Data set entity recognition based on distant supervision

Pengcheng Li, Qikai Liu, Qikai Cheng and Wei Lu
*School of Information Management,
Wuhan University, Wuhan, Hubei, China*

Received 31 October 2020
Revised 6 January 2021
27 March 2021
6 May 2021
15 May 2021
Accepted 27 May 2021

Abstract

Purpose – This paper aims to identify data set entities in scientific literature. To address poor recognition caused by a lack of training corpora in existing studies, a distant supervised learning-based approach is proposed to identify data set entities automatically from large-scale scientific literature in an open domain.

Design/methodology/approach – Firstly, the authors use a dictionary combined with a bootstrapping strategy to create a labelled corpus to apply supervised learning. Secondly, a bidirectional encoder representation from transformers (BERT)-based neural model was applied to identify data set entities in the scientific literature automatically. Finally, two data augmentation techniques, entity replacement and entity masking, were introduced to enhance the model generalisability and improve the recognition of data set entities.

Findings – In the absence of training data, the proposed method can effectively identify data set entities in large-scale scientific papers. The BERT-based vectorised representation and data augmentation techniques enable significant improvements in the generality and robustness of named entity recognition models, especially in long-tailed data set entity recognition.

Originality/value – This paper provides a practical research method for automatically recognising data set entities in scientific literature. To the best of the authors' knowledge, this is the first attempt to apply distant learning to the study of data set entity recognition. The authors introduce a robust vectorised representation and two data augmentation strategies (entity replacement and entity masking) to address the problem inherent in distant supervised learning methods, which the existing research has mostly ignored. The experimental results demonstrate that our approach effectively improves the recognition of data set entities, especially long-tailed data set entities.

Keywords Data set entity recognition, Distant supervision, Scientific literature, Data augmentation, Long-tailed entities, Library automation, Distance learning, Database management

Paper type Research paper

1. Introduction

In recent decades, digital resource management has been an important research topic in the electronic library field (Li and Liu, 2019), but limited attention has been paid to data set resources overall. Data sets play an indispensable role in numerous scientific studies as a valuable digital asset serving to describe research tasks and verify research methods (Tanwani and Farooq, 2009; Vaidhehi, 2014). Identifying the data set used in research is crucial because it could increase the replicability and verifiability of the research. However, the recent increase in cutting-edge technologies, such as deep learning, has made data-driven research a new paradigm (Parish and Duraisamy, 2016). Various data sets for different research tasks have been proposed, resulting in an explosion in the number of data sets and scientific publications. The automatic identification of data set entities in large-scale scientific literature has become an urgent need to conduct further research to better manage the exponential increase in data set resources.

The Electronic Library
Vol. 39 No. 3, 2021
pp. 435-449
© Emerald Publishing Limited
0264-0473
DOI 10.1108/EL-10-2020-0301

Similar studies in this area have specifically dealt with data usage statement extraction, which focuses on extracting claims about how particular data or data sets are obtained, processed or used in an article. Slightly differently, the purpose of data set entity identification is to identify words or phrases that refer to the name of a data set, which is essentially information extraction at the word granularity level. As a classic task in the information extraction field, named entity recognition (NER) has made great progress in the past two decades. Related supporting technologies, such as word segmentation, part-of-speech tagging and syntactic analysis, have also been improved (Palshikar, 2013). With the rapid development of supervised models based on statistical learning, most of the existing studies have transformed information extraction into sequence labelling problems (Yadav and Bethard, 2019). However, while NER research on the traditional seven subclasses (person name, place name, institution name, time, date, currency and percentage) is relatively mature (Ruokolainen *et al.*, 2019), studies on data set entity identification are still in the exploratory stage. Until recently, as far as we know, most existing supervision-based studies on data set entity identification have been conducted on small samples in limited domains and cannot be applied to large-scale scientific literature sets in open domains.

Currently, research on the identification of specific named entities, such as identification of diseases, drugs and genes in the field of medicine (Dong *et al.*, 2016; Li *et al.*, 2015), drug identification in the chemical field (Lamurias *et al.*, 2015) and algorithm identification in the field of computer science (Tuarob *et al.*, 2016), has become a new trend. As for identifying data set entities in scientific literature, although scholars have made some progress after nearly a decade of continuous exploration, the majority of work on the identification and extraction of data set entities has used manual or rule-based methods (Krüger and Schindler, 2019). After complex manual analysis and feature construction, manual or rule-based approaches can perform well on a specific corpus (Maxim *et al.*, 2017). However, these approaches have limitations, such as high manual involvement, poor generalisation and time-consuming engineering, and cannot be applied to large-scale literature in open domains. Hence, supervised learning (Névéol *et al.*, 2011) and rules combined with supervised learning (Duck *et al.*, 2016) are expected to enable the automatic recognition of data set entities. However, the quality and size of a training corpus severely limit the effectiveness of supervised learning-based recognition. Accessible training corpora for data set entity recognition studies are not only small in volume but also limited in domain. The provision of a labelled corpus is a bottleneck for the supervision-based automatic recognition of data set entities under the open domain.

To address the above problems in existing studies, this paper proposes a distant supervision-based approach to identify data set entities automatically from large-scale literature. Although distant supervised learning has been applied in similar studies (Boland and Krüger, 2019), to the best of the authors' knowledge, this is the first time it has been applied in the study of data set entity recognition. The objective is to investigate whether and how a distant supervised NER approach can be used to identify data set entities from large-scale scientific publications. Distant supervision has the advantage of eliminating labour-intensive data labelling using the bootstrapping strategy, but this approach also has inherent drawbacks that have been ignored in most studies. Therefore, a robust vectorised representation and two data augmentation techniques are introduced to improve the method's generalisation and robustness. Moreover, while data sets exist in various domains, the proposed approach requires almost no expert knowledge and can be applied to open domains.

