基于四份区域地质调查报告构建的命名实体识别试验数据集研发

马 凯[1]，田 苗[1]，谭永健[1]，王 曙[4]，谢 忠[2,3]，邱芹军[2,3,*]

1. 三峡大学计算机与信息学院，宜昌 443002；
2. 中国地质大学（武汉）计算机学院，武汉 430074；
3. 国家地理信息系统工程技术研究中心，武汉 430074；
4. 中国科学院地理科学与资源研究所，资源与环境信息系统国家重点实验室，北京 100101

摘 要：区域地质调查报告是全面反映区域地质调查工作成果的重要技术文件。目前全国地质资料馆已经积累了海量的地质成果报告，对其进行信息抽取和挖掘可以充分挖掘现有报告的隐含价值，促进新知识的发现。本文面向自然语言处理领域的命名实体识别任务，构建了基于四份区域地质调查报告的命名实体识别试验数据集，该数据集可以用于训练和测试地质命名实体模型。数据集共包含四份区域地质调查成果报告，对地质时间、地质构造、地层、岩石、矿物和地点六类典型的地质命名实体进行了标注，对数据集分别进行了一致性检验、测试、评估等工作，保证了数据集的质量。数据集大小为 4.84 MB，存储格式为.txt 文本。

关键词：区域地质调查报告；命名实体识别；一致性检验；测试；评估

DOI: https://doi.org/10.3974/geodp.2022.01.11

CSTR: https://cstr.escience.org.cn/CSTR:20146.14.2022.01.11

数据可用性声明：

本文关联实体数据集已在《全球变化数据仓储电子杂志（中英文）》出版，可获取：https://doi.org/10.3974/geodb.2021.09.04.V1 或 https://cstr.escience.org.cn/CSTR:20146.11.2021.09.04.V1.

1 前言

近年来，随着“深时数字地球国际大科学计划”的提出，构建地学知识库，对海量的地学文本数据进行命名实体识别、关系提取等结构化信息抽取工作，实现地质文本知识的深度挖掘已经成为了必然趋势。文本知识挖掘离不开高质量的语料库数据集的支撑。目前国内已有针对地质时间识别[1]与岩石实体识别[2]的单类型、小规模的地质命名实体识别语料库，尚缺少大规模标注的多实体类型的中文地质语料库。本数据集抽取了尼玛区幅 H45C001003 1/25 万区域地质调查报告[3]，治多县幅 I46C003004 1/25 万区域地质调查报告[4]，

收稿日期：2021-08-26；**修订日期**：2021-12-30；**出版日期**：2022-03-25

基金项目：国家自然科学基金（42050101，41871311，U1711267）

***通讯作者**：邱芹军，国家地理信息系统工程技术研究中心，qiuqinjun@cug.edu.cn

数据引用方式：[1] 马凯，田苗，谭永健等. 基于四份区域地质调查报告构建的命名实体识别试验数据集研发[J]. 全球变化数据学报, 2022,6(1): 78–84. https://doi.org/10.3974/geodp.2022.01.11. https://cstr.escience.org.cn/CSTR:20146.14.2022.01.11.

[2] 马凯，田苗，谭永健等. 基于四份区域地质调查报告构建的命名实体识别试验数据集[J]. 全球变化数据仓储电子杂志，2021. https://doi.org/10.3974/geodb.2021.09.04.V1. https://cstr.escience.org.cn/CSTR:20146.11.2021.09.04.V1.

金牛镇幅高桥幅 H50E013003 1/5 万区域地质调查报告[5]，阳春县幅 F49C002003 1/25 万区域地质调查成果报告[6]四份区域地质调查报告中的文本数据，通过预处理和标注、测试、评估等相关流程构建而成。

本数据集主要关注地质时间、地质构造、地层、矿物、岩石和地点 6 种命名实体类型。地球表层的岩层和岩体，在形成过程及形成以后，都会受到各种地质作用力的影响，有些大体上保持了形成时的原始状态，有些则产生了较大形变。它们具有复杂的空间组合形态，即各种地质构造，其中断裂和褶皱是地质构造的两种最基本形式[7]。而地质时间的确立则是我们研究地壳历史的依据。

