Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

基于词形模糊归并的英文同义关系发现研究*

□ 宋培彦 袁旭 / 中国科学技术信息研究所 北京 100038

摘要：同义关系是构建知识组织系统的重要方面。从词汇语义学角度对术语同义关系类型加以界定，采用词形还原工具Norm获得英文术语原形，获得同义词候选集，进而基于词形模糊归并模型确定同义关系。实验结果表明，基于词形的模糊归并有助于提高术语同义关系的构建效率。

关键词：词形还原，术语，同义词，模糊归并
DOI：10.3772/j.issn.1673—2286.2012.12.010

1 引言

同义关系是叙词表、语义网络、本体等知识组织工具最重要的构成要素之一，在信息检索、信息标引、术语服务等许多研究领域有着广泛的应用。面对海量的术语和数字化文献资源，如何有效地对英文术语同义关系准确、快速地进行识别和归并，进而提高构建知识组织系统的效率，是当前需要加强研究的课题。

同义关系本质上是对同一概念的指称。同一个概念可以通过词或词组进行符号化，形成一组意义相同或相近的集合。从词汇学的角度来说，术语作为一种约定俗成的语言符号系统，具有“能指”和“所指”两个方面，人们可以通过特定的符号形式对概念进行指称和交流。语言符号和概念之间存在着某种重要的对应关系，这就为计算机通过术语的形态变化自动进行同义关系发现提供了重要依据。英文术语存在着较为丰富的语言形态变化，通过词形处理获得术语原形并进行归并，有助于发现术语之间隐含的同义关系，为知识组织系统的构建提供必要的支持。

2 研究现状

同义关系在理论方面进行了大量研究。一般认为，同义词是意义相同或意义相似的词。语言学家John I. Saeed的定义是：“同义词是那些意义相同或非常相似，但发音不同的词。”《韦氏同义词新词典》给同义词的界定是：“本词典中，同义词将始终指的是英语中基本意义相同或非常相近的两个或多个词之一。”认知语言学家J. Lyons认为在同义词与近义词之间还有一类词，这类词与另外两类词既有联系又有区别，即把同义词分为三类：绝对同义词、部分同义词、近义词[1]。可见，同义词是一种语义边界较为模糊的词语类聚方法，判断标准带有一定的主观性，常常随着人们对客观事物认识角度、词语使用语境等因素不同，这对计算机进行自动判定带来了很大难度，需要人机结合进行判定。

人们对同义关系的实证研究也取得了很大进展。总体上分为三种方法，一是基于wordnet等语义资源，对词语之间的同义关系进行计算，如Andrew Krizhanovsky提出的基于同义词库提供源数据的语义距离计算方法[2]，这种方法准确率相对较高，但需要事先建立一个语义关系准确、规模适当的术语资源，资源要求较高；二是采用模式匹配方法，制定同义关系的发现规则，从语料库中根据相似度挖掘出同义词，这种方法具有很高的召回率，但准确率相对较低。如M. Batet提出的跨多个本体的相似性估计来判定同义语义[3]；三是词语的特征采用词语聚类的方法，进行粗粒度的归类。如

* 本文系国家“十二五”科技支撑计划“面向外文科技文献的超级科技词表和本体建设研究”（编号：2011BAH10B01）研究成果之一。

D. Mladeni提出的自动化的词形还原模型，并采用聚类的算法对多种自动化处理技术进行测试评价[4]，Staffan Hellberg提出不使用词典方式进行词形归类并记录下词形变化规则的整个流程[5]，这些研究主要是对日常词汇而非术语展开的，该方法对术语同义关系的适用性还需要进一步验证和探讨。如何在脱离同义词词典的基础上，更加快捷、准确地判定术语的同义关系，本文提出了基于词形的模糊归并方法，尝试通过术语原形处理和模糊归并原理建立术语同义关系。