2. Literature review

Recent approaches to identifying data set entities from scientific publications can be divided into three groups: manual-based, rule-based and supervision-based. Manual identification includes all approaches in which humans read a text to identify a target entity. Because full-text content analysis is a time-consuming task, the manual method is often restricted to a limited number of articles or a specific domain. Because of the high reliability of results, manual methods are still active in specific entity recognition research. Zhao *et al.* (2018) used manual methods to count the usage of data sets in scientific publications and found that less than 30% of publications underwent data reuse. Yan and Weber (2018) used content analysis to categorise how open government data are used among different research communities, and provided descriptive statistics on the publication years, publication outlets and open government data sources.

For rule-based approaches, a set of rules created by experts are applied to identify the target entities. The needs of the task determine the complexity of these rules, and regular representation is one of the simplest needs. For example, Maxim *et al.* (2017) used regular expressions to detect unique data set identifiers in PubMed articles. With elaborate engineering, the rule-based approach can perform well on a specific corpus. Several data set citation patterns explored by Kafkas *et al.* (2013), for instance, demonstrated high database recognition precision on structured citation records in medicine-related journal articles. Subsequently, Duck *et al.* (2013) applied local cues and cross-mentioned cues as features to design a rule-based entity recognition system, BioNerDS, to identify data sets and software names in medical science literature. Furthermore, Ghavimi *et al.* (2016) proposed a semi-automatic approach based on special features extracted from data set titles to find data set references and links. Compared to manual methods, rule-based methods enable semi-automatic identification of data set entities across numerous documents. However, rule design relies heavily on expert knowledge, so most rules are only applicable to a specific corpus and are quite limited by the domain.

The supervised-based approach mostly consists of three steps:

(1) constructing a machine learning model;

(2) training this model on parts of a labelled corpus; and

(3) applying and evaluating the fitted model on another portion of the corpus.

Supervision-based recognition tasks are modelled as a sequence-labelling problem. Duck *et al.* (2015) discussed the ambiguity and variability of database and software names in bioinformatics. They then used conditional random fields (CRF) combined with rules to identify databases and software entities automatically. As deep neural networks have gained considerable attention in recent NER studies (Li *et al.*, 2020), scholars have made some neural attempts in the research of data set extraction. Based on a corpus of 5,000 documents provided by the Coleridge Initiative's Rich Context Competition, Prasad *et al.* (2019) explored the feasibility of various neural network models on data mention extraction. Moreover, they applied several joint learning strategies to explore the synergy between data set mention extraction and data set classification tasks. Compared to the previous two approaches, the supervised-based approach, which is free from requiring feature engineering and expert knowledge, is expected to achieve automatic recognition of data set entities from large-scale literature in the open domain.

Overall, the value of data set resources has been realised by researchers across different disciplines. Organisations, such as the Database Systems and Logic Programming Google and the Association for Computing Machinery Digital Library, have provided data set search engines to help researchers quickly find data sets that match their needs. Patra *et al.* (2020) recently built a data set recommendation system to recommend data sets to researchers based on their previous publications. In the past decade of research on data set entity recognition and data usage statement extraction, the manual and rule-based approaches are still the mainstream choices, and relatively few supervised learning-based approaches have been used. One reason might be the lack of large-scale, high-quality training corpora. The recognition effect of supervision-based methods depends heavily on the size and quality of the training data, and the provision of labelled data is commonly a bottleneck in supervision-based research of data set entity recognition. Several methods, such as weakly supervised (Hoffmann *et al.*, 2011) and unsupervised learning (Zhang and Elhadad, 2013), have been proposed to address training corpus acquisition. Zhang *et al.* (2017) proposed an unsupervised approach based on pattern lists to identify data usage at the article level. By applying a bootstrapping strategy to generate text patterns automatically, their method can achieve an F-measure of 85% in determining whether a data usage statement is included in computer science literature.

In this paper, the authors used and implemented a distant supervised method to automate the recognition of data set entities in large-scale scientific literature. Distant supervision is a weakly supervised learning method and has been applied in other similar studies (Boland and Krüger, 2019). To the best of the authors' knowledge, no previous attempt has been made to use a distant supervised approach in data set entity recognition research. To overcome the inherent drawbacks of distant supervision, they further introduce a robust vectorised representation and two data augmentation techniques to improve the recognition performance of data set entities (especially long-tailed data set entities). In addition, they chose academic literature in computer science as the experimental data because most relevant work has explored the medical field, while data set entity recognition in computer science deserves additional effort. Although only computer science literature was used as the experimental data set in this paper, the proposed approach can be applied to any other field because the method requires minimal human involvement.

3. Methodology

3.1 Research design

To overcome the lack of training corpora, this paper uses a supervised learning approach to enable the automatic identification of data set entities from large-scale scientific literature. Specifically, a dictionary was used combined with a bootstrapping strategy to create corpora with noisy labels to apply supervised learning.

In the beginning, the authors selected some common data set names to build their seed dictionary. A set of seed terms was used to label the entity names of the seed words mentioned in the text corpus. Subsequently, the labelled data sets were used to train the NER model constructed. The fitted NER model was applied to the same corpus to identify new data set names appearing in the text. These new data set entity words were added to the original seed dictionary. As the name *bootstrapping* implies, these processes were iteratively repeated until the dictionary size remained constant or met expectations.

Their distant supervised approach requires very few interactions to create the seed dictionary, and the entire weakly supervised learning process can be completed by applying a bootstrapping strategy. However, this approach has some inherent drawbacks. For frequently used data sets, the model might recognise these entities directly by their names rather than statement information. For uncommon data sets, such as long-tailed data set entities, there is a lack of sufficient training corpora for the model to learn contextual features, resulting in poor entity recognition. Long-tailed entities are rare and often relevant only in specific knowledge domains; yet, they are important for retrieval and exploration purposes. Thus, based on the application of distant supervised learning, also introduced was a robust vectorised representation and two data augmentation techniques to enable the effective identification of data set entities in large-scale scientific literature.

