Earth Science Informatics (2020) 13:1393–1410
https://doi.org/10.1007/s12145-020-00527-9

RESEARCH ARTICLE

Automatic spatiotemporal and semantic information extraction from unstructured geoscience reports using text mining techniques

Qinjun Qiu [1,2] · Zhong Xie [1,2] · Liang Wu [1,2] · Liufeng Tao [1,2]

Received: 15 July 2020 / Accepted: 15 September 2020 / Published online: 19 September 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

A large number of georeferenced quantitative data about rock and geoscience surveys are buried in geological documents and remain unused. Data analytics and information extraction offer opportunities to use this data for improved understanding of ore forming processes and to enhance our knowledge. Extracting spatiotemporal and semantic information from a set of geological documents enables us to develop a rich representation of the geoscience knowledge recorded in unstructured text written in Chinese. This paper presents the workflow for spatiotemporal and semantic information extraction, which is a geological document analysis approach that uses automated techniques for browsing and searching relevant geological content. The developed workflow applies spatial and temporal gazetteer matching, pattern-based rules and spatiotemporal relationship extraction to identify and label terms in geological text documents. It offers a representation of contextual information in knowledge graph form, extracts a set of relevant tables and figures, and queries a list of relevant documents by using geological topic information. Here, text mining techniques are used to facilitate the analysis of geological knowledge and to show the effectiveness of text analysis for improving the rapid assessment of a massive number of documents. Furthermore, autogenerated keyword suggestions derived from extracted keyword associations are used to reduce document search efforts. This research illustrates the usefulness and effectiveness of the developed information extraction workflow and demonstrates the potential of incorporating text mining and NLP techniques for geoscience.

Keywords Geoscience document · Knowledge graph · Geological text mining · Natural language processing

Introduction

Publicly large geoscience documents/reports are components of available data sources and offer tremendous challenges and opportunities, as they can enable geology research in a chosen target area. These natural language documents/reports often contain a large amount of explicit and implicit geological knowledge pertaining to ore forming processes or documenting where geological structures occurred (Holden et al. 2019). Research on mathematical geoscience aims to process georeferenced quantitative information/data for information extraction and knowledge discovery (Lima et al. 2017; Wang et al. 2018a, 2018b). Given the content in geological documents, automatically extracting and reviewing the various described geological topics so that this information can support our general knowledge. The ability to extract and obtain information from natural language geological texts offers the opportunity for users to rapidly browse documents/reports rather than read the whole text.

A large portion of the geological documents/reports available today are unstructured natural language descriptions, which range from simple to more complex, and pertain to a variety of geological entities (e.g., siltstone and Potassium siltstone), geological structures (e.g., brittle shear zone, ductile shear zone, and overturned fold), spatial locations (e.g., Inner Mongolia and Xinjiang), spatial relations (e.g., alongside, adjacent, and beside) and other factors (Wu et al. 2017; Qiu et al. 2018a; Qiu et al. 2020). In addition, research in natural language geological documents/reports, which first started with retrieving key information, has now expanded into obtaining

Communicated by: H. Babaie

✉ Liufeng Tao
taoliufeng@cug.edu.cn

[1] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China

[2] National Engineering Research Center of Geographic Information System, Wuhan 430074, China

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

detailed and useful geological information. However, the natural language descriptions for a particular area are often recorded and stored in a variety of different legacy documents/reports, and the content of these reports often consists of significant overlap. This overlap can assist in the recognition and extraction of a subset of data consisting of rich, detailed and anomalous contents for further analysis. Such significant overlap, although very useful, makes automated interpretation very difficult. Choosing the relevant documents/reports from a repository can be done by quickly by browsing the documents/reports or querying the documents/reports using a list of specific keywords based on a given user-friendly search system or platform (Schuhmacher and Ponzetto 2014; Peters and McClennen 2015; Peters et al. 2017). However, the first pass reading of various and diverse reports/documents for analyzing diverse and complex details in the domain of geoscience is rather challenging and time-consuming. In addition, publicly existing query platforms based on generic text (i.e., the Google search engine) are often based on the generalized relations of phrases/words that are automatically built with text knowledge mining models and technologies. Such relations, while very useful, are often applied to generate autocompletions when humans search for documents/reports because they could provide effective and useful semantic information tips and allow for reduced search efforts. For instance, inputting "*Sedimentary Rock*" in a search engine (i.e., the Google search engine) will generate completion matching results based on various formats, such as PDFs. For geologists and professionals, useful information and suggestions may be rock specific, such as those related to rock varieties that typically support sedimentary types, structural features of rocks and rock alteration types. Further, geological documents/reports include a variety of technically detailed information about rock, stratum, geological history and geological structure. Such critical data about geology and its processes contain a large store of geological knowledge and are expected to be useful for mapping and modeling ore forming processes.

Therefore, developing information extraction approaches that identify and obtain data (e.g., geological entity and place names) about spatiotemporal and semantic information from unstructured natural language geological reports, and record/store the obtained information/data for helping to further data investigation and analytics. In this paper, since the corpus we collect is Chinese reports, we only extract and visualize the information of Chinese geological reports. In the future, we will further verify the English report based on our proposed approach. In this paper, we propose a workflow for extracting spatiotemporal and semantic information from geoscience documents/reports with a focus on Chinese documents. The approach presented here is dynamically adopted to extract contextual information (i.e., spatiotemporal and semantic information) by relying on theories of cognitive linguistics using natural language processing. Based on the previously extracted information, this study developed a relevancy network model to represent the association between the spatiotemporal data and topic information. The experimental results demonstrate the transformation of unstructured texts into spatiotemporal and semantic data/information, allowing for further geological interpretation and analysis.

The major contributions of this research are summarized as follows:

(1) This paper developed a workflow for extracting spatiotemporal and semantic information from geological reports with a focus on reports in Chinese.
(2) We developed a relevancy network model to represent the association of the spatiotemporal data and topic information, allowing for the support of autogenerated suggestions for queries and retrieval.
(3) The proposed approach offers a representation of the contextual information in knowledge graph form, a set of extracted relevant tables and figures, and a list of relevant documents based on the geological topic information to assist in fast reading and searching of reports using text mining and NLP techniques.

The remainder of this research is structured as follows: Section 2 provides the related work in automated text analysis and previous research in the geological domain; Section 3 provides details on the workflow for spatiotemporal and semantic information extraction; The experimental results are demonstrated in Section 4; A summary is presented in Section 5.

Related work

Advances in automated text analysis

Natural language processing (NLP), focuses on enabling computers to automate the understanding of human language, has been a significant advancement in automated text analysis (Manning et al. 1999). Specifically, information extraction in various domains from free-form textual documents are often connected to text mining (Wong et al. 2012; Zhang et al. 2015; Peters et al. 2017). To achieve the goal of NLP, the key is to handle and analyze natural language text data using existing models and techniques for further understanding the semantics found in languages. Information extraction uses various NLP techniques including part-of-speech (POS) tagging, which annotates words/terms in a sentence to recognize and extract information and format it into structured form. More complicated and advanced NLP tasks include deciphering language semantics, such as identifying predefined entities or categories (Nadeau and Sekine 2007; Qiu et al. 2019a), e.g.,

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

place names, universities, or organizations (named entity recognition, NER); building significant relations between entities (Yang et al. 2018); and automating machine translation (Young et al. 2018) and language generation (Paulus et al. 2018).

Based on the level of sentence complexity, information extraction can be classified into four categories: (1) NER, which focuses on recognizing predefined entities (Konkol et al. 2015); (2) relation extraction, which focuses on detecting the relations among the recognized entities (Du and Guo 2016; Liu and Elgohary 2017; Zhang and Elgohary 2016; Zhou et al. 2018); (3) event extraction, which focuses on identifying a variety of events from incoming text (here every event simultaneously contains a trigger and some relevant arguments; each event could consist of several entities and their associated relations) (Abraham et al. 2018); and (4) full information extraction, which focuses on extracting and identifying all required data/information demonstrated by a set of sentences using full text mining or data analysis (Zhang et al. 2015). The first three kinds of information extraction (NER, relation extraction) can be regarded as shallow information extraction since they only focus on identifying and extracting partial content in a sentence, while full information extraction could be categorized as deep information extraction since it focuses on extracting an entire set of data from a sentence (Zhang et al. 2015).

Text mining aims to seek structures/patterns from incoming natural language text by a variety of data analytical technologies and NLP models/algorithms (i.e., data mining and statistical analysis) (Harisinghaney et al. 2014). Relevant to this study is the focus on text mining in the geological context.

Previous text mining in the geological context

A larger body of study in the domain of geoscience has explored and described information extraction and specifically, statistical data analysis using text mining (Wang et al. 2015; Qiu et al. 2019a). Some of the well-known work in this domain includes DeepDive (De Sa et al. 2016) and work by Peters et al. (2014) who built a machine learning (ML) platform called the PaleoDeepDive, which extracts and identifies various descriptions and information of paleontological fossils from unstructured text, figures and tables. The data generated by the developed ML system, which contained rates of the genus level turnover and histories of the taxonomic diversity, were used to compare with manually annotated data. Their research incorporates an estimated uncertainty and provides an automated ML analysis in developing a structured database for all automatically produced material. Peters et al. (2017) presented the idea of combining a variety of stratigraphic databases, published open documents, and an ML platform to automatically identify and explore the detailed processes of stromatolites ranging from prevalence and extinction to resurgence in the North American marine environments. Their approach could be used to recognize the previously mentioned occurrence-related words/terms of stromatolites (i.e., prevalence, extinction and resurgence) in geological timescales. Wang et al. (2018a, 2018b) proposed the idea of developing a geological ontology in the timescales relevant to North America, integrating fossil occurrences, and then geospatially visualizing the extracted information.

