

● 吴思竹, 钱 庆, 李丹亚, 李军莲, 胡铁军 (中国医学科学院 医学信息研究所, 北京 100020)

三种词形还原工具对领域词汇的还原效果评估*

摘 要: 文章根据英文超级科技叙词表构建需要, 对三种词形还原工具进行评估。主要选取理、工、农三个领域的的数据, 将词形变形数量比较、工具间处理结果差异比较、抽样评估三种方法相结合, 评估3种词形还原工具对领域专业词汇的词形还原效果。

关键词: 词形还原; 领域数据; 评估

Abstract: Based on the requirement for the construction of the English Super Science & Technology Thesaurus, this paper evaluates 3 lemmatizing tools. The paper selects data mainly from the science, engineering and agriculture fields, and combines the 3 methods of comparison of word form variants, comparison of different results processed with different tools, and sampling evaluation to evaluate the lemmatization effect of the 3 lemmatization tools on the field specialized vocabulary.

Keywords: lemmatization; field data; evaluation

在面向外科技文献的知识组织体系建设中, 超级科技叙词表 (STKOS) 由基础词库、规范概念集和范畴体系三部分构成。它的底层基础词库建设对来自4个领域(理、工、农、医)的多词表数据进行集成和整合, 在多领域术语集成整合策略中, 需要选取适合的原形化工具处理不同领域的术语, 进行词形还原, 作为术语合并的基础。笔者已选取5种词形还原工具使用通用词汇数据进行了测试评估, 选取牛津大学开发的词典分析工具 Word-Simith Tools 中提供的英文的包含词的原形和其变形的对照表作为实验的标准数据, 对5种词形还原工具进行了比较评估^[1]。其中, 专家词典工具^[2] (Specialist NLP Tools) 中的 Norm 工具的还原效果最好, 正确率达到 90.24%。Morph Adorner^[3] 和 Stanford Core NLP^[4] 的词形还原工具的准确率分别达到了 88.93% 和 84.93%, 这说明对于通用词汇, 现有词形还原工具已经取得了很好的处理效果。

但是, STKOS 建设中涉及不同领域中的专业词汇, 其不同于通用词汇, 具有构词复杂、缩写多、外来语多等特点。词形还原工具在处理专业领域词汇时是否依然能达到较好的效果, 值得进一步评估。

本文选取通用英文词汇还原工具评估实验中具有较好还原效果的3种工具, 包括专家词典工具的词形还原工具 Norm, Morph Adorner 和 Stanford Core NLP 的词形还原工具, 基于特定领域数据开展词形还原效果的评估实验。专

家词典工具已被证明能很好地应用于处理医学领域词汇, 被用于医学一体化语言系统 (UMLS) 的构建, 但是其还在其他领域 (如理、工、农), 对词汇的处理效果还未见有效评估。在通用数据的词形还原工具效果评估实验中, 笔者使用标准数据, 能够直接获得工具还原结果的准确率。但是在特定领域中, 没有现成的词汇变形和其原形的标准数据, 并且, STKOS 的建设面向大量领域数据, 要求评估数据也需要具有一定规模数量。因此, 词形还原工具的效果评估具有一定难度。在本文的词形还原效果评估中, 主要选取不同领域的词表, 将工具处理结果的词形变形数量比较、工具间结果差异比较、抽样评估三种方法相结合, 对3种词形还原工具的还原效果进行评估和分析。

1 数据来源和研究方法

1.1 数据来源

词形还原工具效果评估的数据主要选取 STKOS 素材管理系统中已入库和可导出的理、工、农三个领域的英文叙词表, 包括 Eric (理)、Copper (理)、NALT (农)、Eurovoc (农)、USGS (工) 中的词汇。通过对 STKOS 素材管理系统已入库的各领域叙词表进行非优选词和术语数的比值计算和比较后, 选出5个叙词表作为实验数据, 做这样的计算主要是因为该数据后期还要用于同形异义术语的评估, 因此, 选择的非优选词占较大比例的词表作为测试数据, 在本文的词形还原实验中结果并不受此影响。数据是 2012 年 5 月 30 日到 2012 年 7 月 2 日从 STKOS 素材管理系统下载的规范格式的 XML 格式的数据。通过自编 Java 程序对 XML 进行解析, 抽取词表中的优选词和非优