3.2 Data labelling using the bootstrapping strategy

Crowdsourcing provides a method for generating large-scale labelled data, but this process is very expensive and requires annotators to have certain literature reading capabilities and domain-related knowledge. In this paper, the authors adopted a data labelling method based on the bootstrapping strategy combined with a dictionary.

Though their approach is not limited by domain, they chose the computer science field as the source of the experimental corpus because considerable relevant work has been conducted in other fields, such as medicine. In contrast, data set entity recognition in the computer domain deserves more research effort. Specifically, they selected the full text on literature published in the Association of Computational Linguistics (ACL) anthology for two reasons. Firstly, the ACL website provides an interface that allows users to access the full text of literature published each year. Secondly, literature in the ACL anthology primarily comprises empirical studies involving data sets that contain numerous data set entities.

The specific process was conducted as follows. After obtaining the full-text articles in the XML format from the interface provided by the ACL website, they used the Natural Language Toolkit (NLTK; www.nltk.org) to split all the articles into sentence sets. Next, a seed dictionary containing approximately 1,000 common data set entity words was manually constructed and used to label the entity names of the seed words mentioned in the sentence set. Then, the NER model with the labelled corpus was trained, and the fitted NER model was used to identify new data set entities in the sentence set. They added the new data set entities to their original seed dictionary and iteratively repeated this process until the dictionary and corpus reached the desired size.

After the iterative repetition of the above process, they obtained a large number of labelled data sets by applying a dictionary combined with a bootstrapping strategy. The entire process requires only a few human interventions. Although some noisy labels are in these labelled data sets, the authors believe that the interference resulting from these noisy labels would be gradually weakened by more positive example labels as the data volume increases.

3.3 Named entity recognition model based on neural networks

In recent years, empowered by continuous real-value vector representations and semantic compositions through nonlinear processing, deep learning has been used in NER systems, yielding state-of-the-art performance (Li *et al.*, 2020). In particular, the continuous development of pre-trained model techniques, represented by bidirectional encoder representation from transformers (BERT; Devlin *et al.*, 2018), offers the possibility of more robust recognition of various specific entities (Akbik *et al.*, 2019). The NER model adopted in this paper is the classic combination of long short-term memory (LSTM) and CRF, which has shown powerful capabilities in recognising multiple entities from a text (Lample *et al.*, 2016). To improve the recognition of data set entities, they used the BERT network in the embedding layer. Following the general idea behind word embedding, the BERT network model further increases the generalisability of the word vector model. By mining multi granularity feature relations at the character, vocabulary and sentence levels, BERT provides a vectorised representation containing considerable context information to support the identification of data set entities in text.

Figure 1 illustrates the structural details of the NER model used in this paper. For the input sentence, “WordNet groups words together based on their meanings”, the role of the embedding layer is to form the vector representation of the text sequence, allowing the computer to read and compute the underlying semantic information in the text. The authors applied the BERT network in the embedding layer to map the text sequences into a multidimensional vector space. As presented in Figure 1, the final vectorised representation of the text consists of three components: token embedding, segment embedding and position embedding.

After the embedding layer, the vectorised representation of the text is fed into the LSTM layer to further extract the potential semantic information of the sentence. Subsequently, the output of the LSMT layer is input into the CRF layer to calculate the distributional probabilities of whether each word in the sentence is a data set entity. In the output layer, the data set entity word “WordNet” in the text is classified into the category “Entity”, and other non-data set entity words are classified into the category “Other”.

3.4 Data augmentation

Data augmentation, a data-space solution to the problem of limited data, is widely used in computer vision (Shorten and Khoshgoftaar, 2019) and natural language processing tasks (Ferreira and Costa, 2020). Data augmentation encompasses a suite of techniques that enhance the training corpus size, providing solutions to overfitting problems caused by insufficiently labelled corpora. Wei and Zou (2019) proposed an easy data augmentation (EDA) technique to boost the text classification task performance. The EDA technique consists of four data augmentation methods: synonym replacement, random insertion, random swap and random deletion. Their experimental results demonstrated that EDA improves convolutional and recurrent neural network performance, especially on smaller data sets.

The authors obtained a large training corpus with noisy labels by applying a dictionary combined with a bootstrapping strategy. Although the increased training corpus size can

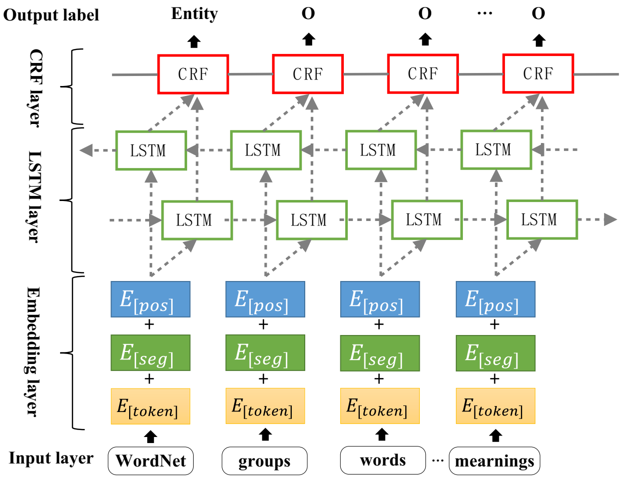


Figure 1.
NER model consisting of BERT, LSTM and CRF

weaken the interference caused by noisy labels, it also exacerbates the training data imbalance. Numerous labelled corpora might be available for NER model training for data set entities that frequently appear, resulting in overfitting during high-frequency data set entity recognition. In such cases, the data set entity is highly likely to be identified by name words rather than semantic information. In addition, for unpopular long-tailed data set entities, data may be insufficient to support the training and fitting of the NER model, resulting in underfitting and poor recognition performance. To address these issues, they introduced two data augmentation strategies: entity replacement and masking:

(1) *Entity replacement*: Replace the data set entity words in the sentences with other entity words, which require the replacement entities to have not appeared in the training corpus.

