基于语义的实体名称规范研究结题报告

刘建华、邹益民、曲云鹏、岳婷

摘要

本研究主要是希望在自动识别出科技数字资源中人物名称、机构团体名称、会议名称等实体名称的基础上，探索一套利用实体名称自身的语言学特征、上下文背景和使用语境等信息，构建基于上述信息判定实体名称的综合模型,自动辨明两个实体名称是否指代同一对象的完整方法，进而构建相应的应用系统，确定在一篇文献中的不同位置或不同文献中出现的两个实体名称的真正指称对象。针对这一目标，研究的研究内容包括四个方面：（1）基于语言学、语义、语用特征的实体名称规范多维特征模型构建；（2）实体名称规范判定的综合评价方法确定；（3）实体名称规范判定涉及的相关信息的获取；（4）基于上述的研究，构建实验系统，形成相应的人物名称、机构团体名称、会议名称等实体的名称规范文档或知识库。其中关键问题包括以下两个方面：（1）实体名称规范中语言学特征、上下文背景、共现、主题聚类等信息的效用；（2）构建综合了语言学特征、上下文背景、共现、主题聚类等信息的语义判断模型，判定两个实体名称的真正指称对象，确定两个实体名称之间是否存在关联。

为了完成项目的研究内容，本项目主要从对国内外的相关研究内容进行了广泛的调研，具体包括：关于实体名称规范的概念、思路、方法和趋势研究，国内外主要的实体名称规范项目，国内外主要的实体名称规范评测会议的内容及成果结论；实体名称规范的关键技术路线研究；实体名称规范中涉及的实体特征研究；实体名称规范的相关资源研究，包括典型的工具及相关的语料库；实体名称规范的典型应用场景研究。最终形成了相应的调研报告。在调研的基础上，总结出实体名称规范的特征模型构建方法及相关算法设计，并基于此开发了相应的系统，在空天领域形成了一个完整的应用，并形成了一个可嵌入第三方调用的软件 jar 包，此外，还将该 jar 包应用到多个领域监测服务中，构建相应的实体名称规范库。本结题报告主要从这三个方面说明项目的开展工作。

第一部分 实体名称规范的相关理论、方法及工具调研

为了全面了解国内外当前在实体名称规范方面的各种研究思路、进展，本项目围绕国内外与实体名称规范相关的理论、方法与工具进行了深入广泛深入的调研，调研首先从实体名称规范的概念、思路、方法和趋势研究入手，明确了实体名称规范的概念、任务划分，然后从传统的思路方法和当前主流的思路与方法两个方面，初步探讨了国内外在实体名称规范方面的发展演化，并进一步探索了实体名称规范的未来研究趋势。在此基础上，项目还从目前国内外典型的几个实体名称规范项目和典型的实体名称规范评测会议入手，分析这些项目和会议的研究内容、研究思路，并将其进行了对比，从而进一步深入了解实体名称规范的主要内容。在概况调研的基础上，项目集中针对实体名称规范的几类关键技术路线和实体名称规范中涉及的实体特征开展了详实而具体的研究，主要围绕着当前主流的基于语义、背景知识等关键方法展开。为了充分利用现有的相关资源，项目还重点筛选了当前几种比较典型的实体名称规范工具及相关的语料资源库进行了分析介绍，一方面为本项目的实验系统开发提供资源储备，另一方面也希望为其他项目提供一些资源上的参考。

项目调研的具体内容可参见“实体名称规范的相关理论、方法及工具调研报告”。

该调研报告中第三章对实体名称规范中涉及的相关特征的分析对于本项目中实验时实际特征的筛选与特征模型的构建有非常重要的参考意义。

第二部分 基于语义的实体名称规范特征集合及算法设计

在“实体名称规范的相关理论、方法及工具调研报告”中，我们已经初步分析了当前相关研究中在实体名称规范中常用的指标。这些特征可归纳为语言学、语义和语用三个方面，分别包括：（1）语言学：词形（字母组合相关性、组合顺序）；词性（是否为同一词性）；单复数形式；位置（是否有前后相关联的位置、结构）；特殊标记（是否有-、()这一类的特殊标记）；语法功能（句中承担的语法角色）；句法功能（对句子的承接作用）；（2）语义：同一语义类别；是否具有重叠的同义词；设定窗口的共现实体类别与名称；实体的关键属性描述是否相同；具有关联关系的实体类别与名称；（3）语用：使用环境（共引、共链）；领域关联。尽管我们总结发现这些特征，但并不是越多特征对名称规范效果越好，从目前的研究、需求来看，针对实体名称所有规范研究主要集中于对缩略语、别称的规范，这里我们主要结合本项目的研究任务重点探讨这两类任务中将重点采用的特征、特征的获取方法及相应的算法设计。

1 基于语义的缩略语规范识别

为了简化研究的任务，本项目将实际的研究过程中，将缩略语的规范识别作为独立的任务进行了研究，这主要是考虑到在目前的相关研究中，研究者们往往也将缩略语的规范识别独立区分。对于缩略语的规范识别其主要包括两个关键的问题：（1）识别出有真实含义的缩略语；（2）将缩略语与规范的全称间构建关联，形成共指对，并用规范全称标识缩略语。第一个任务术语命名实体识别中需要解决的关键问题，在本项目中不进行过多的探讨，本项目主要是基于词典与规则的双重匹配方式进行解决，下文将结合实际文本中的各种现象，分析总结归纳缩略语与规范全称的各种关联特征，从而进一步设计相应的算法。

1.1 缩略语与规范全称的各种关联现象分析

为了实现从现象到本质的规范分析，项目组收集了大量的原始网络资源，通过人工标注的方式标注其中的缩略语及其对应的规范全称。通过分析，从词形组合上看，所有的缩略语基本可以分为：单个分词（所有字母均大写、大小写混合）、多个分词首字母缩写（所有字母均大写）、特殊表达三类。

一、单个分词（所有字母均大写、大小写混合）

（一）大小写混合—多见于项目、机构、会议的缩写中

（1） PrestoSpace---Preservation towards storage and access Standardised Practices for Audiovisual Contents in Europe.

（2）LarKC--large Knowledge Collider

（二）所有字母均大写

（1）仅仅是长字串表达中大写字母的组合

OSDI--Operating Systems Design and Implementation

SOSP--Symposium on Operating Systems Principles

IPTPS--International workshop on Peer-To-Peer Systems

（2）是长字串表达中所有首字母的组合的大写化（包括介词）

ESOP---European Symposium on Programming

（3）是长字串表达中部分首字母的组合的大写化（除去了介词、冠词后剩余的词中首字母的组合，还需要除去特定的几个词，如 ACM、IEEE）

TOIT--ACM Transactions on Internet Technology

TODS--ACM Transactions on Database Systems

（4）需要进行对齐匹配

INFOCOM--International Conference on Computer Communication

DISC---International Symposium on Distributed Computing

Ab3P--Abbreviations Plus Pseudo-Precision

除此之外，还存在一些存在位置上的关联，如全称（缩写），通过括号等连接符的方式将全称、缩写联系在一起。

二、非单个单词：大写字母缩写有空格或者连接符

- SIGSOFT FSE--ACM SIGSOFT International Symposium on the Foundations of Software Engineering
- ESEC/SIGSOFT FSE--European Software Engineering Conference and the ACM SIGSOFT International Symposium on Foundations of Software Engineering
- ECML PKDD（ECML/PKDD）---The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases
- HLT-NAACL---Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting

三、特殊表达情况

- SIGIR--Research and Development in Information Retrieval

通过上述的分析，我们可以看到缩略语与规范全称在词形、词性等语言学特征方面关联存在着比较明显的关联，但是，在在实际的文献中，往往会出现同一个缩略语指代不同内容的情况，具体而言，有以下一些情况：