* 本文为国家“十二五”科技支撑计划项目课题“科技知识组织体系的协同工作系统和辅助工具开发”的研究成果, 项目编号: 2011BAH10B02。

选词 (表 1 为实验数据的基本信息, 包括词表简介、词汇规模等)。通过程序将词汇数据导入到 SqlServer 2005 数据库, 对数据进行标点和符号的过滤, 由于停用词对结果没有影响, 因此没有进行停用词过滤。

表 1 选择的词表及术语数量

	理	理	工	农	农
英文名称	Copper Thesaurus	Educational Research Information Center (ERIC) Thesaurus	USGS Gateway Thesaurus	Eurovoc Thesaurus	NAL Agricultural Thesaurus (NALT) and Glossary
词表类型	叙词表	叙词表	叙词表	叙词表	叙词表
术语数	21974	12232	2119	30362	106177
优选词数	7068	4588	992	14626	62526
非优选词数	14906	7644	1127	15736	43651
主题领域	电子学、工程材料	综合	地质勘探	农学	农学
非优选词/术语数	0.678347	0.624918	0.531855	0.518279	0.411115

1.2 研究方法

由于缺乏用于专业领域词汇进行词形还原工具评估的标准数据, 即词的变形和原形对照表, 所以选取 STKOS 建设中使用的实际领域数据, 尝试通过三类方法相结合评估词形还原工具的处理效果, 从不同角度反映 3 种工具的差异, 最终选择适于处理专业词汇的最优工具。主要的评估方法分阶段包括以下 3 种。

1) 统计和比较三种词形还原工具处理前后的词形变化数量。3 种工具中, MorphAdorner 和 Stanford Core NLP 的词形还原工具返回 1 个原形结果, 而 Norm 工具有可能返回 1 个或多个结果。在结果比较中, 对 Norm 还原结果分为两类进行统计: 第一类, 不考虑 Norm 返回的多个结果, 只要经过还原工具处理, 前后的词形发生了变化, 就认为原形化工具产生了作用, 统计这类词形前后发生变化的词汇数量。第二类, 考虑 Norm 返回的多个原形结果, 即结果集, 将还原后结果集中的每一个结果与还原前的词形进行匹配, 如果二者不同, 进行统计计数, 表示还原工具发生了作用。匹配的方法采用精确匹配。该方法可以通过还原结果整体了解选取的测试数据情况并初步了解各工具的词形还原能力。

2) 对 3 种词形还原工具的处理结果进行两两工具的

结果异同词汇比较, 包括 Norm 和 MorphAdorner, Norm 和 Stanford Core NLP, MorphAdorner 和 Stanford Core NLP 的还原结果比较。通过完全匹配计算两个工具处理结果中相同和不同的词汇原形数量 (结果中的词汇是无重复的)。其中, 对于同一个词, Norm 可能返回包含多个原形结果的结果集, 如果获得的是结果集, 则对其进行切分, 将每个还原结果与另一工具的处理结果进行逐一匹配, 如有一个结果匹配成功, 即认为两个工具获得相同的原形。该方法虽然不能准确获得词形还原工具的准确率, 但是可以通过两两对比获得两个工具的相对还原能力。如果两个工具获得相同的结果数量多, 说明二者的词形还原能力相似, 在获得结果的正确或错误趋向上是相似的, 而如果结果具有差异性的数量较大, 则说明两个工具在处理结果的正确性或错误性趋向可能是相反的。

3) 基于方法 2), 抽样选取两两工具还原结果对比的数据中不同原形的词汇数据进行还原结果正确性判断, 评估工具的词形还原能力的优劣。方法 2) 中对两两工具的还原结果进行了对比, 结果中不同原形的词汇数据反映了两种工具处理结果的区别, 对其进行抽样, 并对抽样数据的准确性进行评估, 能够揭示工具的还原效果。实验数据使用词表整表处理, 数量较大, 因此, 结合抽样评估的方法, 控制人工审核词汇数量, 抽样数据为 1 000 个词 (具有无重复的词形), 词数小于 1 000 个词的直接评估, 如大于 1 000, 则从结果中随机抽取 1 000 个词, 人工进行词形还原, 将其作为标准数据, 分别与 3 种工具的处理结果一一进行完全匹配, 从而获得抽样数据的准确度评估结果。