In recent years, there are some research in the domain of geoscience has discussed and reported Chinese text mining applications. Luo et al. (2018a, 2018b) applied deep learning approaches to extract geological relations based on the attention mechanism. Qiu et al. (2018a, 2018b) presented a Chinese word segmenter that breaks syntactically and semantically a continuous text into meaningful words from geological reports based on BiLSTM model. Work by Wang et al. (2018a, 2018b) built an effective word segmenter using a rule-based method (conditional random fields, CRFs), which can be used to break Chinese geological documents into meaningful words. These developed rules, learned from generic and specific terms, were then applied to identify geologically relevant phrases/words, and the chord and bigram graphs generated from the cooccurrence relation of contextual words were used to visualize the contextual phrases/words and relevant links. Shi et al. (2018) used text mining techniques to extract prospecting information from the Lala copper deposit in China. They trained and classified the geological text based on CNNs. Enkhsaikhan et al. (2018) used a word embedding technique to extract semantic information from geological applications. Their research aimed to calculate semantic-based similarity to determine the similarity of pairs of geological terms as well as their relations based on a newly designed analogy solver. Qiu et al. (2019a) also developed a keyphrase/keyword extraction algorithm in the domain of geoscience using geoscience ontology and modified word embedding. Their model enriched the domain-specific words/terms based on the domain ontology tree by recognizing infrequent but representative keyphrases. Qiu et al. (2019b) proposed a generative model for recognizing and identifying geologically named entities using deep learning. The model used words and related frequencies to construct and generate a training corpus based on a random extraction algorithm. Holden et al. (2019) built a domain-specific text mining system (namely, GeoDocA) in the domain of geoscience by applying machine analysis to automatically accelerate fast reading and searching for documents/reports from a set of exploration reports based on NLP and then visualize this extracted information.

Workflow for spatiotemporal and semantic information extraction

The existing large body of research efforts in the different domain-specific fields (i.e., the domain of geoscience) has

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

been discussed as it relates to deriving various information. Despite the efforts applied to this issue, however, existing information extraction approaches are limited in supporting automatic information extraction.

Most existing information extraction approaches developed recently rely on a rule-based method or a supervised ML-based method. For instance, almost all information extraction efforts in the geoscience domain have applied supervised ML-based methods (Luo et al. 2018a, 2018b; Ma et al. 2018; Shi et al. 2018; Qiu et al. 2018a; Qiu et al. 2019b).

Although ML-based methods that learn from a larger set of training datasets can reduce human efforts needed for the development of hand-crafted pattern matching and design pattern rules, a rule-based method is selected in this study for two primary reasons. First, the rule-based methods tend to achieve higher performance in a specific domain because human expertise often represents more accurate extraction rules and matching patterns (Moens 2006; Liu et al. 2015). The performance of machine learning in a specific domain with a complex task (i.e., information extraction) is usually insufficient and inconsistent (Ireson et al. 2005). Especially in domain-specific applications, the ML-based approaches are limited in identifying and extracting relevant information from highly complex text. A natural choice to extract relevant information is the use of a rule-based method in a specific task; deep information extraction (as mentioned in Section 2.1) is required to recognize and extract all key information, making the task of information extraction more challenging. Second, in this domain-specific application, compared with the ML-based methods, the comprehensive and representative rules/patterns developed manually for the rule-based method could be less than what is needed for labeling a sufficiently large number of training samples.

To address the abovementioned knowledge gaps, in this research, a rule-based information extraction methodology for extracting spatiotemporal and semantic information (STSIE) from a variety of reports/documents in the domain of geoscience is presented. The proposed STSIE methodology consists of four main steps (as seen in Fig. 1): document preprocessing (i.e., sentence splitting, tokenization and part-of-speech tagging), gazetteer creation (i.e., spatial gazetteer and temporal gazetteer), contextual information extraction, and construction of a spatiotemporal topic relevancy network model, followed by evaluation and visualization. Fig. 2 demonstrates where the suggested method of spatiotemporal and semantic information extraction from the incoming inputs occurs and the relevant outputs of the main processing steps (i.e., steps 1–4).

Step 1: Document preprocessing

The main source data applied in this study are PDFs of textual documents written in natural language Chinese from the China Geological Survey and publications from the National Geological Archives. We collected and selected PDFs of 14 geological reports as our study data source from which to extract spatiotemporal and semantic information and build a spatiotemporal topic relevancy network model. Each geological report includes both textual and graphical content and more than 10,000 Chinese characters.

Since text mining and natural language preprocessing techniques/algorithms cannot directly handle the original natural language text, the presence of tables and figures and other irrelevant contents (i.e., acknowledgments and references) can restrain the information extraction performance of the proposed methodology. As such, a natural choice is document preprocessing that segments the existing tables and figures of each geological document/report and then filters out irrelevant contents that contain invalid characters (i.e., numbers and symbols). The goal of preprocessing is to transform the raw data sources in natural language into a useful and readable format and develop a set of defined structures that provide easy labels and further extraction steps to analyze and follow. As described above, the incoming textual data are from the scanned PDFs from the National Geological Archives. With the method for spatiotemporal and semantic information extraction proposed in this research, the scanned textual documents/reports are not further recognized by and cannot be input into the postprocessing if not converted into Text file, Microsoft Word orExtensible Markup Language (XML).

In this research, the contextual information (mainly containing textual content and a set of figures) from the geological documents/reports were identified and extracted using the open source software PDFFigures 2.0 (Clark and Divvala 2016). For a given document/report, this software recognizes and extracts a list of figures in the JPG format (Holden et al. 2019) and outputs a generated JSON file that consists of the following structured document information: a list of section titles and content, section positions and titles, and a set of table and figure positions and captions. An illustrative example of our document preprocessing pipeline is depicted in Fig. 3. For geological documents/reports, a gallery of identified figures often contains explicit and implicit information of relational knowledge about geological factors. For example, the geological maps found in documents show the different concepts or outputs from geological investigations of geological structures, regional lithology, and the stratigraphy of a region. Moreover, highly valuable and analytical data, such as the characteristics of trace elements measured from rock samples that can be used to recognize mineral grades, are illustrated in table form. Therefore, the recognition and extraction of a set of tables and figures play an important role in understanding the content of geological documents/reports. Three main steps are conducted to extract the tables and figures as follows:

(1) recognizing the title labels based on a keyphrase query and a set of comprehensive patter-matching rules;

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

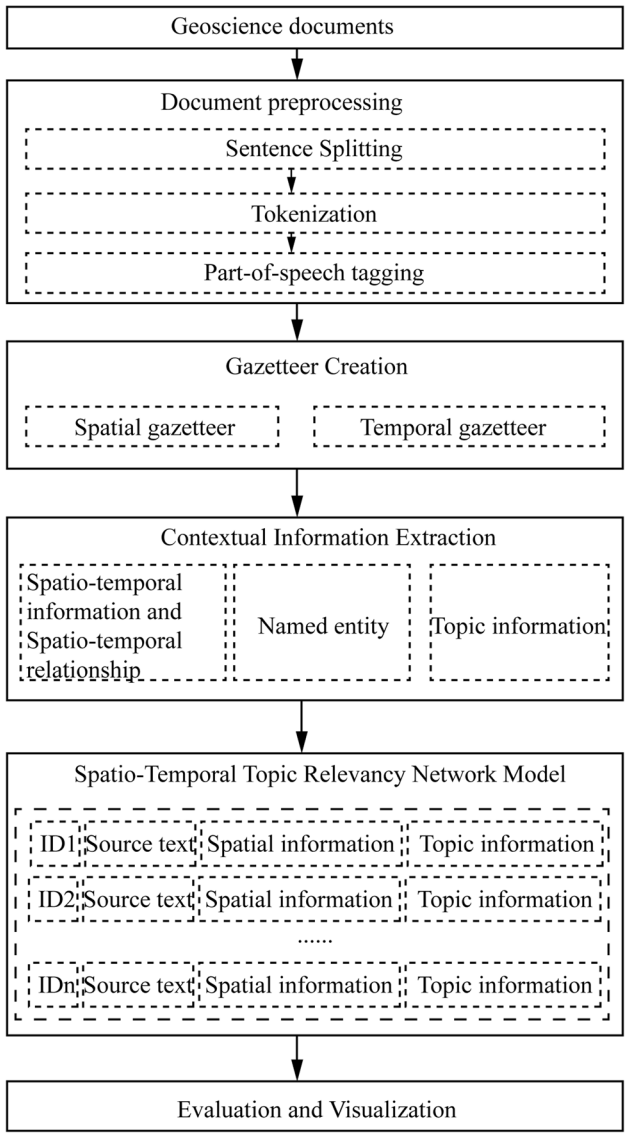


Fig. 1 Proposed spatiotemporal and semantic information extraction algorithm and components. Step 1 preprocesses the input text. Step 2 creates the spatial and temporal gazetteers. Step 3 extracts contextual information (i.e., spatiotemporal information, named entity and topic information). Step 4 develops a spatiotemporal topic relevancy network model. Step 5 evaluates the proposed model and visualizes the data

(2) describing the entire title by differentiating text formats; and
(3) recognizing the regions of different figures in the document and then clustering each figure to the closest title.

Then, remove irrelevant geological textual content, including symbols and numbers, table and figures captions, attachments and appendices, table contents, and bibliographies or references.

Three commonly preprocessing steps were conducted in this study: sentence splitting, tokenization and part-of-speech tagging.

Sentence splitting

Sentence splitting aims to split the input text or content into grammatically correct, meaningful sentences that are ready for further processing and analysis by identifying sentence boundaries, such as sentence-ending characters, e.g., exclamation points, question marks and periods (Qiu et al. 2019a).

