A System Review on Bootstrapping Information Extraction

Hui Fang[1*], Ge Xu[1†], Yunfei Long[2†], Yin Guan[1†], Xiaoyan Yang[1†] and Zhou Chen[1†]

[1*]College of Computer and Control Engineering, Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, No.200, xiyuangong Road, Shangjie Town, Minhou County, Fuzhou, 350108, Fujian, China.
[2]School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park Colchester CO4 3SQ, Colchester, CO2 8JT, Essex, UK.

*Corresponding author(s). E-mail(s): fh_mju@126.com;
Contributing authors: xuge@pku.edu.cn; yl20051@essex.ac.uk; niynaug@foxmail.com; 349622662@qq.com; 1801619490@qq.com;
†These authors contributed equally to this work.

Abstract

Focusing on the supervision problems caused by high-cost and low-quality labeling in information extraction, we provided a detailed overview of the various approaches that were proposed to solve the sub-tasks of bootstrapping information extraction. We summarized current principal approaches and depicted the specific issues addressed in recent research. To provide inspiration and reference for similar studies in terms of mainstream data sources, evaluation specifications and applications, we summarized the relevant datasets, evaluation metrics, and systematic applications of bootstrapping information extraction. In addition, we reflected on the remaining problems of bootstrapping information extraction and highlighted some directions for future work.

Keywords: Bootstrapping Information Extraction, Seed Generation, Pattern Learning, Instance Acquisition, Pattern Evaluation, Instance Evaluation

1 Introduction

Information extraction is broadly viewed as a method for filtering information from large volumes of text [1]. This includes the retrieval of documents from collections and the tagging of particular terms in text. Information extraction is the backbone for knowledge-driven AI systems, where information is evaluated and summarized to form knowledge. One of the main challenges information extraction faces is the supervision problem, such as poor domain scalability, single extraction granularity, and loose supervision signals [1]. Nowadays, the main solutions to this problem include: weakly supervised approaches based on knowledge bases, indirectly supervised approaches from Question Answer(QA), and weakly supervised approaches based on linguistic models, which have the common feature of leveraging the participation of other tasks or other resources to achieve high-quality supervision. Due to the high cost of labeling and the low quality of labeling, machine learning or deep learning models increasingly need to focus on weakly supervised learning approaches, i.e., heuristically using external knowledge bases, patterns/rules, or other classifiers to generate training data. Bootstrapping [2, 3], as a representative of incomplete supervision (a type of weak supervision), refers to a problem setting in which one is given a small set of labeled data and a large set of unlabeled data, and the task is to induce a classifier. If this process continues to iterate, the amount of labeled data that can be used to guide classification will increase accordingly. It has been proved effective in information extraction task, such as semantic lexicon construction, dictionary construction, relation extraction or entity set expansion, etc. The concept of bootstrapping is inherited from the bootstrapping term in statistics [4], which refers to the use of limited sample data to re-establish a new sample that is sufficient to represent the distribution of the parent sample through repeated sampling. This iterative bootstrap process is transferred to the information extraction domain with the hypothesis that a limited number of good relationship instances can refine good relationship patterns, which in turn can usually help find good relationship instances.

The idea of bootstrapping information extraction(BIE) originally comes from the solution to the problem of extracting a relation for a particular type of data from thousands of independent information sources automatically [5].They present a technique that exploits the duality between pattern and relation to grow the target relation starting from a small amount of relation sample. DIPRE(Dual Iterative Pattern Relation Extraction) is a method to extract structured relationships (or tables) from a collection of HTML documents. This method works best in a Web-like environment, where the tuples to be extracted from the table often appear repeatedly in the consistent context of the set document. DIPRE uses set redundancy and inherent structure to extract target relationships and simplify the training process. This idea shown in Figure 1 is later applied in the widely known Snowball system [6], which extends the extraction work to text documents. It produces the selective patterns with high coverage so that they generate correct tuples and identify new tuples. A more important improvement is that the Snowball system provides an evaluation of patterns and tuples based on selectivity. Snowball would only retain tuples and patterns that are considered "reliable enough" for iterative process of the system. These new patterns and tuples generation and filtering strategies can significantly improve the quality of extracted tables. It solves the problem of gradually decreasing correctness and severe semantic drift caused by the gradual enlargement of error patterns. Some subsequent systems follow the bootstrapping method like Snowball, but will add more reasonable descriptions of patterns, restrictions and scoring strategies, or build large-scale patterns based on the extraction results of previous system. For example, in the NELL (Never Ending Language Learner) system [7], given an initial ontology (a few definitions of classes and relationships) and a few samples, it can constantly learn and extract new knowledge from the Web through self-learning. Currently, NELL has extracted more than 3 million tuples of knowledge [7].

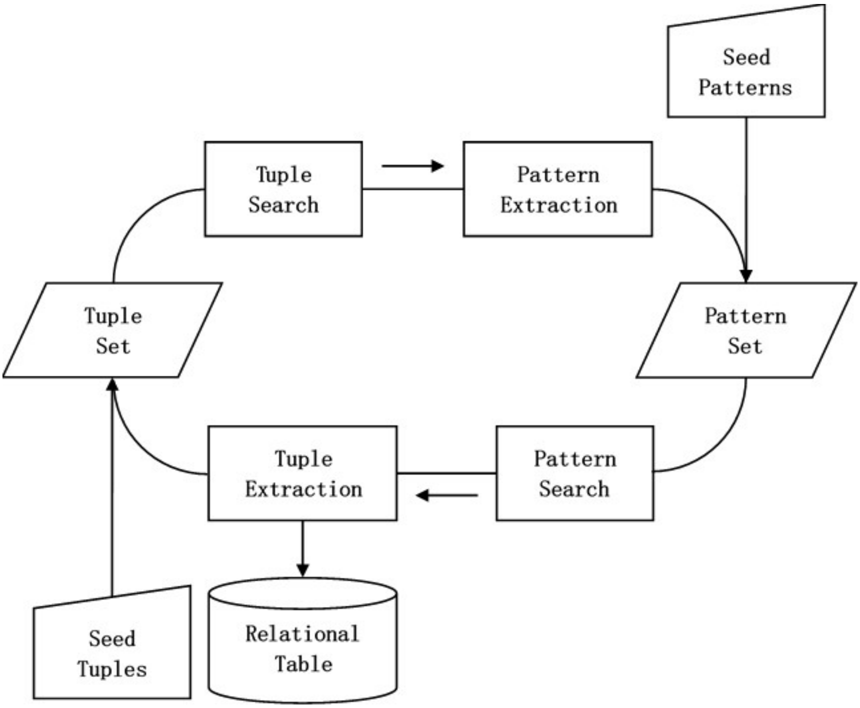

Fig. 1 The Idea of the DIPRE Method

As deep learning approaches are becoming mainstream in NLP, neural network-based snowball has emerged. The neural Snowball approach aims to learn new relations with a small sample of known relations by migrating semantic knowledge over existing relations. Specifically, Neural Snowball [8] uses Relational Siamese Networks (RSN) to learn relational similarity measures between instances based on existing relationships and their labeled data. Subsequently, given a new relation and a small number of labeled samples, RSN is used to accumulate reliable instances from the unlabeled corpus, and these instances are used to train a classifier that can further identify new facts about the new relation. In contrast to traditional bootstrapping, Neural Snowball also makes use of a large-scale labeled dataset. Although the distribution of existing relationships may differ significantly from that of new relationships, the deep learning model can still extract high-level abstraction features to characterize unknown relationships. As a result, Neural Snowball is more expressive and capable of handling more complex relationships. However, the recall growth of the Neural Snowball method is lower than expected, which means that RSN may have overfitted the existing patterns [8].From DIPRE to Neural Snowball, the related work reflects some of the changes in the paradigm of BIE from the traditional symbolic process of extraction to the modern vectorization process, and from heuristic learning methods to neural network-based deep learning methods.

All BIE processes can be summarized into two alternating parts: pattern expansion and instance expansion. The former considers how to efficiently generate high-quality instance templates, while the latter is concerned with how to robustly obtain high-quality instances. In the alternating process, abstract and concrete paradigm occurs between the pattern and the instances. Traditional bootstrapping methods usually use explicit representations, such as rules or symbols. Neural network-based methods, on the other hand, use an implicit representation by means of vectors or features.Nevertheless, the specific strategies and techniques used in the BIE process have not been well organized and summarized.

The current survey on information extraction mainly focuses on open information extraction, information extraction of a specific extraction object, or information extraction of a particular domain. We also find that many journal papers in the IE community are algorithm-centric, with less consideration on the datasets and evaluation methods. In this work, we survey research related to the bootstrapping approaches of IE in general, categorize them, and then summarize the current applications that apply the BIE. The categorization is especially aimed at giving IE and NLP practitioners a perspective on using BIE that is available in diverse forms.A comparison between our work and other major reviews can be found in Table 1.

