图书馆论坛
Library Tribune
ISSN 1002-1167,CN 44-1306/G2

《图书馆论坛》网络首发论文

题目：命名实体识别在数字人文中的应用——基于 ETL 的实现
作者：朱武信，夏翠娟
收稿日期：2019-12-01
网络首发日期：2019-12-16
引用格式：朱武信，夏翠娟．命名实体识别在数字人文中的应用——基于 ETL 的实现[J/OL]．图书馆论坛.
http://kns.cnki.net/kcms/detail/44.1306.G2.20191213.0827.004.html

*本文系国家社会科学基金项目“面向数字人文研究的图书馆开放数据体系构建与服务模式设计研究”（编号：18BTQ027）研究成果。

命名实体识别在数字人文中的应用
——基于 ETL 的实现*

朱武信，夏翠娟

摘 要：近年上海图书馆通过数字人文搭建多个知识服务平台。知识服务平台通过关联数据，以知识图谱、GIS 等多种展示方式提供服务。关联数据提供专业服务对基础数据提出新要求，如数据本体化，具体到人名、地名、时间等实体；再如数据保留关联性，以关联数据形式存储。在新的数据要求与数据量日益增加的背景下，传统通过人力来加工数据的方法，或以简单的实体提取，无法满足需求。为解决此问题，研发命名实体识别工具，以上图的关联数据为词典，结合 HANLP 技术，实现文本的实体挖掘。工具投入使用后，通过工具对数据批量进行实体识别，改进了数据处理流程，缩短了数据加工的周期。

关键词:命名实体识别，关联数据，数字人文，文本标注

引用本文格式：朱武信，夏翠娟.命名实体识别在数字人文中的应用——基于 ETL 的实现[J].图书馆论坛，2020

Application of Name Entity Recognition in Digital Humanities
——Based on ETL Implementation

ZHU Wuxin,XIA Cuijuan

Abstract: In recent years, Shanghai Library has built a number of knowledge service platforms through digital humanities. The knowledge service platform provides services for readers and experts through linked data in various ways such as knowledge graph and GIS. Linked data raises new requirements on basic data while providing professional services. There are two requirements. Firstly, the data must be ontological such as the names of people, places, time and so on. Secondly, the data needs to be preserved and stored in the form of linked data. In the condition of new data requirements and increasing data volumes, the traditional methods of processing data which is manpower or simple entity extraction have been unable to meet the requirements of the Shanghai Library. In order to solve this problem, this paper developed a named entity recognition tool which used HANLP technology to achieve text mining through the linked data of Shanghai Library as a dictionary. After the tool is used, the data is identified by the tool, which improves the data processing flow and shortens the data processing cycle.

Keywords:named entity recognition, linked data, digital humanities, text annotation.

0 引言

命名实体识别 NER（Named Entity Recognition）是自然语言处理 NLP（Natural Language Processing）的一部分内容，是指从文本中提取出命名实体，命名实体是指人名、地名、时间等信息[1]。图书馆用 NER 进行数据挖掘，从摘要、小传、全文中提取大量的命名实体，为构建知识图谱与支持数字人文研究和服务打下了基础。学界在命名实体应用方面作了很多研究，提出了很多新方法，包括规则提取、关系提取、正则提取、神经网络、机器学习等。

上海图书馆（下称“上图”）有大量数字化馆藏资源，其挖掘离不开 NER 技术的推动。在构建数字人文平台初期，使用各种工具与方法进行数据加工，包括 OpenRefine、基于 Python 正则提取、人工等。这些方法解决了一些问题，但也存在不足，一方面识别效率低、人工成本高，另一方面识别的内容仅仅是文本，后续要和其他数据进行关联，还需投入更多人力、物力和时间。为解决上述问题，本研究研发基于数字人文与汉语言处理包 HANLP（Han Language Processing）技术的命名实体识别工具。HANLP 是一个在 github 平台上开放的 NLP 开源工具包，开发语言是 JAVA，提供中文分词、词性标注、命名实体识别、依

存句法分析等功能。本研究主要采用基于 HMM 模型进行分词模型训练、最短路分词与依存句法分析中的基于神经网络的高性能依存句法分析器[2]。

命名实体识别工具的词典源于上图数字人文知识库。选用之有 3 个原因：（1）NER 要提取的实体信息与数字人文所定义的人名、地名、时间、事件不谋而合。（2）2014 年以来上图通过本体建模方法搭建了多个数字人文平台与知识库，有大量的数据基础[3]。上图数字人文平台在功能上分为两类，一是家谱知识服务平台、盛宣怀档案知识库等以提供文献服务为主的文献知识库；二是人名、地名、时间、事件为一体的基础知识库。本研究选择作为词典的知识库指的是上图数字人文基础知识库。（3）上图的数据是关联数据，具有语义性，将其作为词典，则命名实体识别的结果也具有关联数据的特性，可以通过本体获取更多相关信息。