Tokenization

Depending on the information extraction approach, tokenization is needed to split a continuous raw text into a sequence of words or tokens (Qiu et al. 2018a; Zhou and Elgohary 2017). For example, the raw text "*The main sedimentary strata are Triassic strata, accounting for about 80% of the strata.*" is tokenized into "The" "main" "sedimentary" "strata" "are" "Triassic" "strata" "," "accounting" "for" "about" "80%" "of" "the" "strata" "." (here the whitespace tokens are not shown). The goal of this process is to accurately recognize the boundary of sentences and then prepare for

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

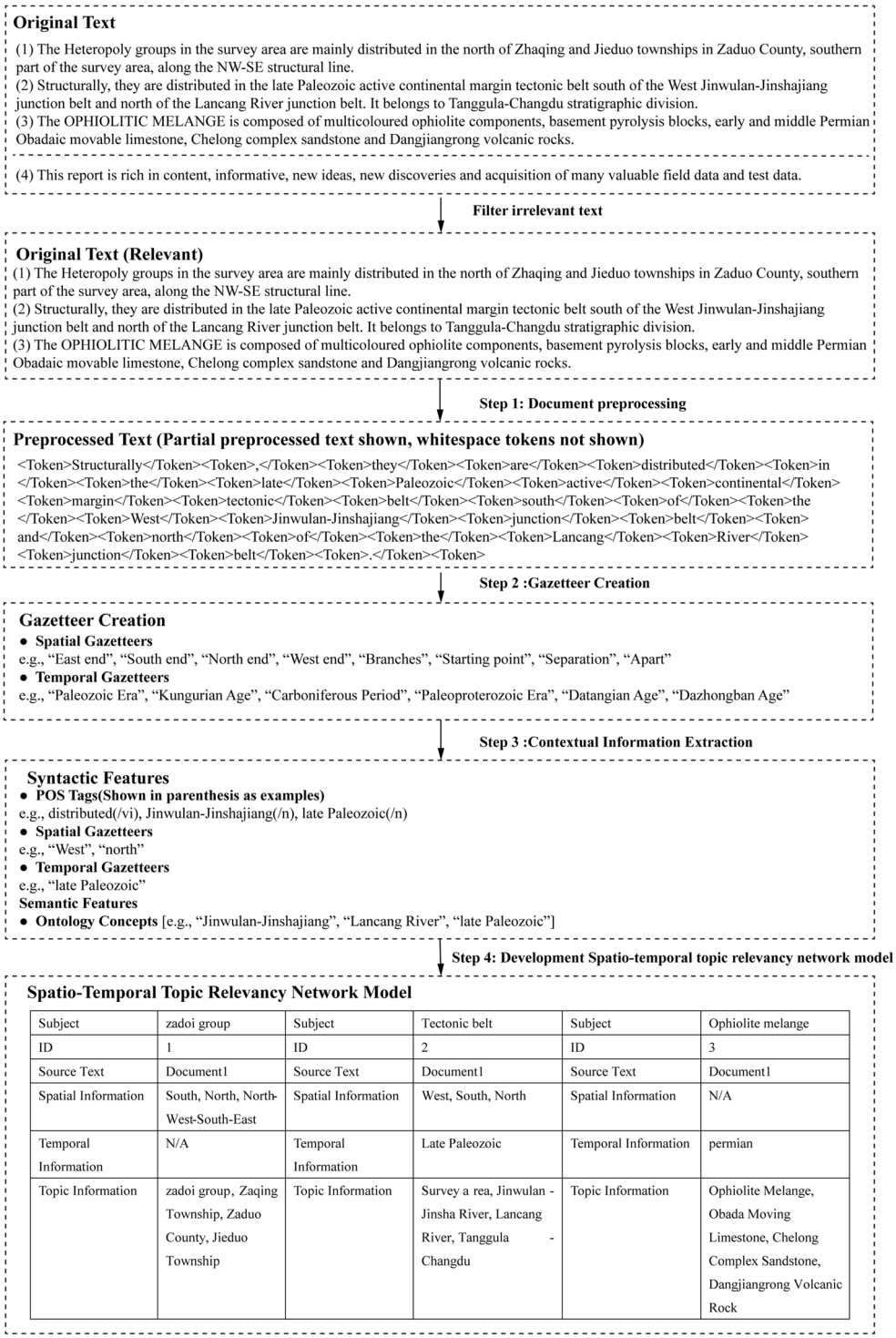


Fig. 2 An illustrative example of the incoming inputs and the relevant outputs of the main proposed methodology steps

further analysis (i.e., the part-of-speech (POS) task) (Holden et al. 2019).

Part-of-speech tagging

Within the information extraction task in the NLP domain, POS tagging represents a process that assigs a relevant tag to each token. A comprehensive list of POS tags can be found in the Penn Treebank (Toutanvoa and Manning 2000; Zhang et al. 2019). As demonstrated in Fig. 4, after performing the POS tagging process, each word in the input corpus is labeled with POS tags computed by the tagging approach. For example, "deep metamorphic rocks" and "colorful ophiolite

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

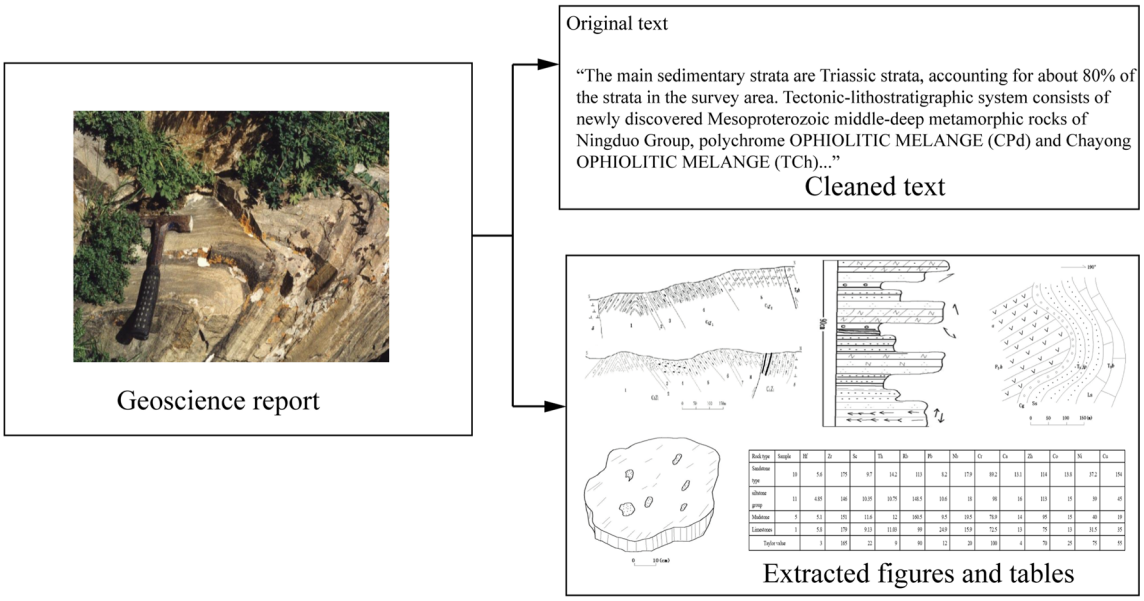


Fig. 3 An illustrative example of a geoscience report split into figures, tables and cleaned (relevant) text

melange" have the same syntactic features (noun class) within a sentence in terms of POS tags.

Step 2: Creation of gazetteers

A gazetteer is a summary of phrases/words that provide a shared and common category (i.e., a list of stratum and geological history).. Two main gazetteers were built and used to assist in the geological entity (i.e., rock, stratum and geological history) extraction from the source geological documents/reports.

In this research, geological named entities (GNEs) refer to the main and important words/terms, which include locations, stratum, geological history, rock and geological structures, while named entity recognition (NER) carries out the task of recognizing these GNEs in the geological documents/reports. With the proposed model, two methods are used to identify and recognize GNEs in the input text:

(1) Based on direct matching, which is a measure of the difference between a new, incoming term (whose interpretation is not known) and an entity (whose class is annotated and predefined), and

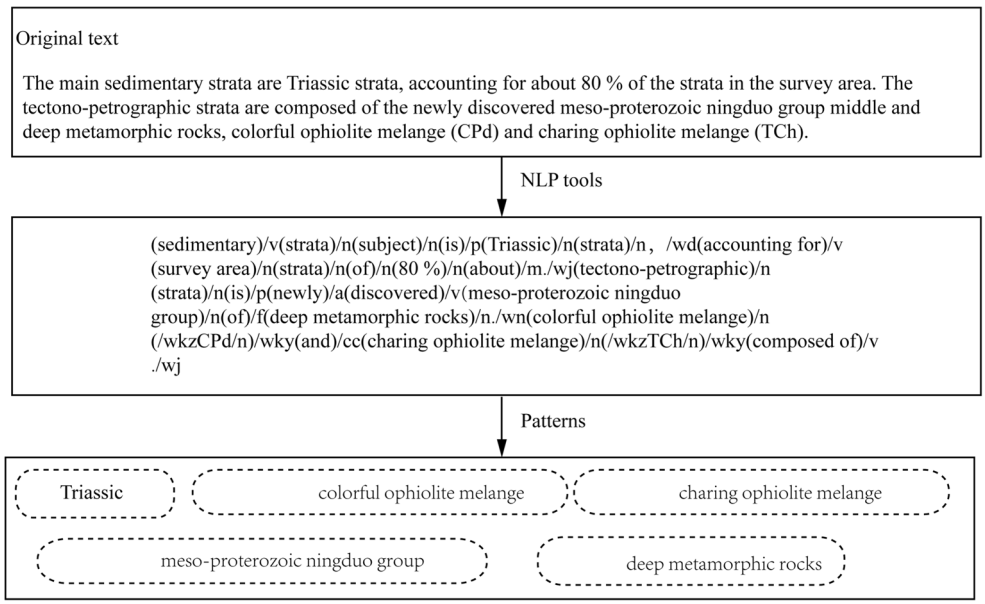


Fig. 4 Example information element instance undergoing POS tagging and named entities extraction

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

(2) Based on the developed set of comprehensive matching patterns using the Java Annotation Patterns Engine (JAPE) (Abraham et al. 2018) (described in the next section).

