面向开放异构知识库的词汇同义关系学习

刘焱灵，吉阳生，顾翀†，崔首领，贾江涛

华为技术有限公司，深圳市龙岗区坂田华为基地， 518219

† 通讯作者，E-mail: guchong@huawei.com

摘要 词汇同义关系识别在文本信息管理、信息检索和自然语言处理等领域中扮演重要作用。识别词汇同义关系的方法主要有两类：基于结构化知识库的匹配方法和基于在线/离线语料库的统计学习方法。基于知识库的方法对词汇关系的整理需要很高的专业技能和昂贵的时间开销。基于语料库的统计方法从大规模的文本语料中学习词汇同义关系，但是其习得的语义关系准确性尚不能令人满意。面向来自因特网的开放异构知识库，本文提出一种从其中提取同义关系的统计方法，可以进一步扩展和补充结构化同义词典知识库；在取得较高准确性的同时，提高词典知识库的同义关系覆盖率。在 CCF 的开放性语义关系评测中，本文提出的方法取得了宏平均 F1 值第三名和微平均 F1 值第二名的成绩。

关键词 词汇同义关系；异构知识库；统计学习；语义挖掘。

Semantic Similarity between Words Learned from Heterogeneous Knowledge Bases

Yiling Liu, Yangsheng Ji, Chong Gu†, Shouling Cui, Jiangtao Jia

Huawei Technologies, Shenzhen, China 518129;

†Corresponding Author, E-mail: guchong@huawei.com

Abstract Recognization of semantic similarity between words plays an important role in text information management, information retrieval and natural language processing. There are two major approaches to recognizing semantic similarity, among which one way is extracting similarity representation based on structured semantic dictionary, while the other way is learning the semantic similarity from a large corpus. Building a semantic dictionary is a time consuming task which also requires much expertise, while the learning method cannot extract precise similarity between words. This paper proposes to expand the semantic dictionary of word similarity by learning from heterogeneous knowledge bases statistically. This method can not only expand the semantic dictionary from the open knowledge bases, but also achieve accurate semantic similarity. In the evaluation of semantic relatedness held by CCF, the proposed system ranked the 3rd place according to the macro average F1 and the 2nd place according to the micro average F1.

Keywords semantic similarity, heterogeneous knowledge bases, statistical learning, semantic mining.

由于自然语言描述事物的多样性，使得同一个事物或概念有多种描述方式。在文本信息中，同一个概念的不同描述形式被视为词汇的同义关系。不同描述方式使得相同概念的同义词汇之间存在“语义鸿沟”。词汇同义现象对于文本挖掘、信息检索和自然语言处理等任务有重要影响。在倾向性文本挖掘中，某一产品的属性可以用不同的词汇进行描述，对该属性的用户意见摘要和倾向性分类，都要求能够对该属性的所有描述形式进行分析和挖掘[1]。在信息检索中，用户输入的检索词是待查询概念的一种描述，而实际存储

的文本中对概念可能使用另一种描述，不同的描述可能导致信息失配[2,3]。在自然语言处理中，机器翻译系统进行训练之后，可以适应概念的某种描述，但是其他的描述形式可能导致翻译系统工作异常[4]。上述任务在日常生活中有着广泛的应用，但是同义词汇引起的语义鸿沟制约着上述应用，因此研究词汇同义关系成为一项重要的研究课题。

目前，词汇同义关系的研究方法主要可以分为两类：基于结构化知识库的匹配方法[5,6,7,8,9,10]和基于语料库的统计学习方法[11,12,13,14,15,16,19]。基于结构化知识库的匹配方法，利用语言学专家整理的同义词词典和其中的语义结构信息，基于词典的语义结构定义所匹配词汇的相似度量。这类方法对于词汇间相似性的度量比较准确，但是词典的构建工作需要耗费昂贵的时间和很强的专业技术，同时该类方法不易进行扩展。基于语料库的统计学习方法，依据大规模语料的统计信息定义词汇相似度，抑或挖掘语料中的词汇同义关系模板。基于统计方法的同义关系识别方法易扩展，但是得到的词汇同义关系往往不够准确。

