IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 March 2024, accepted 23 April 2024, date of publication 29 April 2024, date of current version 6 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3394906

RESEARCH ARTICLE

Cost-Effective Knowledge Extraction Framework for Low-Resource Environments

SANGHA NAM[1] AND EUN-KYUNG KIM[2]
[1]Language AI Laboratory, NCSoft, Seongnam 13494, Republic of Korea
[2]Department of Big Data and AI, Daejeon University, Daejeon 34520, Republic of Korea

Corresponding author: Eun-Kyung Kim (eunkk@dju.kr)

This work was supported by Daejeon University Research Grants (2019).

ABSTRACT Extracting knowledge from texts is crucial for enriching everyday knowledge. Constructing a knowledge extraction environment requires comprehensive processes, such as data generation, data processing, and model and framework design. However, these processes require significant effort in low-resource environments where shared data are not published. Currently, there is no environment that can design an entire knowledge extraction framework and perform step-by-step experiments even with unlimited resources. Thus, this study proposes a method for building a cost-effective knowledge extraction environment. In particular, we present a low-cost, high-quality method for annotating a corpus for knowledge extraction, in which data sharing is unavailable. The dataset collected using this method improves the performance of knowledge-extraction system models. Specifically, the co-reference resolution and relation extraction performance were improved by 10% and 18.9%, respectively. Additionally, the entire knowledge extraction system was evaluated using sequential multitask learning, and the performance was improved by 5% as each trained model was introduced.

INDEX TERMS Crowdsourcing, knowledge base, knowledge extraction, low-resource environment.

I. INTRODUCTION

Recent progress in machine learning and natural language processing (NLP) has advanced the field of knowledge extraction from unstructured text. Knowledge extraction is vital in such areas as information retrieval, natural language understanding, and knowledge management. Various large-scale knowledge bases, such as DBpedia [1], YAGO [2], and Wikidata [3], play a significant role in numerous knowledge-extraction endeavors. Enhancing a factual knowledge base through knowledge extraction requires addressing questions regarding entity identification and their interrelations. There is a substantial body of research on knowledge extraction in both English [4], [5], [6], [7], [8] and Chinese [9], [10], focusing on critical models such as entity linking [11] and relation extraction [12]. These models are crucial and often complemented by entity discovery [6], [13], co-reference resolution [14], [15], [16], [17], and knowledge validation [18], [19] to improve accuracy and comprehensiveness. Advanced deep learning architectures, including LSTM [16] and transformers [17], [20], are increasingly employed throughout the knowledge extraction process. However, these sophisticated models require extensive annotated datasets for optimal performance.

In this context, three primary factors are pivotal for the knowledge extraction framework: the **framework** itself, **constituent models**, and **training data** for each model. The framework integrates various models into a cohesive knowledge extraction system, defining the inputs/outputs for each model to enhance the overall system efficiency. However, the pipeline nature of these systems means that errors can propagate from one model to the next, potentially diminishing the system performance. Thus, there is a critical need for high-quality, labeled data to train these models effectively.

However, compiling a corpus that facilitates the sequential training and evaluation of all models within a knowledge extraction framework remains challenging. This difficulty arises from the disparity in shared data across different sources, even in languages with abundant resources such as

The associate editor coordinating the review of this manuscript and approving it for publication was Ramakrishnan Srinivasan.

English [21], [22], [23]. Furthermore, models often serve diverse purposes, complicating the assessment of overall performance owing to error propagation through various stages.

This study outlines a comprehensive design and its implementation in a knowledge extraction framework in a low-resource setting, emphasizing data collection through practical experiments. We introduce a four-phase crowdsourcing strategy for a unified corpus, covering entity mention detection, entity linking, co-reference resolution, and relation extraction. These phases aim to enhance the recall of discovery models by identifying additional entities within the texts. Moreover, we explore the application of this framework to low resource Korean data, evaluating the experimental outcomes and annotated data requirements, alongside the robustness of the models without annotated data. The findings suggest that model performance depends heavily on the quality and quantity of training data. This method could significantly benefit low-resource knowledge extraction environments, as demonstrated by a Korean case study, showcasing the efficacy of the proposed approach.

II. KNOWLEDGE-EXTRACTION ENVIRONMENT

This section outlines the environment for knowledge extraction that encompasses the framework, models, and data necessary for training each model. We begin by addressing the challenge of compiling a corpus that supports the sequential training and evaluation of models within a knowledge extraction framework, underscoring the need for high-quality and accurately labeled data.

A. COST-EFFECTIVE CORPUS CONSTRUCTION

We developed a dataset from crowdsourced texts of Korean Wikipedia by employing distant supervision (DS) following the methodology proposed by Mintz et al. [24]. DS is a machine learning strategy used to automatically annotate large volumes of unlabeled data. The underlying principle of DS is that if a text segment mentions an entity recognized to have a specific association in a knowledge base, this text can be assigned the corresponding label. DS are frequently utilized in supervised learning contexts to facilitate the training of models for various tasks, including sentiment analysis and named entity recognition. The application of DS for automatic data labeling significantly expedites the annotation process, enabling the creation of substantial high quality datasets for NLP tasks, particularly in languages with scarce resources.

Our approach involved extracting paragraphs eligible for DS annotation, as well as their preceding paragraphs, to serve as primary data. For example, if DS-derived data came from paragraph V of a Korean Wikipedia article, we included paragraphs I-V as the source material for crowdsourcing. This strategy allowed us to generate knowledge extraction datasets at both paragraph and document levels, extending beyond mere sentence-level data. Although platforms sucn as Amazon Mechanical Turk [25] and Figure Eight [26] have seen extensive use for English language crowdsourcing efforts, analogous initiatives for the Korean language have been scarce, posing challenges in engaging native Korean speakers in these tasks.

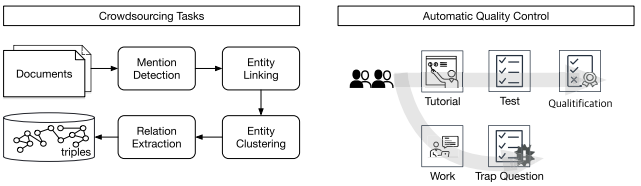

FIGURE 1. Overview of the knowledge extraction and crowdsourcing data collection process.

To overcome this hurdle, we turned to CrowdWorks,[1] a premier Korean crowdsourcing platform renowned for its capacity to recruit and manage crowdsourced labor effectively. Our demanded that participants have specific linguistic capabilities and qualifications. Leveraging the CrowdWorks platform enabled us to enlist registered Korean speakers who possessed the requisite skills, thereby ensuring that the tasks were completed efficiently and effectively.

