面向地学领域的自动语义标注研究*

姚晓娜　祝忠明　王思丽
(中国科学院国家科学图书馆兰州分馆　兰州 730000)

【摘要】如何实现对大量信息资源的自动语义标注是建设基于语义网的数字图书馆的关键问题之一。以地学领域的专业文档为标注对象,基于地球科学术语语义网 SWEET 和文本工程通用框架 GATE 实现自动语义标注,并提出一种依据领域本体中属性的定义域和值域映射 RDF 三元组的方法,实验结果验证该方法的有效性。

【关键词】语义标注　领域本体　SWEET　GATE
【分类号】G250.76

Research on Automatic Semantic Annotation for Geosciences

Yao Xiaona　Zhu Zhongming　Wang Sili
(The Lanzhou Branch of National Science Library, Chinese Academy of Sciences, Lanzhou 730000, China)

【Abstract】How to realize the automatic semantic annotation for a large number of information resources is one of the key issues of the construction of digital library based on the Semantic Web. For the professional documents of geosciences, this paper realizes the automatic semantic annotation based on the Semantic Web for Earth and Environmental Terminology (SWEET) and the General Architecture for Text Engineering(GATE), and proposes a method to map the RDF triples according to the domain and range of a property in domain Ontology. The results of the experiment verify the effectiveness of the method.

【Keywords】Semantic annotation　Domain Ontology　SWEET　GATE

1　引　言

越来越多的研究人员将语义网技术引入数字图书馆领域[1,2],以提升数字图书馆的信息描述、组织和互操作能力。语义网利用形式化的本体对网络上的各种资源进行语义标注,使标注后的资源更适合机器处理[3]。语义标注分为手工标注和自动标注,由于手工标注需要耗费大量人力,无法满足大规模文档的标注需求,因此研究自动语义标注是非常必要的[4]。

目前,大多数研究集中在对一般网页的自动语义标注上,如文献[5]依赖一个预编译好的轻量级本体(Knowledge and Information Management Ontology, KIMO)和知识库,利用自然语言信息抽取技术,将文本中的人名、地名、机构名、日期等7类对象映射为 KIMO 中的实例。文献[6]通过机器学习的方法,利用预先标注好的训练语料库学习标注规则,根据得到的规则对其他文档进行标注。在数字图书馆领域也有对文献资源进行语义标注的研究,如文献[7]依照图书分类法建立知识本体,并对数据库中的图书信息进行标注。但由于图书分类法的层次简单,包含的概念有限,因此在对具体领域的专业文档进行标注时,不能充分表达文本的语义信息。文献[8]对金融领域网页的自动语义标注进行研究,采用依存句法分析技术对主谓宾关系进行识别,并建立与本体概念之间的映射

收稿日期:2013-03-06
收修改稿日期:2013-04-11
*本文系中国科学院国家科学图书馆青年人才前沿领域基金项目"基于学术产出挖掘的用户兴趣建模研究"(项目编号:Y200081001)的研究成果之一。

关系,最终形成 RDF 三元组。但由于实际文本中的句子语法形式多样,单一的主谓宾关系不足以表示概念间的关系。文献[9]提出利用本体学习技术标注关键词,基于语法结构和语义结构的对应性扩展标注范围,并依据 ACM 本体实现了计算机领域专业文档的语义标注。其中,语义结构中的语义距离是按照两个本体概念之间的相关度计算的,但在实际应用中,同一个句子中出现的本体概念并不一定都是相关的。

2　自动语义标注关键技术

(1)领域本体

领域本体在语义标注研究中具有重要作用,它为语义标注生成提供了领域中概念和关系的语义描述[10]。领域本体属于语义 Web 知识过程中的知识元过程,一般由领域专家在本体构建方法指导下完成。目前许多领域都有相应的较为通用的重量级的领域本体,如计算机领域的 ACM 本体、基因领域的 GO 本体等。这些领域本体涵盖了所描述领域的重要概念,语义信息比较丰富,并且已经得到领域专家的广泛认可,因此完全可以利用已有的成熟的领域本体对专业文档进行标注。