针对上述两种方法所面临的问题，本文提出一种从开放异构知识库中提取同义关系的统计方法，可以进一步扩展和补充结构化同义词典知识库；在取得较高准确性的同时，提高词典知识库的同义关系覆盖率。在 CCF 的开放性语义关系评测中，本文提出的方法取得了宏平均 F1 值排名第三和微平均 F1 值排名第二的成绩。

1 面向开放异构知识库的词汇同义关系学习

本系统所面向的异构知识库包括结构化的同义词典，和互联网上的开放文本资源，包括豆瓣、百度百科、百度词典等异构的知识源[17]。异构知识源的融合主要通过在各知识源中挖掘同义关系模式，将非结构化的同义信息转化为与同义词典一致的结构化信息，并经过合并、去重、清洗等步骤，将从互联网上的异构知识源获取的同义关系融入结构化的同义词典中，扩展同义词典的同时，保证了较高的准确性。

1.1 结构化的同义词典

由于哈工大义典收录的词汇较丰富，本文选定义典作为系统使用的结构化同义词典。本系统采用的结构化同义词典定义如下：

定义 1(结构化同义词典）同义词典用 $D = D_S \cup D_R$ 表示，其中 $D_S = \{(w_i, SID_i, S(w_i)) \mid i = 1,...,\mid D_S \mid\}$ 表示同义词义项集合，w_i 表示义项目标词，SID_i 表示同义词概念的 ID 号，$S(w_i)$ 表示词汇 w_i 的同义词集合；$D_R = \{(w_j, RID_j, R(w_j)) \mid j = 1,...,\mid D_R \mid\}$ 表示意义相关词的义项集合，w_j 表示义项的目标词，RID_j 表示相关词概念的 ID 号，$R(w_j)$ 表示词汇 w_j 的相关词集合。

相关词是指与目标词汇的概念相近的词汇，同义词是指与目标词汇的概念等价的词汇，相关词的涵义比同义词更加宽泛。系统中采用的同义词典是同义词义项和相关词义项的并集。

1.2 面向异构知识库抽取同义关系模板

在百度百科和豆瓣等开放知识库中，对于部分词汇和人名、机构名，网页 HTML 源语言对词汇同义关系均有显式的结构标记，这些结构标记对词汇同义关系挖掘的帮助是可靠和可信的；在百度百科和百度词典等知识库中，非结构化的文本信息中同样蕴藏着词汇同义关系的模板。无论是百度词典，还是其他知识库，同义关系模板并不是显而易见的，需要应用统计学习方法进行挖掘。

1.2.1 豆瓣 HTML 模板识别

对豆瓣中词汇同义关系的 HTML 模板挖掘，首先在豆瓣查询框中查询目标词汇，得到返回的网页内容；然后在返回的网页内容中查询目标词汇的同义词；如果查找到，则记录包含同义词的字符串，该字符串被前后 HTML 标签分割；记录该字符串中包含词汇的频率。当同义词典中所有的词汇都统计完毕，包含同义词的字符串中频率最高的词汇或者标签，可以作为同义词的 HTML 模板标签。

记录查询所得字符串中包含的词汇频率，需要对字符串“分词”进而得到其中包含的词汇。由于同义关系的 HTML 模板识别目标是提取网页中的结构化标注信息，因此查询所得字符串中的原始标注信息需要

保留。基于此，字符串中的词汇主要是由空格，制表符等空白符自然分隔而成。豆瓣 HTML 同义模板识别的算法见表 1.

表 1　豆瓣中词汇同义关系 HTML 模板识别算法

Table 1 Douban HTML Pattern Mining for Semantic Similarity Identification

输入：同义词典 *D* 中的查询目标词 *w*，同义词典 *D* 中同义词集合 *S(w)*
输出：模板候选词汇集合 *pattern(w)*
步骤：
1. 将查询词 *w* 输入豆瓣的搜索框，并得到返回页面的 HTML 内容 *page(w)*；
2. 初始化模板候选词汇集合 *pattern(w)*为空；
3. 对于每个 $v \in S(w)$
4. 检索 *v* 是否在 *page(w)*中，如果检索不到，转步骤 7；
5. 抽取 *v* 在 *page(w)*的数据块，并且删除其中的 HTML 标签，得到字符串 *str(v)*；
6. 将字符串 *str(v)*中的内容按照空格、制表符等空白符分割成一组词汇加入模板候选词集合 *pattern(w)*;
7. 转步骤 3；
8. 输出 *pattern(w).*