Fig. 1 illustrates the combined knowledge extraction and crowdsourcing data collection methodology used in this study. The methodology encompasses two principal components. The initial phase involves the crowdsourcing task, during which annotators complete and submit their tasks, followed by an automatic quality control phase that applies various strategies to verify the accuracy and consistency of the annotations. In the initial crowdsourcing phase, annotators are recruited via a crowdsourcing platform and tasked with a series of annotation activities. This phase is organized into four distinct steps: entity/mention detection, entity linking, co-reference resolution, and relation extraction annotation, with the outcome of each step forming the next. This sequential approach yields a more integrated and thorough knowledge extraction process than if the data collection is conducted independently for each step.

In the first step, annotators identified potential entity mentions within the text. The second step involves linking the identified mentions to entities in a knowledge base. The third step required annotators to find pronouns, demonstrative determiners, and antecedents of newly identified entities from the first step. In the final step, annotators analyzed textual cues to ascertain the relationships between entities. The Generalized Inference (GI) protocol, which comprises rules, guidelines, or procedures for effectively synthesizing and analyzing inputs from numerous contributors to make predictions or conclusions, was applied across all tasks. The subsequent subsections further detail the operational specifics of the crowdsourcing work environment for these four phases.

The automatic quality control phase leverages several automated mechanisms such as trap questions, estimation of

[1] https://www.crowdworks.ai/ko/

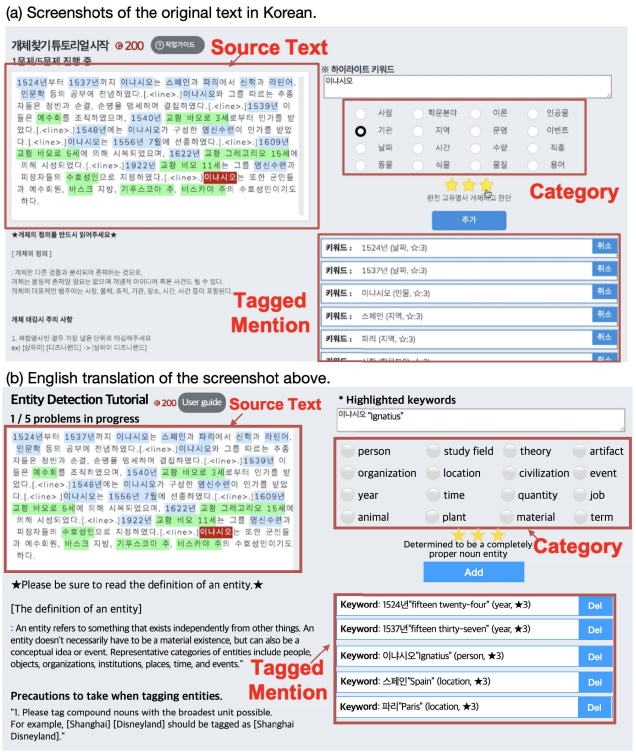

FIGURE 2. Interface for entity detection annotation; all instances of the annotation interface are presented in Korean. Nonetheless, these interfaces are designed to be easily adaptable to additional languages. (Image (a) displays the interface's original homepage screen, whereas image (b) has been translated from Korean into English to facilitate comprehension.)

worker quality scores, and adjudication to oversee and uphold the quality of the annotation work. Prior to starting their tasks, the annotators underwent training with tutorials and practice tests. To assure high-quality annotations, each task, known as a "hit," is distributed among several annotators. Quality control measures include assessing annotator reliability through trap questions, inserted by experts at intervals of every seven to ten hits. The quality assurance process involves excluding the last seven to ten hits of any annotator who fails to meet the predetermined threshold, with repeat offenders removed from the project. Automated tools were introduced at each stage to offer guidance that could lower the complexity level of the task, with comprehensive descriptions provided in each respective phase.

1) ENTITY DETECTION

Within the field of NLP, entity detection involves identifying real-world objects or concepts referred to as entities, such as individuals, organizations, or places mentioned in a text. Fig. 2 illustrates the design of the annotation interface for entity detection. The interface displays the source text by paragraph in the upper-left section. For the Wikipedia content, texts highlighted with Wikipedia-related details (words colored in green) were presented to the annotators. This feature enabled annotators to easily discern which

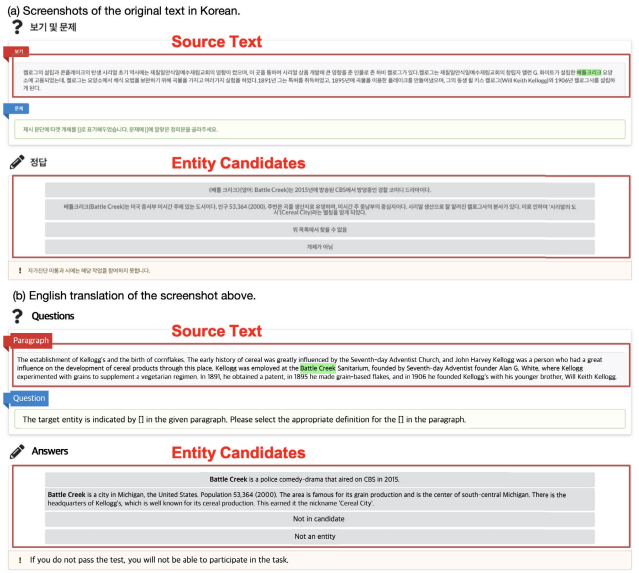

FIGURE 3. Annotation interface for entity linking. (Image (a) displays the original screen of the homepage, whereas image (b) has been translated from Korean into English to enhance understanding.)

mentions needed to be tagged (words highlighted in blue), as they read the text and reviewed the example entities in real time. Upon identifying an entity mention, the annotator chose its category from 16 predefined categories (person, field of study, theory, artifact, organization, location, civilization, event, year, time, quantity, profession, animal, plant, material, and term) commonly utilized for Korean named entity recognition tasks. This procedure assisted annotators in understanding the entity types and applying tagged data in the named entity recognition task. To reduce the chance of overlooking an entity mention, each paragraph was assigned to two annotators. Trap questions are employed in entity detection to continually monitor the quality of the work performed. Assigning two annotators to each task was crucial because even experienced annotators might not detect all entities flawlessly. Furthermore, the same entity can be annotated with varying degrees of detail and as different entity types, depending on the context and the annotator's interpretation. The final dataset was compiled by integrating the contributions of the two annotators.

2) ENTITY LINKING

Entity linking entails associating the mentions of entities within a text with their corresponding entries in a knowledge base. Fig. 3 shows the layout of the annotation interface for entity linking. The interface shows the text with entity mentions highlighted in green at the top of the screen. Below, annotators were presented with a list of abstracts corresponding to potential entity candidates. These abstracts aimed to facilitate a better understanding of the context, thereby increasing the likelihood of correctly matching mentions to entities. Each option set included "*Not in candidate*" and "*Not an entity*" choices along with the

candidate entity abstracts. The "*Not in candidate*" option was selected when an appropriate match could not be found, and "*Not an entity*" was chosen for incorrectly tagged mentions. These options serve not only to ensure the collection of accurate data but also to aid in creating a dataset for the entity discovery task, which involves adding new entities to the knowledge base.