3 词形模糊归并

模糊性是同义词的重要特征。同义术语之间的差异往往是一个连续过渡过程，对计算机来说，术语的同义关系是相对模糊、不易精确定义的，因而需要采用某种方法近似地确定术语对概念的隶属关系。模糊数学中以量化的模型刻画整体模糊程度，为进行术语同义关系归并提供了参考。词形的模糊归并主要语义内涵、语义外延、语义范围及语义强度等判定因素。以模糊度为基础，应用模糊关系模型，可以将一些边界不清、不易定量的因素定量化，实现英文术语归并。

3.1 同义关系划分

根据术语对概念的语义距离，可以将英文术语同义关系划分为三类：绝对同义词群、部分同义词群、相关同义词群。

（1）绝对同义词群

如果术语A与B整体意义完全相同，则术语A与B为绝对同义关系，或称为强同义关系；要满足强同义关系必须满足下列四个条件：核心义等同、外围义等同、选择义等同、义域等同。例如：color与colour仅仅是英美拼写方法的差异，二者本质上是绝对同义的。

（2）部分同义词群

如果术语A与B仅核心义等同，则术语A与B为部分同义关系，或称较强同义关系。部分同义词群之间仅仅核心义等同，其它意义则不同。例如：Library cataloging与catalogues（libraries），后者通过括号形式进行了概念限定，表明两个词语都属于图书馆领域，但由于核心词cataloging与catalogues的词性不同，因而可以视为部分同义。

（3）相关同义词群

如果术语A与B仅核心义中的共性语义成分等同，则词A与词B为近义关系，或称弱同义关系。从语义学的角度来观察，这些术语的核心义中有一共性语义成分。这一共性语义成分使之意义相近，成为相关同义词群。例如：developing country与developed country，属于对偶性概念，可以视为相关同义词。

同义词群的划分主要从词群当中术语的语义内涵、语义外延、语义范围及语义强度4个方面进行界定[6]。语义内涵是指同义词群中术语所表述的概念属性，说明某个对象“是什么”。语义外延是指同义词群中术语所指的对象或所强调的内容不同，重点说明“哪些是”；在基本意义相同的前提下，同义词群中各个术语词义的侧重点有所区别。语义范围（semantic coverage）是指同义词群中一些术语词义范围比较广泛，而其他的术语只是这个术语的词义范围的一个方面或用于特定的学科领域。语义强度（semantic intensity）是说有些同义词群中的术语在语义强度上存在着细微的差别，即一组同义词中各个术语所表示的词义轻重是不同的，也就是所具有的基本词义或概念在程度上有所不同。

同义词的判定可以采用词汇学中的“同形结合”法。设有待检验词语甲与乙，若想知它们的“对象是否同一”，选能与给定两词语“联结”的词“丙”，如果“甲+丙”与“乙+丙”指同样的事物，那么“甲”与“乙”同义。因此在对术语建立同义词群过程中，将术语的词根作为“联结”词进行同义关系传导，通过同一的词根对术语进行归并，形成的集合称为同义词群，在同义词群中的术语之间视为同义关系。如图1所示。

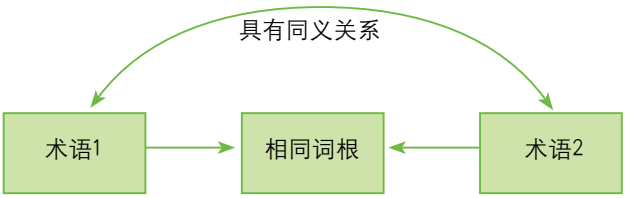

图1 “同形结合”传递原理

采用以上“同形结合法”原理，可以通过词根这种语言表层形式对同义关系进行发现，初步形成候选的同义词群。然后，采用模糊归并模型对候选词群中的语义模糊性进行量化区分，对候选词群进行筛选、归并和优化，推荐出较为准确的同义词。