For the geological documents in this research, some entities (e.g., months, place names and direction indicators) that do not rely on hand-crafted matching rules/patterns depend on the development of a summary of named entities. According to the characteristics of the geological texts, a natural choice for the recognition of these entities is spatial gazetteers, which aim to handle spatial expressions and place names, and temporal gazetteers, which aim to handle temporal expressions and geological history.

One of the important goals in this research is to develop self-tailored gazetteers for guiding the incorporation, development and usage of gazetteers in information extraction. On this note, the JAPE transducer is used because it offers an available platform to construct the self-tailored spatial and temporal gazetteers.

Differing from other efforts that applied existing summaries of temporal taggers and spatial taggers, the geological documents used here include descriptions of spatial and temporal expressions. These include special temporal expressions that describe the sequence of the formation of a geological body or the occurrence of geological events, and include some traditional names that may have been changed or are no longer used and, hence, lack preexisting temporal and spatial taggers. As such, specially designed, self-tailored spatial and temporal gazetteers are considered suitable for this study.

Spatial gazetteer creation

As mentioned above, spatial gazetteers play an increasingly important role in space-time analysis and geographic information exploration. The designed gazetteers focus on recognizing and extracting geographic extents and domain-specific spatial information during the information extraction tasks using NLP models and techniques. Considering our research purpose, the spatial gazetteer refers to the traditional location names for Chinese places and lexical information related to spatial relations. Several Chinese place names recorded in the geological documents/reports may have been changed; therefore, we collected these place names along with relevant alternative names and appended them to the spatial gazetteer.

Frequently, natural language predicative phrases/words related to spatial relations or terms used to describe various spatial relations represent spatial locations and relational knowledge about spatial objects (Du et al. 2015; Du et al. 2015; Du et al. 2017). We used spatial relation terms that are summarized from Chinese text documents as data sources, classified and clustered each term in accordance with its spatial relation type, combination and part-of-speech constraint and, finally, formed a spatial relation vocabulary. A total of 3396 words that represent or reflect spatial relation meanings (i.e., orientation, topology and distance) were collected and developed into indicators for relation extraction.

Some developed spatial relations are shown in Table 1.

Temporal gazetteer creation

Temporal words/phrases in Chinese geological documents can be categorized into two types: domain-general temporal expressions and domain-specific temporal words/phrases.

The domain-general temporal expressions represent a usual and definite data format (i.e., DD-MM-YYYY). There is a need for acquiring all temporal information within geological documents as a set of representative expressions/terms to make recognition, identification and extraction possible. On this note, we have collected and summarized trigger words and expression patterns based on the linguistic characteristics of temporal information in Chinese text. The gazetteer refers to a summary of date formats that could be built with the source JAPE pattern-matching rules and then the development of a JAPE transducer.

It is noted that since there is no relative time (i.e., Friday, Wednesday, spring, morning and last month) in the geological documents, the rules we have created do not include relative time. The domain-general temporal expression gazetteer developed is composed of the date formats illustrated in Table 2.

The domain-specific temporal expressions related to geological time provide a contiguous framework for temporal intervals based on knowledge from stratigraphy, paleontology and chrono stratigraphy, which are used to study the Earth's history (Wang et al. 2018a, 2018b). Geological time plays an important role in studying the rock record and tracing the development and the resultant changes over time.

Ma et al. (2011) encoded a multilingual thesaurus of geological time scale with an extended the Simple Knowledge Organization System (SKOS) model. They implemented methods of characteristic-oriented term retrieval in JavaScript programs for recognizing and translating geological time-scale terms in online geological maps. Cox and Richard (2015) describe an OWL2 ontology (W3C OWL Working Group 2012) derived from the C&R model, and a set of vocabulary instances formalized using this ontology,

Table 1 Details of the developed relations

Relation Terms	Example
topological relation	contain, belong to, attach to, tributary, branch, intersection, flow into
distance relation	front, forward, upward, anterior, posterior
direction relation	distance, apart, apart from, separation

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 2 The domain-general temporal entity gazetteer

No.	Temporal Entity	Pattern
1	Date	June 2019
2	Date	June 23
3	Date	June 23,2019
4	Date	23 June
5	Date	23 June 2019
6	Date	23.06
7	Date	23.06.2019

whose content is based on versions of the timescale published by the ICS as the International [Chrono]stratigraphic Chart (ICC). Although the existing the time Ontology for temporal topology and the geologic time ontologies can be used to construct spatiotemporal gazetteers, Chinese geological ontology that we have obtained may more accurately serve the extraction information from geological report written in Chinese. One possible work is collecting and enriching the obtained ontology for spatiotemporal gazetteers from other data sources in order to obtain satisfactory semantic expressions in geosciences.

Geological time terms are collected from natural-language textual content (e.g., the textual content from Wikipedia articles), published literature, and the geological ontology (Hwang et al. 2012; Qiu et al. 2019b). Finally, a total of 742 geological time terms were collected and used, such as "Zanclean Age", "Mesoproterozoic Era", and "Paleoproterozoic Era". The domain-specific temporal entity gazetteer is illustrated in Table 3.

Step 3: Contextual information extraction

After preprocessing the geological text and developing the gazetteers, a comprehensive set of features were designed for further contextual information extraction (Step 3). The formally defined gazetteers (spatial gazetteer and temporal gazetteer) and the preprocessed geological documents/reports were utilized as the input for assisting contextual information extraction. Based on the input, the goals were to build the features, recognize named entities from the input documents and match them to the gazetteers, label the names of places, and capture the spatiotemporal information.

Table 3 The domain-specific temporal entity gazetteer

No.	Temporal Entity	Example
1	eon	Phanerozoic Eon, Proterozoic Eon, Archean Eon,
2	era	Cenozoic Era, Mesozoic Era, Paleozoic Era, Early Palaeozoic Era
3	period	Quaternary Period, Paleogene Period, Sinian Period, Qingbaikouan Period
4	epoch	Holocene, Pleistocene, Upper Pleistocene, Middle Pleistocene, Lower Pleistocene
5	age	Chattian Age, Rupelian Age, Priabonian Age, Bartonian Age, Lutetian Age, Ypresian Age

The problem of extracting contextual information from the geological documents/reports in this subsection is defined as an NER, a spatiotemporal and relationship task, and a topic information extraction task. These are accomplished using the gazetteers and a comprehensive set of JAPE rules. The order for extracting the above information is very vital because the output results of the NER are the input of the spatiotemporal and relationship pipeline. Fig. 5 demonstrates the core parts of the extraction workflow.

Named entity extraction

The task of NE extraction aims to automatically identify and classify named units/words into a summary of predefined entity categories that are designed and defined by the users who have background knowledge of both geological engineering and the summarization of geological reports/documents within the NLP domain. The defined entity elements (i.e., target information units) in this research contain place names, stratum, geological structures, and rock and data expressions. The named entity extraction pipeline consists of three processing sections: text processing, gazetteer creation and JAPE transducer development. Text processing is a process that handles the text's context for further analysis. Subsequently, JAPE transducers and gazetteer matching are performed. The named entity extraction is performed either by the JAPE transducer or the gazetteer processing based on the input of labeled entities. In this stage, all entities and elements from the incoming geological reports/documents are expected to be matched, annotated and subsequently assigned a relevant class in the information elements. After this processing, the named entities (i.e., place names, stratum, geological structures, and rock and data expressions) are labeled and prepared for the spatiotemporal relationship extraction.

As mentioned above, the named entity extraction can be performed either by the JAPE transducer or gazetteer matching. The predefined named entities for this study consist of place names, stratum, geological structures, and rock and data expressions. In our research, the NE extraction is based on measuring the match in a list of the gazetteer using the gazetteer annotator. Once a match is returned, the word/phrase is assigned a relevant category class.

JAPE represents a finite state machine that performs tagging based on a comprehensive set of regular expressions. It is used here for semantic information extraction, which depends on its ability to match patterns. JAPE grammar contains a list of various terms, each of which includes a rule or pattern. For

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

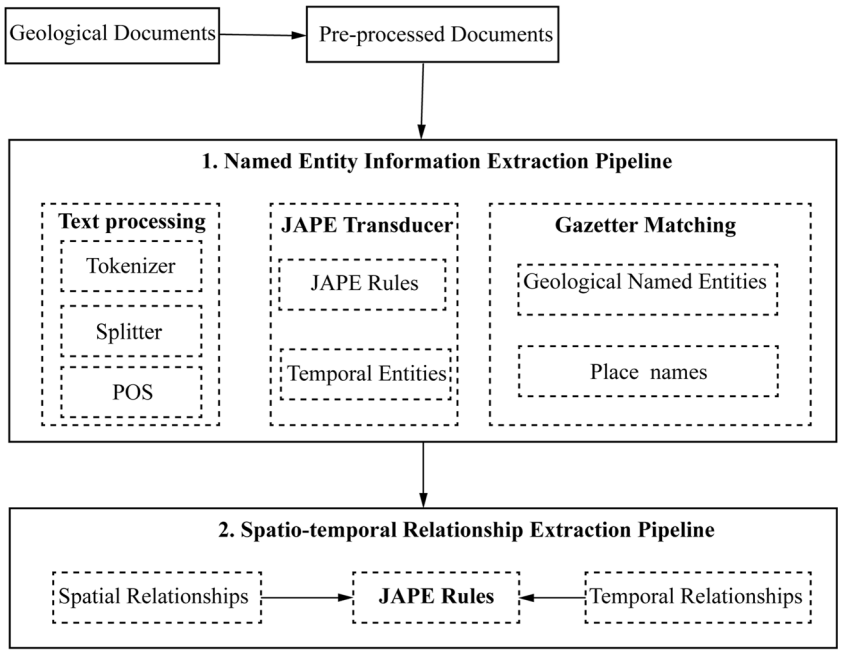


Fig. 5 The information extraction pipeline

example, a developed temporal transducer was applied here to match and extract the domain-general temporal entities based on a summary of seven data patterns (see Table 2), each of which represents a unique rule that is used to describe the pattern. For the named entity extraction task in this stage, we constructed two types of transducers: the temporal transducer for extracting temporal expressions and the geological names transducer for recognizing various names that contain a pattern including phrases from the other named entities (e.g., Sheinwoodian Stage (Age) and Nagaoling Stage (Age)).