（1）同一个缩写是多个不同语义类别的实体名称缩写。

如 ACE 可能是机构 American Council on Exercise 的缩写，也可能是文本处理领域的测评会议 Automatic Content Extraction 的缩写，还可能是航天设备 Advanced Composition Explorer 的名称缩写。

（2）同一个缩写是多个相同语义类别的实体名称缩写

如 NAPP 既可能是联合会 National Association for Patient Participation 的缩写，也可能是 National Association of Photoshop Professionals 的缩写。

这类缩略语的规范时如果仅仅采用语言学的特征就会造成错误。因此，根据实际调研分

析的情况，项目组认为在缩略语的规范识别中，加入上下文共现词、语义类别、共现术语词的领域相似度、与其它对象的关联关系几类背景语义知识可以解决相应的问题。

1.2 选用的特征集合

从上述对现象的表述可以看出，针对缩略语的特征选择通过语言学与语义特征的结合来完成，这里的语义特征主要应该包括如下几类：语义类别，共现背景词、共现术语词的领域相似度、与其它对象的关联关系。而在语言学特征方面需要使用到：词形（字母组合相关性、组合顺序）；位置（是否有前后相关联的位置、结构）；特殊标记（是否有-、()这一类的特殊标记）。

1.2.1 语言学特征组合

从上述的分析可以看出，在语言学特征方面，缩略语与实体全称之间的关联主要体现在词形（字母组合相关性、组合顺序）、位置（是否有前后相关联的位置、结构）、特殊标记（是否有-、()这一类的特殊标记）几个方面，这些特征通过自然语言处理与简单的规则可以标记处理，本项目中重点即考虑这几个方面的语言学特征。

1.2.2 基于语义类别的语义特征

语义类别的语义特征对于判定来源于不同文档的缩略语与实体名称的关联非常有必要。通过对语义类型的标记判别可以将不同类别实体全称对应的同一个缩略语区分开来，如上文中所举之例，ACE 可能是机构 American Council on Exercise 的缩写，也可能是文本处理领域的测评会议 Automatic Content Extraction 的缩写，还可能是航天设备 Advanced Composition Explorer 的名称缩写。在本项目中，主要是通过在命名实体识别过程中，在规则中加入上下文中心标识词来确定缩略语的语义类别，如，ACE 与评测会议关联时，其上下文中往往会出现 Conference 一类的中心标识词，而与 American Council on Exercise 关联时，其出现的全文中往往会出现“Organization”或“Agency”一类的字眼，通过对项目组收集整理的各类实体中心标识词库的匹配，从而在实体识别阶段可确定缩略语的真实语义。

1.2.3 基于共现背景词的语义特征

从上文中可以看到，在语义类别的判定阶段，上下文中的中心标识词起到了很关键的作用，事实上，上下文中的中心标识词也是共现背景词的一种。但这里的共现背景词更多地是指除在实体名称周边出现的其它实词。如“the National Association for Patient Participation promotes and supports patient participation in primary care”一句中，与该机构同时出现的往往有 patient、primary 一类的词，若在“NAPP”出现的页面正文中也经常出现这一类的词，就可以认为这两个名称的关系非常密切，很可能指代同一个实体。共现背景词特征表示如下：

A_contextWord：取缩略语周围的几个词（设定为前后各 4 个窗口词）作为上下文，这

里需要去除连词、介词、形容词、冠词等虚词（项目组收集整理了相应的统一停用词表用于这类词的筛选），仅仅保留实义词。在其可能匹配的全称页面上，判定是否含有这些词中的一个或多个。

F_contextWord：取候选的全称周围的几个词（取词方式与 A_contextWord）相同，然后在其可能的缩略语页面上，判定是否含有这些词中的一个或多个。

1.2.4 基于共现术语词的语义特征

语义类别的语义特征可以很好地区分不同类别的实体缩略语与全称，但在遇到同一语义类别的实体缩略语与全称时，就需要从别的角度进行考虑。与待判定的缩略语在文档中共现的术语词的语义相似度此时就可以发挥很好的作用，在事实上，不同的机构或人员往往在其所从事的研究方向上有所区分，而研究方向则可以通过领域术语得到反映，因此，本项目中充分利用术语识别的结果，构建文档术语集合，通过 WordNet 进行术语集合中术语同义词、上下位类词的语义扩展，从而构建更大集合的术语词袋，进而通过计算词袋的相似度来判定实体缩略语与全称的关联关系。共现术语词的语义特征表示如下：

A_TermCorpus: 随机筛选缩略语所在的文档中的术语词，利用 WordNet 获取其同义词、上位词、下位词集合；

F_TermCorpus: 随机筛选候选实体全称所在的文档中的术语词，利用 WordNet 获取其同义词、上位词、下位词集合；

计算 A_TermCorpus、F_TermCorpus 的相似度，判定其中是否有重合或相似的一个或多个词。

1.2.5 基于共现实体的语义特征

通过领域术语的相关度计算来判定缩略语与全称的领域相关性中涉及到一个很关键的问题，即术语扩展的范围的问题。因为领域内的术语往往是多种多样的，尽管可以通过 WordNet 在一定程度上获取到同义词、上下位类词，但是还是会造成不少关系的缺失。因此，我们进一步考虑通过共现的实体来进行相应的补充。如果某个实体缩略语及其全称在段落内、句内有共同的共现的其它实体集合，如人物、机构、项目、会议等，则可以将其判定为同一个实体。本项目主要通过文本中句内关系的识别，来构建共现实体间的关联关系对，从而为实际的计算提供有效地支撑。共现实体的语义特征表示如下：

A_EntityCorpus: 获取缩略语所在的句中共现的其它实体（包含实体的语义类型），构建共现实体集合；

F_EntityCorpus: 获取实体全称所在的句中共现的其它实体（包含实体的语义类型），构建共现实体集合；

计算 A_EntityCorpus、F_EntityCorpus 的相似度，判定其中是否有重合一个或多个其它实体。

1.3 实验算法流程

针对 1.1 的第一部分所分析的情况，由于有一定规律可循，而且由于篇章内的实体名称在共现背景词、语义类别、共现术语与实体等方面都较难存在差异，因此在篇章内可以依靠提取规范表达的全部或部分首字母、规范表达中的大写字母组合，与候选缩略语进行匹配，

计算相似程度。或者可以通过特殊位置标记来进行获取。由于特殊表达的情况很少，因此可以通过辞典匹配的方式进行归并。而考虑到1.1的第二部分所分析的情况以及篇章间的情况，可能出现语义类型等方面的不同，仅仅依靠语言学特征无法胜任，因此，还需要加入更多的语义背景知识进行综合判定。结合1.2部分对各种语义知识的分析，项目组设计了分别针对篇章内的缩略语识别规范与篇章间的缩略语规范设计了相应的缩略语规范识别算法。