Entity candidates were automatically identified from the KBox knowledge base, which encompasses entities listed on Korean Wikipedia [27]. This provided a comprehensive resource for linking textual mentions to real-world entities. To promote quality control, annotators were encouraged to perform self-assessment of their entity-linking tasks, a practice aimed at preserving the quality of their submissions. Each paragraph of the text was assigned to a single annotator.

3) CO-REFERENCE RESOLUTION

Co-reference resolution, a pivotal task in NLP, involves determining which mentions within a text refer to the same entity or concept. This task shows variable performance across languages, with English achieving a peak accuracy of 73%, whereas Korean trails achieve 58% [16], [28]. The relatively modest success rate in co-reference resolution implies a significant risk of errors being transferred to subsequent stages, such as relation extraction. Additionally, the prevailing guidelines for generating and annotating data for general-purpose co-reference resolution are intricate, incorporating numerous rules that annotators must adhere to [29] and [30]. The complexity of establishing mention boundaries and identifying co-referential relationships poses a substantial challenge for data annotation, especially when carried out by the general public who may not have professional training in this area.

To mitigate these challenges, our approach restricts the types of target mentions to named entities, pronouns, or definite noun phrases that are essential for knowledge extraction tasks. In Korean, a definite noun phrase typically consists of a demonstrative determiner coupled with a noun that collectively serves to reference an antecedent [31]. This focus on named entities, pronouns, and definite noun phrases exclusively is strategic because these elements are crucial for knowledge extraction. Named entities represent the foundational components of knowledge extraction. Pronouns and definite noun phrases are predominantly used to refer back to previously mentioned entities, offering a more intuitive and straightforward framework for determining the range of mentions and their co-references. Consequently, prior research on Korean co-reference resolution has similarly constrained the scope of mentioning these specific elements [31], [32].

Fig. 4 shows the design of the annotation interface for the co-reference resolution. Initially, mentions were extracted based on predefined rules. Named entities were carried over from the entity-linking phase, whereas pronouns and definite noun phrases were automatically identified with a 99% recall

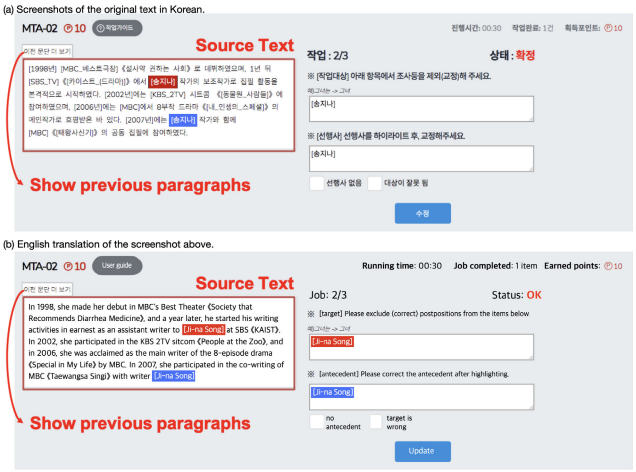

FIGURE 4. Annotation interface for co-reference resolution. (Image (a) displays the original screen of the homepage, whereas image (b) has been translated from Korean into English to enhance understanding.)

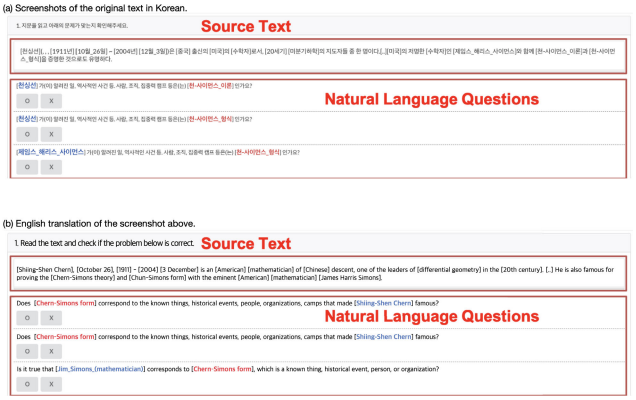

FIGURE 5. Interface for Relation Extraction Annotation. (Image (a) displays the original screen of the homepage, whereas image (b) has been translated from Korean into English to enhance understanding.)

rate, facilitated by straightforward detection rules. Within the annotation interface, automatically identified mentions are highlighted in the upper left section. Annotators contribute by selecting a mention that might serve as an antecedent. Then, choosing between the options "*No antecedent*" or "*Entity error*" to indicate their findings.

If a co-reference resolution is unachievable owing to the absence of an antecedent within the present paragraph, annotators have the option to click on the "*Show previous paragraphs*" button located in the upper left section. This action reveals all the preceding paragraphs, enabling the annotator to locate and tag an appropriate antecedent. This functionality is crucial for facilitating co-reference resolution at the document level, allowing for a more comprehensive understanding of the text structure and entity relationships. Each paragraph was assigned to a single annotator for the evaluation.

4) RELATION EXTRACTION

Relation extraction involves identifying the relationships between pairs of entities within a text. Fig. 5 shows

TABLE 1. Entity detection and linking corpus.

Dataset	CROWD	GOLDSET	AIDA/CoNLL
Document	2,574	434	1,393
Avg. words in dataset	200	39	216
Total number of mentions	152,301	4,186	34,956
Not-in-candidate	10,566	137	7,136
Not-an-entity	1,836	35	-
Empty candidate	17,096	-	-

TABLE 2. Co-reference resolution corpus.

Dataset	CROWD	GOLDSET	CoNLL-2012
Document	2,660	207	3,395
Sentence	55,406	2,493	90,191
Mention of chain (A)	109,484	3,839	187,384
Reference chain (B)	32,172	1,127	40,355
Ratio of A/B	3.403	3.406	4.643

TABLE 3. Relation extraction corpus.

Dataset	CROWD	GOLDSET
Number of relation	113 (a)	76 (subset of a)
Number of true labeled data	141,858	3,190
Number of false labeled data	121,566	0

the configuration of the annotation interface for relation extraction. This phase differs from earlier ones as it focuses on discerning the relationships between entities identified and labeled in preceding tasks rather than grouping co-occurring entities. To gather data for relation extraction, we employed the DS method [24], expanding our dataset from individual sentences to all paragraphs. This expansion enables entity pairs that appear across different sentences to be captured.