(2)命名实体识别

是指识别文本中具有特定意义的实体,主要包括人名、地名、机构名、专有名词等,这是目前信息抽取研究中最有实用价值的一项技术,根据 MUC 的评测结果,英文命名实体任务的 F - 指数(召回率与准确率的加权几何平均值,权重取 1)能达到 90% 以上。命名实体的识别方法主要有三类:基于词典的识别方法、基于规则的识别方法和基于统计机器学习的识别方法,在实际应用中,经常将多种方法综合使用。

(3)本体概念标注

实际上也属于命名实体识别,但是以领域本体为词典,识别文本中包含的领域概念。一般采用关键词匹配的方式进行概念识别,也有通过计算词语间语义距离的方法识别相关概念。由于领域本体中的名称常包含下划线或者以复合词的形式出现,如 Project_Name 和 ClimateChange,因此需要先对领域本体进行文本预处理,过滤下划线等字符,将复合词分割为词组,最终转化为可检索的语义词典,该词典以分词结果为入口,包含对应的类、属性、实例以及 URI 等语义信息。匹配的过程并不依据分词结果,而是采用最大模式匹配法获取文本中的字符串集合。匹配完成后,可以结合分词和词性标注结果对集合进行过滤。

(4)RDF 三元组表示

为了使得语义标注的结果最终能够被计算机处理,通常采用 RDF 三元组(资源 - 属性 - 属性值)对标注结果进行表示。目前许多研究采用依存句法分析技术对句子进行语法结构分析,将得到的依存关系对,如主谓宾等关系映射为 RDF 三元组。但是这种方式比较适合简单的陈述句,而专业文档中的句子语法形式多样,单一的主谓宾关系并不足以表示概念间的关系。此外,依存句法分析的准确率较低,标注结果并不理想。

3　地学领域本体 SWEET

由于标注对象是地球科学领域的专业文档,因此对地学领域的相关本体进行调研,最终选择地球与环境术语语义网(Semantic Web for Earth and Environmental Terminology,SWEET)[11]作为领域本体知识库。SWEET 是由美国 NASA 开发的一个地球科学本体项目,目标在于实现地球科学数据的互操作。SWEET 本体涵盖了地球科学领域的大部分概念,形成一个地球科学概念常识知识库,其中的概念划分为如下 12 个大类[12]:

(1)地球圈层(Earth Realm):描述地球圈层的构成,包括大气层(Atmosphere)、海洋(Ocean)、固体地球(Solid Earth)和相关子领域,如洋底(Ocean Floor)、大气边界层(Atmospheric Boundary Layer)等。

(2)非生命物质(Non - Living Substance):描述自然界的非生命物质,如粒子、电磁辐射和化合物等。

(3)生命物质(Living Substance):描述自然界中的动物和植物,来源于 GCMD 的生物圈(Biosphere)词表。

(4)过程(Process):描述非生物和生物物质变化的过程,如光合作用、冰川过程等。

(5)属性(Property):描述对象(非生物、生物和过程等)的物理或化学性质。

(6)单位(Units):采用 UniData 的 UDUnit 描述,并包含不同单位之间的转换因子。

(7)数值(Numerics):支持对各种数值范围(点、区间、平方等)和数值关系(大于、最大值等)的描述。

(8)时间(Time):基于数值概念,支持对时间范围(期间、季节、世纪和 1996 年等)和时间关系(前、后

等)的描述。

(9)空间(Space):基于数值概念,支持对空间范围(国家、南极洲、赤道和海湾等)和空间关系(上面、北部等)的描述。

(10)现象(Phenomena):用于定义瞬间事件,如飓风、地震、厄尔尼诺、恐怖事件等,常常包含其他的本体概念,如时间、空间、地球圈层、非生命物质和生命物质等。SWEET 本体中还包含最近 20 年发生的现象的实例。

(11)人类活动(Human Activities):用于描述人类参与的、对环境有影响的活动,如废物排放、捕鱼等。

(12)数据(Data):支持对数据集合的描述,包括数据的表示、存储、建模、格式、资源、服务及分布等概念。

SWEET 本体采用 OWL 语言开发,具有开放的网络资源,目前最新的版本是 SWEET2.2。本文采用的是 SWEET2.1 版本,包括 189 个子本体、5 161 个类、1 781个实例以及 576 个属性。