For a total of seven data formats shown in Table 2, some JAPE grammar rules were built. The temporal transducer makes full use of the tagging categories from text processing and gazetteer creation.

Spatio-temporal information and spatiotemporal relationship extraction

After the named entity extraction pipeline, the incoming text is labeled and then assigned to the annotation classes. Information such as "two days later" (which represents a temporal period) and "15 km from the river" (which represents spatial locations) demonstrate different aspects of spatiotemporal relationships. The extraction of spatiotemporal relationships focuses on recognizing and extracting all spatial and temporal expressions/terms for the incoming geological content in natural language.

Natural language spatial relations express relational knowledge about and relative locations of spatial objects (Du et al. 2015). For natural language spatial relation terms/expressions, such descriptions can be used to understand when or where a geological event occurred. For this step, the workflow for the extraction of spatiotemporal relations includes two types of transducers, namely, the spatial relation transducer and the temporal relation transducer. As described before, the process of extracting the spatiotemporal relationships is followed by NER, since it uses the outputs of the NER process (the annotation classes) as input.

For the spatial relation transducer, this operation was done using a labeled set of location entities based on the NER process and subsequently assigned the relation tag from the annotated location entities. We summarized the linguistic characteristics from the Chinese geological reports and defined a total of 9 rules/patterns for the full set of spatial terms/expression and patterns/rules gathered from the reports/documents. The spatial relation patterns/rules are defined and used as follows:

(1) *Rule 1*: Deterministic spatial position relationship, such as East longitude 94 30′; North latitude 33 00′.
(2) *Rule 2*: Between spatial location and spatial location, such as Qinghai and Tibet.
(3) *Rule 3*: Directional indicator accompanied by locations, such as northern Qinghai.
(4) *Rule 4*: Distance accompanied by a location followed by a direction indicator, such as 50 km north of Qinghai.
(5) *Rule 5*: Spatial terms/phrases accompanied by a location(s), followed by other locations, such as areas of Qinghai, Tibet and Xinjiang.
(6) *Rule 6*: Directional indicator accompanied by "of" and location, such as north of Qinghai.
(7) *Rule 7*: "From" accompanied by a location, followed by" via", location," and", and a location, such as from Qinghai via Tibet, Xinjiang and Urumqi to Shihezi.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

(8) *Rule 8*: "about" accompanied by distance, such as 25 km.
(9) *Rule 9*: Deterministic spatial attitude information, such as "position 165°".

The temporal relationship transducer uses similar patterns and rules to those in the above spatial relationship transducer. For example, "between June and July" is extracted and then assigned to an annotation class of temporal relation.

Topic information extraction

Topic information represents significant and meaningful expressions containing one or more terms/words in a document. They highlight the body of the text/document to assist users' fast reading and browsing and have been widely applied in various NLP tasks (Rafieiasl and Nickabadi 2017; Yang et al. 2017).

After above preprocessing, a document from geological reports is assigned multiple keyphrases (as topic information) based on an ontology and a modified word embedding-based method (Qiu et al. 2019a, 2019b).

Step 4: Spatio-temporal topic relevancy network model

The fourth step in the framework is the construction of the model of the Geologic Spatio-Temporal Topic Relevancy Network (GSTTRN). Fig. 6 demonstrates the framework of the GSTTRN.

The traditional spatial association methods, which are based on repeated scanning of databases, are limited in supporting the development of associations because spatial attributes and nonspatial attributes are separated to mine spatial association rules or because they only consider the relationship between the time dimension and spatial dimension. They do not take the nonspatial features contained in spatiotemporal data, such as temporal features and topic features (i.e., mineral categories and geological structures), into account. Most of them are based on similarity calculations or spatial association rules used for mining classical spatial data features (i.e., location, time, attributes).

To address the insufficiencies of the existing spatiotemporal data association as it relates to extracting spatiotemporal and topic features of geological data, we analyzed the association of spatial information (i.e., location and range of survey area), temporal information (i.e., geological age) and topic information (i.e., rock type and geological structure) from geological data. From this analysis, we established a formal description of the spatiotemporal and topic association for geological data in the domain of geoscience, i.e., the GSTTRN.

As mentioned previously, the goal of GSTTRN is to extract and search sematic information by combining the extracted spatial and temporal terms. In the process of information retrieval, the developed model first builds an index file for the geological document, which records the location of paragraphs containing different topic words and the ID of the document. Then, the developed model finds the location of the topics in geological documents through indexed items and

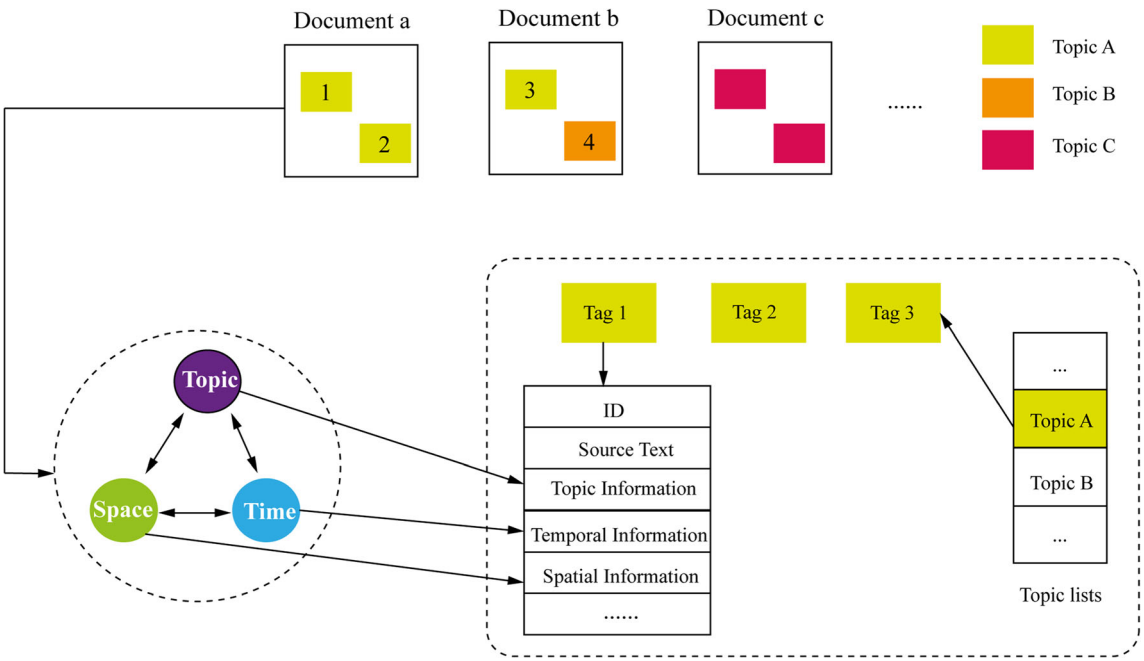


Fig. 6 Geologic spatiotemporal topic relevancy network model

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

obtains the location information from the geological data by inputting query topic words. The attribute information and spatial location information of geological topics are related to structured spatial data; in other words, they establish the relationship between geological document data and structured spatial data. Finally, the model executes the integrated semantic query from space to topic (i.e., keywords) and from topic to space by "searching text by text", "searching graph by text" and "searching text by graph".

Experiments on the analysis and searching of documents/reports

The proposed workflow offers a methodology that can support the visualization of machine fast-reading summaries and subsequently apply the outputs for further searching of geological documents/reports. For the extracted geological topics presented in the above section, it constructs a knowledge graph (Section 4.2) and extracts related tables and figures (Section 4.3) from individual documents/reports, extracts spatiotemporal and semantic information (Section 4.4), and offers autocompletion suggestions depending on keyword cooccurrence information from the developed collective corpus (Section 4.5).

Corpus

A total of 41 geological reports were selected as the corpus that are supported by the China Geological Survey. The reports chosen were considered to be representative because they (1) recorded geological information (i.e., spatiotemporal and semantic information) from different years by different geologists and professionals, (2) were for various types of geological reports that include representative and valuable information, and (3) demonstrated the domain-specific uniqueness and complexities along with various patterns that range from simple to complex.

Knowledge graph of the document

The valuable information stored in geological documents/reports can be correlated based on content words with high frequency (Hovy and Lin 1998). In this subsection, we conducted statistics on the content words with the TF-IDF method for document-level text. TF-IDF is a statistical approach that is introduced into NLP to extract information about word frequency statistics (Jones 1972). In the specific domain, there are some informational terms (i.e., domain-specific words) that occur with a low frequency. The extracted results by TF-IDF method were demonstrated in Fig. 7.

Before we computed the frequency of the content words, the function words/terms and a set of well-informed common

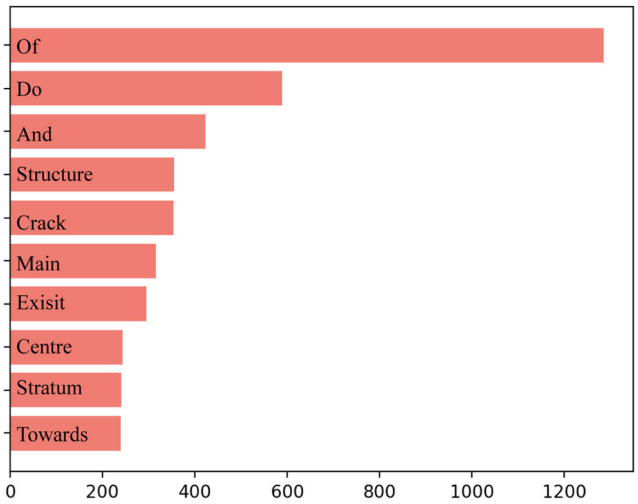


Fig. 7 Top 10 content-words extracted by TF-IDF approach

words were removed based on a word-matching approach with a collective stop-words corpus. The content-word frequency results in the geological document/report are shown in Fig. 8 (each panel represents the statistical results of a geological report) and Fig. 9. The word cloud is composed of content words exceeding the threshold n (in our experiment, n is set to 50) based on the corresponding frequency statistics, which gives a clear visualization of content in the geological document/report.