地层是以成层的岩石为主体，狭义的地层专指已固结的成层的岩石，有时也包括尚未固结成岩的松散沉积物[8]。依照沉积的先后，早形成的地层居下，晚形成的地层在上，这是地层层序关系的基本原理，称为地层层序律[9]。矿物是地壳中的化学元素在各种地质作用下所形成的，具有一定化学成分和物理性质的自然均质体，它们是组成岩石和矿石的基本单位[9]。矿物在地壳中常以集合的形态存在，这种集合体可以由一种，也可以由多种矿物组成，这在地质学中被称为岩石[7]。另外考虑到地点作为出现在文本中的重要的空间参考，数据集把地点也作为一类实体进行了标注。通过对四份地质报告中六类实体的标注，分析了各份报告中实体的统计特征，并对语料库数据集进行了一致性检验、测试、评估等工作，保证了数据集的质量，本数据集可为地质领域的命名实体识别、关系抽取以及知识图谱的构建提供重要的数据基础。

2　数据集元数据简介

《基于四份区域地质调查报告构建的命名实体识别试验数据集》[10]的名称、作者、地理区域、数据年代、数据集组成、数据格式、基金项目、数据出版与共享服务平台、数据共享政策等信息见表 1。

3　数据采集过程与方法

整个数据集的采集主要分为两个步骤，分别为选择代表性区域的地质调查成果报告和对所选择的报告进行命名实体的标注。

数据来源于尼玛区幅、治多县幅、金牛镇幅高桥幅、广东阳春县幅的区域地质调查成果报告中的文本，地理范围共涉及西藏、广东、湖北、青海四个省份。成果报告是全面反映地质勘查工作成果的重要技术文件，在每一份地质成果报告中，都涉及了地质时间、地质构造、地层、岩石、矿物和地名等命名实体，对其进行高效准确的识别和抽取是实现地质知识挖掘的基础。这些命名实体也是本数据集标注的对象，标注好的数据集可以用来训练和检验相关实体识别模型。

本文开发了专用的标注工具，并制定了一套标注规则以规范歧义性实体的标注。数据标注采用半自动方法，采用领域专家与团体交叉标注模式，在软件辅助下通过人工方式开展标注工作。整个标注流程分为以下四个阶段：

（1）制定标注规范阶段：根据地质领域命名实体的语法语义特征结合地质领域专家意见制定标注规范。

表 1 地质领域命名实体识别数据集

条目	描述
数据集名称	基于四份区域地质调查报告构建的命名实体识别试验数据集
数据集短名	NERdata
作者信息	马凯 ABH-2687-2021，三峡大学计算机与信息学院，makai@ctgu.edu.cn 田苗 ABH-2542-2021，三峡大学计算机与信息学院，tianmiao@ctgu.edu.cn 谭永健，三峡大学计算机与信息学院，tanyongjian@ctgu.edu.cn 王曙 P-7465-2019，中国科学院地理科学与资源研究所资源与环境信息系统国家重点实验室，wangshu@igsnrr.ac.cn 谢忠 ABH-2747-2021，中国地质大学(武汉)计算机学院国家地理信息系统工程技术研究中心，xiezhong@cug.edu.cn 邱芹军 ABH-2552-2021，中国地质大学(武汉)计算机学院国家地理信息系统工程技术研究中心，qiuqinjun@cug.edu.cn
数据年代	2020 年　　数据量　4,965 KB
地理区域	金牛镇幅高桥幅、阳春县幅、治多县幅、尼玛区幅
数据格式	.txt
基金项目	国家自然科学基金（42050101，41871311，U1711267）
数据出版与服务平台	全球变化科学研究数据出版系统 http://www.geodoi.ac.cn
地址	北京市朝阳区大屯路甲 11 号 100101，中国科学院地理科学与资源研究所
数据共享政策	全球变化科学研究数据出版系统的“数据”包括元数据（中英文）、通过《全球变化数据仓储电子杂志（中英文）》发表的实体数据集和通过《全球变化数据学报（中英文）》发表的数据论文。其共享政策如下：（1）“数据”以最便利的方式通过互联网系统免费向全社会开放，用户免费浏览、免费下载；（2）最终用户使用“数据”需要按照引用格式在参考文献或适当的位置标注数据来源；（3）增值服务用户或以任何形式散发和传播（包括通过计算机服务器）“数据”的用户需要与《全球变化数据学报（中英文）》编辑部签署书面协议，获得许可；（4）摘取“数据”中的部分记录创作新数据的作者需要遵循 10% 引用原则，即从本数据集中摘取的数据记录少于新数据集总记录量的 10%，同时需要对摘取的数据记录标注数据来源[11]
数据和论文检索系统	DOI, CSTR, Crossref, DCI, CSCD, CNKI, SciEngine, WDS/ISC, GEOSS