数据处理的人工评估规范: 词形还原主要在词形层次进行处理, 不考虑词义。因此, 人工评估结果只从词形层面考虑还原的正确性。数据中的词均为词表中的词, 为名词短语。在构词上, 多数为多词短语, 其中包含的词的词性为名词、动词、形容词、连词、助词等, 词形还原规范是将发生变形或时态变化的词转换为原形, 并且多词短语的变形是将每个发生变形的词均转换为原形。原形, 具体指将动词过去式、过去分词、进行时态等, 转换为动词原形; 名词复数、所有格等转换为名词单数; 副词、形容词比较级、最高级等形式转换为副词、形容词原形。

实验环境为 Intel (R) Core (TM) 2DuoCPU, 2.40GHz, 2.39GHz, 内存 1.98GB, 硬盘 200G, SqlServer 2005 数据库, JDK 1.6, Eclipse 3.7。

2 词形还原结果分析

根据本文的评估方法对还原结果进行评估分析。

2.1 词形还原后词形变化数统计结果分析

使用3个词形还原工具处理领域数据的结果如表2, 分别统计原形化处理完成前后, 词形发生变化的词数、未发生变化的词数、变化词数占总词数的百分比、未发生变化词数占总词数的百分比。

3种工具处理同一词表的结果中, 除了 NAL Agricultural Thesaurus and Glossary 词表数据外, 其他词表的结果中, 处理前后词形发生变化的词数比例较为相似, 特别是 MorphAdorner 和 Stanford Core NLP 处理的结果。如在 Copper 词表数据中, 其原形化后变化词数分别为 31.99%, 33.24%, 无变化词数为 68.01%, 66.76%, 在 ERIC 词表中, 原形化后变化词数为 55.22%, 50.63%, 无变化数为 44.78%, 49.37%。出现这种现象, 一定程度上说明未经还原处理的词表数据中有部分词是原形词, 处理前后没有发生变形, 这些词占有稳定的比例。该表的统计结果基本反映了实验数据表的基本情况。通过各工具原形化处理后无变化词数所占比例可以推断未经处理前各词表中包含是原形本身的词的粗略比例, 如在 Eric, USGS Gateway 词表中约占表内总词数的一半, 在 Eurovoc 词表中约占 70%。

因为3个工具原形化处理后变化词数和无变化词数很稳定。

由于 Norm 处理的结果中可能包含多个结果, 因此结果中可能包含处理前的词形, 如 acetylides, 处理后返回3个结果 acetylis, acetylide, acetylides。在前后结果对比时, 不能判别 Norm 返回的哪个结果是正确的, 只能判断未处理的词是否存在结果集中, 若包含则判断前后无变化。Norm 的处理结果中分别统计了不考虑多个结果和考虑多个结果的原形化前后变化词数和无变化词数, 这种区别并不是很大, 因为只有部分的 Norm 还原结果中包含未处理前的词形。由于 Norm 返回多个结果, 增大前后发生变化的几率, 因此在 Copper, ERIC, USGS, Eurovoc 的词表数据还原结果中, Norm 结果中还原后发生变化的词数比例比 MorphAdorner 和 Stanford Core NLP 大。在 NALT 词表的处理结果中, 3种工具的还原结果存在差异相对较大, 说明对这个词表的数据处理上存在差别。这种方法粗略反映了各词形还原工具处理词变形数量的基本情况, 但无法确切反映词形还原的准确性和有效性。因此, 对各工具的还原能力和准确性, 还要结合其他两个方法。

表2 词形还原前后词形变化数统计表

数据	原形化前后	Norm				MorphAdorner		Stanford Core NLP	
		不考虑多个结果		考虑多个结果		词数	百分比	词数	百分比
		词数	百分比	词数	百分比				
(理) Copper	原形化后变化词数	9437	42.95	9221	41.96	7030	31.99	7305	33.24
	原形化后无变化词数	12537	57.05	12753	58.04	14944	68.01	14669	66.76
	原形化后变化词数	6940	56.74	6854	56.03	6755	55.22	6193	50.63
(理) ERIC	原形化后无变化词数	5292	43.26	5378	43.97	5477	44.78	6039	49.37
	原形化后变化词数	1071	50.54	1053	49.69	1036	48.89	954	45.02
(工) USGS	原形化后无变化词数	1048	49.46	1066	50.31	1083	51.11	1165	54.98
	原形化后变化词数	8326	27.42	7880	25.95	7786	25.64	6198	20.41
(农) Eurovoc	原形化后无变化词数	22036	72.58	22482	74.05	22576	74.36	24164	79.59
	原形化后变化词数	17232	16.23	39183	36.90	24764	23.32	49491	46.61
(农) NALT	原形化后无变化词数	88945	83.77	66994	63.10	81413	76.68	56686	53.39