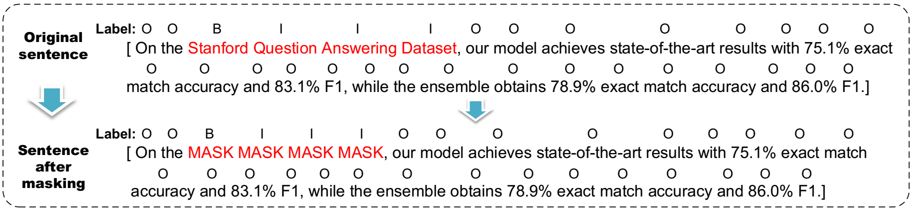
(2) *Entity masking*: Replace the data set entity words in the sentences with "unknown words" with no actual meaning.

Figure 2 illustrates an example of the entity masking strategy, where the data set entity words in a sentence are randomly replaced with other meaningless words. In the specific experiments, they followed the parameter settings of the token mask task in BERT and randomly selected 20% corpus sentences for entity replacement and masking, respectively. Afterward, the entity-replaced or entity-masked corpus was merged with the original corpus to produce an extended training corpus for feature learning and NER model fitting.

3.5 Experiment

After acquiring approximately 200,000 articles published in 2010–2019 from the ACL website, regular expressions were used to extract the required fields from the XML files. Subsequently, the NLTK was applied to complete sentence segmentation (Figure 3). Finally, they created a data set containing 10,747,988 sentences after discarding sentences that were too long or too short.

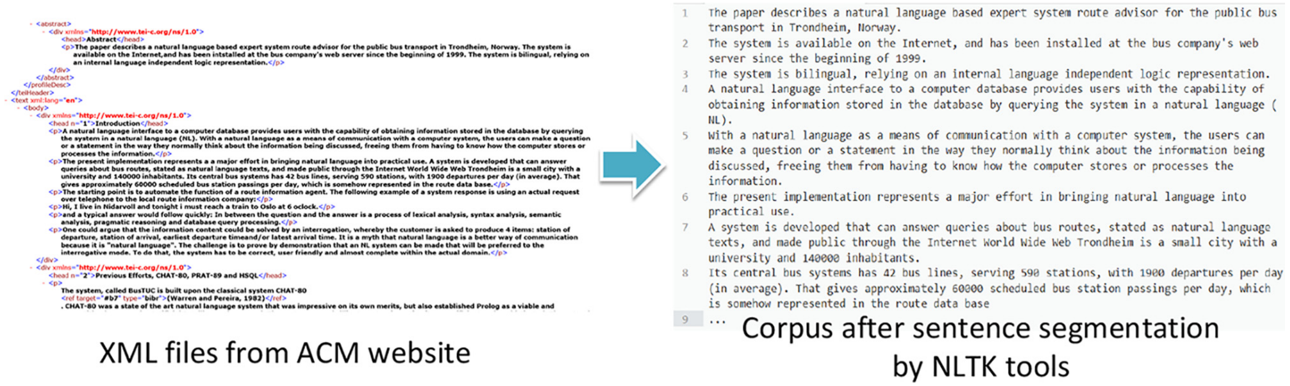
The authors used a dictionary combined with a bootstrapping strategy to create the labelled training data automatically. To construct the seed data set dictionary, they manually collected the common and frequently used data set list. Because the dictionary size did not meet their expectations, they expanded the dictionary through a data set search engine, such as Kaggle and Google, and finally gained a seed dictionary containing approximately 1,000 data set entities. With the automatic labelling of the sentence corpus in an iterative manner, a dictionary with 11,280 data set entities and 70,313 sentences about data set usage were obtained. In addition, they manually selected and labelled 1,000 sentences as gold-standard testing data to measure the effectiveness of their method. The data set entities mentioned in these 1,000 sentences did not appear in the constructed data set dictionary; therefore, it is reasonable to regard these data sets as long-tailed entities. More information about the experimental data is presented in Table 1.


Original sentence
Label: O O B I I I O O O O O O O O
O O O O O O O O O O O O O O O
Sentence after masking
Label: O O B I I I O O O O O O O O O
O O O O O O O O O O O O O O

[On the Stanford Question Answering Dataset, our model achieves state-of-the-art results with 75.1% exact match accuracy and 83.1% F1, while the ensemble obtains 78.9% exact match accuracy and 86.0% F1.]

[On the MASK MASK MASK MASK, our model achieves state-of-the-art results with 75.1% exact match accuracy and 83.1% F1, while the ensemble obtains 78.9% exact match accuracy and 86.0% F1.]