Each relationship (or property) utilized in the DS data collection process was assigned concise English labels as stipulated by the DBpedia ontology schema. To eliminate any potential confusion in interpreting these properties, we presented them in the form of straightforward yes/no questions phrased in natural language. The definitions for each property were derived from Wikidata's descriptions. For example, the property `birthPlace` was rephrased as "the birth location of a person, animal, or fictional character." In this phase, each annotator was tasked with a relation extraction dataset comprising 15 such questions on average.

B. RESOURCES

Ontological knowledge extraction depends on a foundational knowledge base that serve as a critical resource for tasks such as entity linking and relation extraction. This foundational base guarantees that the extracted information aligns with the predefined schema of the knowledge-base. For our purposes, we selected KBox [27], an extension of the Korean DBpedia, as our reference knowledge base.

Table 1 provides basic statistics for the entity detection and linking dataset that we compiled in comparison to the AIDA/CoNLL dataset [33], which is extensively utilized in English-speaking regions. Using the crowdsourcing approach detailed in Section III, we gathered 2,574 document-level samples, yielding a dataset nearly twice the size of the AIDA/CoNLL dataset. To evaluate the effectiveness of our crowdsourcing model, we created 434 expert-reviewed document-level gold standard datasets, featuring fewer words per document than the training dataset. Mentions that lacked appropriate definitions were categorized as "*Not in candidate*," whereas mistakenly identified entities were labeled as "*Not an entity*." An empty candidate was automatically designated for mentions that did not have any identified associated entity candidates.

Table 2 shows the fundamental statistics of the co-reference resolution dataset gathered in this study along with those of the CoNLL-2012 dataset [30], which is extensively utilized in English-language research. This dataset was developed from source texts, wherein entities were identified and linked by annotators to demonstrate the presence of antecedents in the same document. Candidates for pronouns and demonstrative determiners were identified using a pronoun-extraction tool. We amassed 1,480 document-level samples, amounting to slightly less than half the volume of the CoNLL-2012 dataset. Following the completion of the crowdsourcing effort, a gold standard dataset reviewed by four experts was established for model evaluation purposes.

In this context, a **mention chain (A)** denotes the count of mentions that have antecedents, whereas a **reference chain (B)** signifies the total number of distinct entities to which the mentions refer, organized by grouping. The ratio of mention chains to reference chains per document in the CoNLL-2012 dataset was 4.6, compared with 3.3, in our dataset. This suggests that, on average, an entity was mentioned approximately 3.3 times within a document in our model.

Table 3 outlines the basic statistics of the relationship extraction dataset used in this study. We identified 113 extracted relations within the dataset, a figure significantly exceeding those found in TAC-KBP (41) and NYT10 (51), which are commonly used for relation extraction tasks in English. Our dataset facilitated the identification of relationships between entity pairs across different sentences by utilizing paragraph level DS. The source data were derived from Korean Wikipedia and KBox, with an observed average noise level of 49.5%. Notably, the deathPlace and birthPlace relations exhibited exceptionally high noise levels at 97% and 96% respectively, indicating that a model trained solely on DS data might struggle to accurately determine the relationships between these two types of entities. To evaluate the efficacy of the relation extraction model, four experts from Telecommunications Technology Association[2] generated 3,190 gold standard datasets for relation extraction, utilizing both the Korean Wikipedia and the newly compiled corpus.

C. KENet: KNOWLEDGE EXTRACTION FRAMEWORK

The architecture of KENet is shown in Fig. 6. KENet processes textual sources to extract factual knowledge

[2]https://www.tta.or.kr

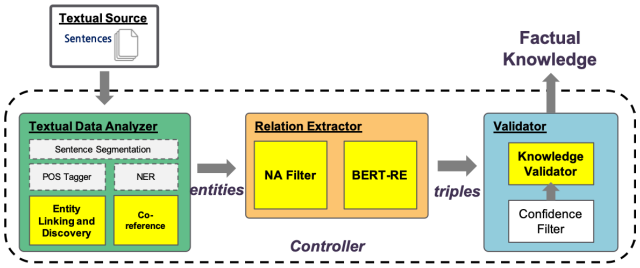

FIGURE 6. Overview of a knowledge extraction framework.

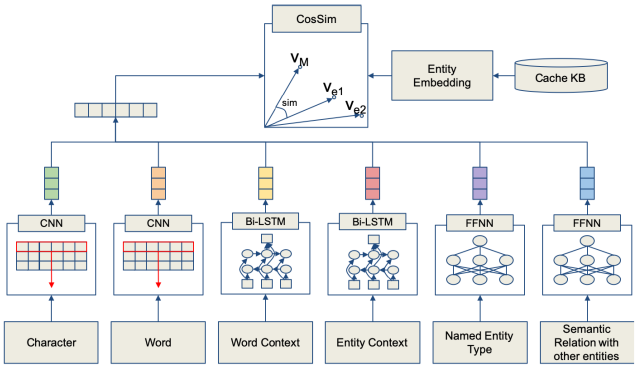

FIGURE 7. Entity encoder for a new entity.

through a series of components: textual data analyzer, relation extractor, and validator. The textual data analyzer comprise an NLP tool, an entity linking and discovery model, and co-reference model. Its primary function is to identify entities within the text and group these entities and noun phrases, including pronouns, by analyzing their anaphoric relations. This analysis is crucial for preparing the input for the subsequent relational extraction phase. The relation extractor component then takes the entity information prepared by the textual data analyzer, forms all possible pairs from these entities, and determines the relationships between each pair. A key feature of this component is the not-a-relation (NA) filter, which evaluates whether a given entity pair's relationship is specified within a predefined dataset of relations. Following the extraction of the relations, the validator or confidence filter–assesses the extracted relations based on their scores. This filtering process ensures that only the most reliable triplets are retained. The final output consists of factual knowledge validated through this rigorous process. A controller seamlessly integrates these components, functioning as a regulatory mechanism that adjusts the threshold for relation extraction based on predefined criteria, and determines the operational status of the model. This architecture ensures a systematic approach for factual knowledge extraction from textual sources.

1) ENTITY LINKING AND DISCOVERY

Entity linking involves associating a specific entity e with a mention m from a knowledge base containing an entity set E. Entity discovery [34] involves performing NIL clustering for mentions that cannot be linked to any entity within a knowledge base. NIL clustering groups mention the same new entity from different sources.