在得到了每个词汇的模板候选词汇*pattern(w)*之后，将同义词典*D*中所有词汇对应的*pattern(w)*进行汇总，其中出现频率较高的模板候选词作为最终的同义关系模板词*pattern(D)*。在豆瓣中，计算一个新词w_{new}的同义词时，在w_{new}的返回网页HTML中，查询*pattern(D)*中每个同义关系模板词；如果查找到，则提取相应的字符串，并将其中的模板词汇和标点符号删除，则剩下的字符为w_{new}的同义词。

1.2.2　百度百科 HTML 模板识别

百度百科中收录的地名、人名和电影名等义项中，提供了大量的同义信息，而且相当一部分以信息表的形式存在。信息表的格式如下：

中文名：	关羽	**出生日期：**	约汉桓帝延熹三年六月
外文名：	GuanYu	**逝世日期：**	建安二十四年冬（西元219年冬）
别名：	关长生，关云长，关公	**职业：**	将领
民族：	汉族	**主要成就：**	阵斩颜良、水淹七军
出生地：	河东郡解县（今山西运城）	**官职：**	前将军

图 1　百度百科信息表

从百科的信息表抽取 HTML 模板词汇的方法与从豆瓣中抽取 HTML 模板词汇的方法类似，不同之处是首先需要识别百度百科中的信息表。百科信息表可以通过网页的 HTML 内容匹配正则表达式来识别。在识别了信息表之后，从信息表的 HTML 表示中抽取同义关系模板词汇，与表 1 中的算法类似。从百度百科中识别 HTML 同义关系模板的算法如表 2 所示。

表 2　百度百科词汇同义关系 HTML 模板识别算法

Table 2 Baike HTML Pattern Mining for Semantic Similarity Identification

输入：同义词典 *D* 中的查询目标词 *w*，同义词典 *D* 中同义词集合 *S(w)*
输出：模板候选词汇集合 *pattern(w)*
步骤：
1. 将查询词 *w* 输入百度百科搜索框，并得到返回页面的 HTML 内容 *page(w)*；
2. 从 *page(w)*中识别百科信息表 *table(w)*；如果识别为空，转步骤 9；
3. 初始化模板候选词汇集合 *pattern(w)*为空；

4. 对于每个 $v \in S(w)$

5. 检索 *v* 是否在 *table(w)*中，如果检索不到，转步骤 8；
6. 抽取 *v* 在 *table(w)*中的数据块，并且删除其中的 HTML 标签，得到字符串 *str(v)*；
7. 将字符串 *str(v)*中的内容按照空格、制表符等空白符分割成一组词汇加入模板候选词集合 *pattern(w)*;
8. 转步骤 4；
9. 输出 *pattern(w).*

在百度百科的同义关系模板识别算法中，步骤 2 是识别百科信息表，其余步骤与豆瓣的同义模板识别算法类似。在百科信息表中，计算新词w_{new}的同义词时，首先需要得到w_{new}的HTML网页内容；然后识别其中是否含有信息表；如果含有信息表，在信息表中查询同义词模板*pattern(D)*中每个同义关系模板词；如果查找到，则提取相应的字符串，并识别其中可能含有的多个同义词。

1.2.3 非结构化文本模板识别

词汇同义关系在因特网上的开放知识库中，如百度百科、百度词典等，更多地以非结构化文本的形式存在。如果能从非结构化的文本中挖掘同义关系模板，将极大地丰富词汇同义关系的知识源，进一步扩展已有的同义词词典资源。

在非结构化文本中识别词汇同义关系模板并非显而易见，然而在无结构的文本中同义关系的出现并非无章可循，与 HTML 模板的识别亦有相似之处。相似之处在于，在豆瓣和百度百科的 HTML 内容和文本内容中，同义关系的出现都伴随着某些指示标记。不同之处在与，文本中使用的指示标记是不受限的自然语言，类型丰富多变，识别的难度较大；而百度百科等知识库的 HTML 内容中，同义关系所使用的指示标记是明确而且受限的，识别相对容易。因此，非结构化文本中挖掘出的同义关系模板，可信度和准确性相对 HTML 同义关系模板而言较低。