2.2 词形还原结果异同词数比较

表3为工具间还原结果异同词数统计结果, 选择 Norm 和 MorphAdorner, Norm 和 Stanford Core NLP, MorphAdorner 和 Stanford Core NLP 两两工具的处理结果进行精确匹配, 并统计两个工具处理同一数据时结果中具有相同原形的词数、具有不同原形的词数、二者所占的比例, 选择的数据已经过词形去重。结果显示, 在5个词表数据的处理中, 3个工具两两比较的结果里具有相同原形的词的数量越多, 说明两个工具处理结果越一致。特别是 Norm 和 MorphAdorner 的结果, 在5个词表中均高于94%, 如 Copper 词表中, 其比例达到96.25%, Eurovoc 中, 比例达到96.67%。Norm 和 Stanford Core NLP 结果中具有相同原形的词在5个词表中达到83%以上, 在4个词表中达到

95%以上, 这说明 Norm 和 MorphAdorner, Norm 和 Stanford Core NLP 的词形还原结果差异性较小。MorphAdorner 和 Stanford Core NLP 的匹配结果也几乎达到80%以上, 但与 Norm 和 MorphAdorner, Norm 和 Stanford Core NLP 相比结果差异性还是较明显, 这点可以从 MorphAdorner 和 Stanford Core NLP 的结果中, 具有不同原形的词数统计值体现出来。在5个数据的比较中, Norm 和 MorphAdorner 的不同原形的词的数量和 Norm 和 Stanford Core NLP 的不同原形的词的数量相比, 除了 Eric 数据中后者稍多外, 在其他数据中 Norm 和 MorphAdorner 结果匹配中的不同原形的词的数量相对较少, 这说明 Norm 和 MorphAdorner 的处理结果的差异性比 Norm 和 Stanford Core NLP 处理结果的差异性小。从结果可基本推知 Norm 和 MorphAdorner 的处理效

果接近，而 Norm 与 Stanford Core NLP 的处理效果相比次之，而 MorphAdorner 和 Stanford Core NLP 的结果差异性相对较大。

表 3 工具间还原结果异同词统计表

数据	对比工具	不同的原形词		相同的原形词	
		词数	百分比	词数	百分比
(理) Copper	Norm 和 MorphAdorner	796	3.75	20437	96.25
	Norm 和 Stanford Core NLP	906	4.27	20334	95.73
	MorphAdorner 和 Stanford Core NLP	3052	14.37	18185	85.63
(理) Eric	Norm 和 MorphAdorner	593	5.12	10992	94.88
	Norm 和 Stanford Core NLP	458	3.95	11127	96.05
	MorphAdorner 和 Stanford Core NLP	1579	13.63	10006	86.37
(工) USGS	Norm 和 MorphAdorner	85	4.43	1833	95.57
	Norm 和 Stanford Core NLP	86	4.48	1832	95.52
	MorphAdorner 和 Stanford Core NLP	229	11.94	1689	88.06
(农) Eurovoc	Norm 和 MorphAdorner	505	3.33	14668	96.67
	Norm 和 Stanford Core NLP	951	6.27	14222	93.73
	MorphAdorner 和 Stanford Core NLP	1911	12.59	13262	87.41
(农) NALT	Norm 和 MorphAdorner	4526	5.56	76811	94.44
	Norm 和 Stanford Core NLP	13711	16.86	67626	83.14
	MorphAdorner 和 Stanford Core NLP	24739	30.42	56598	69.58