The proposed method for entity linking and discovery is divided into three sub steps. Initially, a mention detection model identifies potential entity candidates that could correspond to mentions in the knowledge base. Subsequently, the model searches among these candidates for the entity to be linked and evaluates their linking scores for the best match. For the entity-detection and linking tasks, we adapted the model of Le and Titov [11] for the Korean language. This model examines candidate mentions extracted from document mention M to explore the possible connections between the context and each candidate. Let K represent the number of relations, and let candidate set C_i correspond to mention m_i within M. The model computes the score of each candidate using a local scoring model that incorporates an attention mechanism for context words around m_i to align with $c_{ij} \in C_i$. In the pairwise scoring model, the scores for pairs of entity candidates are calculated by applying K relation matrices for each candidate pair (m_i and m_j). The scores derived from the candidate-context and candidate-candidate assessments were then aggregated to identify the highest-scoring entity set within the document. To tailor this model to the Korean language, we employed the techniques of Le and Titov [11] using 300-dimension GloVe for word embeddings and Gupta et al. [35] for entity embeddings.

Subsequent to linking, entity discovery is executed to identigy entities eligible for registration as new entries. These entities are temporarily stored in a cache knowledge base which is a provisional repository for newly identified entities. An entity migrates to the main knowledge base when its ontological relationship with existing entities is established through relation extraction. The registration of a new entity in the cache includes its representative name, mention of the surface form, entity type, and entity embedding. The process for generating the embedding value utilizes an entity-embedding encoder, as illustrated in Fig. 7.

2) CO-REFERENCE RESOLUTION

Co-reference resolution is the process of identifying when different expressions refer to an entity. This task is crucial in natural language understanding because words such as pronouns, demonstrative determiners, or abbreviations often refer back to previously mentioned entities (antecedents) in various forms. Successful grouping of all expressions that point to the same entity enhances text comprehension and is vital for thorough knowledge extraction, as entity linking alone might not capture all pertinent information. The task involves linking antecedents to a given mention m, where an antecedent y could be either a preceding entity or a dummy antecedent. Dummy antecedents are considered in cases where the text span either does not represent an entity mention or represents one, but does not refer to any prior mentions.

To tailor the co-reference model for the Korean language, modifications were made to the model presented by Joshi et al. [17]. These modifications include breaking the token level input vectors into morpheme-level representations, retraining three word-embedding models (Word2Vec, ELMo, and character embedding) using the Korean Wikipedia corpus and incorporating named-entity recognition for each mention as an additional feature. The key parameters of the adapted model include the use of tje Adam optimizer for optimization, a second order model for word representation, a two-layer feedforward neural network with 250 dimensions each, and an LSTM with a 250-dimension hidden state. The feature and character embedding vector sizes were set to 40 and 24 dimensions, respectively.

3) RELATION EXTRACTION

Relation extraction is the process of identifying the relationship between two entities in a sentence. This task is fundamental to transforming natural language sentences into structured knowledge. For instance, from the sentence "Mark Zuckerberg is the founder of Facebook," a relation extraction system would identify the relationship as `Founder(Facebook, Mark Zuckerberg)`. Each relation is supported by a set S r, classifying two entities into one of the predefined relations R upon receiving an instance x. Relation extraction has been extensively explored, and addressing the **NA** (Not Applicable) problem is crucial from a knowledge extraction standpoint. This involves determining whether the relationship between the two entities in a sentence belongs to the predefined relation set R.

The knowledge extraction framework processes plain text, identifies all entities, and extracts relationships for all possible entity pairs. For instance, in the sentence "The Restol Special Rescue Team aired on Tooniverse and Arirang TV," the relationship between `Tooniverse` and `Arirang TV` should be classified as **NA**, indicating no direct relationship. However, this sentence also allows the extraction of a valid relationship `channel(The Restol Special Rescue Team, Arirang TV)`. Addressing relation extraction with **NA** classification poses challenges, as the relevance of entity pairs grouped in the same sentence varies, requiring a nuanced understanding of the context surrounding each entity pair.

In this study, we developed a model aimed at addressing the challenge of relation extraction, including cases with **NA** relations, as shown in Fig. 8. The core component for handling **NA** relations is a binary classifier based on BERT, finely tuned for the specific task of relation extraction, leveraging the methodology proposed by Soares et al. [12]. During the fine-tuning process, special tokens were inserted before and after an entity was mentioned in the input sentence. Relationship classification relies on the embedding of a special [CLS] token designed to encapsulate the sentence's overall meaning. The role of the **NA** filter is to ascertain whether the relationship between two entities in a given sentence is part of a predefined set of relationships.

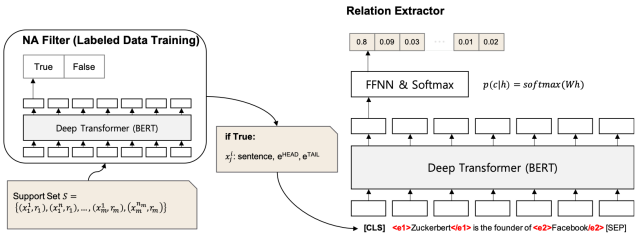

FIGURE 8. Architecture of the relation extractor with NA filtering.

To train this model, we utilize the relation extraction dataset curated via crowdsourcing, as detailed in Section II-B. Performance evaluation of a set of held-out data demonstrated an accuracy rate of 76%. Texts that successfully passed through the NA filter were then processed using the relation extraction model to identify and extract the final triple relationship.

4) KNOWLEDGE VALIDATION

Knowledge validation involves s the confirmation of the accuracy of a given set of facts. In this study, we employed several methods to validate this knowledge:

- **Schema and instance-based domain/range filtering:** This approach validates triples based on constraints defined in an ontology schema. Because some domains and ranges were not explicitly defined in the schema, we supplemented this method with instance-based filtering, which relies on the statistics of the A-box.
- **Positive correlation using knowledge base triple embedding:** Following the method of Kim et al. [36], this technique learns the embedding of knowledge-based triples and validates the target triple by ensuring a positive correlation with other triples within the embedding space.
- **Negative rule mining:** Inspired by Ortona et al. [19], this method learns negative rules between knowledge-bases triples. For example, for a `parent(A, B)` triple, the model validates the relationship based on the learned rule that A's birth date cannot be earlier than that of B's.

III. EXPERIMENT AND ANALYSIS

This section details the findings of our investigation into the development of a cost-effective knowledge extraction framework tailored for low-resource settings. This section is divided into subsections focusing on the different facets of the framework's performance and utility. Initially, we explored how accurately the framework identified correct answers during the knowledge extraction process. Subsequently, we evaluate the performance of our KENet framework, including assessments of individual components, such as entity linking and co-reference resolution. In addition, we scrutinized the crowdsourcing data utilized in this study and assessde the impact of the NA filter on relation extraction.

Finally, we draw comparisons between the results of the relation extraction in English and our findings. Through this comprehensive analysis, we aimed to showcase the efficacy and cost efficiency of our framework in facilitating knowledge extraction in environments with limited resources.

A. IDENTIFICATION OF THE CORRECT ANSWER

To gauge the effectiveness of our knowledge extraction framework, we begin by testing its capability to accurately identify entities across various sentence structures. Specifically, we examined the framework's performance in extracting entity relationships from sentences that fell into four distinct categories.

