Journal of Global Change Data & Discovery. 2022, 6(1): 78-84
www.geodoi.ac.cn

Development of a Named Entity Recognition Dataset Based on Four Regional Geological Survey Reports

Ma, K.[1] Tian, M.[1] Tan, Y. J.[1] Wang, S.[4] Xie, Z.[2,3] Qiu, Q. J.[2,3,*]

1. College of Computer and Information Technology, China Three Gorges University, Yichang 443002, China;
2. School of Computer Science, China University of Geosciences, Wuhan 430074, China;
3. National Engineering Research Center of Geographic Information System, Wuhan 430074, China;
4. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

Abstract: Regional geological survey reports are important technical documents that comprehensively reflect the results of regional geological survey work. At present, the national geological data library has accumulated a large number of geological result reports, and information extraction and mining can fully explore the implicit value of existing reports and promote the discovery of new knowledge. In this paper, a named entity recognition experimental dataset based on four regional geological survey reports is constructed for the task of named entity recognition in the field of natural language processing, which can be used for training and testing geological named entity models. The dataset contains a total of four regional geological survey results reports, which are annotated with six typical categories of geological named entities: geological time, geological formations, strata, rocks, minerals and locations. The dataset is checked for consistency, tested and evaluated separately to ensure the quality of the dataset. The size of the dataset is 4.84 MB, and the data format is .txt.

Keywords: regional geological survey reports; named entity recognition; consistency checking; testing; evaluation

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

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

Dataset Availability Statement:

The dataset supporting this paper was published and is accessible through the *Digital Journal of Global Change Data Repository* at: https://doi.org/10.3974/geodb.2021.09.04.V1 or https://cstr.escience.org.cn/CSTR:20146.11.2021.09.04.V1.

Received: 26-08-2021; **Accepted:** 30-12-2021; **Published:** 25-03-2022

Foundations: National Natural Science Foundation of China (42050101, U1711267, 41871311, 41871305)

***Corresponding Author:** Qiu, Q. J., National GIS Engineering Technology Research Center, School of Geography and Information Engineering, China University of Geosciences (Wuhan), qiuqinjun@cug.edu.cn

Data Citation: [1] Ma, K., Tian, M., Tan, Y. J., *et al.* Development of a named entity recognition dataset based on four regional geological survey reports [J]. *Journal of Global Change Data & Discovery*, 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] Ma, K., Tian, M., Tan, Y. J., *et al*. Named entity recognition dataset for four regional geological survey reports by data mining methodology [J/DB/OL]. *Digital Journal of Global Change Data Repository*, 2021. https://doi.org/10.3974/geodb.2021.09.04.V1. https://cstr.escience.org.cn/CSTR:20146.11.2021.09.04.V1.

1 Introduction

In most recent years, the Deep-time Digital Earth Program (DDE) became an inevitable trend to build a geological knowledge base and conduct structured information extraction work such as named entity recognition and relationship extraction on massive geological text data to realize deep mining of geological text knowledge. Text knowledge mining cannot be achieved without the support of high-quality corpus datasets. At present, there are single-type and small-scale geological named entity recognition corpus for geological time recognition[1] and rock entity recognition[2] in China, but there is a lack of large-scale annotated Chinese geological corpus of multiple entity types. This dataset extracts the regional geological survey report H45C001003 1/250,000 in Nima district[3], the geological report of I46C003004 1/250,000 regional geological survey in Zhiduo county[4], the regional geological survey report H50E013003 1/50,000 in Jinniu township[5], the regional geological survey report F49C002003 1/250,000 in Yangchun county[6], and the geological report of H50E013003 1/50,000 in Jinniu township[7]. The text data from the four regional geological survey reports[6] were obtained by performing preprocessing and related processes such as annotation, testing and evaluation.

This dataset focuses on six named entity types: geological time, geological formations, strata, minerals, rocks and places. The layers and bodies of rock on the Earth's surface are subject to a variety of geological forces during and after their formation, with some remaining largely in the original state they were in when they were formed, while others have undergone deformation. They have a complex spatial assemblage of forms, i.e., various geological formations, of which fractures and folds are the two most basic forms of geological formations[7]. Finally, the establishment of geological time is the basis for our study of the history of the Earth's crust.

