情报科学
第31卷 第4期 2013年4月

基于UMLS专家词典与工具的词形归并算法研究

李晓瑛,李丹亚,胡铁军

(中国医学科学院 医学信息研究所, 北京 100020)

摘　要: 在简述词形归并(原形化处理)基本目标的基础上,着重分析UMLS专家词典的构建方式与词典工具的核心功能,以及Norm原形化工具的处理机制;在此基础上,提出一种词形归并算法及Norm处理缺陷的修正办法,并收集医学词表数据进行算法测试与验证; 此外,本文算法与经典的Porter算法进行了多方面的深入比较研究。

关键词: 词形归并;UMLS专家词典;Norm原形化工具

中图分类号:G254　　**文献标识码**:A　　**文章编号**:1007-7634(2013)04-134-05

Investigation of Algorithm for Lemmatisation Based onUMLS SPECIALIST Lexicon and Lexical Tools

LI Xiao-ying, LI Dan-ya, HU Tie-jun

(Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020*, China)*

Abstract: Based on the brief description of the basic objective of lemmatization (normalization), the mechanism of developing UMLS SPECALIST lexicon and the key performance of the lexical tool, especially the working procedure of Norm, were carefully studied. Then an algorithm for lemmatization and the strategy to correct the defect of Norm processing were proposed, followed by testing and verification using the lexical data from medical vocabularies. Besides, the well-known Porter method was used for algorithm comparison.

Key words: lemmatization; UMLS SPECIALIST lexicon; Norm

1 引　言

在语言学中,词形归并(lemmatisation或lemmatization)是指将一个词与它不同的屈折变形(inflection)[1]组织起来, 以使它们可被视为同一个词项(Term)的过程[2]。就英语而言,由于人称、时态、语态的不同,一个单词会出现多种屈折变形;而词形归并的重点就是查找单词的原形(base form),以合并源于同一单词屈折变形之后的所有单词;为此,词形归并也常被称为原形化处理。

随着信息时代的到来,越来越多的学者开始研究自动词形归并算法,以期提高自然语言处理中的文本处理效率。根据算法的实现原理,词形归并大体可分为两种:基于变形规则的方法[3]与基于词典的方法[4]。前者主要分析单词的各种屈折变形规则,进而借助屈折变形中引入的单词词缀,以还原出单词原形;但这种算法对于不规则变形的单词则束手无措。后者通过借助可提供单词变形列表的词典,如WorldNet,以查找正确的单词原形,从而解决基于规则方法不能处理的不规则变形问题;但是,受词典收录量和学科领域的限制,这种方法亦

收稿日期:2012-04-22

基金项目:国家科技支撑计划(2011BAH10B02)

作者简介:李晓瑛(1982-),女,陕西宝鸡人,博士,主要从事医学知识组织系统建设与自然语言处理研究.

不能解决未收录在词典中单词的词形归并。

为了解决上述两种算法的应用困难，本文通过分析UMLS(Unified Medical Language System)专家词典(SPECIALIST Lexicon)与Norm工具，提出一种变形规则与词典相结合的词形归并算法。算法处理原理为：Norm首先在专家词典内查找单词的原形，以解决不规则变形单词的词形归并工作；对于未登记在专家词典中的单词，Norm则根据变形规则做归并处理；因此，本文算法不仅可处理规则单词的词形归并，亦可应用于不规则单词。此外，针对Norm输出结果的歧义性缺陷，本文通过研究专家词典的结构体系及Norm工具的处理机制，亦提出一种基于词性约束的修正办法。