- **Simple pattern**: A sentence that follows a straightforward pattern, such as "`Mark Zuckerberg`, is educated at `Harvard`." In this case, the relationship between entities is clear and easy to identify, and the correct answer should be *education*.
- **Logical reasoning**: A sentence that requires logical reasoning to infer the relationship between entities, such as "Malan is a `Harvard` professor who taught `Mark Zuckerberg` there." In this case, the system needs to understand the context and use logical reasoning to determine the correct relationship, which should be *education*.
- **Co-reference reasoning**: A sentence that uses co-references to refer to the entities, such as "Eduardo and `Mark Zuckerberg` are close friends and they studied together at `Harvard`." Here, the system needs to correctly identify the co-referential relationship between "Eduardo" and "`Mark Zuckerberg`" and recognize that they are both related to `Harvard` in the correct way, which should be *education*.
- **Commonsense reasoning**: A sentence that requires commonsense reasoning to determine the relationship between entities, such as "`Mark Zuckerberg`, received scholarship from `Harvard`." In this case, the system needs to understand the cultural and societal norms surrounding scholarships and universities to determine the correct relationship, which should be *education*.

By evaluating the system's ability to discern the relationships between entities in these varied contexts, we can determine its overall effectiveness and pinpoint areas for enhancement. The subsequent subsections delve deeper into the performance of each component within our proposed framework and provide insights into its strengths and limitations.

B. PERFORMANCE OF KENet

This subsection describes the experiment that was conducted to assess the performance of the KENet framework. We measured the precision, recall, and F1-score to evaluate its effectiveness. The framework was tested incrementally by introducing sub-models one at a time, starting with entity linking and relation extraction and progressively adding entity discovery, knowledge validation, and co-reference resolution. This approach allowed us to examine the contribution of each component to overall performance.

TABLE 4. Performance of our framework with step-wise addition of submodels.

Model configuration	Precision	Recall	F1-score
EL+RE	0.65	0.57	0.61
ELD + RE	0.65	0.59	0.62
ELD+RE + KV	**0.71**	0.53	0.61
ELD + CR+RE+KV	0.70	**0.63**	**0.67**

TABLE 5. Evaluation results for entity linking.

Model	Precision	Recall	F1 score
KO	0.93	0.91	0.92
EN	0.88	0.98	0.93

The experimental results, summarized in Table 4, indicate a progressive improvement in the performance of the framework with the addition of each sub-model. Incorporation of entity discovery resulted in a 2% increase in recall, whereas the co-reference resolution contributed to a 10% improvement. The inclusion of knowledge validation enhanced accuracy by 6%. The highest performance was achieved when all the sub-models were integrated, yielding a precision of 0.70, recall of 0.63, and F1-score of 0.67. These findings underscore the significant role played by each sub-model within the KENet framework, collectively enhancing the system's efficiency in knowledge extraction. The comprehensive integration of all sub-models facilitates the extraction of factual knowledge of superior quality compared with employing only entity linking and relation extraction.

C. EVALUATION OF EACH MODEL

This section present a detailed assessment of the individual models constituting the KENet framework, designed for extracting high-quality factual knowledge from textual data. We focus particularly on evaluating the performance of two critical components, the entity linking model and the co-reference resolution model, exploring their respective strengths and challenges.

1) ENTITY LINKING

Entity linking involves identifying named entities within a text and associating them with their equivalent entries in a knowledge base. This subsection reviews the performance of the entity-linking model trained using the crowdsourcing data outlined in Section II-B. The evaluation was conducted using a gold standard dataset, providing insights into the model's efficacy and areas for improvement.

Table 5 shows that our model exhibited impressive performance, achieving an F1-score of 92%. This outcome suggests that the model is highly effective at identifying and linking named entities within a text. The evaluation results

TABLE 6. Evaluation results for co-reference resolution.

	MUC			B^3			CEAF			
	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score	Avg. F1 score
KO	0.88	0.65	0.75	0.86	0.60	0.71	0.81	0.64	0.72	0.72
EN	0.83	0.61	0.71	0.81	0.55	0.66	0.77	0.55	0.65	0.67

also indirectly reflected the superior quality of the datasets employed in this study. Furthermore, when comparing our model's performance to that of an English entity linker, which registered a slightly higher F1-score of 93%, our model demonstrated competitive effectiveness.

2) CO-REFERENCE RESOLUTION

Co-reference resolution involves determining all mentions within a text that refer to the same entity in the real world. This subsection details the performance evaluation of our co-reference resolution model and highlight its capabilities and areas of achievement.

Table 6 shows the performance of the co-reference resolution model trained using crowdsourced data. The evaluation highlights that the average F1-score for the Korean model slightly surpasses that of the English model (0.72 vs. 0.67). It should noted that the proposed model incorporates a simplified mention detection system. Overall, the findings indicate that the co-reference resolution model effectively identifies and resolves mentions by referring to the same entities within texts, a critical component for precise knowledge extraction.

D. ANALYSIS OF THE CROWDSOURCING DATA

To evaluate the impact of crowdsourced data, our study engaged in experiments focused on two main tasks: co-reference resolution and relation extraction. For co-reference resolution, we contrasted the performance between models trained on expertly curated data and those trained on crowdsourced data. The expert data used in this comparison shared the same source text as the half-crowd dataset.

The results documented in Table 7 reveal that the models trained with the half-crowd dataset exhibited superior performance, demonstrating an average F1-score improvement of approximately 9% compared to those trained with expert data. This outcome underscores the higher quality of crowdsourced data compared with their expertly curated counterparts. Additionally, our findings showed that combining both datasets did not yield significant performance enhancements beyond those achieved with the half-crowd dataset alone. This observation implies that the existing co-reference dataset, encompassing approximately 1300 documents, is sufficient for model training, suggesting that further enhancements in performance would necessitate refined methodologies.

For relation extraction, we trained models using both crowdsourced and DS data. Given that DS data often include significant noise, achieving a high performance with models can be challenging. To enhance model performance,

TABLE 7. Performance(F1-score) of co-reference resolution model.

Dataset	MUC	B^3	CEAF	Avg.
EXPERT	0.66	0.60	0.61	0.62
1/8 CROWD	0.71	0.63	0.60	0.65
1/4 CROWD	0.73	0.66	0.64	0.68
1/2 CROWD	0.76	0.70	0.67	0.71
CROWD	0.77	0.71	0.67	0.72

TABLE 8. Relation extraction performance: Wikipedia-domain; 49 relations."DS" is added to the name of an architecture model to indicate that the model was trained only on DS data, and "Crowd" indicates that the model was trained using crowdsourced data in addition to DS data.