从非结构化文本中识别同义关系模板，首先需要从目标词汇 *w* 的网页文本内容 *text(w)*中提取目标词和同义词 *v* 的上下文片段，同时对该上下文片段进行分词，因为同义关系的指示标记主要是自然语言中的词汇。然后提取同义词的上下文片段 *context(w,v)*，*context(w,v)*是指(1)从目标词开始到同义词为止的字符串，(2)包含同义词且前后被标点符号所分割的字符串，并且是(1)和(2)中较短的字符串。提取同义词上下文片段中的词语，并形成 *w* 的同义模板示例 *template(w,v).* 同义词典中所有词汇的同义词模板示例组成模板示例集合$T = \{template(w,v) \mid v \in S(w), w \in D_s\}$，并对模板集合 *T* 应用频繁项集挖掘算法，计算出 *T* 中最具代表性的同义关系模板 $P = FrequentPattern(T)$。具体算法步骤如表 3 所示。

表 3 非结构化文本中同义关系模板挖掘

Table 3 Pattern Mining for Semantic Similarity Identification from Text

输入：同义词典 *D*

输出：同义关系模板 $P = FrequentPattern(T)$

步骤：

1. 初始化同义模板示例候选集合 *T* 为空；
2. 对于同义词典 *D* 中每个词语 *w*
3. 初始化模板示例候选集合 *T(w)*为空；
4. 将查询词 *w* 输入相应知识库搜索框，并得到返回页面的 HTML 内容 *page(w)*；
5. 从 *page(w)*中识别文本内容 *text(w)*；如果识别为空，转步骤 13；
6. 对于每个 $v \in S(w)$
7. 检索 *v* 是否在 *text(w)*中，如果检索不到，转步骤 11；

8. 抽取同义词 *v* 在 *text(w)*中的上下文片段 *context(w,v)*;
9. 将 *context(w,v)*中的字符串进行分词，并得到同义模板示例 *template(w,v)*;
10. $T(w) = T(w) \cup \{template(w,v)\}$
11. 转步骤 6;
12. $T = T \cup T(w)$;
13. 转步骤 2;
14. 对 T 应用频繁项集挖掘算法，并计算 $P = FrequentPattern(T)$ 得到最具代表性的同义关系模板.

在非结构化文本中挖掘同义关系模板，基本流程是步骤 6 到步骤 11，计算每一组同义词的同义模板；在得到字典中所有词汇的同义模板集合 T 之后运用频繁项集挖掘算法得到最有代表性的同义模板 P. 在得到 P 之前，需要做一些预处理工作。在词语释义类型的文本中，如果义项解释都是古文，则该义项是古文的义项，这样的情况需要识别以便于决定是否在系统中接收该同义词。在得到 P 之后，还需要进行后处理工作。P 中模板词汇已经失去了词汇间的排序信息，需要从文本中恢复词语顺序；模板词汇中的部分字符需要过滤，如“使”、“的”等。

1.2.4 模板可信度和权重赋值

从豆瓣和百度百科中抽取的 HTML 模板，由于标注信息比较确定，该类型的模板准确性和可靠性都较高；从百度百科和词典的纯文本内容中抽取的同义关系模板，由于标注信息的自由灵活，模板的可靠性降低。因此，应用上述模板对词典中部分词汇进行同义词挖掘，将得到的小规模结果进行人工标注，根据准确率的高低，对规则进行权重赋值。赋值情况部分示例如表 4 所示。

表 4 同义关系模板的权重赋值

Table 4 Importance of Patterns for Semantic Similarity

模板	权值
HTML 模板： 【近义词】 又名： 简称： 别名：	1.0
文本模板： 也说 亦称 俗称 的别名	0.9
文本模板： 谓 泛称 亦称(之称	0.8
文本模板： 犹	0.7
文本模板： 指 的对称	0.6

模板权值可以用来对目标词汇的候选同义词进行排序，根据生成该词汇的模板权重来决定该词汇的权重，从而实现候选词语的排序。

1.3 异构知识库同义关系的融合

在成功抽取豆瓣、百度百科中的 HTML 同义关系模板和百度百科、百度词典中的文本同义关系模板后，借助上述模板，可以持续挖掘新的词汇同义关系，扩充已有的同义词典。由于不同的模板都可以将词汇同义关系输出为统一规定的结构化形式，因此异构知识库中的同义关系可以顺利的融合。