2.3 抽样词形还原结果正确率分析

2.2 节两两工具词形还原结果匹配中，相同词形的统计结果说明两个工具的还原结果相近，对工具还原结果中不同词形差异的正确性进行比较更有助于对比工具还原效果的优劣。因此，对 2.2 节结果中具有不同词形的结果进行数据抽样，并进行人工判断还原词形的正确性。进行抽样评估，主要是因为处理的数据量较大，很难一一判断。因此限制数据的数量，如果两个工具词形还原匹配结果中，不同原形的词数小于等于 1 000 个，取当前结果；如数量大于 1 000 个，则取前 1 000 个词进行人工词形还原。选定数据后，人工词形还原按照 1.2 节中的规范进行，而后，将人工词形还原数据分别与 3 个工具的处理结果进行精确匹配，获得结果如表 4 所示。结果显示，在 3 个领域，5 个数据集的 Norm 和 MorphAdorner，Norm 和 Stanford Core NLP 的词形还原结果比较中，Norm 还原结果的正确率明显高于 MorphAdorner 和 Stanford Core NLP 的结果。而将 MorphAdorner 和 Stanford Core NLP 的还原结果比较发现，前者的效果优于后者，但二者与 Norm 的准确率相比相对较差。这在表 3 中通过 Norm 和 MorphAdorner、Norm 和 Stanford Core NLP 的比较也可以发现，如在 Eurovoc，

Copper, Eric 等数据中，MorphAdorner 的处理结果的正确率均比 Stanford Core NLP 的高。并且，就当前数据也发现在不同领域中词形还原处理的效果差距并不是很大。

表 4 抽样词形还原结果正确率统计表

Copper				
MorphAdorner		Stanford Core NLP		
抽样数	正确率	错误率	正确率	错误率
1000	666	334	144	856
	66.60%	33.40%	14.40%	85.60%
MorphAdorner		Norm		
抽样数	正确率	错误率	正确率	错误率
796	206	590	379	417
	25.88%	74.12%	47.61%	52.39%
Stanford Core NLP		Norm		
抽样数	正确率	错误率	正确率	错误率
906	231	671	574	332
	25.61%	74.39%	63.36%	36.64%
ERIC				
MorphAdorner		Stanford Core NLP		
抽样数	正确率	错误率	正确率	错误率
969	589	411	291	709
	58.90%	41.10%	29.10%	70.90%
MorphAdorner		Norm		
抽样数	正确率	错误率	正确率	错误率
596	91	505	456	140
	15.27%	84.73%	76.51%	23.49%
Stanford Core NLP		Norm		
抽样数	正确率	错误率	正确率	错误率
463	31	432	372	91
	6.70%	93.30%	80.35%	19.65%
USGS				
MorphAdorner		Stanford Core NLP		
抽样数	正确率	错误率	正确率	错误率
228	154	74	48	180
	67.54%	32.46%	21.05%	78.95%
MorphAdorner		Norm		
抽样数	正确率	错误率	正确率	错误率
84	14	70	64	20
	16.67%	83.33%	76.19%	23.81%
Stanford Core NLP		Norm		
抽样数	正确率	错误率	正确率	错误率
85	5	80	78	7
	5.88%	94.12%	91.76%	8.24%
Eurovoc				
MorphAdorner		Stanford Core NLP		
抽样数	正确率	错误率	正确率	错误率
1000	687	303	194	806
	68.70%	30.30%	19.40%	80.60%
MorphAdorner		Norm		
抽样数	正确率	错误率	正确率	错误率
505	77	428	386	119
	15.25%	84.75%	47.42%	23.56%
Stanford Core NLP		Norm		

续表

抽样数	正确率	错误率	正确率	错误率
1000	60	940	911	89
	6.00%	94.00%	91.10%	8.90%
NALT				
	MorphAdorner		Stanford Core NLP	
抽样数	正确率	错误率	正确率	错误率
1000	907	93	51	949
	90.70%	9.30%	5.10%	94.90%
	MorphAdorner		Norm	
抽样数	正确率	错误率	正确率	错误率
1000	222	778	780	220
	22.20%	77.80%	78.00%	22.00%
	Stanford Core NLP		Norm	
抽样数	正确率	错误率	正确率	错误率
1000	49	951	920	80
	4.90%	95.10%	92.00%	8.00%