Stratigraphy is the main body of stratified rocks and in a narrow sense refers exclusively to consolidated stratified rocks, sometimes including loose sediments that have not yet consolidated into rock[8]. This is the basic principle of stratigraphic relationships and is known as stratigraphic law[9]. Minerals are natural homogeneous bodies with certain chemical compositions and physical properties formed by chemical elements in the Earth's crust under various geological actions, and they are the basic units that make up rocks and ores[9]. Minerals are often found in the crust in the form of aggregates, which can be composed of one or more minerals, and are known in geology as rocks[7]. Additionally, considering locations as an important spatial reference that appears in the text, the dataset has annotated locations as a class of entities. Through the annotation of six types of entities in four geological reports, the statistical characteristics of the entities in each report were analyzed, and the corpus dataset was checked for consistency, as well as tested and evaluated to ensure the quality of the dataset. This dataset can provide an important database for named entity recognition, relationship extraction and the construction of knowledge graphs in the field of geology.

2 Metadata of the Dataset

The metadata of the Named entity recognition dataset for four regional geological survey reports by data mining methodology[10] is summarized in Table 1. It includes the dataset full name, short name, authors, year of the dataset, data format, data size, data files, data publisher, and data sharing policy, etc.

Table 1 Metadata summary of the Named entity recognition dataset for four regional geological survey reports by data mining methodology

Items	Description
Dataset full name	Named entity recognition dataset for four regional geological survey reports by data mining methodology
Dataset short name	NERdata
Authors	Ma, K. ABH-2687-2021, School of Computer and Information Science, Three Gorges University, makai@ctgu.edu.cn Miao, T. ABH-2542-2021, School of Computer and Information Science, Three Gorges University, tianmiao@ctgu.edu.cn Tan, Y. J., School of Computer and Information Science, Three Gorges University, tanyongjian@ctgu.edu.cn Wang, S. P-7465-2019, State Key Laboratory of Resource and Environmental Information Systems, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, wangshu@igsnrr.ac.cn, Xie, Z. ABH-2747-2021, School of Computer Science, China University of Geosciences (Wuhan), National GIS Engineering Technology Research Center, xiezhong@cug.edu.cn Qiu, Q. J. ABH-2552-2021, School of Computer Science, National GIS Engineering Technology Research Center, China University of Geosciences (Wuhan) qiuqinjun@cug.edu.cn
Year	2020 Data size 4,965 KB Data format .txt
Geographical area	Jinniu township, Gaoqiao district, Yangchun county, Zhiduo county, Nima district
Foundations	National Natural Science Foundation of China (42050101, 41871311, U1711267)
Data publisher	Global Change Research Data Publishing & Repository, http://www.geodoi.ac.cn
Address	No. 11A, Datun Road, Chaoyang District, Beijing 100101, China
Data Sharing Policy	***Data*** from the Global Change Research Data Publishing & Repository includes metadata, datasets (in the *Digital Journal of Global Change Data Repository*), and publications (in the *Journal of Global Change Data* & *Discovery*). ***Data*** sharing policy includes: (1) ***Data*** are openly available and can be free downloaded via the Internet; (2) End users are encouraged to use ***Data*** subject to citation; (3) Users, who are by definition also value-added service providers, are welcome to redistribute ***Data*** subject to written permission from the GCdataPR Editorial Office and the issuance of a ***Data*** redistribution license; and (4) If ***Data*** are used to compile new datasets, the 'ten per cent principal' should be followed such that ***Data*** records utilized should not surpass 10% of the new dataset contents, while sources should be clearly noted in suitable places in the new dataset[11]
Communication and searchable system	DOI, CSTR, Crossref, DCI, CSCD, CNKI, SciEngine, WDS/ISC, GEOSS

3 Data Collection Process and Methodology

The entire dataset was collected in two main steps: the selection of geological survey result reports for representative areas and the annotation of named entities for the selected reports.

The data are derived from the text in the regional geological survey results reports of the Nima district area, Zhiduo county area, Jinniu town & Gaoqiao area and Yangchun county area of Guangdong, covering a total geographical area of four provinces, namely, Tibet, Guangdong, Hubei and Qinghai. The results reports are important technical documents that comprehensively reflect the results of geological survey work. In each geological results report, named entities such as geological time, geological formations, strata, rocks, minerals and geographical names are involved, and their efficient and accurate identification and extraction is the basis for achieving geological knowledge mining. These named entities are also the objects of this dataset annotation, and the annotated dataset can be used to train and test relevant entity recognition models.

A dedicated annotation tool has been developed, and a set of annotation rules has been developed to standardize the annotation of ambiguous entities. Data annotation is conducted in a semiautomatic way using a cross-annotation model between domain experts and groups and by manual means with the assistance of software. The entire annotation process is divided into four stages as follows.