In summary, this paper has the following findings and contributions:

1. This paper analyzes the main paradigms of bootstrapping information extraction and summarizes their essential components from a methodological perspective.
2. This paper reviews the relevant datasets, evaluation metrics, and systematic applications of bootstrapping information extraction to provide mainstream data sources, evaluation specifications, and inspiration for similar studies.
3. This paper puts forward the challenges faced by bootstrapping information extraction and suggestions for its future research direction.

Table 1 The comparison between our work and related review works

Work	Year	Focus
Cheng et al.[9]	2021	The recognition methods of Chinese NER
Zhou et al.[10]	2022	The methods and evaluation of Neural OIE
Yang et al.[11]	2022	Various extraction techniques based on deep learning
Zhang et al.[12]	2022	The IE methods of Traditional Chinese Medicine text
Landolsi et al.[13]	2023	The methods,datasets and application of medical IE
Abdullah et al.[14]	2023	The methods and application of textual IE

The rest of the survey is organized as follows.Section 2 outlines the overall idea and flow of the overview. We introduce the main BIE methods by focusing on four important phases in Section 3. Commonly used datasets and evaluation metrics for BIE are given respectively in Section 4 and Section 5. Section 6 sorts out typical BIE application systems. We summarize the main prospects and challenges for BIE in Section 7 before final conclusion in Section 8.

2 Materials and strategies

This study provides a systematic review of bootstrap methods in information extraction. Figure 2 illustrates the main stages of the research. First, we survey the source literature from the designated database by keyword search strategy, then we formulate the selection criteria for inclusion and exclusion, and finally we formulate the fundamental issues of the systematic review.The overall review uses the framework proposed by Arksey and O'Malley [15] and is informed by PRISMA guidelines [16].

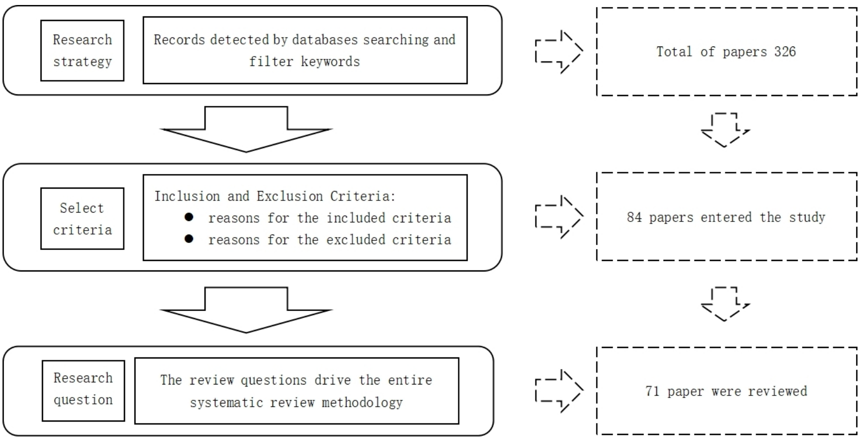

Fig. 2 Research Flow Chart

2.1 Search strategy and database source

In order to systematically sort through all the past research work on BIE, the source we retrieve or search for BIE related literature include Science Direct[1], IEEE Xplore[2], ACM Digital Library[3], SpringerLink[4], Google Scholar[5], and ACL database[6]. We identify bootstrapping-related literature by keywords appearing in the title and abstract of the literature. This study was conducted by restricting the valid basic concepts related to the research object, which were mapped to the corresponding keywords. The main keywords are BIE, pattern-based bootstrapping, bootstrap extraction/mining, seed Self-expansion, semi-supervised information extraction, named entity recognition, relation extraction, pattern matching, etc. As the results, 326 papers were finally collected after our initial search.

2.2 Inclusion and exclusion criteria

Since the existing work related to bootstrapping information extraction has a large time span, multiple sources of heterogeneous data, a variety of study types, and varying study quality, the following inclusion and exclusion criteria are proposed. Table 2 and Table 3 show the details of Inclusion and exclusion criteria.

Table 2 The List of Inclusion Criteria

+ Inclusion criteria
+ Input data must be textual.
+ The study should be innovation or improvement of BIE methodology.
+ The content covered in the paper is the relevant research work of the last decade, except for the information extraction application system
+ Reputable journal or conference paper should be reviewed unless BIE is a section of the dissertation.
+ The proposed method starts with little or almost no manual.

2.3 Research questions

In order to drive the entire review methodology systematically, the research questions focus mainly on the special aspects of BIE method. To be more specific, our review raises the following research questions:

1. How to generate proper and correct seeds for patterns or instances?
2. What are the methods to represent and learn to obtain patterns?

[1] https://www.sciencedirect.com/
[2] https://ieeexplore.ieee.org/Xplore/home.jsp
[3] https://dl.acm.org/
[4] https://link.springer.com/
[5] https://scholar.google.com.hk/?hl=zh-CN
[6] https://aclanthology.org/

Table 3 The List of Exclusion Criteria

- Exclusion criteria
- The input data is obtained with the help of semi-structural features of web pages.
- The study focuses only on the use of previous BIE method.
- The paper does not reveal much obvious information.
- The paper only uses the bootstrapping idea in a purely statistical method study.
- The paper only includes definitions and reviews of the BIE method.
- Short paper that are without model description and experiment.
- The proposed method requires a large amount of annotated data or knowledge base.

3. What methods are used to get instances?
4. How to evaluate patterns and instances?

3 Bootstrapping Information Extraction Methods

BIE methods is referred to as extracting information from a certain information source in a way that starts with a limited set with labels to expand its set size [17]. Due to its minimally supervised, domain-independent, and language-independent nature, the bootstrapping method has its own distinctive features in information extraction, which are respectively reflected in the generation of seeds, the learning of patterns and the evaluation of patterns during pattern expansion, and the acquisition of instances, or the evaluation of instances during instance expansion. Figure 3 illustrates the general principle of the BIE method. Therefore, we review the research work on BIE methods from the above aspects, and try to answer the research questions raised in the previous section.

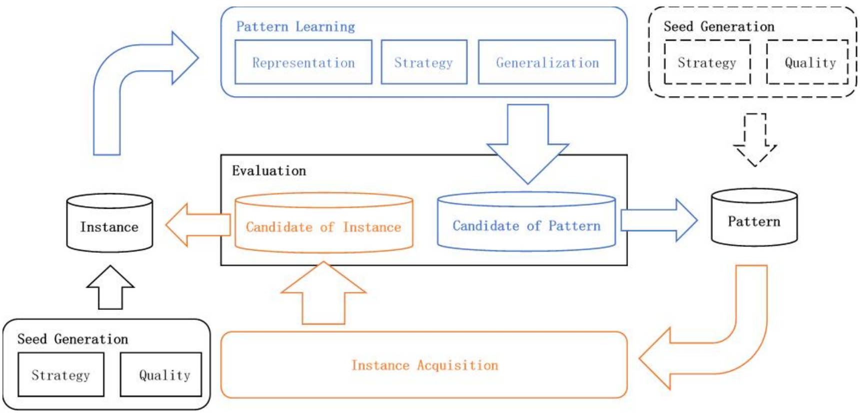

Fig. 3 The General Principle of the BIE Method

3.1 Seed Generation

In the seed generation process, taking into account the appropriateness and correctness of the seeds is the central issue, the seed generation strategy and the seed generation quality are considered. The seed may be either pattern or instance.

3.1.1 Seed Generation Strategy

To avoid confusion, the instances or patterns used to initiate BIE are collectively referred to here as seed, where the pattern seed is designed to obtain high-quality instance seed. The current main strategies for seed generation in the last decade are divided into manual strategy and automatic strategy according to the degree of automation.

The advantage of the manual strategy is that the seed quality is easy to control, but it is labor intensive. In contrast, the automated strategy reduces human involvement and makes seeds easily accessible. However, it is difficult to guarantee the quality of seeds. Nowadays, most studies use automatic strategies, even manual selection often combine automatic strategies to generate seeds for different purposes. For seed pattern, manual construction is the simplest approach, while for seed instance, the co-reference approach is the most straightforward. Yahya et al. developed ReNoun [18], an open information extraction system that extracting facts for noun-based relations by focusing on nominal attributes and on the long tail. ReNoun has been performed in pipeline. Begin by extracting a small number of high-precision facts, ReNoun relied on manually specified lexical pattern that are specifically tailored for noun phrase but are general enough to be independent of any specific attributes. Thanks to such patterns, ReNoun can make the generated seed facts more precise through co-reference that requires the attribute and object noun phrases of a seed fact to refer to the same real-world entity.

In the era of rapid development of web information, there are many available corpus data on web pages. To make full use of the web resources, some researchers have started to obtain seed by searching on the web. For example, in the entity relation extraction model proposed by Wang, they acquire a large corpus of sentences containing company names and relationship patterns on the web through a crawler module to facilitate entity recognition while the process is constantly iterative and keeps acquiring more and more candidate corpus [19].

Another important type of seed is word dictionaries, glossaries, thesauri, etc. Han et al. extracted entity types and relationship types from the internal structure of the thesaurus, and then designed an algorithm for automatic generation of initial seed sets of domain knowledge graphs based on the thesaurus. The experimental results show that the initial seed set obtained by using the thesaurus can achieve a result closer to the manually designed seeds [20]. Considering that there is no readily available lexicons, Tuo et al. in their work on aspect extraction and aspect term expansion, clustered the words of major aspects into categories in which all aspects are included. Aspect categories were selected and the top 30 % of aspect terms for each aspect were chosen as seed terms [21].