首先，对百度词典、百度百科、豆瓣等互联网开放知识库中用爬虫下载的网页，通过上述模板挖掘出若干同义词等价关系，并将等价关系合并形成面向开放知识库的词典

$D_{open}=\{S_i^o=\{w_{i,1},...,w_{i,t_i}\}\,|\,i=1,...,k\}$，其中 $w_{i,j}$ 在 S_i^o 中的排序*j*按照生成该词汇的同义模板权重进行排序。然后，将结构化同义词典*D*与面向开放知识库的同义词典D_{open}进行融合。具体融合过程如表 5 所示。

表 5　异构同义词库的融合算法

Table 5 Heterogeneous Synonyms Fusion Algorithm

输入：结构化同义词典*D*和面向开放知识库的同义词典D_{open}
输出：融合的结构化同义词典 *D*

步骤：

1. 对于每个 $S_i^o \in D_{open}$；
2. 匹配同义词集合 $R=\phi$；
3. 对于每个 $w_{i,j} \in S_i^o$；
4. 对于每个 $(w_k, SID_k, S(w_k)) \in D$；
5. 如果 $w_{i,j} \in S(w_k)$，记录 $R=R\cup\{(w_k, SID_k, S(w_k))\}$；
6. 转步骤 4；
7. 转步骤 3；
8. 如果 $|R|==0$，删除 S_i^o 中排名靠后的同义词，更新 $D=D\cup\{(w_{i,1}, D.total\_sid+1, S_i^o)\}$；
9. 如果 $|R|==1$，更新同义词典 *D* 中的义项 $(w_k, SID_k, S(w_k))=(w_k, SID_k, S(w_k)\cup S_i^o)$；
10. 如果 $|R|>1$，说明 S_i^o 同义集合中有多个概念和多义词，人工处理 R 和 D 的融合；
11. 转步骤 1；
12. 输出融合更新的 D.

如果基于开放知识库的同义词典D_{open}中，同义词项的集合没有歧义，则该义项可以直接与结构化同义词典*D*合并；如果D_{open}中的义项有多义的情况出现，则D_{open}与*D*的合并需要人工干预。

2　基于异构知识库的同义关系计算

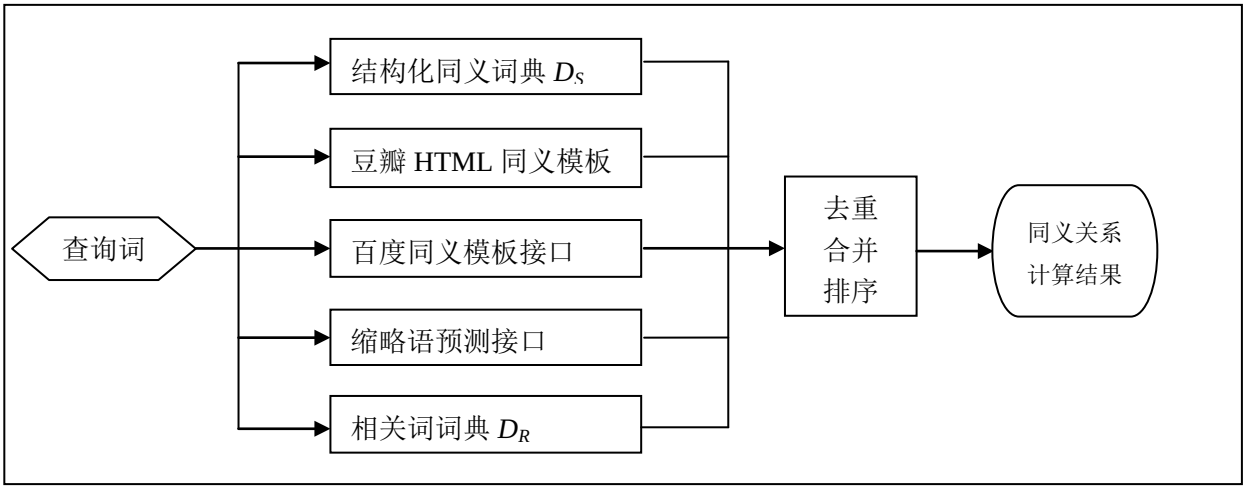

图 2　基于异构知识库的同义关系计算流程

在得到了经过异构知识库融合和扩展的同义词典 D 以及词汇同义关系挖掘模板之后，本系统可以对输入的查询词进行同义关系计算。同义关系计算流程如图 2 所示。由于缩略语是目标词的一类特殊同义词汇，而且缩略语的变化形式多样，百度百科等知识库中很难做到尽数收录[18,20]。另外，组织机构名称一般都有其对应的缩略语，本系统针对组织机构名和其他常用缩略语的全称形式进行处理，得到其缩略语结果[21]。