The content words are the indicators that can be used to reflect the information and knowledge in the reports. The core nodes and relations of the content words reflect the geological information and knowledge. In this research, we apply the knowledge graph (i.e., the bigram graph) to visualize the relations of the content words from geological reports. The knowledge network demonstrates the relation between the

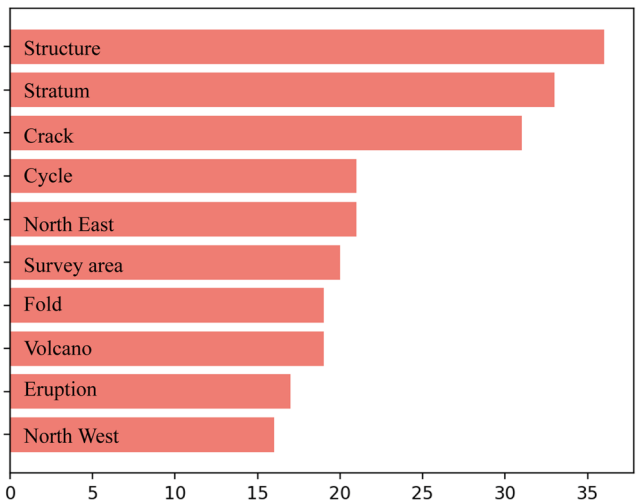


Fig. 8 Content words with high frequency after removing some stop-words

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

(a) (b)

Fig. 9 Word clouds built from extracted words illustrate a visual and brief representation of information stored in geological reports. The font size of word reflects word frequency in the geological report

content words, as shown in Figs. 10 and 11. It contains various information and knowledge related to the domain of geoscience, including topics such as rocks, geology, and data processing, and represents the essential knowledge of a geological report.

As shown in Fig. 10, the black arrows reflect word sequences and point to the latter content words. For example, the word *Structure* attaches to *Gneiss*, *Evolution*, *Unit*, and *Superimposed* directly and clearly. This outcome indicates that the relation between the word *Structure* and other words in geological reports represents an important characteristic.

Contextual-information extraction

Table 3 illustrates the performances in the proposed workflow of spatiotemporal and semantic information extraction from geological documents using the standard evaluation metrics of precision, recall and F1-score (Manning et al. 1999). To calculate these metrics, we applied the following equations: TP, FP, and FN are the number of true positive samples, false positive samples, and false negative prediction samples, respectively.

$$\begin{aligned} P &= \frac{TP}{TP + FP} \\ R &= \frac{TP}{TP + FN} \\ F1 &= \frac{2PR}{P + R} \end{aligned} \tag{1}$$

Since recognized entities (i.e., place name entities) are applied altogether for further analysis in the following step, our

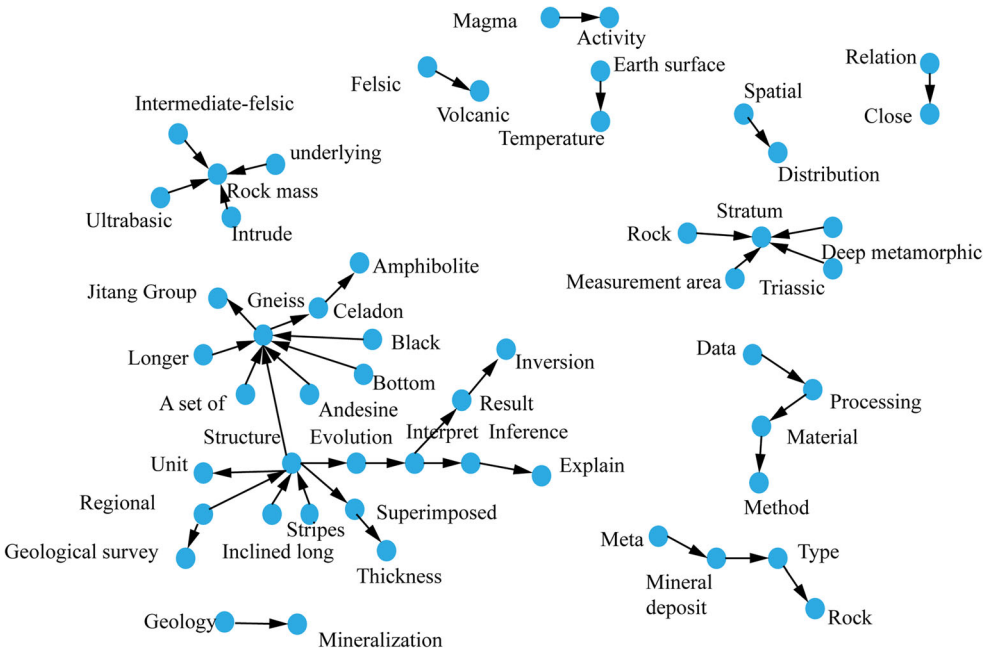


Fig. 10 Bigram graph of content words in a geological report demonstrates the key information. The black arrow represents the word sequence along with the arrow pointing to the other word pairs

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

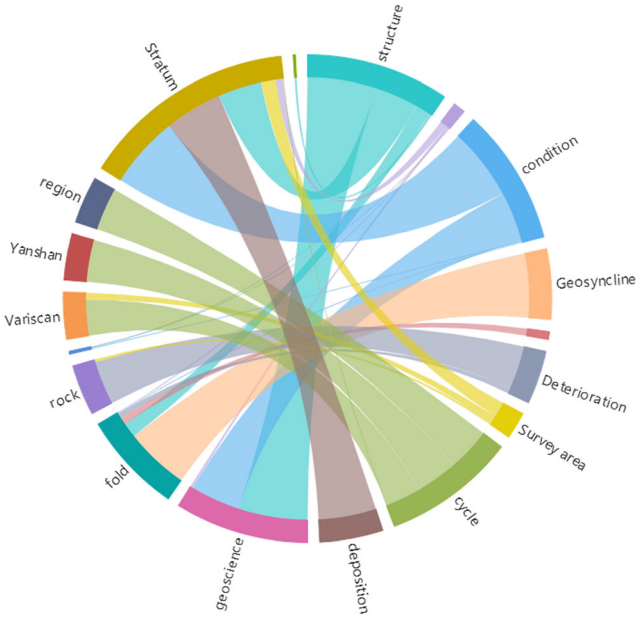


Fig. 11 Chord graph in the geological report

NER evaluation is entity-based: a true positive sample represents the entire entity (i.e., "Xin Jiang" as an entirety compared to single "Jiang") identified as a place.

As seen in Table 4, the experimental results show that the proposed approach for semantic information extraction achieves better performance than spatiotemporal information extraction. The proposed information extraction algorithm obtains an average precision, recall, and F1 of 90.25%, 89.06, and 89.65, respectively.

Figures and tables extraction

As mentioned in Section 3.1, figures and tables are an important information repository, which represent some key information in the geological reports. We extracted a list of figures and tables from the geological reports by recognizing their captions by a method that has been used in Sematic Scholar (Clark and Divvala 2016). The core goal of extracting figures and tables from geological reports is to present them in a format that allows for fast browsing of a list of relevant maps and other tables and figures to recognize the visual content.

Table 4 Performances of spatiotemporal information and semantic information extraction

Category	P	R	F1
Spatial Information	86.42	85.11	85.76
Temporal Information	91.12	90.06	90.59
Semantic Information	93.22	92.01	92.61
Average	90.25	89.06	89.65

Clearly, this format offers an effective approach when a very large number of documents/reports are required to search for a domain-specific map or a set of relevant results.

Figure 12 illustrates the output of the geological document/report "Regional geological survey report 146C 003004". This example suggests that the proposed workflow automatically extracts and identifies a list of figures and tables from the geological documents/reports and lists similar geological reports based on topic information.

Autogenerated keyphrase suggestions

Existing models and techniques browse and search the geological reports/documents in a repository based on word-pattern matching, which limits their support of autogenerated keyphrase suggestions. Further, when geologists or readers query geological reports, they need to not only obtain the relevant text information but also the information in figures and tables.

In this research, the proposed information extraction methodology offers autocompletions using the developed spatiotemporal topic relevancy network model. Fig. 13 demonstrates a user input of 'siltstone' and then the resulting automatically generated suggestions. After selecting the search keywords, the designed tool provides the query results from selected reports (e.g., Inner Mongolia Autonomous Region regional mineral geology survey map instruction L50E017023 delesitai fire station). Report L50E017023 documents exploration in the Inner Mongolia Autonomous Region during 2005. This report also documents/records past relevant exploration in this area and reviews a summary list of domain-general geological content and information that contains the geographical location and a detailed description of relevant rock types in the Inner Mongolia Autonomous Region. A main part of the report presents a list of details of the explorations, including major achievements, ore-forming geological conditions, geophysical, geochemical and remote sensing characteristics, regional mineral and metallogenic regularity and mineral predictions. Exceptionally, this report includes multiple maps and relevant photographs of the area.

The search results ('siltstone' as the input) included in the document are listed in Fig. 14. The developed spatiotemporal topic relevancy network model automatically extracts and annotates elements (i.e., topic, figures and tables) from the geological reports. Once the query is received based on keyword input by users (i.e., geologists or readers), the tool will recommend similar query results based on the annotated topic information. For example, when a user inputs 'siltstone', the sentence-level text will be searched for spatiotemporal information and figures and tables related to 'siltstone', such as southwest angle (spatial information) and clastic rock group (temporal information).