On one hand,lexicons can be considered as semantically orchestrated structured seed sources. They encompass domain-related seed information and provide an effective means for bootstrapping domain information extraction.Yet not every field has well organized lexicons.On the other hand,since web resources are easily accessible and can quickly supplement the corpus, they serve as an open gateway for acquiring instance seeds. But the quality of these seeds remains to be determined [19]. However,obtaining seeds from web resources is still a preferable method for generating seeds if the quality of web resources has been reasonably evaluated, as it saves time and labor costs.

3.1.2 Seed Generation quality

Ensuring seed quality is another significant issue that requires attention in seed generation, particularly concerning automatic strategies. To enhance seed quality, it is often necessary to screen the seeds, primarily based on their characteristics. Much of the relevant work over the past decade has been conducted in a non-supervised manner.

Phi et al. proposed and compared various approaches for automatic seed selection, drawing inspiration from ranking relation instances and patterns computed by the HITS algorithm. They also explored picking cluster centroids using methods such as K-means, Latent Semantic Analysis (LSA), or Non-Negative Matrix Factorization (NMF) [22]. The experimental setup used random seed selection as the baseline comparison method, and the results demonstrated that the relation extraction system utilizing the random method exhibited the poorest average P@50 among all seed selection strategies. The K-means automatic extraction approach demonstrated the best performance, while the performance of the other methods was comparable. Notably, the HITS- and K-means-based approaches displayed a slightly better performance [22]. In a study by Xiong et al., sentiment seeds were selected by arranging the nodes according to their degrees in a semantic graph and manually choosing nodes with evident polarity [23]. In [21], the K-means algorithm was also employed for automatic seed selection, where selecting a large hyperparameter 'k' was deemed necessary to ensure the quality of the seed terms. Additionally, in [24], BONIE examined the presence of seed facts in a large-scale external knowledge base and retained only those that were common. This approach yielded a diverse set of clean facts for further training.

Since the unsupervised approaches [21–24] has no externally supervised signals to guide the learning process, more data and more complex models are needed to effectively learn the structure of the data. Thus these current unsupervised methods, while simple and feasible, are difficult to ensure that higher quality seeds are obtained. Therefore more supervised methods may be explored for quality control of seed generation.

3.2 Pattern Learning

Pattern learning constitutes a fundamental step in BIE, as it determines the accuracy of expansion from a limited number of instances to a larger set. In the realm of pattern learning, it is essential to consider what to learn, how to learn, and how to enhance learning comprehensively and effectively.

3.2.1 Pattern representation

The pattern representation that refers to the input or output of pattern learning may affect the pattern learning framework used for BIE and the pattern generalization ability. Typical pattern representations include the word, part of speech tagging, word sense, parse tree, etc. Please refer to Table 4 for specific information. The earliest symbolic representation is the word itself, often referred to as surface pattern.

Zupon et al. used surface patterns consisting of up to 4 words before/after the targted entities, and argued that such pattern is agnostic to the types of patterns learned, and can be trivially adapted to other types of patterns [29]. Word itself, part of speech(POS) tagging, word sense and parse tree belong to context pattern, which consists of the context information of the target attribute. Ding et al. used a natural language processing toolkit for POS and name entity(NE) recognition tagging and a syntactic parser for parsing, since each event is represented using a frame-like structure to capture the meanings of different events [32].Similar to this work is that the event trigger, each entity mentioned, and the dependency path between them were extracted as event patterns in [33]. Xiong proposed a feasible pattern representation method by constructing a small window of POS tagging pattern for each slot value and then combining all the corresponding "sub-patterns" of the slot values to form a tuple corresponding pattern [34]. Chen et al. developed a 7-dimension tuple as extraction rules generated from a sentence containing entities, relation keywords and a string of several words [35]. Zhang et al. defined a new relation representation named activation force defined dependency pattern, which is the shortest dependency path from entity to attribute value with a trigger word as the semantic anchor [36].In [25],they investigate parse tree path and mixed context patterns, combining three different semantic units in the pattern design.

With the rapid development of machine learning and deep learning models, The vector-based representation of patterns is also becoming a trend. To enrich the pattern representation's semantic meaning, learning more effective features from the input instance is needed to form a high-quality pattern. Neural networks can automatically extract features at a high level for obtaining a better pattern representation. Tandon et al. constructed a tuple graph with includes candidate tuple nodes, pattern set nodes, seed set nodes and relation nodes. The tuple representation allowed it to consider the local graph for each tuple with potentially millions of seeds, patterns and tuples [31]. Zhang et al.

Table 4 The List of Pattern Representation

Pattern type	Representation	Example
Word itself/surface pattern	n-gram	Director_directed[25]
	part of speech(POS) tagging	$\langle Na \rangle$ _ $\langle VC \rangle$[25]
context pattern	word sense	$\langle human \rangle$_$\langle undertake \rangle$[25]
	parse tree	$\langle PER \rangle$ nsubjpass survived agent son appos $\langle PER \rangle$[26]
mixed pattern	surface + context	eat VB:dobj:NN X[27]
	url-text hybrid pattern (utp)	utp = (up, tp, c, f)[28]
	embedding	J = SG + Attract + Repel[29]
vectorized pattern	classifier	RNN[30]
	graph	tuple graph[31]

compared the similarities between pairs of semantic shortest dependency patterns by the bottom-up kernel. They selected the most similar one to update the seed pattern in each iteration of bootstrapping [26].Jianshu et al. applied an augmented dependency tree (dependency information with word and part-of-speech information) as pattern to train and extract in bootstrapping part of our model, then accumulated the root vector for a binary classifier based on matrix-vector recursive neural networks to judge whether the relationship they assumed is correct [30].

In addition to the above, some special pattern representations are designed for specific scenarios or tasks. Zhang et al. proved simple text patterns could also acquire high-quality named entities in specific conditions. They designed URL-text hybrid patterns to guarantee the capability of the patterns from both URL and text aspects by considering the quality of URLs when using text patterns [28].Ziering P argued that bootstrapping methods, known as particularly sensitive to the ambiguity of terms and contexts, benefit from solid semantic coherence in coordination. They introduced "Basilisk Coordination Patterns" that use only coordination and punctuation co-ordinations in Basilisk instantiation [37].

The symbol-based pattern representations [25, 28, 29, 32–37],while explicit and human-readable, require heavy human labor and elaborately design.In contrast, the vector-based pattern representations [26, 30, 31] can be learned automatically by optimization of a training objective instead of handcrafted features. But their shortcoming is that the interpretability of learned embeddings is poor. Therefore, how to combine the respective advantages of symbolic representation and vector representation, such as studying interpretable vectors, will be a possible direction for exploring pattern representation in future.

3.2.2 Pattern learning strategy

Besides pattern representation, how to obtain patterns from existing instances is the main concern of pattern learning. The paradigm used for the strategy of learning pattern can be divided into several stages. The initial paradigm is the stage of using symbolic processing. That is, from Word Segmentation to POS or NER, and then syntactic analysis to form the semi-instantiated pattern.

Yada et al. used a set of initial instances of cue phrases and emotion words as seeds, and acquired functional word sequences between emotion clauses(ECs) and emotion words as cue phrases if the clauses are similar enough to previously learned ECs from a set of (dependency-parsed) sentences [38].In [39],to learn the pattern templates, they first extracted the dependency path connecting the arguments and relation words for each seed tuple and the associated sentence. Then they annotated the relation node in the path with the exact relation word (as a lexical constraint) and the POS (pos tag constraint). Finally they created a relation template from the seed tuple by normalization and replacement. Kozareva et al. use recursive pattern for is-a relation learning. The generated patterns are submitted to the search engine as a web query and all retrieved snippets are kept. If they were not previously explored by the algorithm, they are placed on the relation expression position of pattern and used as seeds in the subsequent verb extraction iteration [40].

In the next stage, vector-based representations begin to emerge. Dalvi et al. argued that an extraction pattern can be treated as a search query over a corpus. They devised an innovative query language that integrates symbolic (boolean) and distributional (similarity-based) search methods. Additionally, they proposed an machine learning-based query suggester capable of refining or broadening the current query to discover additional patterns (queries) that express the target relation [41]. Subsequently, machine learning or deep learning, exemplified by neural networks, is employed as a strategy to learn patterns. In this context, the learning model is regarded as an implicit pattern. Tai et al. introduced a novel kernel function based on the shortest dependency tree, significantly enhancing the reliability of newly acquired patterns. These advancements are attributed to the emphasis on element importance, the flexibility in pattern similarity computation, and the consideration of test pattern length, effectively mitigating uneven distribution of similarity values [42].