本研究结合数字人文与 NER，研发基于关联数据的命名实体识别工具，通过此工具对文本进行数据挖掘，提取相关人名、地名等实体信息，优化上图 ETL 流程。

1 现状调研

关联数据概念于 2006 年由蒂姆•伯纳斯一李提出[4]，互联网上已发布大量数据集。国内外大学、图书馆通过数字人文构建了知识库与知识图谱，比较著名的有哈佛大学中国历代人物传记资料库 (CBDB)、基于维基百科的 DBpedia、OCLC 的虚拟国际规范档、复旦大学的 CHGIS 系统、上海图书馆的数字人文开放平台等。上图在数字人文领域的探索取得较多成果，比如构建家谱知识服务平台，提供人物、地名、时间相关的基础知识平台。2017 年上图搭建人名规范知识库，运用关联数据技术发布了近 130 万人名实体[3]。地名基础数据包含省、市、县共 1，744 条信息；2019 年发布上海地名志信息，包括 2，264 条马路三元组。这些实体已经对外开放服务。

命名实体识别在数字人文中的应用，国外起步较早。1911 年国外就提出命名实体识别的概念。Palo N 研发 DBpedia Spotlight 工具，通过质量测量方法与 DBpedia 的本体进行自动标注文本[5]，验证了利用实体进行命名实体识别的可行性，该工具在互联网上开放给大众使用。2011 年，朱锁玲在关于方志内容的命名实体识别应用研究中指出，由于语言差异，命名实体识别对不同语言有难点与差异；设计了古籍地名识别系统，通过上下文规则识别方法对地方志内的地名进行识别[1]。2014 年，Ferragina P 等发布基于 TagMe 算法，以维基百科实体为基础，实现快速标注文本短语工具，标注结果信息丰富，且与维基百科信息互相关联[6]，但所用知识库仅支持英语。2014 年，Usbec 等提出将 AGDISTIS 方法用于命名实体识别，以标签与 HITS 进行提取[7]。同年， Speck R 研发 FOX 工具，Fox 通过实体关联技术与 EL 算法，实现文本转换提取出 RDF（Resource Descripition Framework）数据，F 值(F-Measure)达 95.23%[7]。张海楠等[9]、Lample[10]在 NER 识别中，提出运用神经网络，通过非监督学习进行识别，以降低人工成本，虽然识别度高，但提取的文本仅是字符串，缺少语义性与关联性。

上述命名实体识别工具识别效果好，方法具有借鉴作用，但无法满足上图所需的场景：一是上述工具的词典与上图所需加工的文本场景不匹配[11]；二是上图需要识别工具根据人名、地名、时间、事件、自定义标签进行识别[12]；三是上图需要识别结果是关联数据，与已有关联数据形成关联。

2 命名实体识别工具需求与设计

2.1 命名实体识别系统需求

上图在众多基础知识库与服务平台实施过程中，通过 OpenRefine 工具与人力处理方法，从大量文本中进行数据加工与实体提取，取得了一定成果，但需要大量人力、时间，

尤其是在处理新数据时，人名、地名实体重复出现，需要再次加工。为改善数据加工，降低成本，加快数据处理速度，快速将识别结果转为关联数据，本研究基于上图基础语义知识库，在 ETL 加工环节增加命名实体识别功能。其主要特征有：对中文文本进行实体识别，命名实体识别词典基于上图数字人文基础知识库；识别实体与上图数字人文知识库的关联数据关联；可识别不同类别的实体，包括人名、地名、机构、姓氏，可自定义新的分类。

2.2 系统架构设计

本研究开发的命名实体识别是基于上图已有关联数据作为识别词典，通过命名实体识别算法对文本中的内容进行识别，识别结果与上图关联数据进行对应。系统架构见图 1。

图 1 命名实体识别系统架构图

（1）输入层。输入层是以需识别的文本为输入参数，通常是文献中的摘要、小传信息。在输入层对识别内容的标签进行预选择，如人名、地名、姓氏，以此根据不同需求进行特定内容的识别。

（2）识别层。识别层是命名实体识别的核心模块，通过关联技术的本体模块与命名实体识别算法模块的结合，实现对输入文本的识别。由于其识别结果是关联数据，一定程度解决部分命名实体识别工具识别结果仅是字符串的问题，具有关系发现的特性。由于是在上图已知的数据源中识别，精准的识别对命名实体消歧起到了改善作用[3]。