2 算法发展基础

2.1 UMLS专家词典

美国国立医学图书馆(National Library of Medicine, NLM)构建专家词典的初衷，是为专家自然语言处理系统提供必要的词典信息[5]。2011版的专家词典收录了涵盖常用英语、NLM测试库及UMLS超级叙词表(Metathesaurus)中生物医学用语在内的约44万个词汇；而这些词汇来自道兰氏图解医学辞典、美国传统词频书、朗文当代高级词典等多种渠道。在专家词典中，每个词汇都作为独立的入口词，连同唯一入口标示符(Entry Unique Identifier, EUI)，记录在“entry”字段中。另外，入口词记录中还登记了词汇的句法、词法及字形信息。专家词典根据句法种类，将单词分为11种类型，即动词(Verb)、名词(Noun)、形容词(Adjective)、副词(Adverb)、助词(Auxiliary)、连词(Conjunction)、代词(Pronoun)、情态动词(Modal)、介词(Preposition)、定语(Determiner)及补语(Complement)，用“cat”字段做标识；由于词的屈折变形存在规则与不规则两种情况，专家词典定义了“variants”字段，以记录这种词法信息；若该字段取值为“reg”，即表明屈折形式遵从变形规则；例如，动词“treat”及名词“virus”在专家词典中的入口登记信息如下：

```
{base=virus
entry=E0064702
      cat=noun
      variants=reg
}
{base=treat
entry=E0061964
      cat=verb
      variants=reg
}
```

而当某个单词的屈折变形式并非来自变形规则时，专家词典不但用“irreg”指明这种不规则变形信息，还列出了该词具体的屈折变形词。比如，专家词典将动词“eat”、名词“woman”分别记录成：

```
{base=eat
entry=E0024352
      cat=verb
      variants=irreg|eat|eats|ate|eaten|eating|
}
{base=woman
entry=E0065549
      cat=noun
      variants=irreg|woman|women|
}
```

其中，“|eat|eats|ate|eaten|eating|”依次为动词“eat”的原形、第三人称单数形式、过去式形式、过去分词及现在分词；而“|woman|women|”为名词“woman”的原形及复数形式。显然，UMLS专家词典这种将单词与它的屈折变形信息相结合的登记方式，为词形归并等自然语言处理带来丰富的词典信息与数据支持。

2.2 Lvg词典工具

除了包含大量词汇及自然语言处理等其它应用所需信息的专家词典，NLM还开发出一种与之配套使用的命令行工具(Command line tool)：词汇变形生成工具(Lexical variant generation, lvg)[6]。这种工具可用于文本模式识别、词汇索引生成等多种应用中。与其它命令行工具类似，lvg从命令窗口或文件中获取输入，处理完毕后又将结果输出到命令窗口或文件中。lvg命令的一般格式为：

lvg –f:命令选项

其中，“lvg”表明将要执行lvg命令，“–f”来自单词“flow”，因为lvg是一种流处理命令，而“命令选项”为用户预期执行的具体操作。2011版的lvg共公布了62种命令选项。表1前两列介绍了几种常用的命令选项及相应的功能[6]，后两列举例说明了这些命令的执行结果，其中多个输出结果已分行列出。

表1 一些常用的lvg命令选项及功能示例[6]

命令	功能	输入项	输出项
A	返回输入项的首字母缩写词	high-altitude headache	HAH
a	返回首字母缩写词的全称	AIDS	acquired immune deficiency syndrome acquired immuno-deficiency syndrome
B	返回输入项中各个单词的原形	Scans, Whole Body	scan whole body
b	返回输入项的原形	ate	eat
E	获取唯一标识符(EUI)	virus	E0064702
i	生成输入项的屈折变形词	virus	virus viruses
N	原形化处理(即Norm)	leaves	leaf leave
s	生成输入项的拼写变形词	acquired immune deficiency syndrome	acquired immune deficiency syndrome acquired immuno-deficiency syndrome
w	按字顺对输入项排序	Scans, Whole Body	Body Scans Whole
y	生成输入项的同义词	eye	ocular ophthalmic optic

2.3 Norm原形化工具

Lvg中有一个非常重要的原形化工具,即Norm,用lvg -f:N表示。对任何输入词汇,Norm将依次进行如下处理:去所有格,用空格代替标点符号,去停用词,小写字母,生成每个屈折变形词的原形,提取每个单词的原形,最后按字顺排序[7]。可见,Norm的处理结果是输入词串的一个原形版本,以使用户忽略字符大小写、屈折及拼写变化、标点符号、所有格、停用词、符号、连字符,及字顺。但是,这些处理(特别是按字顺排序)会导致原形化后的结果失去可读性。