Li et al. proposed two additional views—the semantic relationship view and the morphological structure view—alongside the traditional pattern similarity view. They applied a co-training strategy to merge these perspectives into a minimally supervised learning model. In each view, all pattern candidates were ranked from different angles, and the top-ranked n candidates were selected as accepted patterns [33]. In another study [43], the relationship between the seed verb and the newly acquired verb was represented by two vectors obtained through machine learning.

Similarity between the two relations was calculated using a vector similarity algorithm. When the similarity value exceeded a certain threshold, the new relation verb was obtained, and the corresponding relation pattern was extracted. Shi et al. introduced a probabilistic co-bootstrapping method that more precisely defined the expansion boundary by utilizing both positive and discriminant negative seeds, which are automatically generated during the bootstrapping process [44]. Due to the high cost associated with supervised learning, some studies have adopted semi-supervised learning as a strategy for pattern learning. ReNoun [18] employed distant supervision to learn a set of dependency parse patterns used to extract a greater number of facts from the text corpus. Furthermore, Cheng et al. proposed a novel semi-supervised NER method based on multi-pattern fusion. The approach incorporated soft-matching within the entity internal pattern and obtained an entity external pattern through a bootstrapping process in the training corpus [45].

Obviously, compared with the symbol-based pattern learning method [38–40], the deep learning-based pattern learning method [18, 41–45] can automatically extract deeper and richer features of the pattern, which plays an important role in obtaining better quality patterns in the BIE process. Therefore, it will be more and more widely used in pattern learning strategies. Minimizing the training time and iteration efficiency of deep learning is what it needs to address.

3.2.3 Pattern generalization

As a key component of pattern learning, the generalization of candidate patterns facilitates pattern extension, which can be achieved with the assistance of constructed knowledge bases or similarity calculations. The constructed knowledge base may include a thesaurus, dictionary, and other resources. Makarov P constructed a weighted undirected graph of pattern similarity, where pattern candidates served as nodes, and the edge weights were computed using angular similarity. Following this, a semi-supervised label propagation algorithm was applied, and a verb pattern dictionary was utilized to identify seed patterns [46].In [24], BONIE used WordNet to expand patterns by including all inflections and synset synonyms. Alashri et al. captured contextual synonyms that are not derivable from our corpus by applying WordNet synonyms and hyponyms to the members of concepts, further expanding and generalizing them [47].

Similarity calculation can be performed through clustering, contextual statistics, and other methods.In [48], BREDS generated additional extraction patterns by applying a single-pass clustering algorithm to relationship instances collected in the previous step. Each resulting cluster contains a set of relationship instances represented by their context vectors. Xu introduced the principle of intra-chapter consistency into event extraction based on the structural features of event sentences, using this consistency to reason about other events with homogeneity or relevance, thus expanding the event patterns [49]. Liu proposed a method for obtaining relation extraction patterns based on information gain. The method considers differences in semantic and positional features among different relations, generating corresponding relation extraction patterns for co-occurrence sentences of seed tuples of a certain relation type [50]. Cheng utilized a soft matching method based on the Levenshtein distance to calculate the similarity between the internal patterns of two domain entities, aiming to identify more internal patterns of named entities of a specific category [51].In [52], PACE generated additional candidate patterns solely from the context surrounding known (i.e., seeded or learned) entities by storing known entities along with their respective contexts.

Compared to the similarity computation approach [48–52], the knowledge base-based approach [24, 46, 47] is simple and easy to implement, but requires the existence of a priori knowledge. The similarity computation approach is not subject to this limitation, and vectors encoded by pre-trained models with richer semantic information can be considered for computation in future.

3.3 Instance Acquisition

The ultimate goal of BIE is to obtain instances, and the primary methods for capturing instances include pattern matching and instance distance calculation, among others. Pattern matching in instance acquisition can be achieved through rules, utilizing either the contextual pattern or the surface pattern of the sentence itself, or by referencing an external knowledge base of structured data, such as an ontology.

Thomas et al. proposed the extraction of semantic relations from sentences containing phrasal verbs and conjunctive forms by leveraging the dependency tree structure of the sentences. The proposed system effectively combines the strengths of both Open Information Extraction (OIE) and Ontology-Based Information Extraction (OBIE) techniques to extract domain-specific judicial relations from court opinions by integrating domain ontology [53]. Wu employed both relational type matching and entity pair type matching for pattern matching, ensuring better alignment with most matchable text seeds and achieving a high recall rate for this algorithm [54]. In [47], their algorithm for automatically discovering causal relationships and chains is grounded in the extraction of inter- and intra-sentential patterns. In [18], Yahya M argued that each pattern match against the corpus indicates the potential subject, attribute, and object heads. The noun phrase led by the token matching the vertex is then compared against the set of attributes to which the pattern is mapped.

Differing from previous approaches to acquisition, instance distance calculation is generally conducted by searching for instances that are closely related through graph networks or other statistical methods. Tuo et al. computed the distance from each word to every cluster center for every word in the document of pre-aspect words. They then compared the minimum distance with the inner category distance in the task of aspect terminology expansion. If the minimum distance is less than the inner category distance, the corresponding word term is added to the respective aspect category as an aspect term [21]. Xiong et al. constructed a semantic graph in which sentiment words were represented as nodes and the edge weights indicated the similarity between words. This graph was used to more effectively predict the sentiment polarity of unlabeled candidate sentiment words. They also introduced a global and local point-wise mutual information (GLPMI) method that refined word relevance more precisely through weighted rules [23]. Long et al. proposed a method that utilizes vector similarity calculation in the process of named entity recognition (NER). They then calculated the similarity between the feature vector of the named entity obtained in the previous section and the feature vector of an example containing the named entity. When the similarity reaches a certain threshold, the corresponding named entity can be recognized [55].

Pattern matching methods either require well-designed rules [18, 47, 54] or rely on knowledge hierarchies [53]. In comparison, the methods based on instance distance computation [21, 23, 55] are not subject to these constraints and can essentially continuously obtain higher quality instances by optimizing the similarity computation process.

3.4 Evaluation for Patterns and Instances

The evaluation of patterns or instances plays a pivotal role in ensuring the extraction quality of BIE. If not appropriately handled, it is prone to semantic drift problems or low recall. The evaluation process of BIE may involve ranking or filtering to identify suitable patterns or instances through scoring, comparison, and various other methods.

For instance, Yan et al. argued that pattern evaluation relies on both its direct extraction quality and the extraction quality in subsequent iterations. They employed the Monte Carlo Tree Search (MCTS) algorithm for efficient delayed feedback estimation and applied a prior policy network to eliminate poor patterns, thus reducing the search space in the MCTS [56].

Tai designed a classification model with two kernel functions that were jointly predicted by the combination of two classifiers. This was done to ensure the reliability of the selected relation pattern in the pattern expansion process, along with the accuracy and confidence of the classifiers' classification results. Essentially, the model conducted similarity assessments of patterns by matching kernel functions with classifiers to choose high-quality patterns [57].

Kurihara et al. introduced a scoring method for a confidence measure during the bootstrapping process. After calculating the scores, they extracted phrases in the top N% to serve as new seeds for the next iteration. If the phrases in the current top N% match those from the previous step, the iteration is terminated [58].

Gupta et al. presented a scoring improvement schema that predicted labels for unlabeled entities. This schema utilized various unsupervised features based on contrasting domain-specific and general text, exploiting distributional similarity and edit distances for learned entities [59].

Ziering et al. proposed exploiting linguistic variation between languages to address the problem of gradually decreasing lexicon quality. They introduced a knowledge-lean and language-independent ensemble method [60].

In [61], several scoring functions for similarity-based expansion within a bootstrapping algorithm are applied and compared. They discovered that hypernym/hyponym pairs are automatically and incrementally extracted based on their statistics. Various association measures and graph-based scoring were employed to achieve improved recall.

Identifying appropriate patterns or instances is essentially a search-sorting problem. Current methods focus on either accuracy [59, 60] or recall [56, 58, 61], and few methods [57] are able to balance the two well. There is a need to investigate methods for recognizing patterns or instances that can better satisfy both recall and ranking requirements.

4 Datasets

In this section, we introduce current benchmark datasets related to BIE task. The proposed methods are evaluated on a variety of benchmark data, which we summarized and presented their usage in this section. A list of the datasets is shown in Table 5.

CoNLL [29, 62–64]: CoNLL is constructed for the CoNLL 2003 shared task that concerns language-independent named entity recognition. The CoNLL-2003 named entity data consists of eight files covering two languages: English and German. The English data was taken from the Reuters Corpus . This corpus consists of Reuters news stories between August 1996 and August 1997. The German text data was taken from the ECI Multilingual Text Corpus. This corpus was extracted from the German newspaper Frankfurter Rundshau. CoNLL contains four types of entities: persons, locations, organizations and names of miscellaneous entities.

OntoNotes [29, 62–64]: OntoNotes is from the OntoNotes project, which has created multiple large-scale layers of syntactic, semantic and discourse information in text. The English language comprises roughly 1.7M words, and the Chinese language includes roughly 1M of newswires, magazine articles, broadcast news, broadcast conversations, web data, and conversational speech data. The corpus is tagged with syntactic trees, propositions for most verb and some noun instances, partial verb and noun word senses, coreference, and named entities. The entity type in OntoNotes finally contains 11 entity types without numerical categories.