4 基于 SWEET 和 GATE 的自动语义标注

文本工程通用框架(General Architecture for Text Engineering,GATE)[13]是由英国谢菲尔德大学开发的一个开源文本挖掘软件,提供一系列可重用的自然语言处理的组件和类库,从而能够广泛地应用于各种文本处理任务。本文基于 GATE Java API,结合领域本体 SWEET,设计了一个面向地学领域专业文档的自动语义标注方案,标注过程如图 1 所示:

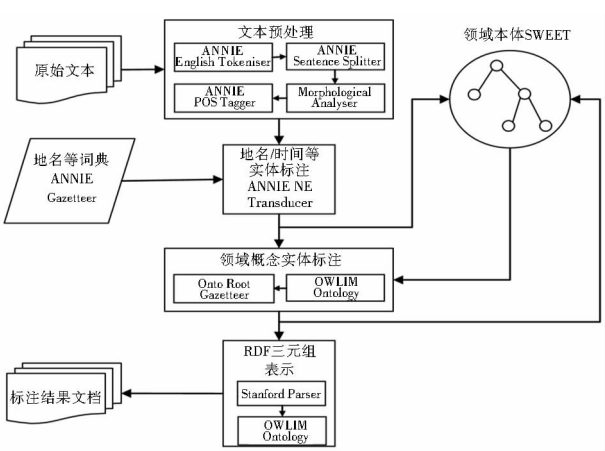

图 1 基于 SWEET 和 GATE 的自动语义标注过程

4.1 文本预处理

首先需要对原始的文本进行分句、分词及词性标注,此外,由于本文处理的是英文文档,还要对过去时态及第三人称单数等形式的词语进行原型标注。本文采用 GATE Java API 实现上述功能,分句使用 ANNIE Sentence Splitter,分词使用 ANNIE English Tokeniser,词性标注使用 ANNIE POS Tagger,原型标注使用 Morphological Analyser。

4.2 命名实体识别和标注

由于时间和空间是地学领域中非常重要的两个概念,因此本文主要对文本中包含的地名、机构名以及时间进行识别。具体实现时,通过 GATE Java API 中的 ANNIE NE Transducer 和 ANNIE Gazetteer 进行命名实体识别,其中,ANNIE NE Transducer 是一个基于 JAPE 规则的识别程序;ANNIE Gazetteer 是一个预定义词典,包含人名、地名等词汇;ANNIE 是一个基于英文的信息抽取组件,对英文中地名、时间等命名实体识别的准确率和召回率都已达到一定的高度。在识别完成后,还需要对识别出的实体进行语义化表示。由于 SWEET 本体中已经包含地理位置、机构以及时间等概念类,因此,只需要将实体表示成相关概念类的实例,并添加到领域本体中即可。

4.3 本体概念标注

以 SWEET 作为领域本体,对文本中的领域概念进行标注:首先采用 GATE 的本体接口 OWLIM Ontology 从本地文件中加载 SWEET 本体;通过 GATE 的本体语义词典工具 Onto Root Gazetteer 将本体转化为可检索的语义词典,并进行领域概念匹配;结合之前词性标注的结果对匹配得到的候选集合进行过滤,如只标注名词、动词及副词等,其中,对于相邻的名词性概念,作为新的复合概念添加到领域本体中;最后使用匹配概念的语义信息直接进行标注。

4.4 RDF 三元组表示

按照 RDF 资源描述框架将标注结果映射为 RDF 三元组,并生成语义标注结果文档。资源、属性和属性值的映射是 RDF 三元组表示的关键问题,但目前基于句法分析的方法并不能很好地确定属性类型。在对 SWEET 本体进行分析之后发现,该本体的属性主要按照不同概念类之间的关联关系进行定义,并且大部分的关联关系跟值域内容直接相关,如 hasTime、hasLocation 及 hasProperty 等,值域分别为 Time、Location 和 Property 类。因此,采用一种根据属性的定义域和值域来映射 RDF 三元组的方法,具体步骤如下:

(1)通过句法分析找出句子的中心词,如果中心词是已被标注的概念,则将中心词作为描述主体,即资源,否则在句法分析树中查找与中心词距离最近的概念作为资源,如果有多个,则选取在句子中的位置最近的那个。

(2)查找失败,则直接返回。

(3)查找成功,则依次将在该句子中出现的其他被标注的概念作为属性值,在 SWEET 本体查找定义域为资源所属类别和值域为属性值所属类别相匹配的属性。

(4)查找成功,生成相应的 RDF 三元组;否则将值域修改为属性值所属类别的父类,依次向上重新匹配,直到顶层类;如果找不到,则将定义域修改为资源所属类别的父类,依次向上重新匹配,直到顶层类。

(5)查找失败,生成属性为 hasRelation 的 RDF 三元组,该属性是本文自定义的属性,表示两个概念之间存在关联关系,定义域和值域可以是 SWEET 本体中的任意概念。

上述句法分析部分主要通过 GATE 的句法分析接口 Stanford Parser 实现,并采用 GATE 的本体接口 OWLIM Ontology 生成 RDF 标注结果文档。

4.5　实例说明

以下通过一个实例对本文的自动语义标注过程进行说明:

"This study analyzes the changes in glacier zones and snow composition of Glacier No. 1 in the Tianshan Mountains of China since 1961."

通过命名实体识别可得到 since 1961 为时间,Tianshan Mountains 和 China 为地名,本体概念标注可将 changes、glacier zones、snow composition 标注成本体中的概念。经过依存句法分析得到句法分析树如图 2 所示:

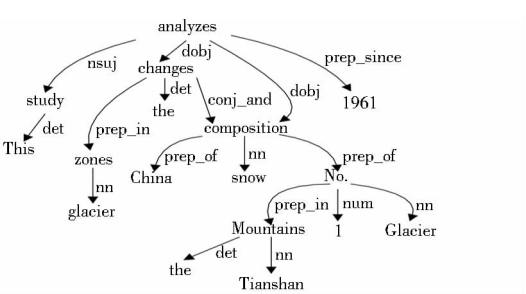

图 2　句法分析树示例

从句法分析树中可以看到,句子的中心词是 analyzes,距离中心词最近的已标注概念有 changes、composition 和 1961,比较它们在句子中的位置,最终确定 changes 为资源对象。对其他的依存关系不做分析,而是对之前识别出的概念依次进行属性匹配,其中,Tianshan Mountains 和 China 属于 Location 类,可找到值域为 Location 的属性 hasLocation;since 1961 属于 Time 类,可找到值域为 Time 的属性 hasTime;glacier zones 是一个基于 zone 的复合概念,属于 Zone 类,该类没有符合条件的属性,因此依据父类查找,可找到值域为 GeometricalObject 的属性 hasGeometricalObject;snow composition 是一个基于 composition 的复合概念,属于 Composition 类,该类没有符合条件的属性,因此需要依据父类查找,可找到值域为 Property 的属性 hasProperty。最终生成的标注结果如下:

```
<rdf:RDF xmlns:rdf = "http://www.w3.org/1999/02/22 - rdf -
    syntax - ns#" >
  <rdf:Description rdf:nodeID = "A1" >
    <rdf:subject rdf:resource = "http://localhost/sweet/individ-
        ual#Change1" / >
    <rdf:predicate rdf:resource = "http://sweet.jpl.nasa.gov/
        2.1/reprSpaceObject.owl#hasGeometricalObject" / >
    <rdf:object rdf:resource = "http://localhost/sweet/individu-
        al#Glacier_Zone" / >
  </rdf:Description >
  <rdf:Description rdf:nodeID = "A2" >
    <rdf:subject rdf:resource = "http://localhost/sweet/individ-
        ual#Change1" / >
    <rdf:predicate rdf:resource = "http://sweet.jpl.nasa.gov/
        2.1/prop.owl#hasProperty" / >
    <rdf:object rdf:resource = "http://localhost/sweet/individu-
        al#Snow_Composition" / >
  </rdf:Description >
  <rdf:Description rdf:nodeID = "A3" >
    <rdf:subject rdf:resource = "http://localhost/sweet/individ-
        ual#Change1" / >
    <rdf:predicate rdf:resource = "http://sweet.jpl.nasa.gov/
        2.1/propSpace.owl#hasLocation" / >
    <rdf:object rdf:resource = "http://localhost/sweet/individu-
        al#China" / >
  </rdf:Description >
  <rdf:Description rdf:nodeID = "A4" >
    <rdf:subject rdf:resource = "http://localhost/sweet/individ-
        ual#Change1" / >
    <rdf:predicate rdf:resource = "http://sweet.jpl.nasa.gov/
        2.1/propSpace.owl#hasLocation" / >
    <rdf:object rdf:resource = "http://localhost/sweet/individu-
```

```
        al#Tianshan_Mountains"/>
    </rdf:Description>
    </rdf:Description>
    <rdf:Description rdf:nodeID="A5">
      <rdf:subject rdf:resource="http://localhost/sweet/individ-
        ual#Change1"/>
      <rdf:predicate rdf:resource="http://sweet.jpl.nasa.gov/
        2.1/reprTime.owl#hasTime"/>
      <rdf:object rdf:resource="http://localhost/sweet/individu-
        al#Since_1961"/>
    </rdf:Description>
</rdf:RDF>
```