3.2 模糊归并模型

在模糊数学中，通常采用模糊度来刻画这一概念的整体模糊程度，具体定义如下：若映射d:F(U)→[0,1]满足条件：

（1）当且仅当A∈F(U)时，d(A)=0

（2）$\forall$ u∈U，当且仅当$A(u)\equiv\frac{1}{2}$ 时，d(A)=1

（3）$\forall$ u∈U，当$B(u)\leqslant A(u)\leqslant\frac{1}{2}$ 时，$d(B)\leqslant d(A)$

（4）$A\in F(U)$，$d(A)=d(A^c)$

称d为F(U)上的一个模糊度，d(A)称为模糊集A的模糊度，此四个条件亦是模糊度的四条公理。定义中条件（1）表明普通集是不模糊的；条件（2）和条件（3）表明，越靠近0.5就越模糊，尤其是当A(u) ≡0.5时，是最模糊的；条件（4）表明模糊集A与其补集A^c 具有同等的模糊度。而在现实的分类问题中，考虑的不仅仅是模糊关系，更重要的是模糊关系的深浅程度。设R是U×V上的一个模糊子集，其隶属函数为R:U×V→[0, 1]，(u, v) →R(u, v) ，确定了U中的元素u与V中的元素v的关系程度，则称R为从U到V的一个模糊关系，记:U→V可见，模糊关系R由隶属函数R:U×V→[0, 1]所刻画，即U×V上的模糊集确定了U到V的模糊关系[7]。

根据模糊隶属函数的概念定义，术语之间具有同义关系即为模糊关系，它的取值为单位区间[0, 1]，也可称为同义词间的相关度，用k来表示。在同义词群中，若术语之间词义相同，并且符合词形变化规则，具有完全同义关系，则k=(0.8,1]亦可视为绝对同义词群；若术语之间词义不完全相同，如在术语中常出现逗号、括号等特殊符号，具有同中有异的情况，其k=(0.5,0.8]，亦可称为部分同义词群；若术语之间具有共性语义成分，存在近义关系，其k=(0,0.5]，亦可称为相关同义词群；若术语之间完全没有同义关系的，则k=0。

模糊归并是在模糊环境下，应用模糊集合论方法对归并过程中所涉及的多因素进行分析做出综合决策的方法模型[8]。具体过程如下：

（1）确定影响因素和判断等级

设$U=\{u_1, u_2, ,,, , u_m,\}$为判断对象的m种影响因素；$V=\{v_1, v_2, ,,, , v_n,\}$为每一影响因素所处的状态的n种判断；其中m为评价因素的个数，由人工制定的具体指标体系决定。

（2）构造判断矩阵和确定权重

首先对影响因素集中的单因素u_i(i=1,2,...,m)作单因素评判，从因素u_i着眼该对象对判断等级v_j(j=1,2,...,n)的相关度为r_{ij}，即得出第i个因素u_i的单因素判断集：

$r_i=\{r_{i1}, r_{i2}, ,,, , r_{im}\}$

从而m个因素的判断集构造出总的评价矩阵R，即每一个对象确定了从U到V的模糊关系R：

$$R=(r_{ij})_{m\times n}=\begin{bmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{nm} \end{bmatrix}$$

其中r_{ij}表示从因素u_i着眼，该评判对象能被判断为v_j的相关度(i=1,2, ..., m; j=1,2, ..., n)；r_{ij}具体表示为第i个因素u_i在第j个判断等级v_j上的频率分布。

权数是表征因素相对重要性大小的量度值。影响因素集中的各个因素在“判断目标”中有不同的地位和作用，即各影响因素在综合判断中占有不同的比重；因此引入U上一个模糊子集$A=(a_1, a_2, ..., a_m)$，即为权重，其中$a_i\gg0$，且$\sum a_i=1$。对于权数可以采用数学的方法——层次分析法（Analytic Hierarchy Process，简称ＡＨＰ）。层次分析法是指在处理复杂的决策问题时定性与定量相结合的一种方法，人们对每一层次中各因素的相对重要性给出的判断，并用权重进行表示。尽管该方法在人工处理过程中难免掺杂主观性，但是逻辑性相对完整，可以对确定“权数”进行量化处理，尽量排除主观成分、符合客观现实情况。

