以上的各种结果进行合并，并且进行过滤，比如词语中含有怎么、咋样、什么、多少、价格、批发、回收等特征词时要进行过滤。最后将每一个词语的同义词集中重复的词语只保留一份。

3 评测指标

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

3.1 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：

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。

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：

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

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 宏平均

宏平均以每个词为一个计算单位，每个词的评价指标计算公式如下：

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

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

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

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。

4 实验

4.1 实验数据

本文采用的语料是 NLP&&CC 2012 的词汇语义关系的评测数据集作为本实验的训练语料，本语料包含有同义关系词表和上下位关系词表，其中同义关系词表中有 9455 个词语，上下位关系词表中有 9286 个词语。对所有结果求并集，并进行人工标注，最后生成 778 个同义词语的标准答案和 256 个下位词语的标准答案，据此进行评测。

4.2 实验结果

表 1 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
ZZU1	0.2975	0.6423	0.3598	0.2530	0.6792	0.3687
ZZU2	0.3256	0.6961	0.3927	0.2541	0.7072	0.3738
MAX	0.3588	0.6961	0.3984	0.3025	0.7072	0.4106

从表格 1 数据可观察到，宏平均召回率和微平均召回率都取得了最大值，但其宏平均准确率、微平均准确率、宏平均 F1 值都要低于最大值，其原因主要在于对同义词进行抽取过程中抽取特征词不完善以及抽取出来的词语出现一些噪音，比如在进行同义词抽取中特征词“释义”后面会出现具体的解释而不是同义词，例如“美景”的百科内容中释义后面的内容是“优美的景色”、“美丽的景色”等。此外在进行特征词抽取过程中没有基于上下位内容边界进行抽取，方法是先找特征词的位置，然后抽取其后面的以逗号、冒号或句号结束的内容，此方法会降低准确率，例如“美景”百科中有“故宫，又名紫禁城，位于北京市中心，是世界上规模最大的宫殿建筑群。今天人们称它为故宫，意为过去的皇宫。”，此时根据特征词“又名”抽取的内容是“紫禁城”，根据特征词“称它为”抽取的内容是“故宫”。以上抽取的词语明显不是目标词的同义词。

表 2 下位关系评测结果

	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
ZZU1	0.5603	0.3321	0.3742	0.6492	0.3518	0.4563
ZZU2	0.6119	0.5988	0.5605	0.6233	0.5045	0.5576
MAX	0.6119	0.5988	0.5605	0.7827	0.5045	0.5596

从表格 2 数据可观察到，宏平均准确率、宏平均召回率、宏平均 F1 值、微平均召回率都取得了最大值，但微平均准确率、微平均 F1 值都要低于最大值，其主要原因在于百度百科标签、互动百科、维基百科等内容都是由不同背景知识的人而非语言专家进行的发布，所以在准确性不是很高。

5 结论

主要介绍了郑州大学自然语言处理实验室在NLP&&CC 2012任务中语义关系所做的工作。包括同义词关系的抽取和上下位关系的抽取。采用的是基于 WEB 和词典的语义关系抽取，其中同义关系词部分的结果取得了较好的效果，但仍存在不足，由于在进行同义词的抽取中特征词的不完备性，下一步我们应该对特征词进一步完善，同时在下一步的抽取过程中考虑上下文内容的同义词抽取。

参考文献

[1]于江生，俞士汶.中文信息概念的结构[J].中文信息学报，2002，16(4)：12-20.

[2]哈尔滨工业大学.《同义词词林》扩展版 [EB/OL].(2010-10-15)[2012-07-24].
http://ir.hit.edu.cn/phpwebsite/index.php?module=pagemaster&PAGE_user_op=view_page&PAGE_id=162.

[3] 百度百科[EB/OL].(2012-07-21)[2012-07-21]. http://baike.baidu.com.

[4] 有道词典[EB/OL].(2012-07-23)[2012-07-23]. http://dict.youdao.com.

[5] 曹馨宇,曹存根.从 Web 获取部分整体关系语料的方法[J].中文信息学报, 2012, 25(5)：17-23.

[6] 互动百科[EB/OL].(2012-07-25)[2012-07-25]. http://www.hudong.com.

[7] 维基百科[EB/OL].(2012-07-28)[2012-07-28].http://www.wikipedia.org.

[8] 百度[EB/OL].(2012-07-29)[2012-07-29].http://www.baidu.com.