通过对3个领域、3种工具、5个数据集的统计,结合2.1、2.2节的结果,基本能够获得3种工具在不同领域中的词形还原效果。综合来看, Norm 的还原效果最好, MorphAdorner 次之, Stanford Core NLP 的效果相对较差。这也在一定程度说明 Norm 返回多个结果增加了获得正确结果的可能性,提升了正确率。在结果分析中,也分析了各词形还原工具在处理词形方面的优势和不足,并对其进行了归纳和整理,以便进一步优化各工具,在具体应用中提高词形还原的准确率。单纯从统计结果来看,不同领域间的词在词形还原上并没有很明显的区别,但是和之前利用通用数据进行的词形还原效果评估相比,发现领域数据的处理效果不及通用词汇的处理效果,这说明专业术语相对复杂,使工具准确率降低。各工具也存在一些具体的问题: MorphAdorner 主要在名词复数、动词过去式的还原方面存在一些问题,如 angora goat breed (原形为 angora goat breed), 被还原为 angora goat bree, burros (原形为 burro) 被误还原为 burros, autoantibodies (原形为 autoantibody) 误还原为 autoantibodies, bone (原形为 bone) 被误还原为 bon, branches (原形为 branch) 误还原为 branche 等。Norm 具有较好的效果,它除了能够较好地处理规范变形外,还能较好地处理一些特殊变形,如 bee lice 正确变形为 bee louse, bronchi 被正确变形为 bronchus, 但是其也会对非规范变形进行错误还原。Stanford Core NLP 中的错误主要对多词术语进行原形化时存在的问题,它仅对多词术语中的最后一个词原形化,而没有对每个词都进行原形化,如 corrugated strip (原形为 corrugate strip) 误还原为 corrugated strip, degassing annealing (原形为 degas anneal) 误还原为 degas annealing, 因此,产生大量错误。此外,它对特殊变形的处理也不是很好,如 ammannia (原形为

ammannia) 误还原为 ammannium, ammi (原形为 ammi) 误还原为 ammus。它也不能很好地处理动词 ing 形式: 如 tracking (原形为 track) 误还原为 tracking, flow brazing (原形为 flow braze) 误还原为 flow brazing 等,由此,导致准确率相对偏低。

3 结束语

通过领域数据对3种词形还原工具进行还原效果的比较分析,得出专家词典工具的词形还原工具 Norm 取得较好的处理效果,优于 MorphAdorner 和 Stanford Core NLP 的结论。分析其主要原因是采用规则和词典相结合的方法: ①涵盖多词典,包括美国传统词频书、《朗文当代高级词典》、《道兰氏图解医学词典》、《韦伯斯特医学词典》、《柯林斯词典》等,提供较好的词汇原形来源。②各领域数据中包括大量通用词, Norm 对通用词处理效果好。③对不规范变形词处理效果较好,能够正确地将其还原,如评估实验中其能将 aquaria home 还原为 aquarium home, 而其他两个工具在不规范变形词的处理上存在较多错误。④ Norm 词形还原生成多个结果,提升了获得准确结果的可能性。本次评估中,由于测试数据量较大,人工数据审核难免存在一些误差,如审核规范的严格性、完整性等。由于抽样选择的数据中存在非英文拼写,或为外来语的引用,如 arbres a cames 等,导致原词形难以确认。但总体的工具评估方法基本能够反映3种词形还原工具在特定领域内的还原能力,专业领域词汇由于其构词的特殊性与通用词汇相比,词形还原工具处理效果稍差。本文也分析了各词形还原工具在处理领域数据中的不足,希望能够有助于促进对词形还原工具的改进。□

参考文献

- [1] 吴思竹, 钱庆, 胡铁军, 等. 词形还原方法及实现工具的对比研究 [J]. 现代图书情报技术, 2011 (3): 27-34.
- [2] Specialist NLP tools [EB/OL]. [2012-09-05]. <http://specialist.nlm.nih.gov/>.
- [3] MorphAdorner [EB/OL]. [2012-09-05]. <http://morphadorner.northwestern.edu/morphadorner/>.
- [4] Stanford CoreNLP [EB/OL]. [2012-09-05]. <http://nlp.stanford.edu/software/corenlp.shtml>.

作者简介: 吴思竹, 女, 1981年生, 博士, 助理研究员。
钱庆, 男, 1970年生, 硕士, 研究员。
李丹亚, 女, 1954年生, 研究员, 硕士生导师。
李军莲, 女, 1972年生, 硕士, 副研究馆员。
胡铁军, 男, 1949年生, 研究员, 硕士生导师。

收稿日期: 2012-11-08