Figure 2.
Masking data entities in a sentence



```
- <abstract>
  - <div xmlns="http://www.tei-c.org/ns/1.0">
      <head>Abstract</head>
      <p>The paper describes a natural language based expert system route advisor for the public bus transport in Trondheim, Norway. The system is
        available on the Internet,and has been intstalled at the bus company's web server since the beginning of 1999. The system is bilingual, relying on
        an internal language independent logic representation.</p>
    </div>
  </abstract>
 </profileDesc>
</teiHeader>
- <text xml:lang="en">
  - <body>
    - <div xmlns="http://www.tei-c.org/ns/1.0">
        <head n="1">Introduction</head>
        <p>A natural language interface to a computer database provides users with the capability of obtaining information stored in the database by querying
          the system in a natural language (NL). With a natural language as a means of communication with a computer system, the users can make a question
          or a statement in the way they normally think about the information being discussed, freeing them from having to know how the computer stores or
          processes the information.</p>
        <p>The present implementation represents a a major effort in bringing natural language into practical use. A system is developed that can answer
          queries about bus routes, stated as natural language texts, and made public through the Internet World Wide Web Trondheim is a small city with a
          university and 140000 inhabitants. Its central bus systems has 42 bus lines, serving 590 stations, with 1900 departures per day (in average). That
          gives approximately 60000 scheduled bus station passings per day, which is somehow represented in the route data base.</p>
        <p>The starting point is to automate the function of a route information agent. The following example of a system response is using an actual request
          over telephone to the local route information company:</p>
        <p>Hi, I live in Nidarvoll and tonight i must reach a train to Oslo at 6 oclock.</p>
        <p>and a typical answer would follow quickly: In between the question and the answer is a process of lexical analysis, syntax analysis, semantic
          analysis, pragmatic reasoning and database query processing.</p>
        <p>One could argue that the information content could be solved by an interrogation, whereby the customer is asked to produce 4 items: station of
          departure, station of arrival, earliest departure timeand/or latest arrival time. It is a myth that natural language is a better way of communication
          because it is "natural language". The challenge is to prove by demonstration that an NL system can be made that will be preferred to the
          interrogative mode. To do that, the system has to be correct, user friendly and almost complete within the actual domain.</p>
      </div>
    - <div xmlns="http://www.tei-c.org/ns/1.0">
        <head n="2">Previous Efforts, CHAT-80, PRAT-89 and HSQL</head>
      - <p>
          The system, called BusTUC is built upon the classical system CHAT-80
          <ref target="#b7" type="bibr">(Warren and Pereira, 1982)</ref>
          . CHAT-80 was a state of the art natural language system that was impressive on its own merits, but also established Prolog as a viable and
```

```
1  The paper describes a natural language based expert system route advisor for the public bus
   transport in Trondheim, Norway.
2  The system is available on the Internet, and has been installed at the bus company's web
   server since the beginning of 1999.
3  The system is bilingual, relying on an internal language independent logic representation.
4  A natural language interface to a computer database provides users with the capability of
   obtaining information stored in the database by querying the system in a natural language (
   NL).
5  With a natural language as a means of communication with a computer system, the users can
   make a question or a statement in the way they normally think about the information being
   discussed, freeing them from having to know how the computer stores or processes the
   information.
6  The present implementation represents a major effort in bringing natural language into
   practical use.
7  A system is developed that can answer queries about bus routes, stated as natural language
   texts, and made public through the Internet World Wide Web Trondheim is a small city with a
   university and 140000 inhabitants.
8  Its central bus systems has 42 bus lines, serving 590 stations, with 1900 departures per day
   (in average). That gives approximately 60000 scheduled bus station passings per day, which
   is somehow represented in the route data base
9  ...
```

Figure 3.
Original corpus required for NER model training

With further analysis of the labelled corpus obtained, they found that most of the labelled data were concentrated in commonly used data set entities. The top 25 data set entities with the highest frequency in the labelled corpus are presented in Figure 4. The results in Figure 4 indicate that numerous uncommon data sets in scientific literature are distributed in the tail, with a few popular data sets distributed in the head. These long-tailed data sets are infrequently used but are indispensable and extremely large, and identifying long-tailed data set entities from scientific literature is of great theoretical and practical importance.

The imbalance in the training data leads to two problems. First, for common data set entities, the model may appear to be overfitted. A large volume of homogeneous training data would reinforce the name features of the data set entities, resulting in recognition of the entity more often via the name words rather than the semantic information. Second, for long-tailed data set entities, no sufficient training corpus exists for the model to learn the contextual features of the entity words, resulting in underfitting. They first used BERT-based vectorised representation to assist in semantic feature extraction from the text to address these problems. Subsequently, they adopted two data augmentation methods, entity replacement and entity masking, to enhance the data set entity recognition performance. By replacing entity terms in sentences with other words, the NER model is forced to learn contextual information about entity words, which leads to improvements in the generalisation and robustness of data set entity recognition.

Obtained XML documents	219,829	Scale of the manually constructed test data set	1,000
Sentences after NLTK split	10,747,988	Sentences containing the description of the data set	70,313
Size of the seed data set dictionary	1,000	Size of the expanded data set dictionary	11,280

Table 1. Overview of experimental data

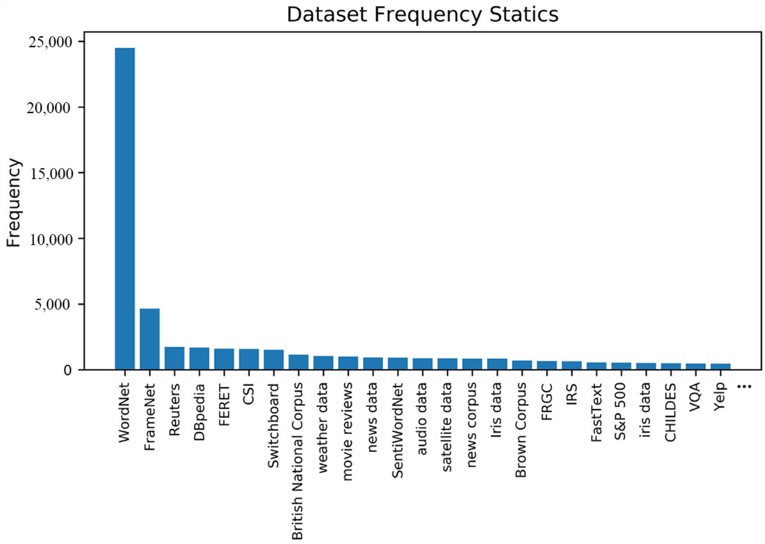


Figure 4. Frequency distribution of the data set entities matched in the corpus

Specifically, according to the size of the labelled corpus obtained, the training, validation and testing sets were divided into a ratio of 90:5:5. They also manually labelled 1,000 sentences as gold-standard testing data to measure the effectiveness of the method. As the data set entity words in the gold-standard testing data never appeared in the model fitting process, these gold-standard testing data can be used to test the recognition performance on the long-tailed data set entities. Finally, they randomly selected 20% of the corpus sentences for entity replacement and masking.