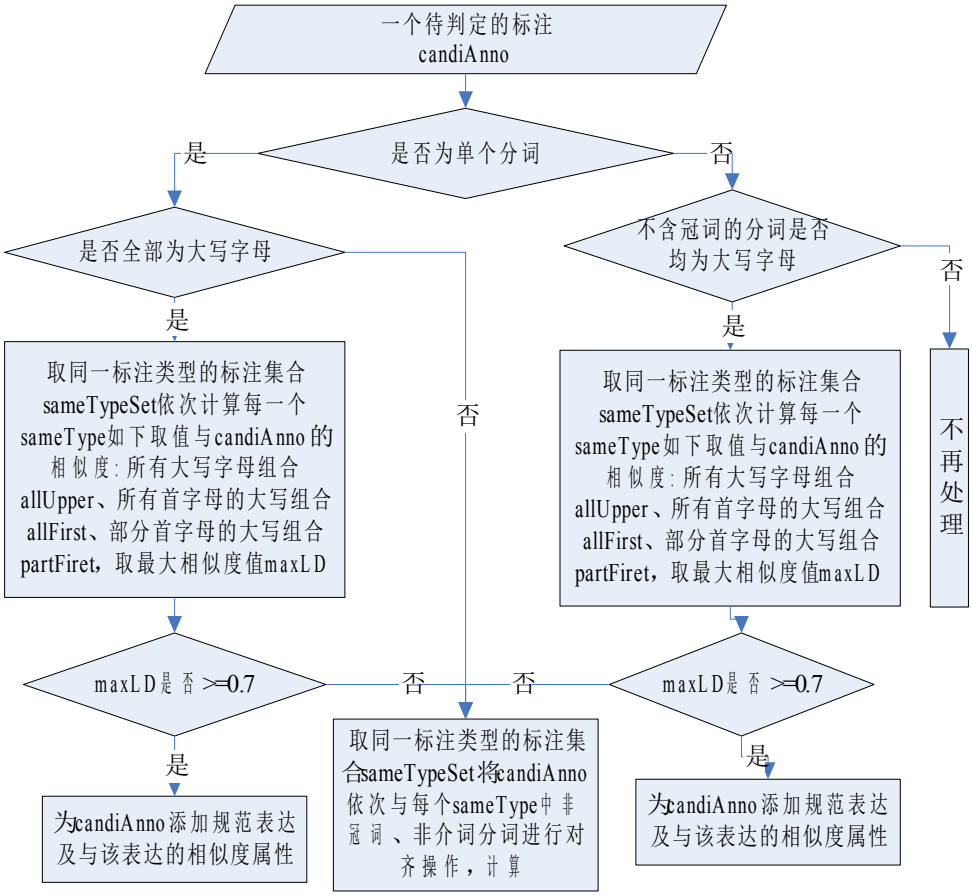


图1 篇内缩略语处理的部分算法

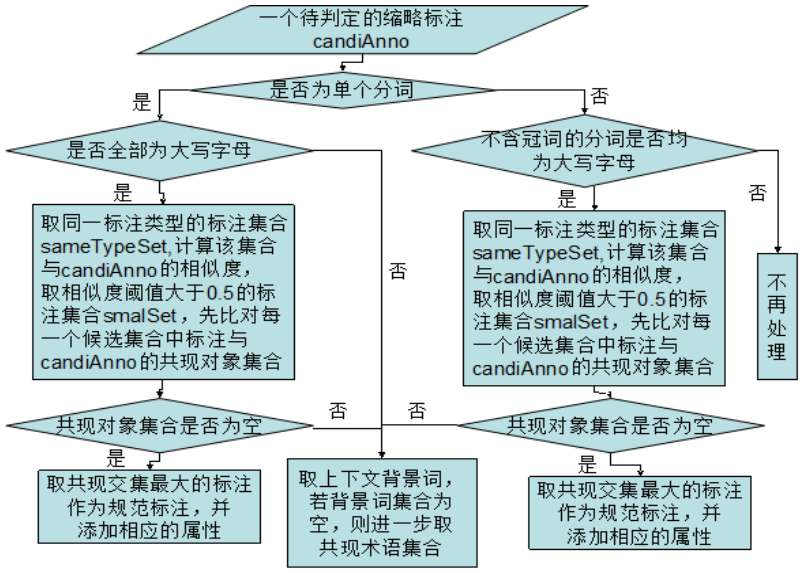


图2 篇章间缩略语处理的部分算法

2 针对别称及同一名称指代实体的规范问题

2.1 别称规范的各种关联现象分析

别称及同一名称指代实体的规范主要指的是两个不同表达的名词是否指代同一个实体及两个相同表达的名词是否指代同一个实体的情况。在这里，由于缩略语共指的问题在第一部分已经有所阐述，这里就不在说明。

1、人名名词共指：“explains Dr Erich Rome, coordinator of the MACS project…….. says Rome”.名词共指，Dr Erich Rome---Rome.

2、同一名称的歧义：Present Bush。两篇文章中均出现 Present Bush 时，两个 Bush 是否指代同一个 Bush 总统就需要阐明。

2.2 选用的特征集合

由于名称共指的情况比较复杂，而且从语言学上比较难以找到特别明显的特征，因此，在这类任务的处理上将更侧重于语义背景知识的使用。

2.2.1 语言学特征组合

在语言学特征上，不同别称之间的关联主要体现在词性、词的句法功能以及词中关键词的重叠部分。这类特征通过浅层的文本处理，包括自然语言处理、句法分析等，就可以获取此类信息，在这里不再做更为详细的赘述。

这里的语义特征主要应该包括如下几类：语义类别，共现背景词、共现同义词、实体的关键属性描述是否相同、具有关联关系的实体类别与名称；而语言学特征则包括：词性、词的位置、句法功能、语法功能几类。

而针对同一名称的歧义，由于其往往涉及多篇文章，而其由于在表达、语义类型等方面均无差异，因此需要考虑更多文章内的更多背景信息，要充分从语义和语用的角度充分进行判断，语义方面的特征主要包括：共现背景词、共现同义词、实体的关键属性描述是否相同、具有关联关系的实体类别与名称，而语用功能则要充分考虑共引、共链等信息，同时还要结合领域相关性方面进行考虑。

2.2.2 基于语义类别与属性的语义特征

与缩略语不同，这里的实体名称在标注过程中往往可以加注更多的属性信息，如对人物而言，可以利用词典或规则，标识出人物的性别、职称等信息，而对机构而言，则可以标识出机构的性质，如实验室、公司、科研团队、咨询机构等，因此，本项目在实体识别阶段，构建了大量的规则与词典，尽可能多地识别出标注内容的属性信息，以便获取更多的实体属性信息。此外，项目还借助 Yago 等开放关联数据对识别出的实体进行了属性信息的补充标识，以便扩充实体的属性信息。

2.2.3 基于共指模板的语义特征

通常而言，在文本中还存在一类表达特定语义关系的模式，即通过某些特殊表达将某两个实体进行关联的模板，这种基于模板的方式往往在关系识别过程有着很广泛的应用。比如通过 Hearst 模式表达式获取上下位类的关系。本项目通过对语料的分析，发现存在着如“also called”、“which is named”、“i.e”等一类的模板可以明确标识出共指对，因此，项目组将这一类的模板进行整理收集，转换成相应的规则，在实际判定中，即判断待判定的两个实体名称是否符合这样的模板，若符合，则直接生成共指对，若不符合，则需要进一步通过其它特征进行判定。具体过程可设计如下：

（1）将待判定的两个实体名称与相应的模板一起构建相应的检索式，即构建"candiAnno1""模板""candiAnno2",比如构建“strawberry bud weevil”“also called”“strawberry clipper weevil”，提交给 Yahoo! Search API。这里需要注意检索策略的问题，通过实验表明，选用“完整包含”的高级检索策略[52]，其返回结果的相关度大大高于普通检索。

（2）将 Yahoo! Search API 返回的检索结果下载到本地，过滤检索摘要，保留其中包含检索式的完整句子。若存在完全匹配的此检索结果，则生成两者的共指对，若无，则不处理。

2.2.4 基于共现实体和基于共现术语集合的语义特征

共指模板是一种能准确定位共指对的语义方法，但毕竟在实际的文本中，此类表达少之又少，模板也非常有限，因此，在进行实体名称规范时，还需要进一步考虑共现实体的语义特征。这里的特征与在缩略语中的处理类型，共现实体的语义特征表示如下：