（2）标注工具开发阶段：根据标注规范和策略开发标注管理工具（图 1）。

（3）预标注与一致性检验阶段：该数据集所采用的标注方法是 BIOES 标注法，其标签类型的定义如表 2 所示。首先，我们对语料库进行预标注，并在标注过程中去除英文单词、特殊符号图表等无关信息，然后根据预标注的结果进行一致性检验，对不一致的情况进行讨论分析，确定标注结果，本阶段的工作共进行了四轮迭代。

（4）语料库的评估及测试阶段：在标注好的地质领域命名实体语料库数据集上进行多个命名实体识别模型的训练和测试，最终根据测试结果对数据集进行分析和评估（表 3、图 2）。

4 数据结果与验证

4.1 数据集组成

《基于四份区域地质调查报告构建的命名实体识别试验数据集》归档在一个.txt 文件中。原始数据来自四份区域地质调查报告的文本，分别为尼玛区幅、治多县幅、金牛镇幅高桥

幅、广东阳春市幅等四份报告。

图 1　标注工具及界面展示图

表 2　标签类型定义

定义	全称	备注
B	Begin	实体片段的开始
I	Inside	实体片段的中间
E	End	实体片段的结束
S	Single	单个字的实体
O	Other	其他不属于任何实体的字符（包括标点等）

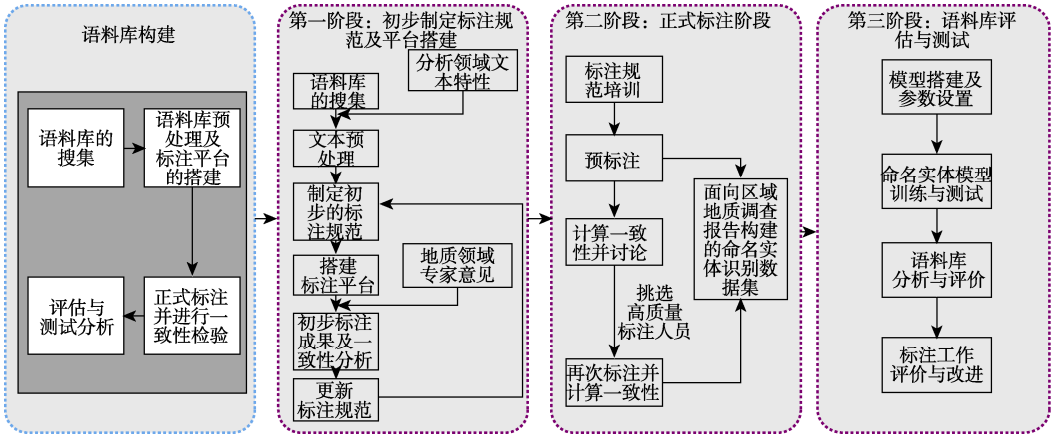

图 2　语料库数据集标注、评估、测试流程图

4.2　数据结果

地质命名实体是地质调查报告文本中重要的知识表达载体，数据集标注出的实体类型为地质时间、地质构造、地层、岩石、矿物和地点 6 种。每种类型实体对应的关键词列入表 3。

数据集共标注句子 10,803 句，已标注字数为 100,106，未标注字数为 598,406。其中尼

玛区幅区域地质调查报告共标注句子1,526句,已标注字数为20,615,未标注字数为67,107；治多县幅区域地质调查报告共标注句子 3,294 句，已标注字数为 32,764，未标注字数为205,158；金牛镇幅高桥幅区域地质调查报告共标注句子3,074句，已标注字数为23,126，未标注字数为 176,885；阳春县幅区域地质调查报告共标注句子 2,909 句，已标注字数为23,601，未标注字数为149,256。其具体数目如表4、表5所示，数据集中各类实体数量如图3所示。

4.3 数据结果验证

数据集标注完成后，通常需要分析其标注一致性。标注一致性常采用Kappa值[12]及F值[13]两类评价指标。其中Kappa值是一个用于一致性检验的指标，常用于情感分类的语料库构建，其计算是基于混淆矩阵[14]，取值为–1到1之间，通常大于0；而在命名实体识