To investigate the effects of vectorised representation on data set entity recognition, they set up four groups of control experiments, as follows:

1. LSTM + CRF: LSTM + CRF is the most widely used neural network in the current NER research and performs well in recognising various entities. In this paper, they used the GloVe (Pennington *et al.*, 2014) vectorised representation as the text feature input for LSTM + CRF and took the result as the benchmark.
2. LSTM + CRF + Char: LSTM + CRF + Char adopts both the GloVe word vector and self-trained character vector as inputs. For character-level word embedding, each character of a word is associated with a vector, and they ran the bi-LSTM over the sequence of character embeddings and concatenated the final states to obtain a fixed-size vector.
3. BERT + LSTM + CRF: The BERT network generated the multigrained dynamic representation of the input text, and the final vectorised representation of the text consisted of token embedding, segment embedding and position embedding. In the BERT + LSTM + CRF model, they did not reuse the additional GloVe word vector and character-level word embedding.
4. SCIBERT + LSTM + CRF: SCIBERT is a pre-trained BERT using a total of 1.14 million scientific papers in the biomedical (82%) and computer science (12%) directions, and may be more suitable for natural language processing tasks in the scientific paper direction.

4. Evaluation

4.1 Metrics

To measure the performance of the data set entity recognition task, they used the precision, recall and F1-measure as the evaluation metrics for the experiment. In this paper, precision reflects how many of the results predicted as data set entities are correct, whereas recall reflects how many data set entities can be correctly identified. The F1-measure reflects the overall recognition performance and is calculated as follows:

$$\mathrm{F1} = 2 \times \frac{\mathit{Precision} \times \mathrm{Recall}}{\mathit{Precision} + \mathrm{Recall}}$$

4.2 Results and discussion

The identification performance of the distant supervised approach with the bootstrapping strategy on the data set entities is illustrated in Table 2. The data set labelled with the bootstrapping strategy comprised the rule-based testing data, and gold-standard testing data were their manually labelled data set.

When the authors compared the results of LSTM + CRF with those of LSTM + CRF + Char, they found that the recognition effect decreased in all metrics after adding the character-level vector feature inputs, which is somewhat different from their expectations. In general, a richer semantic representation should improve the model's ability to capture semantic information and lead to better recognition performance. They speculate that this is because of the mutual interference between the self-trained character-level vectors and the pre-trained GloVe lexical-level vectors. The combination of these two vectorised representations instead restricted the model learning and fitting, leading to a decline in the effectiveness of data set entity recognition.

In addition, LSTM + CRF and BERT + LSTM + CRF performed similarly in the rule-based test data but exhibited a significant difference in their performance using the gold-standard testing data. For the gold-standard testing data, the original precision decreased slightly (6%) after applying the BERT network, but both the recall and F1-measure values improved significantly, with recall improving by 38% and the F1-measure improving by 17%. When using SCIBERT to replace the original BERT, the results of SCIBERT + LSTM + CRF are further improved in terms of precision, recall and F1. These results indicate that the text vectorisation provided by the BERT network enables a considerable improvement in the recognition of data set entities, and the application of the BERT network can effectively enhance the generalisability of the NER model. Meanwhile, compared to the original BERT, SCIBERT is more suitable for data set entity recognition task in the direction of scientific papers, which achieves the best performance on both rule-based testing data and gold-standard testing data.

In contrast to the results using the rule-based testing data, the performance of the four models using the gold-standard testing data indicated different degrees of decline. Among them, BERT + LSTM + CRF and SCIBERT + LSTM + CRF had the smallest decline, demonstrating the usefulness of the BERT network in robustly identifying data set entities from the scientific literature. Subsequently, they found that although all models experienced a decline in value for all three indicators, the extent of the decline varied across indicators. In particular, the recall declined significantly more than the other two indicators, with LSTM + CRF showing the most pronounced decline in recall from 85.71 to 44.56 (48%). Meanwhile, the most obvious decrease in precision is LSTM + CRF + Char, which decreases from 79.21 to 60.4 (24%). Additionally, the low recall means that their NER model can only identify partial data set entities (e.g. common data sets), but is less capable of identifying long-tailed data set entities that are not commonly used.

To alleviate the data imbalance caused by bootstrapping-based data labelling, they introduced two data augmentation methods to improve the recognition performance on data set entities (especially long-tailed data set entities). To examine the utility of data augmentation techniques, they used SCIBERT + LSTM + CRF with the best performance on the gold-standard testing data as the baseline for the following experiments and applied entity replacement and masking. The bottom half of Table 2 lists the experimental results

Experimental results of Dataset entity	Rule-based testing data			Gold-standard testing data		
recognition on different models	Precision	Recall	F1	Precision	Recall	F1
LSTM + CRF	86.41	85.71	86.06	*74.37*	44.56	55.73
LSTM + CRF + Char	79.21	75.36	77.24	60.40	47.89	53.42
BERT+ LSTM + CRF	86.09	86.21	86.15	69.89	61.55	65.46
SCIBERT + LSTM + CRF	*88.70*	*88.60*	*88.65*	71.34	*62.65*	*66.71*
SCIBERT + LSTM + CRF_Replace	89.58	86.33	87.92	78.75	68.08	73.03
SCIBERT + LSTM + CRF_Mask	*91.00*	*86.89*	*88.90*	*80.76*	*71.22*	*75.69*

Note: Italic values indicate the highest number in each column

Table 2. Experimental results of data set entity recognition

after applying entity replacement and masking. After randomly copying 20% of the samples from the corpus, SCIBERT + LSTM + CRF_Replace replaces the data set entity words in the samples with other data set entity words, and SCIBERT + LSTM + CRF_Mask replaces the data set entity words in the samples with words that have no real meaning.