A_EntityCorpus: 获取缩略语所在的句中共现的其它实体（包含实体的语义类型），构建共现实体集合；

F_EntityCorpus: 获取实体全称所在的句中共现的其它实体（包含实体的语义类型），构建共现实体集合；

计算 A_EntityCorpus、F_EntityCorpus 的相似度，判定其中是否有重合一个或多个其它实体。

与之相似的还有共现术语词的语义特征。

2.3 实验算法流程

针对上述的几个问题，本项目主要考虑从特征出发，项目组设计了相应的算法。图 3 展示了在利用语义特征部分的算法流程。

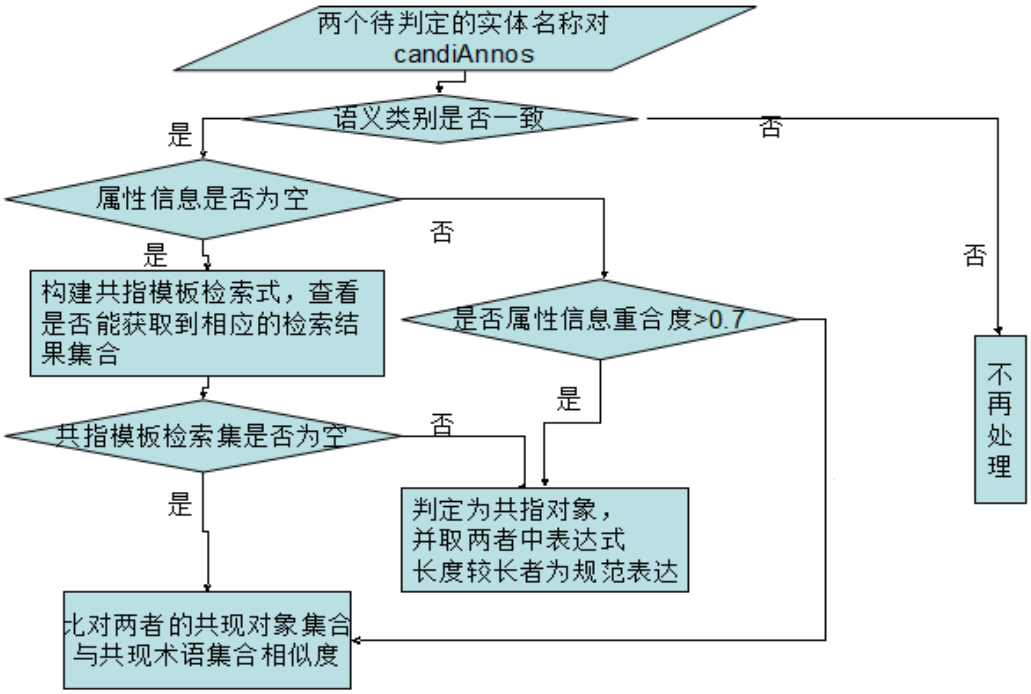


图 3 基于语义信息的实体规范算法流程

第三部分 实验系统的开发和应用展示

在上述调研分析和算法设计的基础上，本项目结合需要获取的特征、系统的可扩展开发便利性以及项目成员对开源系统的熟悉程度，以 Sheffield 大学提供的 GATE 为核心的基础组件，加入 Stanford Parser、Yago 等开源工具和开源语料，开展相应的实验系统的开发，形成了一个从基础的命名实体识别到实体名称规范的完整工具包，并将该工具包应用到了实际的监测服务系统中，实现了两个实验应用：（1）自动识别多领域的人物、机构、会议、地理名称、奖项等实体，并将识别出的实体名称进行自动规范关联，便于进行重要机构、重要人物、重要会议等的计算与判定。此外，按照本体的思想，以三元组的方式构建了一批实体名称规范库，便于其它系统直接利用。（2）自动识别并判定空天领域空天任务实验与设施的全称与缩略语的关联，构建完整的 ISS 实验与设施的完整描述语料库，通过空天领域的监测系统呈现不同学科、不同方向、不同任务编号所关联的 ISS 实验与设施的统计结果，一方面便于用户直观地了解当前各种 ISS 实验与设施的分布变化发展趋势，另一方面，也为进一步空天领域的其它应用提供了语料库。

1 实体名称规范工具包的开发

为验证第二部分提出的“基于语义的实体名称规范的特征模型及相关算法设计”的可行性和有效性，本项目选用人工智能领域的相关文本资源进行实验，这一节将主要阐述实验的技术框架、数据选择及实验的流程。

1.1 实验技术框架

经过比较，项目组认为，谢菲尔德大学研究开发的通用文本工程框架（General Architecture for Text Engineering，GATE）[1]能很好地满足本实验的需求。

（1）本着简化语言工程系统开发流程的目标，GATE 构建了“算法＋数据＋图形用户界面＝应用”的基本结构，按照此结构选用面向对象的编程语言和基于 JavaBean 组件的软件开发方式，开发出一个核心库和一系列可重用语言工程组件（a Collection of REusable Objects for Language Engineering，CREOLE）。每个 CREOLE 组件包括语言、处理和可视化三类资源，资源参数存储于 creole.xml 文件中。用户可根据具体的技术解决方案，快速灵活地定制、修改、扩展各组件。实验即要实现各技术解决方案实现相应的处理组件，这一框架可为实验提供很好的基础。

（2）GATE 中本身包含了很好的自然语言处理底层基础，包括基础的分词、分句、词性、词形标注等，同时提供了与其他第三方软件的良好接口，如 Stanford Parser、WordNet 等，在实际实验中，项目组需要综合利用这些工具来获取对应的语言学或语义特征，这样为实际的开发提供了很多便捷，项目组仅需要根据实验结果对相应的组件进行适当的修改即可，无需在底层开发上投入更多精力。同时，GATE 中提供了一种模版匹配机制，即 JAPE 规则，这些规则一方面可用于辅助标注命名实体，另一方面也可以用于识别出实体间的关联关系，如概念定义式、共指式等简单的模式。依赖于这样的机制，实验中涉及模式构建的部分都只需要关注基于特征的模式构建方式，而无需考虑规则的转换方式。

（3）GATE 提供标注的方式是将所有标注信息包括在一个 AnnotationSet 类中，该类包括“type”和“feature”两个属性，前者提供标注的主要类型，后者包括了处理过程中不断添加的实体特征信息，这些保存的特征信息将成为实体规范过程中的重要参考依据。

（4）GATE 采用 JAVA 语言开发，项目组成员对此比较熟悉，便于快速入手。

项目开展中，除了应用了 GATE 外，还加入了斯坦福大学开发的 Stanford Parser[2]工具作为句法分析工具，基于此，构建句内的命名实体的共现关系对，形成两个命名实体间的关系关联，用于构建基于共现的命名实体间关系的实体名称关系确认。为了获取更多的语义知识，项目中也引入了 WordNet[3]工具，一方面用于获取相同语义类别的实体名称特征词，另一方面用户获取上下位类、同义等关系。除此之外，项目组还将 Yago[4]开放关联数据作为语料引入到了项目中，用于辅助实体的属性识别与匹配。

1.2 实验数据选择

为了检验技术解决方案，同时也为了获取方案中的最佳特征组合，项目组首先选择了适当的小量实验数据（100 个网页新闻），主要是来源于 ScienceDaily 网站 Computers & Math5 栏目下的人工智能部分。作为互联网中在线杂志和科学、技术及医学相关的 Web 门户领导者之一，ScienceDaily 包含了大量科技文献、科技新闻报道等非结构化文本资源。这些文本中分布着大量的科研要素及其关系实例（如图 4）。选择这一资源，一方面可以提供科研要素及其关系实例的训练语料，另一方面可也作为实验数据，检验技术解决方案的可行性和有效性，有较好的实验意义。由于网络数据在机器自动处理过程中会产生比较多的噪音，对后续的实体识别和实体规范产生影响，因此，项目组人工预处理了这 100 个网页新闻，主要是将网页新闻的主体部分（标题、摘要、附录信息）手工提取出来，并加注相应的标点符号，以用于后续的实验处理。