表3 实体类型及其关键词

实体类型	关键词
地质时间（GTM）	早期的冥古宙、太古宙和远古宙（远古宙在中国含有一个震旦纪） 之后显生宙的古生代、中生代和新生代。古生代分为寒武纪、奥陶纪、志留纪、泥盆纪、石炭纪和二叠纪；中生代分为三叠纪、侏罗纪和白垩纪；新生代分为古近纪、新近纪和第四纪
地质构造（GST）	褶皱、节理、断层、劈理、向斜、背斜、地垒、地堑
地层（STR）	岩石地层单位：群、组、段、层 年代地层单位：宇、系、统、阶、时带 生物地层单位：延限带、组合带、富集带、谱系带、间隔带
岩石（ROC）	岩浆岩、沉积岩、变质岩、火山岩、浮岩、玄武岩、花岗岩、安山岩、粗面岩、响岩、火山碎屑岩、橄榄岩、糜棱岩、碎裂岩、角岩、板岩、千枚岩、片岩、片麻岩、大理岩、石英岩、角闪岩、片粒岩、榴辉岩、混合岩等
矿物（MIN）	橄榄石、辉石、闪石、云母、长石、石英、铬铁矿、金刚石、透辉石、透闪石、石榴子石、符山石、硅灰石、硅镁石、黑云母等
地点（PLA）	多数为县、村、区、乡等 例如灵乡、大冶灵乡、大治县、鄂城县、简庄村、库里南村等

表4 命名实体数量统计表

	数量	占比
地质时间（GTM）	1,864	7.99%
地质构造（GST）	1,359	5.82%
地层（STR）	3,016	12.92%
岩石（ROC）	9,827	42.09%
矿物（MIN）	4,924	21.09%
地点（PLA）	2,355	10.09%

表5 各地质调查报告中命名实体数量统计表

	地质时间	地质构造	地层	岩石	矿物	地点
尼玛区幅	215	428	950	2,282	680	742
治多县幅	931	360	953	2,828	1,956	677
金牛镇幅高桥幅	194	275	668	2,615	871	473
阳春县幅	524	296	445	2,102	1,417	463
总计	1,864	1,359	3,016	9,827	4,924	2,355

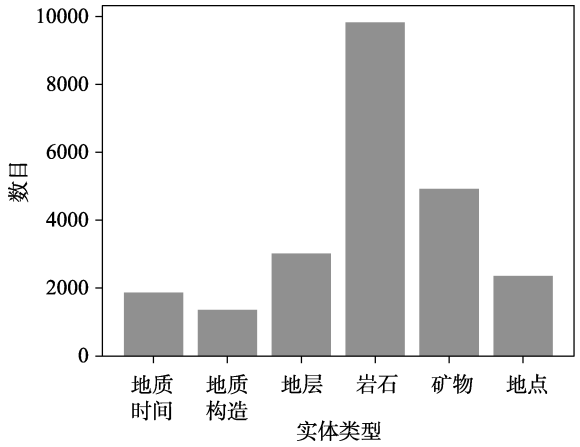

图 3　六类数据统计图

别语料库标注中，由于未标注的文本只能当做负例，因此无法统计。在负例较多且难以统计情况下，可直接采用 F 值评价，这种情况下 F 值往往与 Kappa 值比较接近。

本数据集标注一致性采用 F 值来评价，其具体的评价方法为：将其中的一个标注者视为标准，然后计算另一个标注者的准确度及召回率，最后计算 F 值。计算公式如下所示：

$$P = \frac{A1和A2一致的标注结果总数}{A2的标注总数} \tag{1}$$

$$R = \frac{A1和A2一致的标注结果总数}{A1的标注总数} \tag{2}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{3}$$

本次数据集共标注四轮，在每一阶段完成后都进行了标注一致性检验，具体结果如表 6 所示。可以看到，当三轮标注完成后，一致性检验结果全部在 0.85 以上。而文献[15]指出，当标注一致性达到 0.8 时，可以认为语料库的一致性是合格的。这表明我们标注的地质领域命名实体识别数据集在一致性上是可靠的。