（3）模糊合成和作出等级判断

R中不同的行反映了对象从不同的单因素判定各等级模糊子集的相关度。用模糊权向量A将不同的行进行综合，得到该被评对象从总体上来看对各等级模糊子集的隶属程度，引入V上的一个模糊子集B，一般令B=A*R(*为算子符号)，称为模糊评价，又称为决策集，即$B=(b_1, b_2, ..., b_n)$，b_j表示对象具有评语v_j的程度。

4 实例分析

采用开源的英文词形处理工具对多个术语进行词形还原，根据生成的同一词根对术语进行归并，形成候选同义词群；然后，采用模糊归并模型对同义词群的模糊性进行多变量的量化和分类，形成相对完整、科学的度量与判断。

4.1 词形还原处理

词形还原处理可以采用多个开源工具。国外已

经开发了多个词形处理工具，如CST's Lemmatiser[9]、NLTK[10]、MorphAdorner[11]、Stanford CoreNLP[12]等，各词形还原工具其功能和用途各不相同，通过对比实验发现现有词形还原工具Norm工具能较好地解决英文叙词表词形规范问题，其准确率达到90.24%[13]。

Specialist NLP Tools（自然语言处理专业工具）是由The Lister Hill National Center for Biomedical Communications（李斯特山国家医学交流中心）为辅助词汇处理和进行文本分析而开发的一个开源工具，包含了多种处理自然语言的工具。其中Norm可以对词根进行还原处理。Norm工具基于Java环境运行，可以对词形规范化进行处理；它能忽略字母大小写、所有格标记及各种符号，并能对词形、拼写、词序的变化进行处理，返回词的多个原形，为进行同义关系归并提供更多的线索，此工具是开源工具，可以在Specialist NLP Tools官网中进行下载[14]。

4.2 归并过程

在术语同义关系模糊归并中，多个术语指向同一个词根，以此为中介建立同义关系。下面，以术语ZIRCONIA与ZIRCONIUM为例，由人工根据概念的语义关系进行模糊归并分析。

（1）影响因素和判断等级的确定

通过对同义词模糊归并过程分析，得同义词模糊归并的影响因素，其层次结构示意图如图2所示。

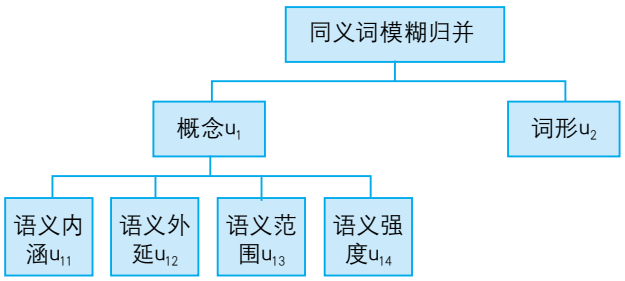

图2 同义词模糊归并影响因素层次结构图

因素集U分为两层：第一层为$U=\{u_1, u_2\}$；第二层为$u_1=\{u_{11}, u_{12}, u_{13}, u_{14}\}$。根据相似系数k的取值范围，将同义关系强度粗略分为4个层级，$V=\{$优v_1，良v_2，中v_3，差$v_4\}$。相对应的K的取值范围：优$v_1\in(0.8, 1]$，良$v_2\in(0.5, 0.8]$，中$v_3\in(0.1, 0.5]$，差$v_4\in(0, 0.1]$。

（2）判断矩阵和权重的确定

各因素权重由多位专家（评价者）填写咨询表，再通过层次分析法分析得出权重值：$A_1=(0.59, 0.18, 0.16, 0.07)$；$A=(0.75, 0.25)$。