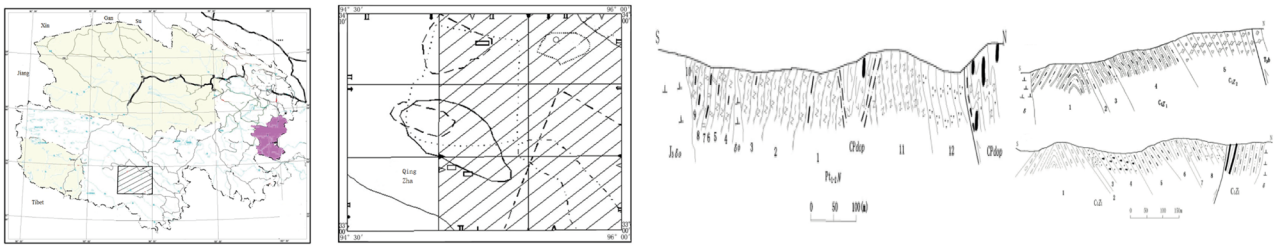
Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Regional geological survey report
146 C 003004

Abstract

I46 C 003004 (zhiduo county), I46 C 004004 (ziduo county) 1:250,000 regional geological survey (joint survey)

Lists of figures and tables



Rock type	Sample	Cu	Pb	Zn	Cr	NI	Co	Ba	Sr	Rb	V	Mn	Au	Ta	Nb	Zr	Th	Ag	Ti
Gneisses	11	17	306	110	116	48	2.4	419	102	214	152	135	1.05	1.2	15	223	20	0.017	8352
Schists	12	19	33	61	75	33	10	686	129	117	99	435	1.28	4.7	25	252	11.5	0.06	4211
Marble type	3	105	25	72	54	66	27	335	478	44	72	1082	1.25	13.4	7.5	83	2	0.21	5401
Quartzite	17	6.8	21	30	41	85	6.5	653	112	61.5	58	269	1.65	0.95	8.5	16.3	8.9	0.09	2553
Taylor value		55	12.5	70	100	75	25	425	375	90	135	950	0.43	2	20	165	9	70	5700

Rock type	Sample	Hf	Zr	Sc	Th	Rb	Pb	Nb	Cr	Ca	Zh	Co	Ni	Cu
Sandstone type	10	5.6	175	9.7	14.2	113	8.2	17.9	89.2	13.1	114	13.8	37.2	154
siltstone group	11	4.85	146	10.35	10.75	148.5	10.6	18	98	16	113	15	39	45
Mudstone	5	5.1	151	11.6	12	160.5	9.5	19.5	78.9	14	95	15	40	19
Limestones	1	5.8	179	9.13	11.03	99	24.9	15.9	72.5	13	75	13	31.5	35
Taylor value		3	165	22	9	90	12	20	100	4	70	25	75	55

Most similar documents by topic information

- Regional geological and mineral survey in the area of aljin mountain gulkou spring, ruoqiang county, xinjiang
- Regional geological survey report of yangchun city, guangdong province (F49C002003)
- Regional geological survey report of zengjiangda county, zengchangdu county, nangqian county

Fig. 12 An example of extracting the figures and tables from geological reports. The proposed workflow can automatically recommend similar reports based on topic information

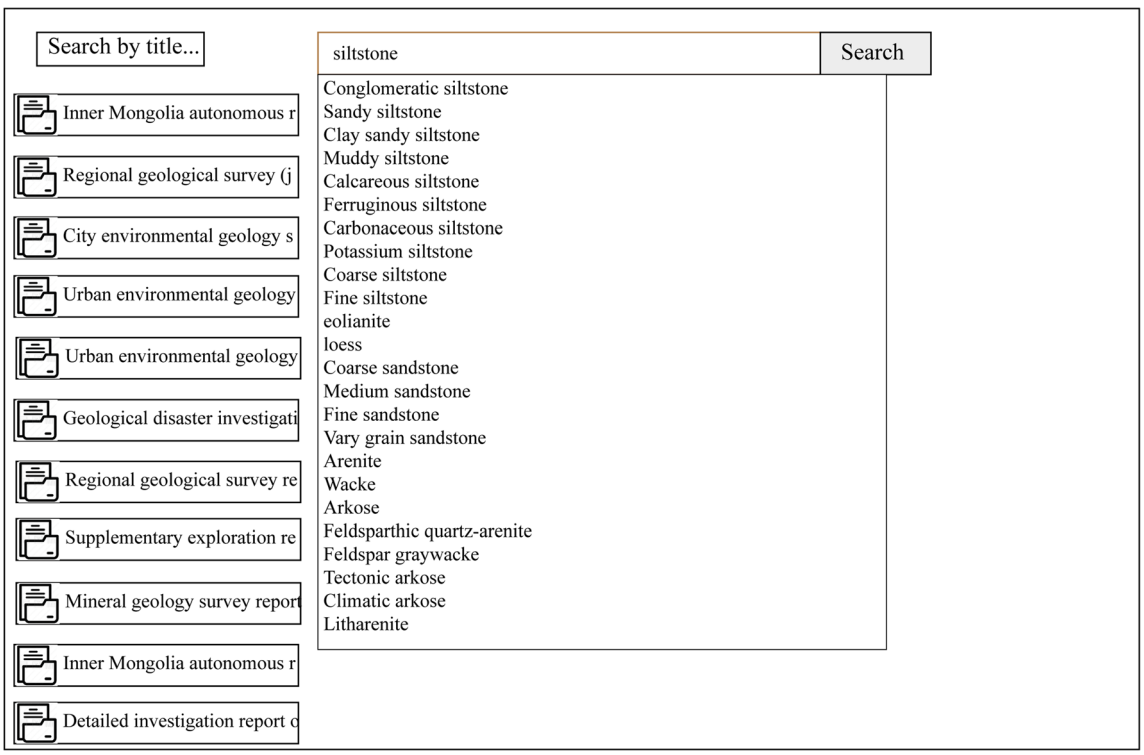


Fig. 13 Autogenerated keyphrase suggestions search based on the input entity of 'siltstone'

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

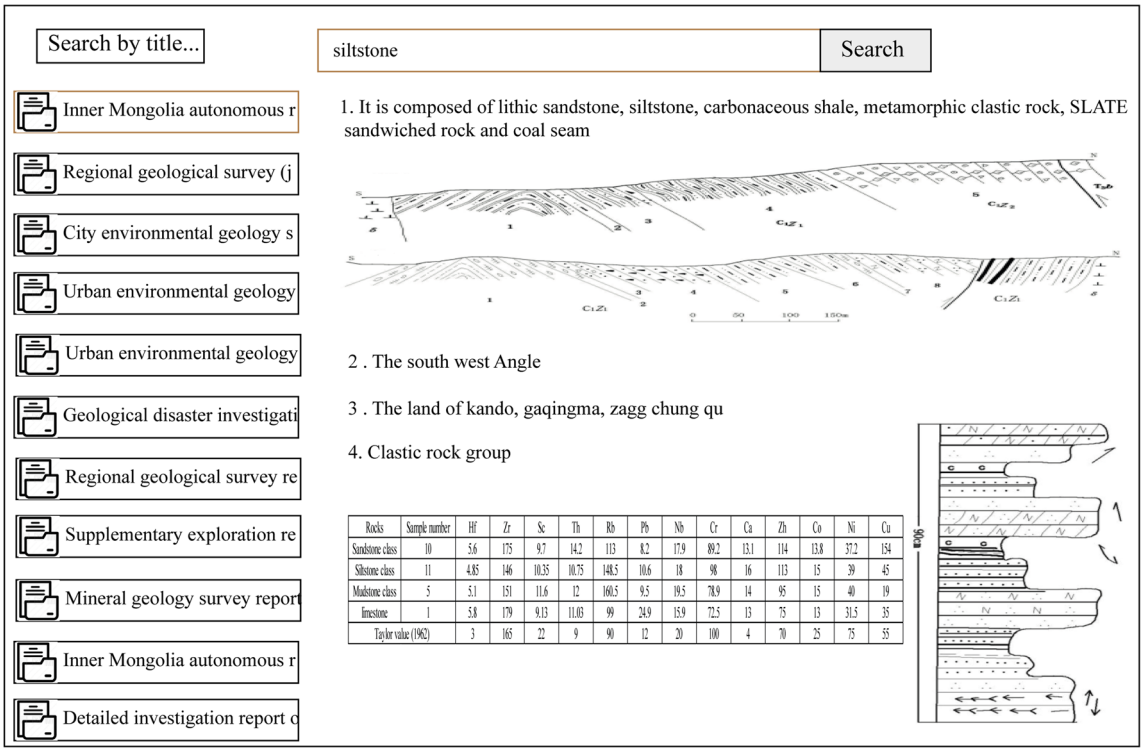


Fig. 14 Autogenerated keyphrase suggestions search results based on the input entity of 'siltstone'

Conclusions and future work

In this study, a workflow designed to automatically extract relevant spatiotemporal and semantic information from geological reports using text mining and NLP techniques is presented. The proposed information extraction approach allows for adapting the existing text mining and NLP techniques and offers a valuable tool for fast browsing documents/reports/literature and their geological information and content. Based on a customized set of domain-specific extracted geological information, the cooccurrence relations in a knowledge graph pattern are an effective and representative way to analyze, explore and rapidly read the content of the reports. In addition, the proposed methodology can also identify figures and tables from a selected document/report, extract a visualization of the figures and tables and recognize similar documents/reports based on the geological topic, providing a useful and effective way to leverage textual data to search geological content and information.