TREC KBP 2012 SSF [30, 65]: The TREC KBA 2012 SSF corpus includes information about various entities and add any new information to respective infoboxes from a 2008 snapshot of Wikipedia. There are only 42 slots that pertain to general information about persons and organizations.

TAC KBP 2013 ESF [26, 36, 66]: The TAC KBP 2013 ESF corpus includes 2.3 million news docs and 1.5 million Web pages and other docs from 2009 to 2012, and includes 1 million docs from Gigaword, 1 million web docs, and about 100,000 docs from web discussion fora in 2013.

Google Web 1T corpus [31, 44, 56]: Google Web 1T contains a large scale of ngrams compiled from a one trillion words corpus. Google has published a dataset of raw frequencies for n-grams (n = 1, . . . , 5) computed from over 1,024G word tokens of English text, taken from Google's web page search index. In compressed form, the distributed data amounts to 24GB.

EPO [37, 60]: The patent data are distributed by the European Patent Office between 1998 and 2008. The patent description is the main part of a patent. Most European patents provide their claims (the part of a patent defining the scope of protection) in German, English and French.

TACRED [67, 68]: The TAC Relation Extraction dataset is a large-scale crowd-sourced relation extraction dataset following the TAC KBP relation schema. The corpora are collected from all the prior TAC KBP shared tasks. It has more than 100,000 relation mentions with relations categorized into 42 classes.

ACE 2005 [33, 49]: This dataset was released by the Language Data Consortium (LDC) in 2005. The dataset consists of entities, relations, and event annotations for various types of data, including English, Arabic, and Chinese

Table 5 The List of Datasets

Data	Languages	Recognition Type	Annotate	Source
CoNLL	English,German	enity	partially annotated	newpaper
OntoNotes	English,Chinese	entity	annotated	various
TREC KBP 2012 SSF	English	relation	annotated	various
TAC KBP 2013 ESF	English	relation	annotated	news docs,web docs
Google Web 1T corpus	English	entity	unannotated	web pages
EPO	German,English,French	lexicon	unannotated	Patent
TACRED	English	relation	annotated	TAC KBP
ACE 2005	English,Arabic,Chinese	event	annotated	various
PFR	Chinese	entity	annotated	newspaper

training data, to develop automatic content extraction techniques to support the automated processing of human language in textual form. The ACE corpus addresses the identification of five subtasks: entities, values, temporal expressions, relations, and events.

PFR [45, 51]: The People's Daily Annotated Corpus (version 1.0, referred as the PFR corpus) is an annotated corpus produced by the Institute of Computational Linguistics, Peking University and Fujitsu Research and Development Center Ltd. with the permission of the News Information Center of People's Daily. The corpus is annotated with more than 6 million bytes of Chinese articles for word separation and lexical annotation. It is used as raw data in many studies and papers.

5 Evaluation Metrics

Usually the evaluation of the method needs to be compared with other baseline methods on the basis of designed evaluation metrics. We summarize the typical evaluation metrics for BIE methods in the Table 6. it can be found that the performance evaluation of BIE methods often requires a combination of multiple metrics on different subtasks to achieve both a comprehensive and objective evaluation.

6 Bootstrap information extraction application systems

The BIE method has been applied in many scenarios, and the BIE application system built on this basis has further advanced the research and development of BIE. Below we have collected and described some representative system applications and pointed out their advantages and disadvantages.

DIPRE [5]:The DIPRE system was designed to enable the extraction of structured data from large-scale HTML documents. Using this system one only needed to give a small number of initial relation seeds (e.g. $\langle Mao, 1893\rangle$, etc.) as input for the entity relation to be processed, and its method can automatically obtain the five-tuple description pattern and rich relation instances corresponding to that entity relation. However, the disadvantage was that it relies on HTML tags and needs more evaluation of new patterns and tuples, resulting in noisier extraction results and lower recall of extraction results.

Snowball [6]:Based on DIPRE, the Snowball system defined a five-tuple relational description schema representation with weights, annotates sentences using named entity recognition techniques, extracted only the relations between named entities, and given a well-established schema containing evaluation and filtering criteria for tuples. The effectiveness of the method was also verified on a news corpus of size 300,000 pieces. However, since Snowball used a heuristic-based approach to obtain strict and complex rules, it limited the generalization ability and thus generated low recall.

Table 6 Evaluation criteria used in the papers on BIE

Method Name	Evaluation Method	Formula
Precision[21, 23, 25–27, 30, 31, 33, 35, 37, 38, 42, 45, 47, 48, 52, 55, 61, 67–79]	It measures the accuracy of the information extraction method.	$\frac{True\quad Positive}{True\quad Positive+False\quad Positive}$
Recall[21, 23, 25–27, 30, 31, 33, 35, 37–39, 43, 45, 49, 53, 56, 68–73, 75–79]	It measures the rate of completeness of the information extraction method.	$\frac{True\quad Positive}{True\quad Positive+False\quad Negative}$
F1[20, 23, 28, 29, 32, 36–38, 42, 45, 48, 52, 55, 67–72, 74–77, 79]	It measures the combined performance of the information extraction methods, i.e., the summed average of precision and recall.	$\frac{2\times Precision\times Recall}{Precision+Recall}$
Precision at top n (P@N)/Rank Precision@N[22, 27, 44, 56, 80, 81]	It refers to the percentage of correct entities among the top N entities in the ranked list. Usually N can be taken as 5, 10, 20, 50 and 100, etc.	$\frac{Correct\quad Instances\quad In\quad Top\quad N}{N}\times 100\%$
Average Precision(AP)/AUC-PR[22, 44, 59]	It's the average of Precision values under different Recall, which is equivalent to the area under the precision-recall curves.	$\int_0^1 Precision(Recall)dRecall$
MAP/MAP@n[20, 27, 44, 56, 63, 80]	It refers to taking the mean value of AP for all categories. MAP is usually calculated for topN, where N can be taken as 10, 20 and 50, etc.	$\frac{\sum_{c=0}^{Categories} AP(c)}{Number\quad of\quad Categories}$
Throughput/Cumulative Precision-Throughput Curve/Precision-Yield Curve[18, 24, 29, 39, 41, 62–64]	Throughput or yield is the number of extraction objects and precision is the proportion of extraction objects that were correct.	None
Precision versus Iteration /P@Iter.K[63, 64, 80, 82]	It means the change in precision after K-th expansion iterations. Usually K can be taken as 1, 10 and 20, etc.	None

Basilisk [83]: The Basilisk system started with an unannotated corpus and seed words for each semantic category. Then, the system iteratively extracted words for each semantic category and assumed a semantic category of words by heavily extracting collective information from the pattern context. It relied heavily on the quality of the vocabulary seeds and the richness of the representation of dependencies between extractions.

KnowItAll [84]:The KnowItAll system automatically extracted domain-independent factual information from the Web, with inputs such as "scientist", "city", "movie", etc. The input was category concept information such as "scientist", "city", "movie", etc., and the output was a collection of instances under a specific category. This system had high extraction accuracy but low recall. The main bottleneck to KnowltAll's scalability was the rate at which it can issue search-engine queries.

URES [85]:The relation extraction system URES (an Unsupervised Web Relation Extraction System) started from the seed set of relation, further generalized the patterns based on sequential patterns using the best matching dynamic programming algorithm to obtain Soft Pattern, and finally uses Soft Pattern matching to identify new relation instances. The system was experimented on five relation types and finally obtains about 90 % accuracy. However, the degree of generalization of the pattern was reflected in the selection of scores and thresholds in unit matching in best matching, and the system did not give detailed explanatory notes and comparative experiments.

Espresso [86]:The Espresso system applied information theory to evaluate the reliability of patterns and candidate instances, and combined web-based knowledge extension techniques to extend the instances for iterative extraction of binary semantic relations under weak supervision. Experimental results showed that the exploitation of generic patterns substantially increases system recall with small effect on overall precision. However, the system did not take into account the selectional constraints on generic patterns, so the extraction effect in NLP applications was yet to be tested.

TEXTRUNNER [87]:The TEXTRUNNER system was a fully implemented Open IE one based on self supervised method. It was demonstrated that it has the ability to extract massive amounts of high-quality information from a nine million Web page corpus, and also have shown that TEXTRUNNER is able to match the recall of the KnowITALL state-of-the-art Web IE system, while achieving higher precision. The problem of detecting synonyms as well as multiple mentions of entities had not been well addressed.

AERTEWM [88]:The AERTEWM(Automated Entity Relation Tuple Extraction Using Web Mining) system proposed to use seed set and keywords as input, which solved the circular dependency problem to a certain extent, and based on this, an improved pattern acquisition and iteration strategy was proposed to extract relational tuples from the Web using web mining techniques, with a good average accuracy of 98.42 %, which can better meet the practical application requirements of information extraction. The final number of tuples extracted was influenced by the inaccuracy of NE identification and the NE designation problem.