After applying data augmentation techniques, the performance of the NER model decreased for the rule-based testing data but significantly improved with the gold-standard testing data. The opposite effect confirms the existence of the data imbalance problem caused by bootstrapping-based data labelling. For high-frequency data sets, too many training samples exist for a single data set entity. The model may learn more about the name words of entities than the contextual information, and the strong dependence on the name feature might cause overfitting. For low-frequency data sets, such as long-tailed data set entities, there are insufficient training corpora for the model to learn and fit the features; thus, underfitting occurs. In this paper, they introduced the bootstrapping strategy to solve the problem of lacking labelled data under supervised learning, but such a method suffers from data imbalance. To this end, they subsequently adopted the data entity replacement and masking to:

- augment the data set to improve the robustness of the training model; and
- force the model to learn the contextual information of the data set entities information to improve the generalisation ability of the model.

Moreover, further comparison revealed that entity replacement and masking can improve the recognition of long-tailed data set entities, but the mask strategy had a higher degree of improvement than the replacement strategy regarding the precision, recall and F1-measure. By replacing the data set entity words in sentences with other meaningless words, the entity masking mechanism could force the NER model to learn context information rather than the entity words to a greater extent, thus making the model more generalised and robust. In the above experiments, they followed the token mask task in BERT and implemented entity replacement and masking with a probability of 20%. To investigate the influence of this probability value on data set entity recognition further and compare the utility of entity replacement and entity masking, they increased the preset probability values from 20% to 50% and 100%. The experimental results are presented in Table 3.

The precision, recall and F1 values of entity replacement and entity masking all decreased when the probability value was increased from 20 to 100. Furthermore, in contrast to entity replacement, the increased probability value caused a greater decrease in the recognition performance using entity masking. In addition, the decrease in precision for entity masking was significantly higher than the decrease in the recall. The precision rate of entity masking decreased by 10%, which is significantly higher than that of the recall rate

Data augmentation strategies	Precision	Recall	F1
Baseline	71.34	62.65	66.71
Entity replacement (20%)	78.75	68.08	73.03
Entity replacement (50%)	77.90	67.30	72.21
Entity replacement (100%)	75.46	67.29	71.14
Entity masking (20%)	*80.76*	*71.22*	*75.69*
Entity masking (50%)	77.64	70.17	73.72
Entity masking (100%)	72.48	66.89	68.44

Table 3.
Impacts of different data augmentation strategies on data set entity identification

by 6%. The authors suspect that too much entity masking might have caused underfitting in the NER model. In general, the results in Tables 2 and 3 indicate that the data enhancement strategy can enhance the generalisability of NER. However, excessive entity replacement or entity masking may lead to underfitting in the NER model, which is more obvious in the entity masking strategy. When the probability value was set to 20%, the entity replacement and masking methods yielded the greatest improvement in identifying data set entities, and the improvement effect of the entity masking method was higher than that of the entity replacement method.

5. Conclusion and future work

The results of this study contribute to the field of information extraction, especially digital resource management. In this paper, the authors proposed a method based on distant supervised learning to recognise data set entities automatically in scientific papers. To the best of the authors' knowledge, this is the first attempt to apply distant learning to the study of data set entity recognition. The proposed approach can overcome the provision problem with training corpora and automatically identify data set entities from large-scale literature in an open domain, compared to existing studies. To improve the generalisability and robustness of their approach and enhance the recognition performance on long-tailed data set entities, they introduced BERT in the embedding layer for text vectorisation. Subsequently, they used two data augmentation techniques to address the data imbalance problem caused by bootstrapping-based automatic data labelling.

The multilayer stacked self-attention mechanism enables BERT to learn contextual interaction information in the text regardless of space and distance. By continuously acquiring information about the text location, vocabulary and grammar, the BERT network can provide multilevel, multigrained, and vectorised representations of the text, which helps the NER model better learn the potential semantic information. Moreover, although automatic data labelling based on the bootstrapping strategy has been widely used to overcome the lack of training corpora, most studies have ignored the inherent problem of automatic labelling: poor recognition because of imbalanced training samples. To this end, they introduced two data augmentation methods: entity replacement and masking. The experimental results reveal that both methods effectively improve the recognition of data set entities, especially long-tailed data set entities.

Data set identification is important for managing data resources and supporting various research scenarios, such as connecting the research task with data sets used to answer the question "which <Dataset> can be used for which <Task>". Moreover, it is essential for evaluating the scientific influence of data sets through reuse frequency and exploring research hotspots by analysing the distribution range and new-born speed of data sets. Because the experimental results on the gold-standard testing data were not ideal, they believe that there is still room to improve the performance of data set entity recognition. After an iterative bootstrapping strategy, they obtained 70,000 training samples, but the training corpora for certain data set entities were very sparse. Although their method can be applied to the open domain, they only chose articles in the computer science field for the experiments. Finally, their work focused on identifying data set entity words in papers, but the data resource acquisition requires access to link addresses in the text.

In future work, the authors will apply their approach to more fields, such as medicine, chemistry and sociology, and diversify their data sources (e.g. journals, patents and conference data) to cover as many common and long-tailed data sets as possible. In addition, they will further explore the identification of data set links and matches between data set entities and links, contributing to sharing and reusing data set resources.

References

Akbik, A., Bergmann, T. and Vollgraf, R. (2019), "Pooled contextualized embeddings for named entity recognition", *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Long and Short Papers), Vol. 1, pp. 724-728.

Boland, K. and Krüger, F. (2019), "Distant supervision for silver label generation of software mentions in social scientific publications", *BIRNDL@ SIGIR*, pp. 15-27.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018), "BERT: pre-training of deep bidirectional transformers for language understanding".

Dong, X., Qian, L., Guan, Y., Huang, L., Yu, Q. and Yang, J. (2016), "A multiclass classification method based on deep learning for named entity recognition in electronic medical records", paper presented at the New York, NY Scientific Data Summit (NYSDS '16), *IEEE*.

Duck, G., Kovacevic, A., Robertson, D.L., Stevens, R. and Nenadic, G. (2015), "Ambiguity and variability of database and software names in bioinformatics", *Journal of Biomedical Semantics*, Vol. 6 No. 1, pp. 1-11.