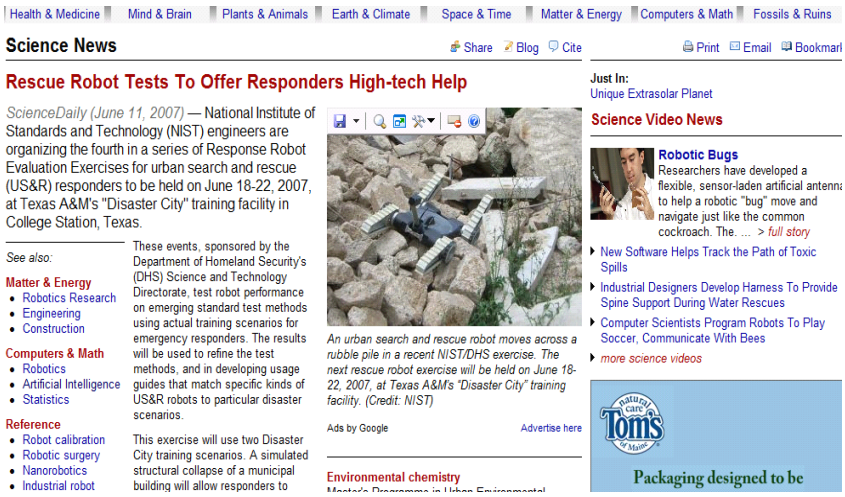


图 4　ScienceDaily 新闻示例

1.3 实验流程

实验中，项目组基于第二部分文档中提出的特征模型和算法方案，进行了实际的实验操作。实验中，项目组主要参考了现有研究中对各种语义背景信息使用评价的结果，对实验中所涉及的几种语义知识设定不同的贡献比重，从而构成最终的实验结果。具体实验中，分词、分句和词性标注这几个自然语言处理步骤均借助 GATE 实现，这将为后续识别提供词性、词形、词串等相关的语言学特征。

为了辅助语义关联的识别，项目组还收集整理了不同的人物职称、机构标识词（如 School、University、Institute、Lab 等），构建了相应的词典库，在这些中心标识词的基础上，利用 WordNet 进一步进行扩展，收集整理其同义词，扩展中心标识词语料库。此外，项目组还收集整理了相应的共指模板，如<i>also called<j>、<i>is named<j>等，依靠这类共指模板来构建共指模板关系对，从而直接生成相应的共指对。其中，包括通过语料的训练获取到的科研要素中心词 156 个，上下文特征词 103 个，模板对 10 个。为了充分利用 GATE 的 JAPE 规则，项目组通过对大量语料的半自动化学习，将部分上下文特征明显的关联关系词直接作为筛选规则写入了 JAPE 规则中，在实体识别阶段，直接判定其是否共指。这一部分主要用于篇章内的小部分共指识别，在这一部分，主要构件了 28 条规则。

除直接在实体识别阶段加入共指关联标识属性外，我们将实体识别过程中识别出的实体语言学特征（词性、句法角色）、语义类别、属性（如人物的性别、职称）、与其它实体的关系、上下文特征背景实词（主要留存与该实体位于同一句内，在该实体前后 4 个的窗口实词，介词、冠词等词通过停用词表进行筛除，选择 4 个窗口词主要是参考了 Wei Jiang[6]等人的研究）存储在 SQL Server 数据库中。

1.4 实验结果的界面展示

```
641 <Conference---2008 International Conference on Information and Knowledge Engineering|held|Location---Las Vegas>
642 <Conference---2008 International Conference on Information and Knowledge Engineering|held|Location---USA>
643 <Conference---2008 International Conference on Information and Knowledge Engineering|held|Date---17th July>
644 <Date---2008|held|Location---Las Vegas>
645 <Date---2008|held|Location---USA>
646 <Date---2008|held|Date---17th July>
647 <University---School of Computing|experimented|ResearchInstitute---the Royal Academy of the Spanish Language>
648 <ResearchInstitute---The Burn Center|employing|Person---Hero>
649 <Date---June 25|presents|System---Intelligent Tutoring Systems>
650 <Person---Marian|present|Date---2008>
651 <Person---Marian|present|Conference---the 2008 IEEE International Workshop on Computer Vision and Pattern Recognition>
652 <ResearchLab---the Machine Perception Laboratory|present|Conference---the 2008 IEEE International Workshop on Computer Vision and Pattern Recognition>
653 <Person---Stewart|present|Conference---the 2008 IEEE International Workshop on Computer Vision and Pattern Recognition>
654 <Date---Saturday, June 28|present|Date---2008>
655 <Date---Saturday, June 28|present|Conference---the 2008 IEEE International Workshop on Computer Vision and Pattern Recognition>
656 <System---Intelligent Tutoring Systems|accepted|Conference---CVPR 2008>
657 <System---Intelligent Tutoring Systems|accepted|Date---2008>
658 <System---Intelligent Tutoring Systems|accepted|Conference---Workshop on Human Communicative Behavior Analysis>
659 <Project---CHIL|received|Foundation---the EU&quot;s Sixth Framework Programme>
660 <Project---CHIL|funding|Foundation---the EU&quot;s Sixth Framework Programme>
661 <ResearchInstitute---GCI|in|Location---Ga.>
662 <Person---Hai Nguyen|at|ResearchInstitute---GCI>
663 <Person---Hai Nguyen|at|ResearchInstitute---Georgia Canines for Independence>
664 <Company---IBM|be|ResearchLab---Los Alamos National Laboratory>
665 <Company---IBM|housed|ResearchLab---Los Alamos National Laboratory>
```

图 5　中间识别出的实体间关联关系示例

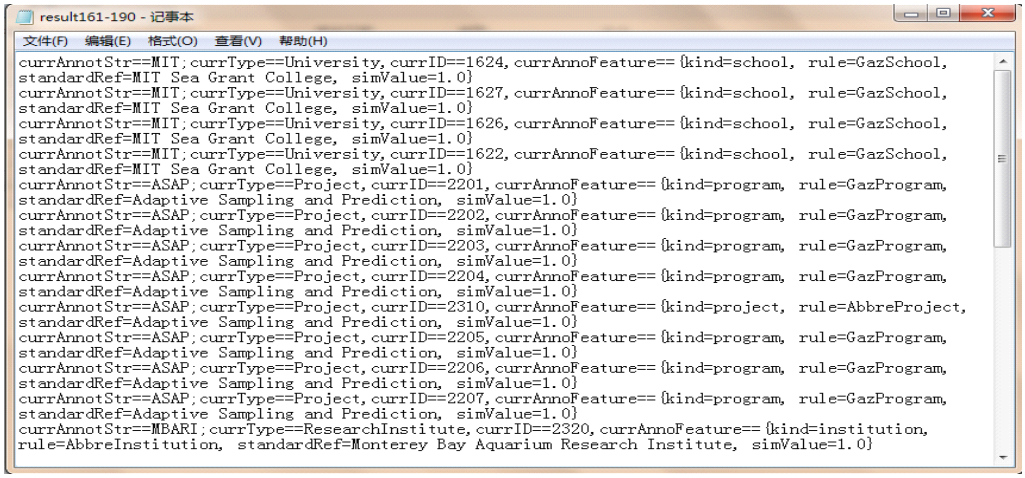


图 6 某篇测试文档的实体名称规范结果

1.5 实验结果的分析

在上文展示了实验结果基础上，本节将从总体统计及与人工标引结果对比两方面，展示实验识别情况。本实验从 100 篇去噪后的新闻报道中共识别出 2104 个科研实体名称，其中包含了 105 个缩略语。这些识别出的实体中主要包括人物、实验室、高等院校、科研机构、科研项目、科研基金组织等几类。在基础语言学特征上，项目组分别从缩略语的规范和其它别称的规范两个方面，进行了规范测试结果的讨论如表 1（说明：这里测试过程中排除了自动的命名实体识别中识别不准确及未识别出的实体数量，仅考虑在已经识别出的科研实体名称中规范的结果。即计算中，R 代表查全率，P 代表查准率，V 表示实验标识出的规范关联数量，W 表示经过人工判断后正确的规范关联数量，Y 表示文档中应该有的准确的规范关联数量。查全率的计算方法是：R=W/Y；查准率计算方法为：P= W/V；）。