O-CRF [89]:The O-CRF system was a CRFbased Open IE one that can extract different relationships with a precision of 88.3 % and a recall of 45.2 %. This system was a compromise between Traditional Information Extraction and Open IE to build the O-CRF system, which solved the problem of known categories but fewer categories and limited corpus size, and was also applicable to the case where the categories were unknown and the Web was needed as a corpus. The O-CRF still failed to locate the various ways in which a given relation was expressed, which makes its recall slightly low.

StatSnowball [90]:The StatSnowball system was an improved version of Snowball, which used Markov Logic Networks (MLNs) to learn the weights of Pattern, and used probabilistic methods to evaluate and select Pattern, instead of heuristic rules. And the system was highly scalable to solve both Traditional Information Extraction and Open IE problems. Empirical results showed that StatSnowball can achieve a significantly higher recall without sacrificing the high precision during iterations with a small number of seeds. Because of the limited learning and inference capability of MLN, statsnowball would also be limited in pattern learning, which affects the recall rate. Currently, the system is released with a Chinese version of Microsoft Human Cube Relationship Search and an English version of EntityCube.

NELL [7]: The NELL system proposed a never-ending learning framework that takes advantage of the ever-growing nature of information on the Web by running a system on a computer that continuously extracts information from the Web to populate the knowledge base, enabling the knowledge base to grow. After running for 67 days, this implementation populated a knowledge base with over 242,000 facts with an estimated precision of 74 %. However, the system lacked self-reflection to decide what to do next and did not interact enough with humans.

7 Prospects and Challenges

Despite more than a decade of BIE research, and the continuous emergence of new information extraction tasks, many problems and challenges still need to be solved in BIE research.

Firstly, there needs to be more in-depth research and study on seed quality control. Although some studies have discussed the selection strategy for seeds [21–24], they have yet to make comparisons without delving into the essence of the selection strategy, which has important implications for guiding BIE. Especially when obtaining seed resources from the web, low-quality seeds can directly lead to extraction failure. We should study the mechanism of seed selection strategy and explore more binding selection methods, such as self Supervised learning-based.

Secondly, the analysis in [91] has shown that semantic drift [92] is an inherent property of iterative bootstrapping algorithms and poses a fundamental problem. They have shown that iterative bootstrapping without pruning corresponds to an eigenvector computation and thus as the number of iterations increases the resulting ranking will always converge towards the same static ranking, regardless of the particular choice of seed instances. Existing solutions simply discard bad patterns. However, such methods are sacrificing recall in exchange for high precision. A portion of the research tends to rely on external resources [56] or internal constraints [59, 93, 94] that make it possible to avoid semantic drift by guaranteeing that the quality of patterns or instances is not degraded when they are generated. However, since heuristic constrains algorithms are invented for different problems, each algorithm has its own scope of application and requires a lot of expert effort. The more effective exploitability of external auxiliary resource data may be a potential research direction. Unlike constrains, the structured or regular nature of auxiliary resource data makes the semantic constraints on patterns or instances more effective and more directed to resource-rich directions. Liang et al. argued that most of these patterns and instances can be kept as long as being applied selectively, guided by prior knowledge [95]. It's worth noting that external or auxiliary knowledge like event trigger knowledge or constructed knowledge graph may bridge the seen patterns or instances and the unseen ones, thus enabling the mutual match between pattern and instance. Therefore, a key for knowledge-aware information extraction is to consider such knowledge resources as evolving side information and keep them up-to-date.

Thirdly, existing BIE methods put too much emphasis on automation, and better extraction performance, especially in the recall, might be obtained with proper expert intervention. In recent years, several works has emerged focusing on human-in-loop information extraction paradigm [96–99]. In [99], Rahman et al. presented a semi-structured interview-based study to understand IE work practices, identified several challenges with the existing IE workflows and proposed a set of design considerations, based on cognitive engineering principles, for developing human-in-the-loop IE tools. Nevertheless, further research is still needed on maximizing bootstrapping gains with minimal manual effort and conducting effective subjective and objective evaluations of the human-in-the-loop paradigm.The human-in-the-loop paradigm on IE is also well suited for data labeling tools in specialized domains where there is a severe lack of labeled data. On moving from large-scale manual to semi-automatic labeling(e.g.,Label Studio[7]), it is reasonable to expect that BIE will be able to fulfill its potential for fast and accurate recognition therein.

Fourthly, there need to be more BIE application systems for the other low resource language including Chinese. As the Chinese language example mentioned in [100], since Chinese differs significantly from English in many aspects, such as word formation, syntax, semantics, and tense, the general pattern-matching method, which extracts better results on English, is relatively less

[7] https://labelstud.io/

effective when dealing with Chinese text. Therefore, the lexical-semantic pattern matching technique is more suitable for the Chinese entity relationship extraction task. Meanwhile, with the popularity of deep learning and neural networks, research on explicit pattern-based BIE application systems that require tedious feature engineering has received less attention in the last decade. Given that end-to-end approaches both require large supervised corpora and have problems in explainability and high computation costs[11], BIE approaches with parameterized implicit patterns in incomplete supervision [101–104], has the potential to obtain large amounts of supervised data quickly and iteratively under limited supervision, significantly reducing the cost of manual annotation, especially when high-quality supervised corpora are lacking in many domains. The research and application of BIE should have its value [105–109]. BIE's contribution to information extraction has much potential to be explored in the future.

8 Conclusion

We have provided a systematic overview of diverse methods proposed in the realm of BIE. Our comprehensive review covers the four core stages of the BIE process: seed generation, pattern learning, instance acquisition, and the evaluation of patterns and instances. This summarize the underlying principles and implementation details of each BIE step, shedding light on their collective impact.

Generally, seed generation is progressively reliant on external sources. In the pattern learning stage, a discernible trend towards distributed representation of patterns and advanced pattern learning strategies is observed. Pattern generalization and instance acquisition, inherently similar, are achieved through matching or distance calculation methodologies. A multitude of filtering and ranking techniques are applicable to both pattern and instance evaluation.

Furthermore, we survey the datasets and metrics employed to assess the performance of proposed BIE methods, along with illustrating their application and system integration. Lastly, we highlight persistent challenges, encompassing seed quality control, semantic drift, manual intervention, BIE application, and outline prospective directions for future endeavors.

Acknowledgments. This research is supported by the following projects: Central Leading Local Project "Fujian Mental Health Human-Computer Interaction Technology Research Center",under the authorization number 2020L3024.

Declarations

Conflict of interest Authors declare that they have no conflict of interest.
Data Availability Statement Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

References

[1] Chen, M., Huang, L., Li, M., Zhou, B., Ji, H., Roth, D.: New frontiers of information extraction. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts, pp. 14–25. Association for Computational Linguistics, Seattle, United States (2022). https://doi.org/10.18653/v1/2022.naacl-tutorials.3. https://aclanthology.org/2022.naacl-tutorials.3

[2] Riloff, E., Jones, R., *et al.*: Learning dictionaries for information extraction by multi-level bootstrapping. In: AAAI/IAAI, pp. 474–479 (1999)

[3] Abney, S.: Bootstrapping. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 360–367 (2002)

[4] Abe, N.: Query learning strategies using boosting and bagging. Proc. of 15ˆ¡ th¿ Int. Cmf. on Machine Learning (ICML98), 1–9 (1998)

[5] Brin, S.: Extracting patterns and relations from the world wide web. In: The World Wide Web and Databases: International Workshop WebDB'98, Valencia, Spain, March 27-28, 1998. Selected Papers, pp. 172–183 (1999). Springer

[6] Agichtein, E., Gravano, L.: Snowball: Extracting relations from large plain-text collections. In: Proceedings of the Fifth ACM Conference on Digital Libraries, pp. 85–94 (2000)

[7] Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E.R., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: Twenty-Fourth AAAI Conference on Artificial Intelligence (2010)

[8] Gao, T., Han, X., Xie, R., Liu, Z., Lin, F., Lin, L., Sun, M.: Neural snowball for few-shot relation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 7772–7779 (2020)

[9] Cheng, J., Liu, J., Xu, X., Xia, D., Liu, L., Sheng, V.S.: A review of chinese named entity recognition. KSII Transactions on Internet & Information Systems **15**(6) (2021)

[10] Zhou, S., Yu, B., Sun, A., Long, C., Li, J., Yu, H., Sun, J., Li, Y.: A survey on neural open information extraction: Current status and future directions. arXiv preprint arXiv:2205.11725 (2022)

[11] Yang, Y., Wu, Z., Yang, Y., Lian, S., Guo, F., Wang, Z.: A survey of information extraction based on deep learning. Applied Sciences **12**(19),

9691 (2022)

[12] Zhang, T., Huang, Z., Wang, Y., Wen, C., Peng, Y., Ye, Y., et al.: Information extraction from the text data on traditional chinese medicine: A review on tasks, challenges, and methods from 2010 to 2021. Evidence-Based Complementary and Alternative Medicine **2022** (2022)

[13] Landolsi, M.Y., Hlaoua, L., Ben Romdhane, L.: Information extraction from electronic medical documents: state of the art and future research directions. Knowledge and Information Systems **65**(2), 463–516 (2023)