(1) Formulation of annotation specifications: formulation of annotation specifications

based on the syntactic-semantic characteristics of named entities in the geological field combined with the opinions of experts in the geological field.

(2) Markup tool development phase: development of markup management tools based on markup specifications and strategies (Figure 1).

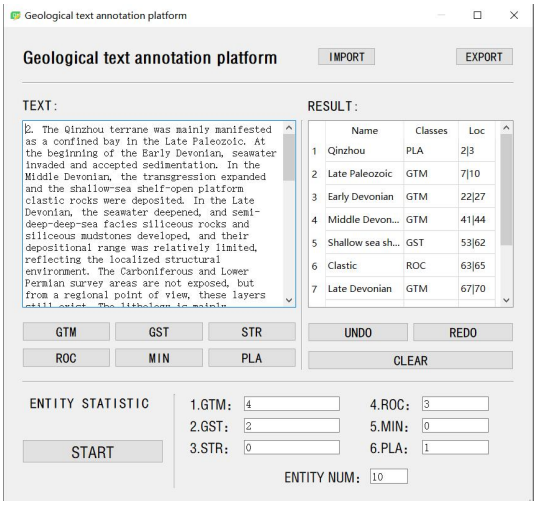

Figure 1 Illustration of the annotation tools and interface

(3) Pre-annotation and consistency check phase: the annotation method used for this dataset is the BIOES annotation method, and the definition of its label types is shown in Table 2. First, we pre-annotated the corpus and removed irrelevant information such as English words and special symbol charts during the annotation process. Then, we conducted a consistency check based on the pre-annotation results, discussed and analyzed the inconsistencies, and determined the annotation results. Four iterations of this phase of the work were carried out.

(4) Corpus evaluation and testing phase: Multiple named entity recognition models were trained and tested on the annotated geological domain named entity corpus dataset, and the dataset was eventually analyzed and evaluated based on the test results (Table 3, Figure 2).

Table 2 Label type definitions.

Definition	Full name	Remarks
B	Begin	The start of the entity segment
I	Inside	The middle of an entity segment
E	End	End of entity segment
S	Single	Single-word entities
O	Other	Other characters that are not part of any entity (including punctuation, etc.)

4 Data Results and Validation

4.1 Dataset Composition

The named entity recognition test dataset constructed based on four regional geological survey reports is archived in a .txt file. The original data were obtained from the texts of four regional geological survey reports, namely, the Nima district frame, the Zhiduo county frame, Jinniu town & Gaoqiao area frame, and the Guangdong Yangchun city frame.

4.2 Data Results

Geological named entities are important carriers of knowledge expression in the text of geological survey reports, and the dataset is labeled with six types of entities: geological time, geological formations, strata, rocks, minerals and locations. The keywords corresponding to each type of entity are included in Table 3.

A total of 10,803 sentences were annotated in the dataset, with 100,106 annotated words and 598,406 un-annotated words. A total of 1,526 sentences were marked in the regional geological survey report of the Nima area, totaling 20,615 marked words and 67,107

unmarked words. A total of 3,294 marked sentences, 32,764 marked words and 205,158 unmarked words were included in the regional geological survey report for the Zhiduo county area. There are total of 3,074 marked sentences with 23,126 marked words and 176,885 unmarked words in the regional geological survey report for the Gaoqiao district area of the Jinniu township area. The Yangchun County Regional Geological Survey Report contains a total of 2,909 marked sentences, with 23,601 marked words and 149,256 unmarked words. Their exact numbers are shown in Table 4 and Table 5, and the number of entities of each type in the dataset is shown in Figure 3.

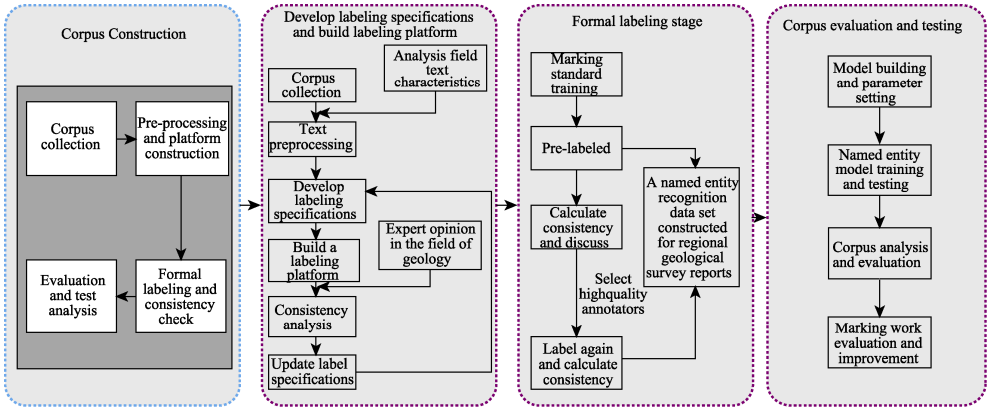

Figure 2 Flow chart of corpus dataset annotation, evaluation and testing.