Ongoing research efforts in information extraction are focusing on the development of a more accurate information extraction methodology. The existing ontologies about domain-specific topics in geosciences, such as geological time scale, geological structure, and rock deformation can lead to innovative functions in smart geoscience data, and will provide solid support to geoscience researchers in data discovery and analysis. We will focus on developing a knowledge graph which differs from the graph demonstrated in this paper that is based on keyword co-occurrences only. More sophisticated text mining techniques are required to provide meaningful connections between keywords in the knowledge graph. Such extension can assist in identifying and extracting geological information and knowledge to establish a robust search framework in geological environments.

Acknowledgments We would like to thank the anonymous reviewers for carefully reading this paper and their very useful comments. This study was financially supported by the National Natural Science Foundation of China (U1711267, 41671400, 41871311, 41871305), the National Key Research and Development Program (2018YFB0505500, 2018YFB0505504).

Author contributions Conceived and designed the experiments: Qinjun Qiu, Liufeng Tao and Zhong Xie; Performed the experiments: Qinjun Qiu, Liufeng Tao, and Zhong Xie; Analyzed the data: Qinjun Qiu, Liufeng Tao, and Zhong Xie; Wrote the paper: Qinjun Qiu, Liang Wu, Zhong Xie and Liufeng Tao.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

References

Abraham S, Mas S, Bernard L (2018) Extraction of spatio-temporal data about historical events from text documents. Trans GIS 22(3):677–696

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Clark, C, Divvala, S, (2016). 2.0: mining figures from research papers. In: IEEE/ACM joint conference on digital libraries (JCDL) IEEE, pp. 143–152

Cox S, Richard SM (2015) A geologic timescale ontology and service. Earth Sci Inf 8(1):5–19

De Sa C, Ratner A, Re C, Shin J, Wang F, Wu S, Zhang C (2016) DeepDive: declarative Knowledge Base construction. International conference on management of data 45(1):60–67

Du S, Guo L (2016) Similarity measurements on multi-scale qualitative locations. Trans GIS 20(6):824–847

Du S, Feng C, Guo L (2015) Integrative representation and inference of qualitative locations about points, lines, and polygons. Int J Geogr Inf Sci 29(6):980–1006

Du S, Wang X, Feng C, Zhang X (2017) Classifying natural-language spatial relation terms with random forest algorithm. Int J Geogr Inf Sci 31(3):542–568

Enkhsaikhan, M, Liu, W, Holden, EJ, Duuring, P, (2018). Towards geological knowledge discovery using vector-based semantic similarity. In: proceedings of the international conference on advanced data mining and applications. Springer, Cham, pp. 224–237

Harisinghaney, A, Dixit, A, Gupta, S, Arora, A, (2014). Text and image based spam email classification using KNN, Naïve Bayes and reverse DBSCAN algorithm. In: proceedings of international conference on optimization, Reliabilty, and information technology (ICROIT). IEEE, pp. 153–155

Holden E, Liu W, Horrocks T, Wang R, Wedge D, Duuring P, Beardsmore T (2019) GeoDocA - fast analysis of geological content in mineral exploration reports: a text mining approach. Ore Geol Rev 111:102919

Hovy, E, Lin, CY, (1998). Automated text summarization and the SUMMARIST system. In: proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998(TIPSTER "98). Association for Computational Linguistics, Stroudsburg, PA, pp. 197–214

Hwang J, Nam KW, Ryu KH (2012) Designing and implementing a geologic information system using a spatiotemporal ontology model for a geologic map of Korea. Comput Geosci 48:173–186

Ireson, N, Ciravegna, F, Califf, ME, Freitag, D, Kushmerick, N. and Lavelli, A (2005). Evaluating machine learning for information extraction. International conference on machine learning

Jones KS (1972) A statistical interpretation of term specificity and its applications in retrieval. J Doc 28(1):11–21

Konkol M, Brychcín T, Konopík M (2015) Latent semantics in named entity recognition. Expert Syst Appl 42(7):3470–3479

Lima, LA, Gornitz, N, Varella, LE, Vellasco, MM, Muller, K and Nakajima, S (2017). Porosity estimation by semi-supervised learning with sparsely available labeled samples. Computers & Geosciences, 33–48

Liu, K and Elgohary, N (2017). Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports. Automation in construction, 313-327

Liu W, Chung BC, Wang R, Ng JQ, Morlet N (2015) A genetic algorithm enabled ensemble for unsupervised medical term extraction from clinical letters. Health information science 3(1):1–14

Luo X, Zhou W, Wang W, Zhu Y, Deng J (2018a) Attention-based relation extraction with bidirectional gated recurrent unit and highway network in the analysis of geological data[J]. IEEE Access 6: 5705–5715

Luo, X, Zhou, W, Wang, W, Zhu, Y and Deng, J (2018b). Attention-Based Relation Extraction With Bidirectional Gated Recurrent Unit and Highway Network in the Analysis of Geological Data. IEEE Access, 5705–5715

Ma X, Carranza EJ, Wu C, Der Meer FD, Liu G (2011) A SKOS-based multilingual thesaurus of geological time scale for interoperability of online geological maps. Comput Geosci 37(10):1602–1615

Ma K, Wu L, Tao L, Li W, Xie Z (2018) Matching descriptions to spatial entities using a Siamese hierarchical attention network. IEEE Access 6:28064–28072

Manning, CD, Manning, CD and Schütze, H (1999). Foundations of statistical natural language processing. MIT press

Moens, MF (2006). Information extraction: algorithms and prospects in a retrieval context (Vol. 21). Springer Science & Business Media

Nadeau, D, Sekine, S, (2007). A survey of named entity recognition and classification. Linguisticae Investigationes 30 (1), 3–26 Publisher: John Benjamins publishing company

Paulus, R, Xiong, C and Socher, R (2018). A deep reinforced model for abstractive summarization. International conference on learning representations

Peters SE, McClennen M (2015) The Paleobiology database application programming interface. Paleobiology 42:1–7

Peters SE, Zhang C, Livny M, Re C (2014) A machine reading system for assembling synthetic paleontological databases. PLoS One 9(12): e113523

Peters SE, Husson JM, Wilcots J (2017) The rise and fall of stromatolites in shallow marine environments. Geology 45(6):487–490

Qiu Q, Xie Z, Wu L (2018a) A cyclic self-learning Chinese word segmentation for the geoscience domain. Geomatica 72(1):16–26

Qiu, Q, Xie, Z, Wu, L and Li, W (2018b). DGeoSegmenter: A dictionary-based Chinese word segmenter for the geoscience domain. Computers & Geosciences, 1–11

Qiu, Q, Xie, Z, Wu, L and Li, W (2019a). Geoscience keyphrase extraction algorithm using enhanced word embedding. Expert Systems With Applications, 157–169

Qiu, Q, Xie, Z, Wu, L and Tao, L (2019b). GNER: a generative model for geological named entity recognition without labeled data using deep learning. Earth and Space Science

Qiu, Q, Xie, Z, Wu, L and Tao, L (2020). Dictionary-based automated information extraction from geological documents using a deep learning algorithm. Earth and Space Science, 7, e2019EA000993. https://doi.org/10.1029/2019EA000993

Rafieiasl, J and Nickabadi, A (2017). TSAKE: a topical and structural automatic keyphrase extractor. Applied soft computing, 620-630

Schuhmacher, M, Ponzetto, SP, (2014). Knowledge-based graph document modeling. In: proceedings of the 7th ACM international conference on web search and data mining, pp. 543–552

Shi, L, Jianping, C and Jie, X (2018). Prospecting information extraction by text mining based on convolutional neural networks–a case study of the Lala copper deposit, China. IEEE access, 52286-52297

Toutanvoa, K and Manning, CD (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. Empirical methods in natural language processing: 63-70

Wang, R, Liu, W, McDonald, C, (2015). Using word embeddings to enhance keyword identification for scientific publications. In: Databases Theory and Applications. Springer, pp. 257–268

Wang C, Ma X, Chen J (2018a) Ontology-driven data integration and visualization for exploring regional geologic time and paleontological information. Comput Geosci 115:12–19

Wang C, Ma X, Chen J, Chen J (2018b) Information extraction and knowledge graph construction from geoscience literature. Comput Geosci 112:112–120

Wong W, Liu W, Bennamoun M (2012) Ontology learning from text: a look back and into the future. ACM Comput Surv 44(4):20

Wu, L, Xue, L, Li, C, Lv, X, Chen, Z, Jiang, B, Guo M and Xie, Z (2017). A knowledge-driven geospatially enabled framework for geological big data. ISPRS Int J Geo Inf, 6(6)

Yang S, Lu W, Yang D, Li X, Wu C, Wei B (2017) KeyphraseDS: automatic generation of survey by exploiting keyphrase information. Neurocomputing 224:58–70

Yang, D, Wang, S, Li, Z, (2018). Ensemble neural relation extraction with adaptive boosting. In: proceedings of the 27th international joint conference on artificial intelligence. IJCAI'18 AAAI press,

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

pp. 4532–4538. http://dl.acm.org/citation.cfm? Id=3304222.3304400

Young T, Hazarika D, Poria S, Cambria E (2018) Recent trends in deep learning based natural language processing. Ieee. Computational intelligenCe magazine 13(3):55–75

Zhang J, Elgohary N (2016) Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking. J Comput Civ Eng 30(2):04015014

Zhang, Y, Chen, M, Liu, L, (2015). A review on text mining. In: proceedings of the 6th IEEE international conference on software engineering and service science (ICSESS) IEEE, pp. 681–685

Zhang F, Fleyeh H, Wang X, Lu M (2019) Construction site accident analysis using text mining and natural language processing techniques. Autom Constr 99:238–248

Zhou, P and Elgohary, N (2017). Ontology-based automated information extraction from building energy conservation codes. Automation in construction, 103-117

Zhou P, Xu J, Qi Z, Bao H, Chen Z, Xu B (2018) Distant supervision for relation extraction with hierarchical selective attention. Neural Netw 108:240–247

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com