3 实验评估

3.1 实验数据

在系统构建中采用的**异构知识库**资源包括哈工大整理的义典、百度词典、百度百科、豆瓣等互联网上的开放资源。

系统的**评价数据集**采用 2012 年 CCF 自然语言处理与中文计算会议（NLP&CC 2012）中文微博情感分析&词汇语义关系抽取评测所提供的数据，待检索词汇数量为 9455 个。该数据集在本文中用 NLPCC2012 来指代。

3.2 词汇同义关系模板和扩展的同义词典

词汇同义关系模板有两类：HTML 同义关系模板和文本同义关系模板。HTML 同义关系模板从网页的 HTML 内容中，通过特定的同义关系标注抽取；文本同义关系模板从网页的纯文本内容中，通过语言描述的频繁模式抽取。表 6 中列出了部分 HTML 和文本模板。

表 6　同义关系模板示例

Table 6 Templates for Semantic Similarity

HTML 同义关系模板	"中文名：", "别名：", "本名：", "别名昵称：","封爵：", "谥号：", "爵位：", "别称：", "公司名称：", "庙号：", "中文名称：", "其它译名：","粤语名 ：", "简称：", "外文名：", "近义词：", "同音词：", "中文学名：", "定义："
文本同义关系模板	简称[“《为：](.*?)[”》)）(，、；。]|简称(.*?)[，.))(，、；。] 全称应为(.*?)[，。]|全称[是|为|、](.*?)[)(，。]|全称(.*?)[)(，。A-Z]

扩展的同义词典规模在表 7 中展示。从表中可以看出，经过异构数据库中的同义关系扩展以后，同义词典的规模具有很大应用前景。相关词汇的扩充通过从百度百科中直接抽取相关词条来扩充。

表 7　扩展后的同义词典规模

Table 7 Size of Expanded Synonym Dictionary

同义词汇数量	同义义项数量	相关词汇数量	相关义项数量
102882	17345	38096	3907

3.3 系统评估结果

为了评估本文提出方案的效果，我们设计了两套系统 System1 和 System2，对 NLPCC2012 数据中列出的词汇进行同义关系计算，并对结果的正确率、召回率和 F1 值进行评估。其工作流程为：

(1) System1：如果同义词典D_S、豆瓣HTML模板或者缩略语预测接口可以计算得到同义词，则同义计算的结果为上述三个接口输出的并集；否则，输出百度同义模板接口同义关系的计算结果；如果百度百科模板接口没有结果输出，则输出相关词词典D_R中的查询结果。

(2) System2: 如果同义词典D_S、豆瓣HTML模板、百度同义模板接口或者缩略语预测接口可以计算得到同义词，则同义计算结果为上述四个接口输出的并集；否则，输出相关词词典D_R中的查询结果。

System1 和 System2 的差别主要在于百度同义模板所计算得到的同义词，而百度同义模板中文本同义模板的数量较多，覆盖范围较大。所以两个系统的差别可以体现出文本同义模板对同义关系计算的影响。

System1 和 System2 在数据集 NLPCC2012 上的评估结果在表 8 中列出。

表 8　参评系统在 NLPCC2012 数据集上的实验结果

Table 8 Evaluation of participant systems on the data set NLPCC2012

	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
System1	0.3641	0.5176	**0.3664**	0.2754	0.5829	**0.3740**
System2	0.3305	0.5506	0.3635	0.2615	0.6102	0.3662
南京师范大学	0.3588	0.6041	**0.3968**	0.3025	0.6358	**0.4100**
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**
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
华侨大学	0.0382	0.0111	0.0151	0.2996	0.0115	0.0221
哈尔滨工业大学	0.3225	0.3885	0.2842	0.2303	0.3676	0.2832