Architecture	Precision	Recall	F1 score
PCNN-DS	0.62	0.55	0.58
PCNN-Crowd	0.77	0.58	0.66
GAN-DS	0.70	0.64	0.67
GAN-Crowd	0.76	0.70	0.73
RL-DS	0.80	0.67	0.73
RL-Crowd	0.82	0.73	0.77
BERT-DS	0.88	0.76	0.82
BERT-Crowd	0.89	0.78	0.83

various studies have incorporated crowdsourced data with DS data [21], [37]. We evaluated the performance of four models: Piecewise Convolutional Neural Networks (PCNN) [38], Generative Adversarial Networks (GAN) [39], Reinforcement Learning (RL) [40], and BERT [12] using a combination of DS and crowdsourced data. Our objective was to identify the conditions under which these models exhibit high performance when solely utilizing DS data and to ascertain the necessary quantity of crowdsourced data in relation to the volume of DS data.

To assess the performance of the models, we focused on 49 relations identified through crowdsourcing annotations, each with no more than 50% noise. According to the results shown in Table 8, the BERT model demonstrated resilience by tolerating up to 50% noise in the DS data provided that the noise level for each relation does not exceed this threshold. This led us to explore two key questions: what outcomes would training on the top 49 noisy relations yield? What results can be expected from utilizing a broader range of relations, irrespective of their noise levels? To address the first query, we evaluated the performance based on the top 49 noisy relations using the BERT model, given its superior performance in preliminary tests. The findings shown in Table 9 reveal a significant performance decline of approximately 40%. This indicates that training models on relations with less than 50% DS noise could potentially mirror the performance achievable with crowdsourced data,

TABLE 9. Performance of relation extraction: Wikipedia-domain; top 49 noisy relations.

Architecture	Precision	Recall	F1 score
BERT-DS	0.57	0.28	0.37
BERT-Crowd	0.56	0.30	0.39

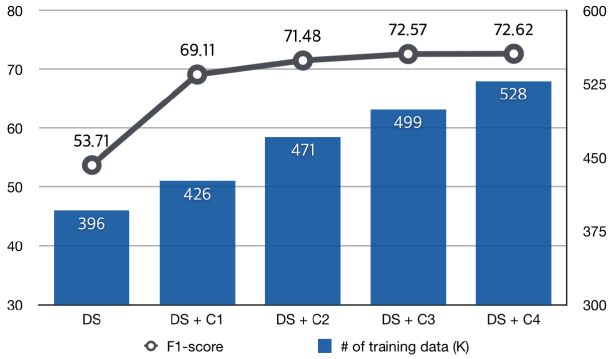

FIGURE 9. Performance comparison of relation extraction between Wikipedia and News Domain for 76 Relations. Cn denotes crowdsourced data chunk.

thereby highlighting the importance of selecting relations with manageable noise levels for training.

To address the second question, we evaluated the model's performance when incorporating 76 relations, of which 27 had noise ratios that exceeded 50%. The results depicted in Fig. 9 show that training the model exclusively with DS data yielded an F1-score of approximately 54%. However, this score improved significantly to 73% when the crowdsourced data were integrated into the training set. With 400,000 DS data samples, a noticeable performance enhancement was observed with the addition of up to approximately 100,000 crowdsourced data samples. This suggests that the optimal amount of crowdsourced data needed is approximately 25% of the DS data volume, indicating the substantial impact of incorporating high-quality crowdsourced data to complement the DS data for relation extraction tasks.

E. EFFECT OF THE NA FILTER

In our experiment, we explored the impact of implementing an NA (Not Applicable) filter on the performance of knowledge validation by comparing scenarios with and without the filter application. The primary function of the NA filter is to detect and discard instances in which no relationship exists between two entities within an extracted triple. To assess the efficacy of the NA filter, we annotated a dataset from a knowledge-validation standpoint, where the dataset comprised randomly sampled triples extracted by a relation extractor and subsequently tagged by experts. A triple was deemed true if the annotator could verify its accuracy through evidence on Wikipedia, and it was marked as an error in the absence of such evidence. Of the 1,759 analyzed triples, 290 were tagged as true and 1,469 were identified as errors.

TABLE 10. Performance of knowledge validation models with and without NA filtering.

Model configuration	Precision	Recall	F1-score	ERR(%)
EF	0.81	0.33	0.46	32.61
EF+NA	0.88	0.74	0.81	74.40
RF	**0.91**	0.58	0.71	57.86
RF+NA	**0.91**	0.83	0.87	83.19
EF+RF	0.86	0.74	0.80	74.20
EF+RF+NA	0.88	**0.90**	**0.89**	**90.27**

We compared several knowledge validation models: EF [36], which is a model for learning the probability of a positive correlation between triples based on knowledge based embedding, and RF [19], which is a negative rule-mining model. The performance of these models was quantified using the Error Reduction Ratio (ERR). The results, displayed in Table 10, indicate a significant improvement in knowledge validation performance by up to approximately 90% upon application of the NA filter. Furthermore, the most substantial performance enhancement was observed when the NA filter was employed alongside all tested combinations of knowledge validation models. These findings highlight the capability of tje NA filter to effectively screen out non relational instances between two entities, thereby augmenting the overall accuracy of knowledge validation.

F. RELATION EXTRACTION COMPARISON WITH THE ENGLISH LANGUAGE

Research on knowledge extraction, particularly relation extraction, has been predominantly conducted in English, a high-resource language. Numerous datasets have been published to support this research, as listed in Table 11. The NYT-10 dataset, for example, is derived from the New York Times corpus and Freebase using named entity recognition and surface matching for entity linking. However, it faces a long-tail issue, with fewer than 100 instances for 30 out of 53 relations. The SemEval-10 dataset consists of expertly tagged data from a web-base corpus, making it costly and relatively small. In addition, it diverges somewhat from the task of extracting factoid triples from a knowledge base. The FewRel dataset collects distant-supervision data from Wikipedia and Wikidata, with noise filtered through crowdsourcing.

Our evaluation results were compared with those of existing studies. The Matching the Blanks [12] approach achieved a 94.27% accuracy on the 10-way-5-shot FewRel dataset. Given the similarity of the FewRel dataset to ours, we trained our model on FewRel in a supervised manner, achieving an F1-score of 89.7%, demonstrating its applicability beyond the Korean language.

IV. BACKGROUND AND RELATED WORK

This section provides an overview of the critical research areas relevant to our study, namely entity linking, co-reference resolution, relation extraction, and crowdsourcing.

TABLE 11. Comparison of relation extraction data set and evaluation result.