[14] Abdullah, M.H.A., Aziz, N., Abdulkadir, S.J., Alhussian, H.S.A., Talpur, N.: Systematic literature review of information extraction from textual data: Recent methods, applications, trends, and challenges. IEEE Access (2023)

[15] Arksey, H., O'Malley, L.: Scoping studies: towards a methodological framework. International journal of social research methodology **8**(1), 19–32 (2005)

[16] Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Group*, P.: Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. Annals of internal medicine **151**(4), 264–269 (2009)

[17] Canisius, S., Sporleder, C.: Bootstrapping information extraction from field books. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 827–836 (2007)

[18] Yahya, M., Whang, S., Gupta, R., Halevy, A.: Renoun: Fact extraction for nominal attributes. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 325–335 (2014)

[19] Wang, Q.: Research on entity relationship extraction based on convolutional neural network. Master's thesis, Nanjing University (2017)

[20] Qichen, H., Yawei, Z., Zheng, Y., Lijun, F.: Automatic algorithm for initial seed set generation of domain knowledge graph based on syllogism table. Chinese Journal of Informatics **32**(8), 1–8 (2018)

[21] Tuo, J., Yan, S., Li, B., Wang, H., You, X.: Aspect extraction and aspect terms expansion in chinese reviews using cluster semi-supervised expansion model. In: 2017 4th International Conference on Information Science and Control Engineering (ICISCE), pp. 212–217 (2017). IEEE

[22] Phi, V.-T., Santoso, J., Shimbo, M., Matsumoto, Y.: Ranking-based

automatic seed selection and noise reduction for weakly supervised relation extraction. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 89–95 (2018)

[23] Xiong, G., Fang, Y., Liu, Q.: Automatic construction of domain-specific sentiment lexicon based on the semantics graph. In: 2017 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), pp. 1–6 (2017). IEEE

[24] Saha, S., Pal, H., *et al.*: Bootstrapping for numerical open ie. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 317–323 (2017)

[25] Chen, P.-Y., Lee, Y.-H., Wu, Y.-H., Ma, W.-Y.: Iexm: Information extraction system for movies. In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 189–193 (2017)

[26] Zhang, C., Xu, W., Gao, S., Guo, J.: A bottom-up kernel of pattern learning for relation extraction. In: The 9th International Symposium on Chinese Spoken Language Processing, pp. 609–613 (2014). IEEE

[27] Vechtomova, O.: A semi-supervised approach to extracting multiword entity names from user reviews. In: Proceedings of the 1st Joint International Workshop on Entity-Oriented and Semantic Search, pp. 1–6 (2012)

[28] Zhang, C., Zhao, S., Wang, H.: Bootstrapping large-scale named entities using url-text hybrid patterns. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp. 293–301 (2013)

[29] Zupon, A., Alexeeva, M., Valenzuela-Escárcega, M., Nagesh, A., Surdeanu, M.: Lightly-supervised representation learning with global interpretability. In: Proceedings of the Third Workshop on Structured Prediction for NLP, pp. 18–28 (2019)

[30] Jianshu, J., Guang, C., Chunyun, Z.: A bootstrapping and mv-rnn mixed method for relation extraction. In: 2014 4th IEEE International Conference on Network Infrastructure and Digital Content, pp. 117–120 (2014). IEEE

[31] Tandon, N., Rajagopal, D., de Melo, G.: Markov chains for robust graph-based commonsense information extraction. In: Proceedings of COLING 2012: Demonstration Papers, pp. 439–446 (2012)

[32] Ding, H., Riloff, E.: Human needs categorization of affective events using labeled and unlabeled data. In: Proceedings of the 2018 Conference of

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1919–1929 (2018)

[33] Li, P., Zhou, G., Zhu, Q.: Minimally supervised chinese event extraction from multiple views. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) **16**(2), 1–16 (2016)

[34] Feng, X.: Research and application of chinese comparative sentence elements extraction technique. Master's thesis, Beijing University of Posts and Telecommunications (2016)

[35] Chen, C., He, L., Lin, X.: Rev: extracting entity relations from world wide web. In: Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication, pp. 1–5 (2012)

[36] Zhang, C., Zhang, Y., Xu, W., Ma, Z., Leng, Y., Guo, J.: Mining activation force defined dependency patterns for relation extraction. Knowledge-Based Systems **86**, 278–287 (2015)

[37] Ziering, P., van der Plas, L., Schuetze, H.: Bootstrapping semantic lexicons for technical domains. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp. 1321–1329 (2013)

[38] Yada, S., Ikeda, K., Hoashi, K., Kageura, K.: A bootstrap method for automatic rule acquisition on emotion cause extraction. In: 2017 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 414–421 (2017). IEEE

[39] Schmitz, M., Soderland, S., Bart, R., Etzioni, O., *et al.*: Open language learning for information extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 523–534 (2012)

[40] Kozareva, Z.: Learning verbs on the fly. In: Proceedings of COLING 2012: Posters, pp. 599–610 (2012)

[41] Dalvi, B., Bhakthavatsalam, S., Clark, C., Clark, P., Etzioni, O., Fader, A., Groeneveld, D.: Ike-an interactive tool for knowledge extraction. In: Proceedings of the 5th Workshop on Automated Knowledge Base Construction, pp. 12–17 (2016)

[42] Tai, L., Qin, S., Guo, F.: A pattern learning method based on kernel function. In: Proceedings of the 2017 2nd International Conference on Communication and Information Systems, pp. 324–328 (2017)

[43] FengYingHui: Research on information extraction techniques for tibetan

cultural field. Master's thesis, Central University for Nationalities (2016)

[44] Shi, B., Zhang, Z., Sun, L., Han, X.: A probabilistic co-bootstrapping method for entity set expansion (2014)

[45] Cheng, Z., Zheng, D., Li, S.: Multi-pattern fusion based semi-supervised name entity recognition. In: 2013 International Conference on Machine Learning and Cybernetics, vol. 1, pp. 45–50 (2013). IEEE

[46] Makarov, P.: Automated acquisition of patterns for coding political event data: Two case studies. In: Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 103–112 (2018)

[47] Alashri, S., Tsai, J.-Y., Koppela, A.R., Davulcu, H.: Snowball: extracting causal chains from climate change text corpora. In: 2018 1st International Conference on Data Intelligence and Security (ICDIS), pp. 234–241 (2018). IEEE

[48] Batista, D.S., Martins, B., Silva, M.J.: Semi-supervised bootstrapping of relationship extractors with distributional semantics. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 499–504 (2015)

[49] Xia, X.: Research on semi-supervised chinese event extraction. PhD thesis, Suzhou: Soochow University (2014)

[50] Liu, Y.: The information gain based binary entity relationship extraction on web corpus. PhD thesis, East China Normal University (2014)

[51] Cheng, Z.: Research on named entity recognition and relation extraction facing to domain-oriented knowledge base construction. PhD thesis, Harbin: Harbin Institute of Technology (2014)

[52] McNeil, N., Bridges, R.A., Iannacone, M.D., Czejdo, B., Perez, N., Goodall, J.R.: Pace: Pattern accurate computationally efficient bootstrapping for timely discovery of cyber-security concepts. In: 2013 12th International Conference on Machine Learning and Applications, vol. 2, pp. 60–65 (2013). IEEE

[53] Thomas, A., Sivanesan, S.: An adaptable, high-performance relation extraction system for complex sentences. Knowledge-Based Systems **251**, 108956 (2022)

[54] Wu, Z.: Research and application on content understanding algorithm for conditional semi-structured text. Master's thesis, South China University of Technology (2019)

[55] Long, L., Yan, J., Fang, L., Li, P., Liu, X.: The identification of chinese named entity in the field of medicine based on bootstrapping method. In: 2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems (MFI), pp. 1–6 (2014). IEEE

[56] Yan, L., Han, X., Sun, L., He, B.: Learning to bootstrap for entity set expansion. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 292–301 (2019)

[57] Tai, L.-t.: Research on entity relation extraction algorithm based on semi-supervised machine learning. PhD thesis, Beijing University of Posts and Telecommunications (2018)

[58] Kurihara, K., Shimada, K.: Trouble information extraction based on a bootstrap approach from twitter. In: Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, pp. 471–479 (2015)

[59] Gupta, S., Manning, C.D.: Improved pattern learning for bootstrapped entity extraction. In: Proceedings of the Eighteenth Conference on Computational Natural Language Learning, pp. 98–108 (2014)

[60] Ziering, P., van der Plas, L., Schütze, H.: Multilingual lexicon bootstrapping-improving a lexicon induction system using a parallel corpus. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp. 844–848 (2013)

[61] Yildirim, S., Yildiz, T.: Automatic extraction of turkish hypernym-hyponym pairs from large corpus. In: Proceedings of COLING 2012: Demonstration Papers, pp. 493–500 (2012)

[62] Yan, L., Han, X., He, B., Sun, L.: End-to-end bootstrapping neural network for entity set expansion. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 9402–9409 (2020)

[63] Yan, L., Han, X., He, B., Sun, L.: Global bootstrapping neural network for entity set expansion. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 3705–3714 (2020)