5 实验及结果分析

本文从 Web of Science 数据库获取 50 篇地学领域文献的元数据(标题、摘要、关键词等),先在 GATE 可视化界面中使用人工方法进行标注,再使用基于 GATE Java API 开发的自动语义标注程序进行标注,并将两组标注结果进行对比分析。

首先是实体识别的结果。本文采用查准率、查全率和综合评价指标 F1 - Measure 三个指标对标注结果进行分析。按照概念类别分别计算各指标,结果如表 1 所示:

表 1 实体识别结果

概念类别	查准率 P(%)	查全率 R(%)	F1 - Measure(%)
地球圈层	82.11	78.20	80.11
非生命物质	70.79	44.62	54.73
生命物质	77.43	29.01	42.21
物理过程	70.02	52.69	60.13
物理属性	80.15	65.21	71.91
单位	91.00	19.97	32.75
数值	76.00	16.08	26.54
时间	90.20	89.37	89.78
空间	82.77	67.64	74.44
现象	74.34	34.00	46.66
人类活动	76.99	44.25	56.20
数据	89.65	20.43	33.28
平均值	80.12	46.79	59.08

可以看出,本文的自动语义标注方法查准率较高,平均值可以达到 80.12%,但是查全率较低,平均值只有 46.79%。主要原因在于:SWEET 本体包含的主要是地学的上层概念,某些类别,如生命物质的下层概念较少,这需要对 SWEET 本体进行扩充或者结合相关地学词典进行识别;另外,自然语言与本体中对同一概念的表述方式不同,如 kilometer 常缩写为 km,这就需要定义相应的映射规则。最终的综合评价指标 F1 - Measure 值为 59.08%,说明本文的方案具备一定的可用性。

在对 RDF 三元组关系的标注结果进行分析时,将 50 篇文档分为 5 组,每组 10 篇,分别采用本文方法和基于主谓宾关系的方法进行标注,并对标注数量进行对比,结果如图 3 所示:

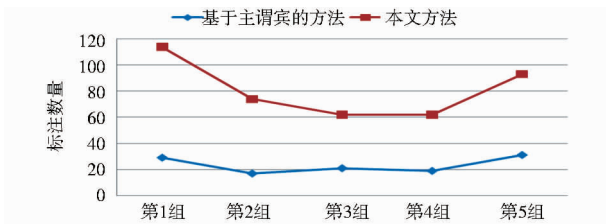

图 3 标注数量比较

通过对比可以看出,本文方法在标注数量上有显著提高。本文方法和基于主谓宾的方法对概念间的关系定义不同,无法从标注的准确性方面对两者进行比较分析。但由于本文方法的 RDF 三元组是根据领域本体中属性的定义域和值域进行映射的,识别出的关系来自于地学领域本体,具有地学领域的特性,因此与主谓宾的方法相比,本文方法更适合地学领域,能有效地识别地学领域专业文档中的实体关系。