Duck, G., Nenadic, G., Brass, A., Robertson, D.L. and Stevens, R. (2013), "BioNerDS: exploring bioinformatics' database and software use through literature mining", *BMC Bioinformatics*, Vol. 14 No. 1, pp. 1-13.

Duck, G., Nenadic, G., Filannino, M., Brass, A., Robertson, D.L. and Stevens, R. (2016), "A survey of bioinformatics database and software usage through mining the literature", *PLoS One*, Vol. 11 No. 6, p. e0157989.

Ferreira, T.M. and Costa, A.H.R. (2020), "DeepBT and NLP data augmentation techniques: a new proposal and a comprehensive study", *Brazilian Conference on Intelligent Systems*, Springer, Cham, pp. 435-449.

Ghavimi, B., Mayr, P., Lange, C., Vahdati, S. and Auer, S. (2016), "A semi-automatic approach for detecting dataset references in social science texts", *Information Services and Use*, Vol. 36 Nos 3/4, pp. 171-187.

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L. and Weld, D.S. (2011), "Knowledge-based weak supervision for information extraction of overlapping relations", *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 541-550.

Kafkas, Ş., Kim, J. and McEntyre, J.R. (2013), "Database citation in full text biomedical articles", *PLoS ONE*, Vol. 8 No. 5, p. e63184.

Krüger, F. and Schindler, D. (2019), "A literature review on methods for the extraction of usage statements of software and data", *Computing in Science and Engineering*, Vol. 22 No. 1, pp. 26-38.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C. (2016), "Neural architectures for named entity recognition".

Lamurias, A., Ferreira, J.D. and Couto, F.M. (2015), "Improving chemical entity recognition through h-index based semantic similarity", *Journal of Cheminformatics*, Vol. 7 No S1, pp. 1-9.

Li, Y. and Liu, C. (2019), "Information resource, interface, and tasks as user interaction components for digital library evaluation", *Information Processing and Management*, Vol. 56 No. 3, pp. 704-720.

Li, J., Sun, A., Han, J. and Li, C. (2020), "A survey on deep learning for named entity recognition", *IEEE Computer Architecture Letters*, Vol. 1, pp. 1-1.

Li, L., Jin, L., Jiang, Z., Song, D. and Huang, D. (2015), "Biomedical named entity recognition based on extended recurrent neural networks", *International Conference on Bioinformatics and Biomedicine (BIBM '15)*, *IEEE*, pp. 649-652.

Maxim, G., Hoifung, P. and Bill, H. (2017), "Wide-open: accelerating public data release by automating detection of overdue datasets", *PLoS Biology*, Vol. 15 No. 6, p. e2002477.

Névéol, A., Wilbur, W.J. and Lu, Z. (2011), "Extraction of data deposition statements from the literature: a method for automatically tracking research results", *Bioinformatics*, Vol. 27 No. 23, pp. 3306-3312.

Palshikar, G.K. (2013), "Techniques for named entity recognition: a survey", *Bioinformatics: Concepts, Methodologies, Tools, and Applications*, IGI Global, pp. 400-426.

Parish, E.J. and Duraisamy, K. (2016), "A paradigm for data-driven predictive modeling using field inversion and machine learning", *Journal of Computational Physics*, Vol. 305, pp. 758-774.

Patra, B.G., Roberts, K. and Wu, H. (2020), "A content-based dataset recommendation system for researchers: a case study on gene expression omnibus (GEO) repository", *Database*, Vol. 2020.

Pennington, J., Socher, R. and Manning, C.D. (2014), "Glove: Global vectors for word representation", *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '14)*, pp. 1532-1543.

Prasad, A., Si, C. and Kan, M.Y. (2019), "Dataset mention extraction and classification", *Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications*, pp. 31-36.

Ruokolainen, T., Kauppinen, P., Silfverberg, M. and Lindén, K. (2019), "A Finnish news corpus for named entity recognition", *Language Resources and Evaluation*, Vol. 54 No. 1, pp. 247-272.

Shorten, C. and Khoshgoftaar, T.M. (2019), "A survey on image data augmentation for deep learning", *Journal of Big Data*, Vol. 6 No. 1, pp. 1-48.

Tanwani, A.K. and Farooq, M. (2009), "The role of biomedical dataset in classification", paper presented at the Conference on Artificial Intelligence in Medicine in Europe.

Tuarob, S., Bhatia, S., Mitra, P. and Giles, C.L. (2016), "AlgorithmSeer: a system for extracting and searching for algorithms in scholarly big data", *IEEE Transactions on Big Data*, Vol. 2 No. 1, pp. 3-17.

Vaidhehi, V. (2014), "The role of dataset in training ANFIS system for course advisor", *International Journal of Innovative Research in Advanced Engineering*, Vol. 1 No. 6, pp. 249-253.

Wei, J. and Zou, K. (2019), "EDA: Easy data augmentation techniques for boosting performance on text classification tasks".

Yadav, V. and Bethard, S. (2019), "A survey on recent advances in named entity recognition from deep learning models", *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2145-2158.

Yan, A. and Weber, N. (2018), "Mining open government data used in scientific research", *International Conference on Information*, Springer, Cham, pp. 3003-3313.

Zhang, S. and Elhadad, N. (2013), "Unsupervised biomedical named entity recognition: experiments with clinical and biological texts", *Journal of Biomedical Informatics*, Vol. 46 No. 6, pp. 1088-1098.

Zhang, Q., Lu, W., Yang, Y., Chen, H. and Chen, J. (2017), "Automatic identification of research articles containing data usage statements", *Knowledge Discovery and Data Design Innovation-Proceedings of the International Conference on Knowledge Management (ICKM '17)*, World Scientific, Vol. 14, p. 67.

Zhao, M., Yan, E. and Li, K. (2018), "Data set mentions and citations: a content analysis of full-text publications", *Journal of the Association for Information Science and Technology*, Vol. 69 No. 1, pp. 32-46.

Corresponding author
Qikai Cheng can be contacted at: chengqikai0806@163.com