分层进行单因素模糊评判，即单独从上述各个因素出发，对模糊归并进行评判，分别得到单因素评判集为：$r_1=(0.8, 0.2, 0, 0, 0]$，$r_2=(0.7, 0.2, 0.1, 0, 0]$，$r_3=(0.8, 0.2, 0, 0, 0]$，$r_4=(0.8, 0.1, 0.1, 0, 0]$；得到

$$R=(r_{ij})_{4\times5}=\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix}=\begin{bmatrix} 0.8 & 0.2 & 0 & 0 & 0 \\ 0.7 & 0.2 & 0.1 & 0 & 0 \\ 0.8 & 0.2 & 0 & 0 & 0 \\ 0.8 & 0.1 & 0.1 & 0 & 0 \end{bmatrix}$$

（3）模糊合成和结果等级判断

由此可得$B_1=A_1\circ R_1=(0.78, 0.19, 0.03, 0, 0)$；$B_2=R_2=(0.5, 0.3, 0.2, 0, 0)$，则$B=A\circ R=A\circ\begin{bmatrix} B_1 \\ B_2 \end{bmatrix}=(0.75, 0.25)\circ\begin{bmatrix} 0.78 & 0.19 & 0.03 & 0 & 0 \\ 0.5 & 0.3 & 0.2 & 0 & 0 \end{bmatrix}=(0.725, 0.225, 0.05, 0, 0)$

由此可知，术语ZIRCONIA与ZIRCONIUM的模糊归并结果72.5%隶属于评价等级优，即两者的相关关系在(0.8, 1]之间，即可判定语ZIRCONIA与ZIRCONIUM为绝对同义词群，因此可以将术语ZIRCONIA与ZIRCONIUM进行概念归并。

术语除了是词、词组，还有一部分是缩略语，缩略语是指通过一定的方式，将原来结构较长、较繁的语词改成结构较为简略的形式而形成的语词[15]；词形还原工具对于缩略语处理还存在不足，在归并过程中，缩略词可以通过同义词典等工具进行概念归并。

4.3 归并测试结果

根据以上的模糊归并方法，利用统计原理，从语料库中抽取了146个工程技术叙词表，去重后，共有373287个术语；其中一个词根对应两个及其以上的术语共有77824（20.8%）个；295463（79.2%）的词是一个术语唯一对应一个词根。归并形式如图3所示。

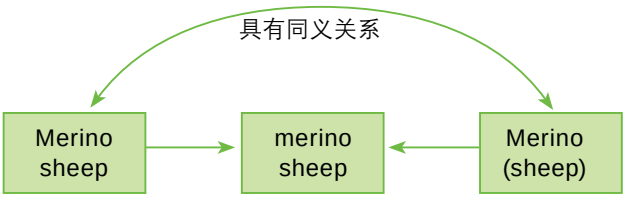

图3 归并过程实例图

其中选取测试数据2036个，以同义词词典作为正确性参照，如果在同义词典中出现，则视为同义词。最终所统计数据的正确率如表1所示。

表1 正确率范围分布

K(相关关系)	K>0：相关同义	K≥0.5：部分同义	K≥0.8：绝对同义
P(正确率)	99.7%	93.5%	88.3%

由此可以得出，相关关系大于0.8的正确率是88.3%，因此可以采用词形模糊归并的方法对叙词表中的部分术语进行概念归并。

5 结语

同义词模糊归并可以辅助用于知识组织工具的构建。本文建立了基于词形还模糊归并模型，帮助用户对词形相同的术语进行同义关系归并；然后，利用工程技术叙词表中的术语进行模型验证。实验证明，基于词形模糊归并方法可以发现术语同义关系，辅助进行概念归并。本方法的优点在于不用依靠专业词典，简单易行，能够以量化的指标对模糊的同义关系进行判定。同时，本方法也有一定的局限性，例如其只能对相似形态进行处理，无法处理异形同义词。在知识组织构建过程中，可以作为同义词处理的辅助手段，为编制人员提供参考。