表 1 对缩略语规范加入特征及实验结果

任务	查全率	查准率
缩略语规范	90.8%	87.5%
其它别称规范	72.1%	66.7%

1.6 实验结果与解决方案的关连分析

从上述比较的结果可以看到，对缩略语的规范识别，实验和人工标注结果比较接近，查全率和查准率较高。分析出现差异的部分与技术解决方法的关系，主要存在以下几种情况：

（1）对缩略语语义类型标注的错误导致缩略语规范的识别错误。例如“Johns Hopkin”可为人名，可为学校名，在本次识别中，由于存在错误的语义标注的情况，导致相应规范关联的遗漏和错误。

（2）部分词长过长的实体全称识别断裂导致的规范关联对的识别错误。本实验中存在部分词长过长的实体全程由于实验在未能完整识别，如“Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting”，从而导致关联计算时阈值偏低而被排除，下一步研究中还需要针对此类问题进行进一步的探讨。

受这些情况影响，缩略语的规范识别受到了一定影响，但实验结果已经能在一定程度上反映缩略语识别规范方案的可行性和有效性。

与缩略语的规范识别相对比，其它别称的规范识别查全率和查准率都偏低，除缩略语中存在的原因外，在别称规范识别中还存在的问题还包括以下几个：

（1）实体关联关系识别的错误导致规范识别的关联错误。

（2）领域术语集合的相似度计算导致规范识别的关联错误。由于文本中自动识别出来的领域术语在不同文档中不一致，在通过领域术语关联性计算两个实体名称的关联时，主要是通过计算两个文档中领域术语的语义相似度来获取相应的阈值，在语义相似度计算中产生的误差导致了规范识别的关联错误。

与当前的相关研究相比，本项目提出的设计方案所取得的查全率和查准率基本接近了当前的主流测试效果，但是由于评测标准文本的不同（主流的评测主要是利用参加国际评测会议时提供的测试文本进行查全率和查准率的对比，而本文是通过项目组人工标注的文本进行评测），在实际的效果对比上无法也欠缺一些真实的比对。

2 该工具包在多个监测领域的应用展示

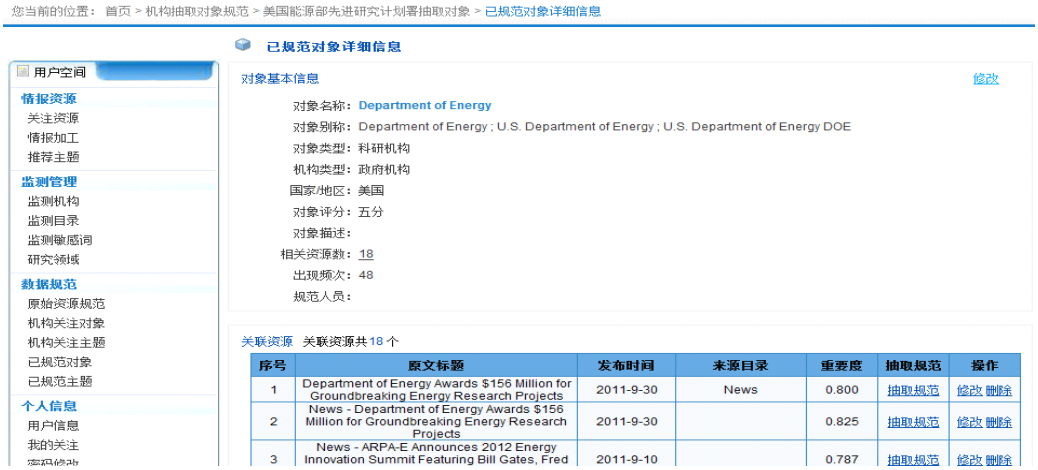
该工具包形成后，项目直接将其应用于科技自动监测服务项目中。在科技自动监测服务项目中，该工具包一方面用于标识实体名称及其自动规范，最终支持从统计计算的角度获取重要对象、热点对象等。另一方面，为人工规范提供参考。具体内容参加图 7、图 8 和图 9 所示。

fileObjectID	objectType	objectValue	standardValue	abbreValue	recordId	startOffset	endoffset
375785	Foundation	The National Science Foundation	National Science Foundation	National Science Foundation	12723	529	560
375787	Person	Crystal Johnson	Crystal Johnson	NULL	12723	614	629
375790	University	Louisiana State University	Louisiana State University	NULL	12723	656	682
375792	Person	Phillip Taylor	Phillip Taylor	NULL	12723	1745	1759
375809	Person	Josh Chamot	Josh Chamot	Josh Chamot	12723	3947	3958
375811	Person	Cheryl Dybas	Cheryl Dybas	NULL	12723	4054	4066
375813	Foundation	The National Science Foundation	National Science Foundation	National Science Foundation	12723	4102	4133
375824	Person	Sir William Wakeham	Sir William Wakeham	NULL	12724	208	227
375827	Person	Sir William Wakeham	Sir William Wakeham	NULL	12724	1179	1198
375935	Project	Research Area Science and Society ...	Research Area Science and Society ...	NULL	12728	1792	1853
375937	University	University Hospital of Navarre	University Hospital of Navarre	NULL	12729	216	246
375938	University	University Hospital of Navarre	University Hospital of Navarre	NULL	12729	465	495
375939	Association	American Society of Brachytherapy	American Society of Brachytherapy	NULL	12729	562	595
375940	Person	Doctor Rafael Mart	Doctor Rafael Mart	Doctor Rafael Mart	12729	1218	1236
375942	University	University Hospital of Navarre	University Hospital of Navarre	NULL	12729	2462	2492
375943	Person	Doctor Mart	Doctor Rafael Mart	Doctor Rafael Mart	12729	2737	2748
375945	Person	Crystal Smith-	Crystal Smith-	NULL	12730	21	35
375950	University	BS University of North Carolina	BS University of North Carolina	NULL	12730	157	188

图 7 领域监测服务中自动规范存储的部分数据

序号	对象名称	对象类型	相关资源数	导入本体实例	操作
1	US Department of Transportation	科研机构	1	导入本体实例	修改 删除
2	Host University	高等院校	1	导入本体实例	修改 删除
3	Leibniz Prize	重要奖项	3	导入本体实例	修改 删除
4	Coordinating University	高等院校	2	导入本体实例	修改 删除
5	U.S. Department of Commerce	科研机构	125	导入本体实例	修改 删除
6	CAISE	会议	9	导入本体实例	修改 删除
7	The University of Connecticut	高等院校	3	导入本体实例	修改 删除
8	Australia-Indonesia Joint Symposium	会议	2	导入本体实例	修改 删除
9	Australian Space Science Conference	会议	3	导入本体实例	修改 删除
10	Japan Symposium on Earth Systems Science	会议	1	导入本体实例	修改 删除
11	The National Academies Forum	会议	6	导入本体实例	修改 删除
12	National Youth Science Forum	会议	5	导入本体实例	修改 删除
13	Brazil Workshop on Biotechnology Innovations	会议	2	导入本体实例	修改 删除
14	Australia Symposium on Sustaining Global Ecosystems	会议	2	导入本体实例	修改 删除
15	Australia Symposium on Remote Sensing Technologies and Sustainability	会议	2	导入本体实例	修改 删除
16	Australian National University	高等院校	340	导入本体实例	修改 删除