值得提出的是,对于个别特殊的单词,原形化后可能会有多种形式。这是因为一些英文屈折变形词具有多个原形。如表1所示,"leaves"可能是名词"leaf"的复数,亦或动词"leave"的第三人称单数形式。在这种原形化可能出现歧义的情况下,Norm将在不做任何筛选的前提下直接返回所有的原形。但是在某些特殊的应用中,用户要求输出结果与输入项的个数之间必须是一对一的关系,而非Norm命令所表现出的一对多。此时,应考虑使用Norm的另一个版本LuiNorm(lvg命令为-f:N3)[8]。LuiNorm与Norm的不同之处为,对任何输入而言,前者的输出是唯一的。在lvg中,这种处理歧义的屈折变形词还原问题的技术称为标准化(canonicalization)。标准化过程预先计算出单词的原形,接着将屈折变形后可能为同一单词的所有原形聚合为一类,最后从这一类原形中选出一个原形,并作为类中所有屈折变形词还原时的代表。lvg选择代表词时采用了以下的规则:

① 标准化代表首先从专家词典中;

② 标准化代表不包含非ASCII编码的字符;

③ 标准化代表为类中字符长度最短的;

④ 标准化代表的字顺优先。

虽然标准化提供了一种歧义的处理方式,但也会引入不精确的处理结果。例如,单词"left"、"leave"及"leaf"因屈折变形词"leaves"而处于同一类,它们的标准化代表都是"leaf";使用LuiNorm归并"left"时得到结果"leaf",而Norm处理后为"left"与"leave";但是"left"与它的标准化代表"leaf"其实并没有关系。因此,对同一输入而言,LuiNorm与Norm的输出并不一定相同;而除了输入与输出的个数之间要求必须为一对一关系的特殊应用之外,一般情况下都应使用Norm。

3 基本原理与分析、验证

3.1 基本原理

如上所述,UMLS不仅提供了具有丰富词典信息的专家词典,还包含了Norm原形化处理工具。考虑到Norm处理结果的非唯一性,本文提出一种基于UMLS专家词典与工具的新颖词形归并算法;算法的基本思想为,两个待归并词分别经过Norm原形化后的(一个或多个)输出结果中,只要有一个结果完全相同,那么这两个词便可看作同一个词项,且应具有相同的词项唯一标识符和规范形式。算法可用伪代码表示为:

if any Norm(Str1) in Norm(Str2)

then TUI(Str1) = TUI(Str2)

and Norm_Term = Norm(Str1)∪Norm(Str2)

其中,Str1与Str2分别表示两个待归并词,Norm()为lvg的Norm工具(即命令lvg -f: N),Norm(Str1)与Norm(Str2)代表这两个词原形化之后的结果;TUI()表示词项唯一标识符(Term Unique Identifier, TUI), Norm_Term为Str1与Str2词形归并之后共同的规范词项集合。算法执行时,如果词Str1经Norm原形化处理后的结果中有一个(或多个)词与Str2的Norm原形化处理结果中的一个(或多个)词

完全相同，那么Str1与Str2便可认为是同一个词项，且它们具有相同的词项唯一标识符TUI，而这个词项的规范形式Norm_Term是最初两个词Str1与Str2分别经过Norm原形化后结果的并集。例如，“left”经Norm处理后的结果为“left”与“leave”，所以Norm(left) = {left, leave}；“leaves”经Norm处理后得到“leaf”、“leave”，Norm(leaves) ={leaf, leave}；由于“left”原形化结果中的“leave”与“left”原形化结果中的“leave”相同，因此“left”与“leaves”就是同一个词项，且应分配相同的词项唯一标识符，而它们共同的规范词项集合为{left, leave, leaf}。