为了实现同义词关系自动化发现，可以将同义关系的判定因素划分为3个方面进行探索。一是根据术语来源的规范性赋予权值，例如经过专业人士编制审核的叙词表、专业词典的权值较高，而一般用户自定义词则权值较低；二是根据术语作为关键词在文献库中的出现频率，高频词优先归并；三是根据术语的共现关系，在上下文中共现频率高的术语更有可能作为同义词。这样，计算机可以对术语外部因素的权值进行综合判断，把原型化以后的词语分别归入绝对同义词、部分同义词或者相关同义词其中之一。如何将模糊模型加以改进，形成计算机自动进行术语归并的新方法，是下一步需要重点研究的任务。

参考文献

[1] CRUSE D A. 词汇语义学[M].世界图书出版公司,2009:84-109.
[2] KRIZHANOVSKY A. Related Terms Search Based on WordNet / Wiktionary and its Application in Ontology Matching [C]// Proceedings of the 11th Russian Conference on Digital Libraries RCDL'2009, Petrozavodsk, Russia,2009.
[3] BATET M. Semantic similarity estimation from multiple ontologies [J]. Business Media, 2012.
[4] MLADENIC D. Automatic word lemmatization [J/OL]. [2012-08-20]. http://www.sdjt.si/dogodki/sdjt02/zbornik/sdjt02-26mladenic.pdf.
[5] HELLBERG S. Computerized Lemmatization without the Use of a Dictionary: A Case Study from Swedish Lexicology [J].Computers and the Humanities,1972, 6(4):209-212.
[6] 史俊.英语同义词聚合的语义差别分析及其教学[J].浙江理工大学学报,2006.23(4):540-544.
[7] 杨纶标,高英仪,凌卫新.模糊数学原理及应用[M].广州:华南理工大学出版社,2011:25-53.
[8] 杜栋,等.现代综合评价方法与案例精选[M].北京:清华大学出版社,2008:34-61.
[9] CST's Lemmatiser [EB/OL]. [2012-08-20]. http://cst.dk/online/lemmatiser/uk/.
[10] NLTK [EB/OL]. [2012-08-20]. http://www.nltk.org/.
[11] MorphAdorner [EB/OL]. [2012-08-20]. http://morphadorner.northwestern.edu/morphadorner/.
[12] StanfordCoreNLP [EB/OL]. [2012-08-20]. http://nlp.stanford.edu/software/corenlp.shtml.
[13] 吴思竹,钱庆,胡铁军,等.词形还原方法及实现工具比较分析[J].现代图书情报技术,2012(3):27-34.
[14] Specialist NLP Tools [EB/OL]. [2012-08-20]. http://specialist.nlm.nih.gov.
[15] 王昊.基于层次模式匹配的命名实体识别模型[J].现代图书情报技术.2007(5):62-67.

作者简介

宋培彦（1980-），男，博士，中国科学技术信息研究所信息资源中心助理研究员。E-mail: songpy@istic.ac.cn
袁旭（1987-），女，硕士，中国科学技术信息研究所见习研究员。

Research on the Discovery of Relationship between the Synonyms Based on the Morphology Fuzzy Merger

Song Peiyan, Yuan Xu / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: Synonymy is an important aspect of building knowledge organization system. According to lexical semantics, this paper proposes the type of relationship between the synonyms and uses the Norm Lexical Tools which can obtain the prototype of English terminology to find the relationship between the synonyms based on the morphology fuzzy merger. The result of experiment indicates that fuzzy merger method based on morphology can improve the efficiency of construction in the relationship of synonyms.

Keywords: Lemmatization, Terminology, Synonyms, Fuzzy Merger

（收稿日期：2012-09-03）