图 8 科技政策与战略领域已确认规范的部分实体界面展示



序号	原文标题	发布时间	来源目录	重要度	抽取规范	操作
1	Department of Energy Awards $156 Million for Groundbreaking Energy Research Projects	2011-9-30	News	0.800	抽取规范	修改 删除
2	News - Department of Energy Awards $156 Million for Groundbreaking Energy Research Projects	2011-9-30		0.825	抽取规范	修改 删除
3	News - ARPA-E Announces 2012 Energy Innovation Summit Featuring Bill Gates, Fred	2011-9-10		0.787	抽取规范	修改 删除

图 9 已确认规范的对象详细信息页面展示

3 空天领域 ISS 实验与设施的自动识别与规范应用展示

空天领域的 ISS 实验与设施对于分析空天领域研究的发展趋势有重要参考作用，在空天领域的监测中主要需要完成的是 ISS 实验与设施的全称与缩略语的关联识别任务。项目组针对该任务进行了 ISS 实验与设施的自动识别与规范，在识别与规范过程中应用到的相关的研

究人员、资助机构及其相应的属性信息、学科信息、研究方向信息都得到了保存，构建了一个完整的规范应用库，从而为空天领域 ISS 实验与设施的应用展示提供了很好的支撑。图 10 到图 13 分别展示了自动识别与规范后的 ISS 实验与设施；涉及的人员及其属性信息、学科与研究方向信息等；基于这些实体名称所开展的实际应用。通过该实验，我们共处理了 1042 个 ISS 实验与设施，与 ISS 关联的人员共 2402 个，机构共 1907 个，学科共 6 个，研究方向 48 个。

type	abbName	fullName
experiment	Mirsupio	Evaluation of a Multi Purpose Bag
experiment	Biosfera	Investigation of a Closed Ecological System
experiment	CPCF-2	Commercial Protein Crystallization Factory - 2: Produ...
experiment	DOSMAP	Dosimetric Mapping
experiment	MACE-II	Middeck Active Control Experiment-II
experiment	Popular_Mechanics	Commercial Promotion of Popular Mechanics Journal
experiment	Torso	Organ Dose Measurement Using the Phantom Torso
experiment	Education-SEEDS	Space Exposed Experiment Developed for Students
experiment	BRIC	Biological Research in Canisters
experiment	Biotube	Biotube
experiment	CGBA-KCGE	Commercial Generic Bioprocessing Apparatus - Kidn...
experiment	CGBA-SM	Commercial Generic Bioprocessing Apparatus - Syn...
experiment	GN2	Kennedy Space Center Gaseous Nitrogen Freezer
experiment	CellCult	Cell Culturing
facility	ABS	Autonomous Biological System
facility	Actiwatch	Actiwatch
facility	Actiwatch_Spectrum	Actiwatch Spectrum System
facility	ADF	Avian Development Facility
facility	ADSEP	ADvanced Space Experiment Processor
facility	AEM	Animal Enclosure Module
facility	APCF	Advanced Protein Crystallization Facility
facility	ABS	Autonomous Biological System

图 10 ISS 实验与设施

939	Principal Investigator	William Johnson	Ph.D.	California Institute of Technology	Pasadena,CA
939	Co-Investigator(s)/Collaborator(s)	Marios Demetriou	Ph.D.	California Institute of Technology	Pasadena,CA
939	Co-Investigator(s)/Collaborator(s)	William Kaukler	Ph.D.	University of Alabama - Huntsville	Huntsville,AL
939	Co-Investigator(s)/Collaborator(s)	Chris Veazey		California Institute of Technology	Pasadena,CA
779	Principal Investigator	Paul Todd			
871	Principal Investigator	Catharine Conley	Ph.D.	Ames Research Center	Moffett Field,CA
872	Principal Investigator	Sharmila Bhattacharya		PhD.,Biomodel Behavior and Performance Lab,NASA A...	Moffett Field,CA
872	Co-Investigator(s)/Collaborator(s)	Deborah Kimbrell	Ph.D.	University of California - Davis	Davis,CA
873	Principal Investigator	Millie Hughes-Fulford	Ph.D.	University of California - San Francisco	San Francisco,CA
873	Co-Investigator(s)/Collaborator(s)	Augusto Cogoli	Ph.D.	Swiss Federal Institute of Technology	Space Biology,Zurich Switzerland
874	Principal Investigator	Joshua Zimmerberg	Ph.D.	National Institutes of Health	Bethesda,MD
875	Principal Investigator	Jeanne L. Becker	Ph.D.	National Space Biomedical Research Institute	Houston,TX
876	Principal Investigator	Timothy G. Hammond		M.B.B.S.,Durham Veterans Affairs Medical Center	Durham,NC
877	Principal Investigator	Albert Sacco	Jr.,Ph.D.	Northeastern University	Boston,MA
878	Principal Investigator	Alexander Mc			
879	Principal Investigator	Adriana Zagari	Ph.D.	University of Naples	Naples,Italy
880	Principal Investigator	Lode Wyns	Ph.D.	Free University	Brussels,Belgium
881	Principal Investigator	Juan Manuel Garcia Ruiz	Ph.D.	University of Granada	Granada,Spain
881	Co-Investigator(s)/Collaborator(s)	Dario Castagnolo		MARS Center,Napoli,ItalyE. Manas,University of Granada	Granada,Spain
882	Principal Investigator	Hiroaki Tanaka	Ph.D.	Japan Space Forum	Tokyo,Japan
882	Co-Investigator(s)/Collaborator(s)	Ari Yamanaka		Japan Space Forum	Tokyo,Japan
883	Principal Investigator	Lawrence De			

图 11 ISS 关联人员信息描述部分数据

issID	firstDomain	secondDomain
997	Technology Development and Demonstration	Radiation Measurement and Dosimetry
996	Technology Development and Demonstration	Imaging Technology
995	Technology Development and Demonstration	Communication and Navigation
994	Technology Development and Demonstration	Communication and Navigation
993	Physical Science	Materials Science
992	Physical Science	Fluid Physics
991	Physical Science	Fluid Physics
990	Physical Science	Fluid Physics
989	Human Research	Radiation Impacts on Humans
988	Human Research	Neurological and Vestibular Systems
987	Human Research	Neurological and Vestibular Systems
986	Human Research	Integrated Physiology and Nutrition
985	Human Research	Integrated Physiology and Nutrition
984	Human Research	Integrated Physiology and Nutrition
983	Human Research	Bone and Muscle Physiology
982	Human Research	Bone and Muscle Physiology
981	Human Research	Bone and Muscle Physiology
980	Human Research	Bone and Muscle Physiology
979	Educational Activity and Outreach	Student Developed Experiment
978	Educational Activity and Outreach	Educational Demonstration
977	Educational Activity and Outreach	Educational Demonstration
976	Educational Activity and Outreach	Educational Demonstration