3.2 分析

仔细分析可发现，上述规范词项集合{left, leave, leaf}中的三个元素之间并非都存在关联；具体而言，第一个元素“left”与第二个元素“leave”均作为动词时存在关联，而第二个元素“leave”与第三个元素“leaf”因名词“leaves”关联；但第一个元素“left”与第三个元素“leaf”之间却无任何联系。

从本质上分析，上述具有歧义的处理结果，源于Norm原形化过程中忽略了单词的词性。而本文认为，只要对Norm原形化结果加以词性的约束，便会修正这种歧义性缺陷。再者，UMLS专家词典中已记录了词性信息（“cat”字段），因此借助词性对Norm处理过程加以修正是可行的。例如，将“left”的原形化结果归为两类：形容词“left”、动词“leave”，而将“leaves”处理成名词“leaf”、动词“leave”；那么在词形归并“left”与“leaves”时，因这两个单词都具有动词词性，便可获得它们共同的动词原形：“leave”，此时原形化结果已完全精确而无歧义。

从上面的分析过程可看出，对少数具有多种词性（通常也具有多种词义）的单词而言，Norm原形化结果具有一定的歧义性，亦即Norm原形化工具存在一定的缺陷。然而，NLM却并未采用任何修正措施来处理这种歧义性缺陷。这是因为NLM开发Norm工具的初衷，是让UMLS超级叙词表整合过程忽略词项的词形变化，为下一步的概念归并奠定基础[7]。而上面提到的歧义性问题，完全可在概念归并过程中通过词义来鉴别，因为概念归并的目标便是聚合各种拼写变形及屈折变形同义词、异形同义词（如“renal”与“kidney”），而区分同形异义词[9]。因此，本文建议在使用Norm原形化工具时，应依据处理结果的歧义性缺陷是否会对具体应用的最终效果产生影响而决定是否采取词性约束的缺陷修正办法。

3.3 实现与验证

本文提出的基于UMLS专家词典与工具的词形归并算法已用Java高级编程语言实现，且运行在Eclipse集成开发环境下，而其中的Norm工具与专家词典均来自UMLS官方网站上公布的最新版本[6]。算法成功运行后，来自<<医学主题词表>>(Medical Subject Headings, MeSH)[10]及<<国际系统医学术语集-临床术语>>(Systematized Nomenclature of Medicine - Clinical Terms, SNOMED CT)[9]的医学词汇作为测试数据，对算法进行了验证。而为了验证算法的正确性和可靠性，本文采用本算法的词形归并结果与UMLS超级叙词表整合MeSH、SNOMED CT词表时所生成的原形词项集合的字符精确匹配方法，作为评价指标。测试结果表明，本算法的词形归并结果与UMLS超级叙词表归并的原形词项是一致的。表2列出部分数据测试结果。其中，“Cardiac Disease”与“Diseases, Cardiac”经Norm处理后为同一个结果“cardiac disease”，这是最简单、最理想的情况；但是，“Mycological Typing Technique”、“Techniques, Mycological Typing”与“Typing Techniques, Mycological”各自经过Norm后均得到两个结果“mycological technique type”、“mycological technique typing”且完全相同，而“ceratitis capitata”的两个Norm结果“capitaton ceratitis”、“capitata ceratitis”与“capitatas, Ceratitis”的两个Norm结果“capitata ceratitis”、“capitatas ceratitis”不完全相同；可见，即使输入数据有多个不完全相同的Norm原形化处理结果，本算法都能正确对其进行词形归并，以实现它们可被视为同一个词项的目标。

表2 部分测试数据及处理结果

测试数据	处理结果
Cardiac Disease Diseases, Cardiac	cardiac disease
Mycological Typing Technique Techniques, Mycological Typing Typing Techniques, Mycological	mycological technique type mycological technique typing
ceratitis capitata capitatas, Ceratitis	capitaton ceratitis capitata ceratitis capitatas ceratitis