[64] Yan, L., Han, X., Sun, L.: Progressive adversarial learning for bootstrapping: A case study on entity set expansion. arXiv preprint arXiv:2109.12082 (2021)

[65] Ji, J.: A grammar and dependency information based relation extraction system for streaming data. Master's thesis, Beijing University of Posts

and Telecommunications (2015)

[66] Sijia, C.: Research on entity relationship extraction. Master's thesis, Beijing University of Posts and Telecommunications (2014)

[67] Tang, Z., Surdeanu, M.: Interpretability rules: Jointly bootstrapping a neural relation extractorwith an explanation decoder. In: Proceedings of the First Workshop on Trustworthy Natural Language Processing, pp. 1–7 (2021)

[68] Lin, H., Yan, J., Qu, M., Ren, X.: Learning dual retrieval module for semi-supervised relation extraction. In: The World Wide Web Conference, pp. 1073–1083 (2019)

[69] Deepika, S., Geetha, T.: Pattern-based bootstrapping framework for biomedical relation extraction. Engineering Applications of Artificial Intelligence **99**, 104130 (2021)

[70] Zhuang, Y., Jiang, T., Riloff, E.: Affective event classification with discourse-enhanced self-training. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 5608–5617 (2020)

[71] Li, Z., He, Y., Gu, B., Liu, A., Li, H., Wang, H., Zhou, X.: Diagnosing and minimizing semantic drift in iterative bootstrapping extraction. IEEE Transactions on Knowledge and Data Engineering **30**(5), 852–865 (2017)

[72] Wu, W., Li, H., Wang, H., Zhu, K.Q.: Semantic bootstrapping: A theoretical perspective. IEEE Transactions on Knowledge and Data Engineering **29**(2), 446–457 (2016)

[73] Phi, V.-T., Matsumoto, Y.: Integrating word embedding offsets into the espresso system for part-whole relation extraction. In: Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers, pp. 173–181 (2016)

[74] Bhutani, N., Jagadish, H., Radev, D.: Nested propositions in open information extraction. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 55–64 (2016)

[75] He, Y., Grishman, R.: Ice: Rapid information extraction customization for nlp novices. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 31–35 (2015)

[76] Rondon, A., Caseli, H., Ramisch, C.: Never-ending multiword expressions learning. In: Proceedings of the 11th Workshop on Multiword

Expressions, pp. 45–53 (2015)

[77] Ye, F., Shi, H., Wu, S.: Research on pattern representation method in semi-supervised semantic relation extraction based on bootstrapping. In: 2014 Seventh International Symposium on Computational Intelligence and Design, vol. 1, pp. 568–572 (2014). IEEE

[78] Zhang, C., Niu, Z., Jiang, P., Fu, H.: Domain-specific term extraction from free texts. In: 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, pp. 1290–1293 (2012). IEEE

[79] Qadir, A., Riloff, E.: Ensemble-based semantic lexicon induction for semantic tagging. In: * SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pp. 199–208 (2012)

[80] Momtazi, S., Moradiannasab, O.: A statistical approach to knowledge discovery: Bootstrap analysis of language models for knowledge base population from unstructured text. Scientia Iranica **26**(Special Issue on: Socio-Cognitive Engineering), 26–39 (2019)

[81] Zhao, H., Feng, C., Luo, Z., Tian, C.: Entity set expansion from twitter. In: Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval, pp. 155–162 (2018)

[82] Wang, C., Wang, F.: A bootstrapping method for extracting sentiment words using degree adverb patterns. In: 2012 International Conference on Computer Science and Service System, pp. 2173–2176 (2012). IEEE

[83] Thelen, M., Riloff, E.: A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pp. 214–221 (2002)

[84] Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Unsupervised named-entity extraction from the web: An experimental study. Artificial intelligence **165**(1), 91–134 (2005)

[85] Rosenfeld, B., Feldman, R.: Ures: an unsupervised web relation extraction system. In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pp. 667–674 (2006)

[86] Pantel, P., Pennacchiotti, M.: Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In: Proceedings of the 21st

International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pp. 113–120 (2006)

[87] Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open information extraction from the web. Communications of the ACM **51**(12), 68–74 (2008)

[88] LI, W.-g., LIU, T., LI, S.: Automated entity relation tuple extraction using web mining. ACTA ELECTONICA SINICA **35**(11), 2111 (2007)

[89] Banko, M., Etzioni, O.: The tradeoffs between open and traditional relation extraction. In: Proceedings of ACL-08: HLT, pp. 28–36 (2008)

[90] Zhu, J., Nie, Z., Liu, X., Zhang, B., Wen, J.-R.: Statsnowball: a statistical approach to extracting entity relationships. In: Proceedings of the 18th International Conference on World Wide Web, pp. 101–110 (2009)

[91] Komachi, M., Kudo, T., Shimbo, M., Matsumoto, Y.: Graph-based analysis of semantic drift in espresso-like bootstrapping algorithms. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 1011–1020 (2008)

[92] Curran, J.R., Murphy, T., Scholz, B.: Minimising semantic drift with mutual exclusion bootstrapping. In: Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, vol. 6, pp. 172–180 (2007). Citeseer

[93] Zhang, Y., Shen, J., Shang, J., Han, J.: Empower entity set expansion via language model probing. arXiv preprint arXiv:2004.13897 (2020)

[94] Huang, J., Xie, Y., Meng, Y., Shen, J., Zhang, Y., Han, J.: Guiding corpus-based set expansion by auxiliary sets generation and co-expansion. In: Proceedings of The Web Conference 2020, pp. 2188–2198 (2020)

[95] Liang, J., Feng, S., Xie, C., Xiao, Y., Chen, J., Hwang, S.-W.: Bootstrapping information extraction via conceptualization. In: 2021 IEEE 37th International Conference on Data Engineering (ICDE), pp. 49–60 (2021). IEEE

[96] Alba, A., Coden, A., Gentile, A.L., Gruhl, D., Ristoski, P., Welch, S.: Multi-lingual concept extraction with linked data and human-in-the-loop. In: Proceedings of the Knowledge Capture Conference, pp. 1–8 (2017)

[97] Gentile, A.L., Gruhl, D., Ristoski, P., Welch, S.: Explore and exploit.

dictionary expansion with human-in-the-loop. In: European Semantic Web Conference, pp. 131–145 (2019). Springer

[98] Kirsch, B., Niyazova, Z., Mock, M., Rüping, S.: Noise reduction in distant supervision for relation extraction using probabilistic soft logic. In: Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II, pp. 63–78 (2020). Springer

[99] Rahman, S., Kandogan, E.: Characterizing practices, limitations, and opportunities related to text information extraction workflows: A human-in-the-loop perspective. In: CHI Conference on Human Factors in Computing Systems, pp. 1–15 (2022)

[100] Deng, B., Fan, X., Yang, L.: Entity relation extraction method using semantic pattern. Jisuanji Gongcheng/ Computer Engineering **33**(10), 212–214 (2007)

[101] Pengfei, L., Zheng, Y., Chunning, W., Yueqin, Z., Wei, L.: Research on the geological entities business relation extraction based on the bootstrapping method. Transformations in Business & Economics **21**(2) (2022)

[102] Yang, C., Xiao, D., Luo, Y., Li, B., Zhao, X., Zhang, H.: A hybrid method based on semi-supervised learning for relation extraction in chinese emrs. BMC Medical Informatics and Decision Making **22**(1), 169 (2022)

[103] Li, Y., Yu, X., Liu, Y., Chen, H., Liu, C.: Uncertainty-aware bootstrap learning for joint extraction on distantly-supervised data. arXiv preprint arXiv:2305.03827 (2023)

[104] Novotnỳ, V., Luger, K., Štefánik, M., Vrabcová, T., Horák, A.: People and places of historical europe: Bootstrapping annotation pipeline and a new corpus of named entities in late medieval texts. arXiv preprint arXiv:2305.16718 (2023)

[105] Sheikhpour, R., Berahmand, K., Forouzandeh, S.: Hessian-based semi-supervised feature selection using generalized uncorrelated constraint. Knowledge-Based Systems **269**, 110521 (2023)

[106] Doumari, S.A., Berahmand, K., Ebadi, M., et al.: Early and high-accuracy diagnosis of parkinson's disease: Outcomes of a new model. Computational and Mathematical Methods in Medicine **2023** (2023)

[107] Menhour, H., Şahin, H.B., Sarıkaya, R.N., Aktaş, M., Sağlam, R., Ekinci, E., Eken, S.: Searchable turkish ocred historical newspaper collection 1928–1942. Journal of Information Science **49**(2), 335–347 (2023)

[108] Yurtsever, M.M.E., Özcan, M., Taruz, Z., Eken, S., Sayar, A.: Figure search by text in large scale digital document collections. Concurrency and Computation: Practice and Experience **34**(1), 6529 (2022)

[109] Omurca, S.I., Ekinci, E., Sevim, S., Edinc, E.B., Eken, S., Sayar, A.: A document image classification system fusing deep and machine learning models. Applied Intelligence **53**(12), 15295–15310 (2023)