评测结果中，本文提出的 System1 系统在宏平均 F1 值指标上排名第 3，在微平均 F1 值指标上排名第 2；System2 的结果稍差，但是仍然取得了较好的名次。从 System1 和 System2 的对比中看出，引入百度百科的文本同义关系模板，对系统计算效果有小幅度下降的影响；但是，文本同义关系模板可以使得更多的同义关系被挖掘，从而融合到同义词典中。

4　总结和展望

词汇同义关系计算在文本信息管理、信息检索和自然语言处理等领域中是一个重要的研究课题。面向来自因特网的开放异构知识库，本文提出一种从中提取同义关系的统计方法，通过从异构内容中挖掘同义关系模板，进而可以利用该模板从不同的知识库中计算同义词。本文提出的方法，不仅能进一步扩展和补充结构化同义词典知识库，还可以在取得较高准确性的同时，提高词典知识库的同义关系覆盖率。在 CCF 的开放性语义关系评测中，本文提出的方法取得了宏平均 F1 值第三名和微平均 F1 值第二名的成绩。

参考文献

[1] 王素格, 李德玉, 魏英杰, 宋晓雷。基于同义词的词汇情感倾向判别方法。中文信息学报，2009 年第 5 期。

[2] 曹晶。同义词挖掘及其在概念信息检索系统中的应用研究。硕士学位论文，东北大学。

[3] 陆勇，侯汉清。用于信息检索的同义词自动识别及其进展。南京农业大学学报(社会科学版), 2004 年第 3 期。

[4] Zhifei Li, David Yarowsky: Unsupervised Translation Induction for Chinese Abbreviations using Monolingual Corpora. ACL 2008: 425-433.

[5] Alexander Budanitsky, Graeme Hirst. Semantic distance in WordNet: An experimental, application- oriented evaluation of five measures. Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, USA ,2001.

[6] 梅立军，周强，臧路，陈祖舜。知网与同义词词林的信息融合研究。中文信息学报，2005 年第 1 期。

[7] 章成志。一种基于语义体系的同义词识别研究。淮阴工学院学报，2004 年第 1 期。

[8] 田久乐，赵蔚。基于同义词词林的词语相似度计算方法。吉林大学学报(信息科学版)，2010 年第 6 期。

[9] Yuhua Li, Zuhair A. Bandar, David Mclean. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. IEEE Trans. on Knowl. and Data Eng., Piscataway, NJ, USA, 2003, 4: 871-882.

[10] Dekang Lin. An Information-Theoretic Definition of Similarity. International Conference on Machine Learning, Morgan Kaufmann 1998: 296-304.

[11] Mehmet Ali Salahli. An Approach for Measuing Semantic Relatedness Between Words via Related Terms. Mathematical and Computational Applications, 2009, 1: 55-63.

[12] Patrick Pantel, Dekang Lin. Discovering Word Senses from Text. Proceedings of the eighth ACM international conference on Knowledge discovery and data mining, New York, NY, USA, 2002: 613-619.
[13] Danushka Bollegala, Yutaka Matsuo, Mitsuru Ishizuka. Measuring Semantic Similarity between words via Search Engines. Proceedings of the 16th international conference on World Wide Web New York, NY, USA, 2007: 757-766.
[14] P.D. Turney. Mining the Web for Synonyms: PMI-IR vs LSA on TOEFL. Proceedings of the 12th European Conference on Machine Learning, 2001: 491-502.
[15] 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 年 02 期。
[16] 宋宇轩。基于搜索日志和点击日志的同义词挖掘的研究和实现。北京交通大学硕士学位论文，2011.
[17] 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 年 1 期。
[18] 谢丽星，孙茂松，佟子健，王灿辉。基于用户查询日志和锚文字的汉语缩略语识别。中国计算机语言学研究前沿进展（2007-2009），2009.
[19] 石静，邱立坤，王菲，吴云芳。相似词获取的集成方法。中国计算语言学研究前沿进展（2009-2011），2011.
[20] 王厚峰。汉语缩略语自动处理研究现状。中文信息学报，2011 年 5 期。
[21] Jing-shin Chang. A Preliminary Study on Probabilistic Models for Chinese Abbreviations. Proceedings of the Third SIGHAN Workshop on Chinese Language Learning, 2004.