表 6　语料库一致性评价结果

实体类型	第一轮	第二轮	第三轮	最终
地质时间（GTM）	92.4%	97.6%	96.4%	97.2%
地质构造（GST）	85.1%	85.8%	91.3%	92.2%
地层（STR）	74.3%	83.4%	91.6%	86.1%
岩石（ROC）	76.3%	84.8%	88.7%	91.5%
矿物（MIN）	94.1%	93.6%	95.8%	98.4%
地点（PLA）	73.4%	83.6%	84.2%	85.2%

5　讨论和总结

区域地质调查报告是指在选定地区的范围内，在充分研究和运用已有资料的基础上，采用现代地质理论和方法进行全面系统的综合性的地质调查研究工作，一般按照国际分幅

的图幅进行。其内容包括调查区的一般自然、经济地理情况，以及该地区在各个时期的地质构造、地层、岩石以及矿物的情况。区域地质调查报告作为基础地质资料的重要载体，对其中的地质命名实体进行标注具有非常重要的现实意义。本数据集基于四份区域地质调查报告的文本进行构建，对地质时间、地质构造、地层、岩石、矿物和地点六大类典型的命名实体进行了标注，为开展地学领域的命名实体识别、关系抽取及知识图谱研究提供了数据基础。

作者分工：邱芹军、马凯、谢忠、王曙对数据集的开发做了总体设计；谭永健、田苗采集和处理了六类实体数据；谭永健、田苗设计了模型和算法；谭永健、田苗做了数据验证；谭永健、田苗撰写了数据论文等。

利益冲突声明：本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献

[1] 刘文聪, 张春菊, 汪陈等. 基于 BiLSTM-CRF 的中文地质时间信息抽取[J]. 地球科学进展, 2021, 36(2): 211−220. DOI:10.11867/j.issn.1001-8166.2021.017.

[2] 张雪英, 叶鹏, 王曙等. 基于深度信念网络的地质实体识别方法[J]. 岩石学报, 2018, 34(2): 343–351.

[3] 卢书炜, 杜凤军, 任建德. 尼玛区幅 H45C001003 1/25 万区域地质调查报告[DS]. 全国地质资料馆, 2002. DOI:10.35080/n01.c.93307.

[4] 王毅智, 刘生军, 祁生胜等. 治多县 I46C003004 1/25 万区域地质调查地质报告[DS]. 全国地质资料馆, 2006. DOI:10.35080/n01.c.105419.

[5] 李雄伟, 吴兵, 施彬等. 金牛镇幅 H50E012003 高桥幅 H50E013003 1/5 万区域地质调查报告[DS]. 全国地质资料馆, 2009. DOI:10.35080/n01.c.123962

[6] 洪裕荣, 郭良田, 刘辉东等. 阳春县幅 F49C002003 1/25 万区域地质调查成果报告[DS]. 全国地质资料馆, 2004. DOI:10.35080/n01.c.122045.

[7] 吴泰然, 何国琦. 普通地质学[M]. 北京: 北京大学出版社, 2003.

[8] 全国地层委员会. 中国地层指南及中国地层指南说明书[M]. 北京: 地质出版社, 2001.

[9] 宋春青, 邱维理, 张振春. 地质学基础[M]. 北京: 高等教育出版社, 2005.

[10] 马凯, 田苗, 谭永健等. 基于四份区域地质调查报告构建的命名实体识别试验数据集[J/DB/OL]. 全球变化数据仓储电子杂志, 2021. https://doi.org/10.3974/geodb.2021.09.04.V1. https://cstr.escience.org.cn/CSTR:20146.11. 2021.09.04.V1.

[11] 全球变化科学研究数据出版系统. 全球变化科学研究数据共享政策[OL]. https://doi.org/10.3974/dp.policy.2014.05(2017 年更新).

[12] Carletta, J. Assessing agreement on classification tasks: the Kappa statistic [J]. *Computational Linguistics*, 1996, 22(2): 249–254.

[13] Hripcsak, G., Rothschild, A. S. Agreement, the f-measure, and reliability in information retrieval [J]. *Journal of the American medical informatics association*, 2005, 12(3): 296–298.

[14] Tang, W., Hu, J., Zhang H., Pan, W., *et al*. Kappa coefficient: a popular measure of rater agreement [J]. *Shanghai archives of psychiatry*, 2015, 27(1): 62.

[15] Artstein, R., Poesio, M. Inter-coder agreement for computational linguistics [J]. *Computational Linguistics*, 2008, 34(4): 555–596.