Table 3 Entity types and their keywords

Type of entity	Keywords
Geological time (GTM)	The early Plutonic, Archaic and Paleogene (the Paleogene contains an epoch in China) followed by the Palaeozoic, Mesozoic and Cenozoic of the Eocene. The Paleozoic is divided into the Cambrian, Ordovician, Silurian, Devonian, Carboniferous and Permian; the Mesozoic into the Triassic, Jurassic and Cretaceous; and the Cenozoic into the Palaeocene, Neogene and Quaternary
Geological formations (GST)	Folds, joints, faults, cleavage, oblique, back-slope, basement, graben
Stratigraphic (STR)	Lithostratigraphic units: groups, formations, sections, layers Stratigraphic units of chronostratigraphy: Uranian, Systematic, Tertiary, Order, Temporal Zone Biostratigraphic units: extensional, combined, enrichment, spectral, interval zones
Rocks (ROC)	Magmatic, sedimentary, metamorphic, volcanic, pumice, basalt, granite, andesite, rough facies, rattles, volcanic clastic, peridotite, vesicular, fractured, hornblende, slate, schist, schist, gneiss, dacite, quartzite, hornblende, gneiss, garnets, mixed rocks, etc.
Minerals (MIN)	Olivine, pyroxene, amphibole, mica, feldspar, quartz, chromite, diamond, tremolite, tremolite, garnet, fushanite, wollastonite, magnesite, black mica, etc.
Location (PLA)	Most are counties, villages, districts, townships, etc. For example, Lingxiang, Daye Lingxiang, Daji county, Echeng county, Jianzhuang village, Kulinan village, etc.

4.3 Data Validation

After a dataset has been annotated, it typically needs to be analyzed for annotation consistency. Two types of evaluation metrics are often used for annotation consistency: kappa value[12] and *F* value[13]. The Kappa value is a consistency check metric commonly used in corpus construction for sentiment classification and is calculated based on the confusion matrix[14], which takes values between –1 and 1, usually greater than 0. In

named entity recognition corpus annotation, un-annotated text cannot be counted because it can only be treated as negative examples.

Table 4 Statistics of the named entities

	Number	Percentage		Number	Percentage
Geological time (GTM)	1,864	7.99%	Rocks (ROC)	9,827	42.09%
Geological formations (GST)	1,359	5.82%	Minerals (MIN)	4,924	21.09%
Stratigraphic (STR)	3,016	12.92%	Location (PLA)	2,355	10.09%

Table 5 Statistics on the number of named entities in each geological survey report

	Geological time	Geological formations	Stratigraphic	Rocks	Minerals	Location
Nima district	215	428	950	2,282	680	742
Zhiduo county	931	360	953	2,828	1,956	677
Jinniu town & Gaoqiao area	194	275	668	2,615	871	473
Yangchun county	524	296	445	2,102	1,417	463
Total	1,864	1,359	3,016	9,827	4,924	2,355

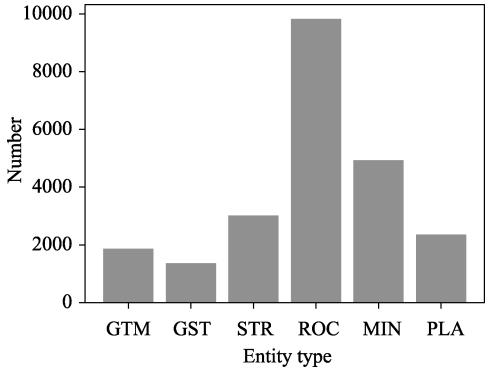

Figure 3 Statistical chart of six categories of data
Notes: GTM, Geological time; GST, Geological formations; STR, Stratigraphic; ROC, Rocks; MIN, Minerals; PLA, Location

In cases where there are more negative examples that are difficult to count, *F* values can be used directly for evaluation, in which case the *F* values is often closer to Kappa values.