图 12 ISS 关联学科与方向部分数据

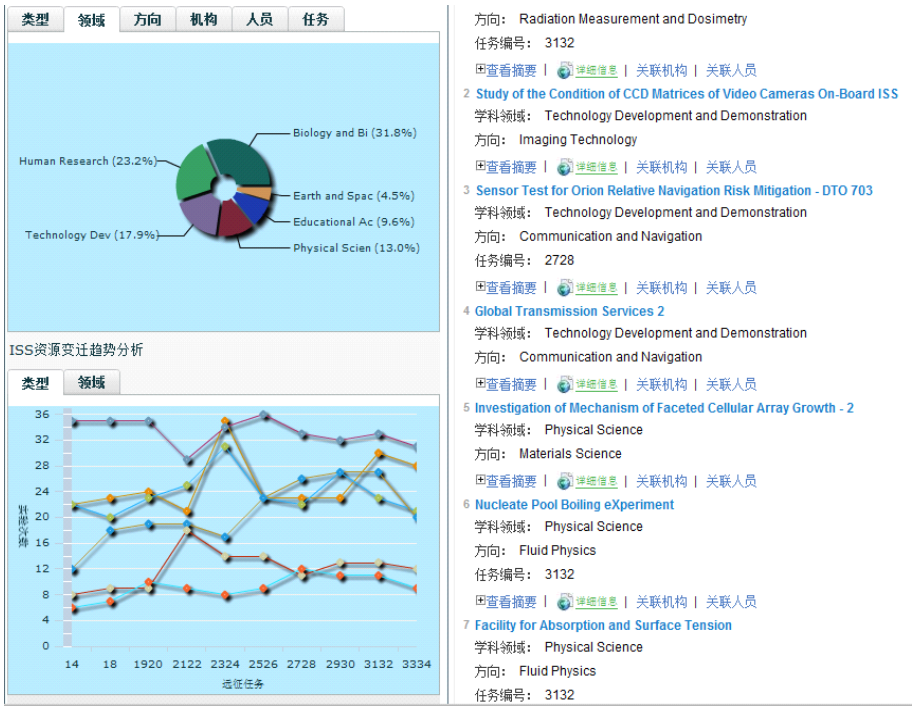


图 13 基于 ISS 实验与设施的监测应用

4 项目构建的实体名称规范库展示

为了向其它服务提供有效的实体名称规范库语料支撑，项目组还基于多个领域的科技监测服务的应用，构建了相应实体名称规范库。目前库中主要包含了人员、机构、会议、国家地理名称等多种类型的实体名称及其相应的属性描述、规范表达、别称表达等内容，共含有3221 条数据，随着监测应用的进一步推广，该数据量将不断得到累积，从而为其它服务提供支撑。图 14 展示了该实体规范库中的部分内容。

sourceAgency	normName	alias	objectCategory	objectType	country	description
Research Councils UK	Arts & Humanities Research Council	AHRC	Organization	ResearchInstitute	United Kingdom	AHRC Established in April 2005, the Arts and Humanit...
Research Councils UK	Biotechnology & Biological Sciences Research Cou...	BBSRC	Organization	ResearchInstitute	United Kingdom	BBSRC was established by Royal Charter in 1994 by i...
Research Councils UK	Medical Research Council	MRC	Organization	ResearchInstitute	United Kingdom	The Medical Research Council is a publicly-funded or...
Research Councils UK	Engineering & Physical Sciences Research Council	EPSRC	Organization	ResearchInstitute	United Kingdom	EPSRC is the main UK government agency for fundin...
Research Councils UK	Economic & Social Research Council	ESRC	Organization	ResearchInstitute	United Kingdom	ESRC are the UK's largest organisation for funding re...
Research Councils UK	Natural Environment Research Council	NERC	Organization	ResearchInstitute	United Kingdom	Since 1968, the North American Electric Reliability Co...
Research Councils UK	Science & Technology Facilities Council	STFC	Organization	ResearchInstitute	United Kingdom	The Science and Technology Facilities Council is an i...
Research Councils UK	RCUK Shared Services Centre Ltd	SSC;Shared Services Centre	Organization	OtherOrganization	United Kingdom	The UK´s seven Research Councils, working togethe...
Research Councils UK	Rick Rylance	Professor Rylance	Person	HeadOfAgency	United Kingdom	Research Councils UK Executive Group Chairman; Ar...
Research Councils UK	Ian Diamond	Professor Diamond	Person	HeadOfAgency	United Kingdom	the Chief Executive of UK Economic and Social Rese...
Research Councils UK	John Savill	John Savil	Person	HeadOfAgency	United Kingdom	CEO of the British Medical Research Council, Vice-Ch...
Research Councils UK	John Chisholm	John Chisholm	Person	HeadOfAgency	United Kingdom	Chairman of the British Medical Research Council
Research Councils UK	Paul Boyle	Professor Boyle	Person	HeadOfAgency	United Kingdom	UK Economic and Social Research Council Chief Exe...
Research Councils UK	Douglas Kell	Professor Kell	Person	HeadOfAgency	United Kingdom	UK Biotechnology and Biological Sciences Research ...
Research Councils UK	David Delpy	David Delpy	Person	HeadOfAgency	United Kingdom	UK Engineering and Physical Sciences Research Co...
Research Councils UK	Michael Sterling	Professor Sterling	Person	HeadOfAgency	United Kingdom	UK Science and Technology Facilities Council Presid...
Research Councils UK	John Womersley	Professor Womersley	Person	HeadOfAgency	United Kingdom	CEO of the UK Science and Technology Facilities C...
Research Councils UK	Edmund Wallis	Mr Wallis	Person	HeadOfAgency	United Kingdom	Chairman of the UK National Environment Research ...
White House	President Barack Obama	Barack H. Obama	Person	NationalLeader	America	Barack H. Obama is the 44th President of the United ...
Research Councils UK	Schools Policy Statement	SPS	Event	Project	United Kingdom	RCUK believe that working with schools can provide ...
Research Councils UK	Efficiency 2011-15: Ensuring Excellence with Impact	NULL	Event	Document	United Kingdom	This new report, Efficiency 2011-15: Ensuring Excelle...

图 14 实体规范库中的部分数据内容

5 总结

通过多个实际应用的测试，本项目提出的基于语义的实体名称规范方法具有可行性和有效性，但本项目中的实体名称规范与实体名称识别过程紧密结合在一起，是一个完整的文本处理流程，需要存储应用到在实体识别过程中的多种信息，对于规范文档建设、情报分析中

单一的对某两个实体的规范场景应用有所限制，这也是本项目目前实际应用中的一个限制之处，需要在下一步的研究中进一步拓展考虑对不同应用场景的普适性。另外，本项目对多种类型语义知识的深入融合利用挖掘还不够，对各类型语义知识在实际环境中各自的贡献效果没有分别进行实际的实验测评，主要是直接参考了现有相关研究中对各种语义知识的利用结果来设定相应的阈值，实体规范的准确率尚需要进一步提高。

第四部分 本研究的不足与下一步工作

1 本研究的不足

本研究在开展过程中，主要基于监测项目开展了多种实证应用研究，但对于其他的应用场景考虑不足，因此方法的实用性还有待进一步提升。在进一步考虑多种实证应用场景的基础上，才能更有效地设计出比较独立的技术、方法、流程，从而为其它第三方应用提供更好的共享资源。

此外，由于研究精力的限制，本研究着重于英文实体名称的规范方法研究，在中文方面未有进一步的深入探索，这也限制了本研究的进一步应用。

2 下一步的工作

下一步工作中，将在本研究的基础上，进一步基于我馆当前的规范文档构建、情报数据清洗等实证应用工作，进一步深化方法，提升研究的实用性与普适应，在提升方法的效率基础上，进一步设计出独立、灵活的可供第三方调用的资源库和软件包。同时，下一步研究中，本项目将进一步扩展对中文实体名称规范的研究，从而进一步提升研究的实用性。

参考文献

1 The GATE User Guide.[EB/OL]. [2011-02-12]. http://gate.ac.uk/sale/tao/tao.pdf

2 The Stanford Parser: A statistical parser.[EB/OL].[2011-02-12].
http://nlp.stanford.edu/software/lex-parser.shtml

3 What is WordNet? [EB/OL].[2011-02-12]. http://wordnet.princeton.edu/

4 YAGO2：A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. [EB/OL].[2012-02-12]. http://www.mpi-inf.mpg.de/yago-naga/yago/downloads.html

5 Computers & Math News [EB/OL]. [2011-12-12]. http://www.ScienceDaily.com/news/computers_math/

6 Wei Jiang, Yi Guan, Xiao-long Wang. Improving feature extraction in named entity recognition based on maximum entropy model, In: Proceedings of the fifth International Conference on Machine Learning and Cybernetics, Dalian[J], 2006(8): 13-16