（3）输出层。输出层包括识别结果的展示与下载，当识别完成后，会展示文本的识别结果，展示结果添加了关联数据的 URI。通过 URI，此文本与上图的数据形成关联，通过上图 API 接口获取更多的内容信息。

2.3 识别词典设计

本研究使用的识别词典主要来自上图，包括上图的人名规范库、地理名词库、上海年华事件库，3 个知识库分别对应关联数据中的人名、地名、事件[12]。使用上图关联知识库的主要原因包括：（1）上图知识库数据由语义网 RDF 框架组成，通过三元组形式构建本体。正因为以本体作为词典进行识别，识别结果也是本体。（2）上图人名规范库的人名本体有 130 万个，其来源于上图馆藏。因为上图搭建了大数据级别的人名关联数据，所以能作为命名实体识别的词典。（3）上图知识库是开放的，提供通用 API 接口，支持 JSON、XML 等格式，调用方便，兼容性好[11]。关联数据的特征是每个本体都有一个 URI 标示，数据以三元组形式进行描述。将本体作为识别词典，当识别的实体与本体形成关联，则能通过关联数据的本体结构，获取文本之外的信息。例如，识别出一个人名实体，则通过关联数据可以获取此人的籍贯、朝代、年龄信息。通过关联获取的信息，一方面丰富了识别

内容，另一方面也为识别结果的消歧提供了依据。

2.4 命名实体识别功能设计

本研究命名实体识别流程见图 2。下文结合样例对上述过程进行说明。

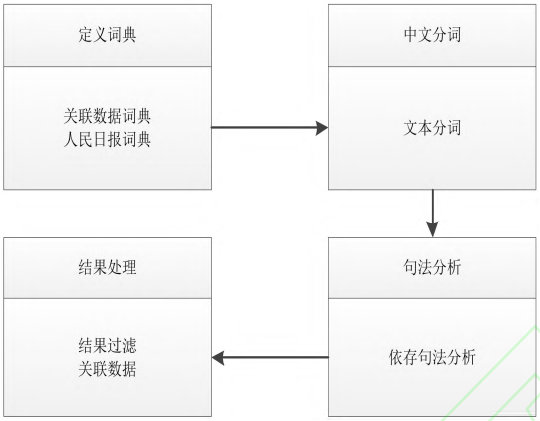

图 2 命名实体识别流程图

（1）定义词典。工具在识别前，首先引入 2 部词典作为语料：1998 年的人民日报语料库、上图关联数据词典。上图关联数据词典包含人名、机构、姓氏，其中人名词典近 130 万人名、姓氏 607 个、机构 42 个。

（2）中文分词。中文分词通过 HANLP 提供的基于隐马尔可夫模型 HMM-Bigram 模型对输入文本进行分词。使用 HANLP 的主要原因是其对命名实体识别、机器学习算法进行封装，使用便捷。例如，“长江剧场位于黄河路 35 号，原名卡尔登大戏院”这段话，通过分词得到的结果是“长江剧场/名词……卡尔登大戏院/名词”。

（3）句法分析。句法分析使用 HANLP 提供的基于神经网络的高性能依存句法分析，依存句法分析是对文本的内容进行关系标注。关系有 15 种，包括主谓关系、动宾关系、间宾关系等[2]。引入句法分析主要是为了做过滤操作。例如，“长江剧场位于黄河路 35 号”，在未引入词典的情况下，通过中文分词会提取到“长江，剧场”，通过依存句法分析对语句进行分析，得到“长江”与“剧场”是定中关系，算法会排除“长江”这个识别结果。

（4）结果处理。结果处理包含结果过滤与数据关联。结果过滤主要是将中文分词的结果与句法分析的结果进行过滤，让实体结果进一步准确。数据关联是将识别的结果与上图本体一一匹配与关联。识别实体通过上图 API 接口获取更多相关信息[13]。

3 命名实体识别工具的实现

3.1 命名实体识别实现

（1）关联数据词典转换实现。关联数据的词典转换方法实现见图 3。图 3 是以人名为例，首先从 RDF 数据中提取人名的名称与 URI，将其以词典的要求进行转换，再通过 HANLP 提供的自定义词典方法将命名实体添加到词典。提取人名的作用是作为词典的语料，URI 的作用是保留关联数据的关联性，最终的识别结果可以通过 URI 来获取关联数据的其他信息。

图 3 关联数据词典转换实现图

（2）命名实体识别实现。伪代码见图 4。方法以文本 A 为输入，先加载关联数据词典，通过词典识别方法对输入文本进行命名实体识别，得到基于词典的命名实体识别结果 B。通过依存句法分析 A 获得结果 C，结果 C 主要记录的是定中名词与状中名词，以此在结果 B 中排除状中名词与定中名词，最终生成的就是命名实体识别实体。

```
Begin
输入 识别文本A
Do 引入词典
Do B = 词典识别(A)
Do C = 依句法分析(A)
For(i in B){
i not in C
D.put(i)
}
Return D
End
```