6 结 语

语义标注的准确性依赖于本体的成熟度。因此,在对某一研究领域的专业文档进行语义标注时,所采用的领域本体直接决定了标注的效果。本文采用的地学领域本体 SWEET 涵盖地球科学领域的大部分概念,将其应用到地学领域专业文档的自动语义标注时也取得较好的效果。在进行 RDF 三元组表示时,由于目前基于依存句法分析的标注效果不好,本文提出一种根据领域本体中属性的定义域和值域来确定属性的映射方法,并通过实验验证了该方法的有效性。目前本文方法尚处于实验阶段,标注效率还有待提高,基于属性的映射方式也存在一定的局限性。如何提高标注效率并进行系统性的应用,以及如何突破现有的局限性,将是笔者下一步的工作。

参考文献:

[1] 张晓林. Semantic Web 与基于语义的网络信息检索[J]. 情报学报,2002,21(4):413 - 420. (Zhang Xiaolin. Semantic Web

and Semantic – based Networked Information Retrieval[J]. *Journal of the China Society for Scientific and Technical Information*, 2002,21(4):413 – 420.)

[2] 侯集体,程慧荣. 近年来国外关于语义 Web 的数字图书馆研究进展[J]. *图书情报工作*,2011,55(3):37 – 40,115. (Hou Jiti, Cheng Huirong. Review of Researches on Digital Library Based on Semantic Web Abroad[J]. *Library and Information Service*,2011, 55(3):37 – 40,115.)

[3] Dill S, Eiron N, Gibson D, et al. SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation[C]. In: *Proceedings of the* 12*th International Conference on World Wide Web*. New York:ACM,2003:178 – 186.

[4] 沙丽华. 面向领域文档的语义标注方法研究[D]. 长春:吉林大学,2009. (Sha Lihua. Research on Semantic Annotation for Domain Documents[D]. Changchun: Jilin University,2009.)

[5] Popov B, Kiryakov A, Kirilov A, et al. KIM – Semantic Annotation Platform[C]. In: *Proceedings of the* 2*nd International Semantic Web Conference*, Florida, USA. 2003:834 – 849.

[6] Vargas – Vera M, Motta E, Domingue J, et al. MnM: Ontology Driven Semi – Automatic and Automatic Support for Semantic Markup[C]. In: *Proceedings of the* 13*th International Conference on Knowledge Engineering and Knowledge Management*. Springer – Verlag,2002:379 – 391.

[7] 田欣. 基于知识本体的图书馆语义检索系统模型研究[J]. 情报杂志,2006,25(6):78 – 81. (Tian Xin. Research on the Model of Semantic Search System in Library Based on Knowledge Ontology [J]. *Journal of Information*,2006,25(6):78 – 81.)

[8] 荆涛,左万利,孙吉贵,等. 中文网页语义标注:由句子到 RDF 表示[J]. *计算机研究与发展*, 2008, 45(7):1221 – 1231. (Jing Tao, Zuo Wanli, Sun Jigui, et al. Semantic Annotation of Chinese Web Pages: From Sentences to RDF Representations[J]. *Journal of Computer Research and Development*, 2008, 45(7): 1221 – 1231.)

[9] 魏墨济,于涛. 基于领域本体的专业文档语义标注方法[J]. *计算机应用*,2011,31(8):2138 – 2142. (Wei Moji, Yu Tao. Professional Literature Annotation Method Based on Domain Ontology [J]. *Journal of Computer Applications*, 2011, 31(8): 2138 – 2142.)

[10] Sánchez D, Isern D, Millan M. Content Annotation for the Semantic Web: An Automatic Web – based Approach[J]. *Knowledge and Information Systems*,2011,27(3):393 – 418.

[11] SWEET Ontologies[EB/OL]. [2013 – 03 – 06]. http://sweet.jpl.nasa.gov/sweet.

[12] Raskin R, Pan M, Mattmann C. Enabling Semantic Interoperability for Earth Science Data[EB/OL]. [2013 – 03 – 06]. http://esto.nasa.gov/conferences/estc2004/papers/a5p1.pdf.

[13] GATE. A General Architecture for Text Engineering[EB/OL]. [2013 – 03 – 06]. http://gate.ac.uk.

(作者 E – mail:yaoxn@llas.ac.cn)