3.4 比较

Porter算法[11]是词形归并领域中最经典的方法之一，自1980年问世以来，已被广泛运用于自然语

言处理的各种应用中。但该算法与本文算法却有较大不同。首先,从算法的基本发展思想来比较,Porter算法是一种基于变形规则的原形化方法,而本文算法结合使用了英语单词的各种变形规则以及词典信息,可解决不规则变形单词的原形化问题;例如,“left”的Porter处理结果为“left”,而经本算法处理后得到“left”与“leave”;显然,后者因提供了两种原形而显得更准确。其次,从处理对象来比较,前者主要应用于单词,而后者可适合于单词及由多个单词组成的复合词;例如,理论上应被归并为同一个词项的两个词“Cardiac Disease”、“Diseases, Cardiac”,经前者处理后分别得到单词“cardiac”与“diseas”、“diseas”与“cardiac”;这种离散的结果因忽略了由字顺不同引起的变形问题,使得在计算机上执行的自动词形归并工作变得异常困难;但经本算法处理后却能得到预期的唯一结果“cardiac disease”。最后,从处理结果来看,前者并不能正确给出一些单词的原形,例如,将“disease”、“leaves”还原成“diseas”、“leav”;但本算法的处理结果“disease”、“leaf”与“leave”是正确的。可见,不论从算法发展思想,亦或处理对象、处理结果,本文算法较之Porter算法都有较大改进,所以更适合应用于词形归并中。

4 结 语

本文提出一种基于UMLS专家词典与工具的词形归并算法,且已通过医学词表数据的测试与验证。同时,本文深入分析了Norm原形化处理结果的歧义性产生根源,提出针对这种歧义缺陷的修正办法。此外,通过与经典的Porter算法进行多方面比较,发现本文算法更适合用于词形归并的应用中。可见,词形归并的结果可使研究者忽略词汇的各种拼写及屈折变形,为自然语言处理等应用中鉴别同原形同义奠定了基础。今后,本文提出的词形归并算法将进一步结合同原形异义、异形同义的词义鉴别[12]及语义关系等方法而应用于概念归并(即词表整合)过程中。

参考文献

1 霍恩比, 李北达.牛津高阶英汉双解词典[M]. 北京:商务印书馆,1997:764.
2 维基百科. Lemmatisation[EB/OL]. http://en. wikipedia.org/wiki/Lemmatisation,2011-07-14 .
3 Jursic M, Mozetic I, Erjavec T, et al. LemmaGen: Multilingual Lemmatisation with Induced Ripple-Down Rules [J]. Journal of Universal Computer Science, 2010, 16(9):1190-1214.
4 Jacob Perkins. Python Text Processing with NLTK 2.0 Cookbook[M]. Birmingham: Packt Publishing,2010: 28-30.
5 Browne AC, McCray AT, Srinivasan S. The SPECIALIST Lexicon [EB/OL]. http://lexsrv3.nlm.nih.gov/ Specialist/Docs/Papers/2000/techrpt.pdf,2000-06-09.
6 U.S. National Library of Medicine. Lexical Tools[EB/OL].http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/current/docs/userDoc/tools/lvg.html,2011-04-12.
7 U.S. National Library of Medicine. Norm [EB/OL].http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/current/docs/userDoc/tools/norm.html,2011-07-12.
8 U.S. National Library of Medicine. LuiNorm [EB/OL]. http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/current/docs/userDoc/tools/luiNorm.html,2011-04-12.
9 U.S. National Library of Medicine. SNOMED CT Browser[EB/OL].http://www.nlm.nih.gov/research/umls/Snomed/snomed_browsers.html,2011-04-12.
10 U.S. National Library of Medicine. MeSH Browser [EB/OL].http://www.nlm.nih.gov/mesh/MBrowser.html,2011-08-22.
11 Porter[EB/OL].http://tartarus.org/martin/PorterStemmer/,2011-10-25.
12 Huang KC, Geller J, Halper M, et al. Using WordNet synonyms substitution to enhance UMLS source integration [J]. Artif. Intell. Med, 2009, 46(2):97-109.

(实习编辑:赵红颖)