图 4 命名实体识别伪代码

3.2 NER 工具效果

NER 工具见图 5。图中展示的是以年华人名为词典进行识别的结果。年华人名主要包含出身年月介于 1840—1950 年，近 7 万多名人。以此人名为词典对输入文本进行人名识别。 选用年华人名词典的主要原因是以其时间与输入事件的时间吻合，通过此词典提高结果的 准确率与召回率。实体识别结果如图所示，以橙色标注的是识别出的实体，其中的数字代 表的是识别对应的个数，通过单击实体，可以跳转到上图人名规范库的对应实体，从而获 取此实体更详细的信息。

图 5 实体识别功能展示图

3.3 实体识别工具效果对比

分别对上图命名实体识别工具、人工、Bosonnlp 工具进行比较，共用 10 组数据。综合来看，工具在降低少量准确率的前提下，可以对文本进行快速处理，这是人工无法比拟的。上图识别工具识别的结果是关联数据，其丰富性、关联性、可挖掘性远胜于人工与 Bosonnlp 所识别的结果。

表 1 实体识别效果对比

	人力	上图识别工具	Bosonnlp
所用时间	耗时长	耗时短	耗时短
识别结果	准确	较准确	较准确
识别结果类型	文本	关联数据	文本

4 结论及展望

上图研发的命名实体识别工具在 ETL 数据处理过程中起到了很大作用，弥补了上图没有命名实体识别的短板。其主要特色包括：（1）实现了基于数字人文词典的命名实体识别，识别的实体不再是简单的字符串，而是关联数据。关联技术与命名实体识别技术形成互补，命名实体识别可以在更多文本中挖掘关联数据，关联数据增强了识别结果的质量。（2）命名实体识别加强了 ETL 功能，数据处理得到改善。在大量文本中，通过工具可以快速识别其中的实体，在其识别的基础上加入部分人工，可以更高效率地获得高质量数据。

本研究实现的命名实体识别也有需要改进的地方：（1）基于已知数据进行挖掘，把不在词典中的命名实体过滤了，在今后的功能设计中应引入新的工作流来处理这些被过滤的命名实体。既能对这些命名实体进行发现，又能转换成关联数据。（2）在中文词性分析上有欠缺，对文本中挖掘实体的词性分析还需要重新梳理，缩小范围，以提高实体的准确度。

参考文献

[1] 朱锁玲.命名实体识别在方志内容挖掘中的应用研究[D].南京：南京农业大学, 2011.

[2] He H. HanLP: Han Language Processing[J]. URL: https://github. com/hankcs/HanLP, 2014.

[3] Xia C, Liu W. Name Authority Control in Digital Humanities: Building a Name Authority Database of Shanghai Library[J]. International Journal of Librarianship, 2018, 3(1): 21-35.

[4] 刘炜.关联数据：概念，技术及应用展望[J].大学图书馆学报, 2011（2）: 5-12.

[5] Frontini F, Brando C, Ganascia J G. Semantic web based named entity linking for digital humanities and heritage texts[C]. 2015.

[6] Ferragina P, Scaiella U. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)[C]//Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010: 1625-1628.

[7] Usbeck, Ricardo, et al. "AGDISTIS-graph-based disambiguation of named entities using linked data." International Semantic Web Conference. Springer, Cham, 2014.

[8] Speck R, Ngomo A C N. Named Entity Recognition using FOX[C]//International Semantic Web Conference (Posters & Demos). 2014: 85-88.

[9] 张海楠，伍大勇，刘悦，等.基于深度神经网络的中文命名实体识别[J].中文信息学报, 2017, 31(4): 28-35.

[10] Lample, Guillaume, et al. "Neural architectures for named entity recognition." arXiv preprint arXiv:1603.01360 (2016).

[11] Ritter, Alan, Sam Clark, and Oren Etzioni. "Named entity recognition in tweets: an experimental study." Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2011.

[12] Cuijuan X. The opening and application of Chinese historical geography data in digital humanities projects of libraries[J]. Journal of Library Science in China, 2017, 43(2): 40-53.

[13] Mendes P N, Jakob M, García-Silva A, et al. DBpedia spotlight: shedding light on the web of documents[C]//Proceedings of the 7th international conference on semantic systems. ACM, 2011: 1-8.

作者简介：朱武信（通信作者，ORCID：0000-0003-3874-1018，wxzhu@libnet.sh.cn），上海图书馆系统网络中心助理工程师；夏翠娟，上海图书馆系统网络中心高级工程师。

收稿日期：2019-12-01

（责任编辑：沈丽霞）