This dataset annotation consistency is evaluated using the *F* value, and the specific evaluation route is as follows: one of the annotators is considered the standard, then the accuracy and recall of the other annotator is calculated, and finally the *F* value is calculated. The calculation formula is shown below.

$$P = \frac{Total\ number\ of\ consistent\ labelling\ results\ for\ A1\ and\ A2}{Total\ number\ of\ markings\ for\ A2} \tag{1}$$

$$R = \frac{Total\ number\ of\ consistent\ labelling\ results}{Total\ number\ of\ markings\ for\ A1} \tag{2}$$

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

Table 6 Results of corpus consistency evaluation

Type of entity	First round	Second round	Third round	Final total
Geological time (GTM)	92.4%	97.6%	96.4%	97.2%
Geological formation (GST)	85.1%	85.8%	91.3%	92.2%
Stratigraphic (STR)	74.3%	83.4%	91.6%	86.1%
Rocks (ROC)	76.3%	84.8%	88.7%	91.5%
Minerals (MIN)	94.1%	93.6%	95.8%	98.4%
Location (PLA)	73.4%	83.6%	84.2%	85.2%

The dataset was annotated in four rounds, the annotation consistency test was conducted

after each stage was completed, and the specific results are shown in Table 6. When the three rounds of annotation were completed, the consistency test results were all above 0.85. The literature[15] notes that when the annotation consistency reaches 0.8, the consistency of the corpus can be considered to be satisfactory. This indicates that our annotated named entity recognition dataset for the geological domain is reliable in terms of consistency.

5 Discussion and Conclusion

The regional geological survey report refers to rich information about the selected area. By using modern geological theories and methods to attract the data regarding general physical and economic geography of the survey area, as well as the geological formations, strata, rocks and minerals in different periods. From annotated names in these reports, the dataset based on the texts can be efficiently constructed and reused in digital format.

Author Contributions

Qiu, Q. J., Ma, K., Xie, Z., and Wang, S. designed the algorithms of dataset. Tian, M., Tan, Y. J. contributed to the data processing and analysis. Tian, M., Tan, Y. J. wrote the data paper.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Liu, W. C., Zhang, C. J., Wang, C. H., *et al*. Chinese geological time information extraction based on BiLSTM-CRF [J]. *Advances in Earth Sciences*, 2021, 36(2): 211–220. DOI: 10.11867/j.issn.1001–8166.2021.017.

[2] Zhang, X. Y., Ye, P., Wang, S., *et al.* A deep belief network-based method for geological entity identification [J]. *Journal of Petrology*, 2018, 34(2): 343–351.

[3] Lu, S. W., Du, F. J., Ren, J. D. Report on the regional geological survey of the Nima area H45C001003 1/250,000 [DS]. DOI:10.35080/n01.c.93307.

[4] Wang, Y. Z., Liu, S. J., Qi, S. S., *et al*. Geological report of the I46C003004 1/250,000 regional geological survey in Zhiduo county [DS]. National Geological Data Library, 2006. DOI: 10.35080/ n01.c.105419.

[5] Li, X. W., Wu, B., Shi, B., *et al*. Report on the H50E013003 1/50,000 regional geological survey of the Jinniu Township area H50E012003 Gaoqiao area [DS]. National Geological Data Library, 2009. DOI: 10.35080/n01.c.123962.

[6] Hong, Y. R., Guo, L. T., Liu, H. D., *et al*. Report on the results of the Yangchun county F49C002003 1/250,000 regional geological survey [DS]. DOI: 10.35080/n01.c.122045.

[7] Wu, T. R., He, G. Q. General Geology [M]. Beijing: Beijing University Press, 2003.

[8] National Commission on Stratigraphy. A Guide to the Stratigraphy of China and a Manual for the Stratigraphy of China [M]. Beijing: Geological Press, 2001.

[9] Song, C. Q., Qiu, W. L., Zhang, Z. C. Fundamentals of Geology [M]. Beijing: Higher Education Press, 2005.

[10] Ma, K., Tian, M., Tan, Y. J., *et al*. Named entity recognition dataset for four regional geological survey reports by data mining methodology [J/DB/OL]. *Digital Journal of Global Change Data Repository*, 2021. https://doi.org/10.3974/geodb.2021.09.04.V1. https://cstr.escience.org.cn/CSTR:20146.11.2021.09.04.V1.

[11] GCdataPR Editorial Office. GCdataPR data sharing policy [OL]. https://doi.org/10.3974/dp.policy. 2014.05 (Updated 2017).

[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.