Dataset	Train	Test	Relation	Model	Precision	Recall	F1-score
NYT-10 [41]	566,190	170,866	53	Att-CapNet [42]	30.8	63.7	41.6
				BLSTM+C2SA-dot [43]	-	-	40.1
				MultiTask [44]	55.0	40.0	46.3
SemEval-10 [45]	8,000	2,717	9	Matching the Blanks [12]	-	-	89.5
FewRel [37]	54,000	1,600	80	SN-L+CV (unsupervised) [46]	48.9	74.0	52.6
				BERT-Ours (supervised)	89.7	89.7	89.7
Ours	141,858	3,190	76	BERT-Ours	74.5	74.2	74.4

A. ENTITY LINKING

Entity linking is the task of connecting an entity e to a mention m in a natural language text, where e belongs to entity set E in a given knowledge base K, and m belongs to a mention set M. The objective of entity linking is to link a mention to its corresponding entity in a knowledge base. For instance, in the sentence "`Steve Jobs is Apple's founder`," the mention set M is represented by [`Steve Jobs`, `Apple`]. The entity linking task is to link the mention "`Apple`" to "`Apple_(company)`" rather than to"`apple_(fruit)`". Historically, statistics-based machine learning methods were widely used in entity linking, such as calculating the similarity between words close to a mention or calculating the word distribution of a document containing the candidate entity [47]. Recent entity linking models incorporate advanced machine learning techniques, such as word embedding, context word embedding, and entity embedding, implemented using the output of the entity description [35]. Other approaches include modeling a mention set by assuming latent relations between entities [11] and modeling a mix of jointly learning mention detection and entity linking [48].

B. CO-REFERENCE RESOLUTION

Co-reference resolution is a critical task in NLP intended for grouping expressions that refer to the same entity. In many cases, a word mentioned earlier in a text (an antecedent) is referred to in different forms later on (e.g., a pronoun, demonstrative determiner, or abbreviation). To extract knowledge from text, it is necessary to determine whether such expressions refer to the same entity. Co-reference resolution complements entity linking and is essential for extracting all the relevant information from a text. For example, consider the sentences "Gordon Moore, who majored in electrical engineering. He was born in the United States of America." Without co-reference resolution, only the triple "*Field*(`Gordon_Moore`, `Electrical_Engineering`)" can be extracted, as entity linking cannot extract the triple " *birthPlace*(`Gordon_Moore`, `United_States`)" without identifying which entity the pronoun "he" refers to.

Recent research has focused on deep-learning-based models for co-reference resolution, using state-of-the-art models achieving high F1-scores. For instance, Lee et al. [16] achieved an F1-score as high as 73% with an English co-reference resolution model. The model comprises two parts. First, it identifies representations of all possible mentions in a document using Bi-LSTM and computes a mention score that indicates the likelihood of a candidate mention being an actual mention. It then computes an antecedent score, which indicates the anaphoric relation between two spans, and completes the co-reference resolution task by combining the mention and antecedent scores. The problem of co-reference occurring with a word in between (singular or plural) was addressed by higher-order span representation using an attention mechanism. To reduce computational load, they employed a coarse-to-fine method.

C. RELATION EXTRACTION

Relation extraction is the task of identifying the relationship between two entities in a sentence. For instance, a relation extraction system could extract "`Founder`(Facebook, Mark Zuckerberg)" from the sentence "Mark Zuckerberg is the founder of Facebook." Traditional approaches have largely depended on human intervention through the creation of handcrafted rules and the manual tagging of training data for pre-specified relations.

DS learning has been employed for relation extraction in numerous studies since its introduction by Mintz et al. [24]. Many studies have used the DS approach to reduce the cost of creating handcrafted training data. However, a statistical analysis of the DS data from Wikipedia-DBpedia collected in this study revealed that it contained 49% noise, for example, "*Founder*(`Steve Jobs`, `Apple`)" from the sentence "Steve Jobs argued with Wozniak, the co-founder of Apple."

Deep neural networks (DNNs) have become the primary focus of relation extraction research, and various DNN based models have been proposed. Among these, convolutional neural networks (CNNs) are most commonly used. CNN-based relation extraction offers the advantages of enabling model learning without human feature selection and faster processing than other DNN architectures such as LSTM and GRU. Consequently, learning in CNN-based relation extraction models involves providing pre trained word embeddings as input vectors and identifying the most informative n-gram words within the model. Many studies [38], [49], [50] have

incorporated various features into the input vector to enhance the performance. For example, the piecewise-CNN (PCNN) model [38] adds position embeddings to calculate the relative distance between two entities for each word in a sentence, and extends the max pooling layer to the piecewise max pooling layer. To date, most studies [51], [52], [53], [54], [55] have focused on addressing noisy data issues using multi-instance and multi-labeling (MIML) approaches.

Recent research on relation extraction has concentrated on generative adversarial networks (GANs) or reinforcement learning with complex architectures. Wu et al. [39] proposed a model for predicting bag representation from a bag of sentences, which are sets of sentences containing the same target entities, wherein the relation extractor serves as an agent. To develop models that directly address the noisy data problem in DS learning, Feng et al. [56] and Qin et al. [40] introduced architectures comprising a sentence selector to eliminate noise-labeled sentences from the training text and a relation extractor. Through reinforcement learning, the sentence selector, which acts as an agent, is trained to maximize the reward output from the relation extractor, thereby improving its performance by filtering noisy sentences from the training data. Notably, some studies [21], [57], [58], [59] have attempted to enhance relation classifier performance by collecting crowdsource-based training data for relation extraction and using them along with the training data for DS learning.

D. CROWDSOURCING

Crowdsourcing-based machine-learning research has gained considerable attention [60], [61]. Machine learning models for NLP tasks using data generated by the general public with common sense-level knowledge have been demonstrated to be equivalent or even superior to models that use expert generated data. This has motivated many researchers to generate massive training data for various NLP tasks using crowdsourcing approaches and to use them for machine learning models.

Bontcheva et al. [62] and Demartini et al. [23] conducted crowdsourcing-based research on entity linking. They collected webpages to extract entity mentions and presented a candidate set for each entity mention to the crowd workers. Workers were asked to select one entity from each candidate. Bontcheva et al. [62] proposed a method for providing an abstract (definition) for each entity candidate. In a study on crowdsourced co-reference resolution, Chamberlain et al. [22] conducted and annotation in two steps: presentation of a coreferent mention within a document (e.g., pronoun, demonstrative determiner, or abbreviation) and antecedent annotation by one crowd worker; this was followed by quality control conducted by two other crowd workers to verify whether the correct antecedent was annotated. Liu et al. [21] improved the classification performance of a relation extraction model using data collected according to the GI protocol, which is a crowdsourcing scheme that they designed. The GI protocol consisted of three phases: tutorial, weed-out, and annotation. In the tutorial phase, workers participate in the annotation training and receive immediate feedback through the same user interface. During the weed-out phase, workers are assessed using simple questions and are disqualified if they fail to answer them correctly. In the annotation phase, workers are provided with batches of gold question sets created by an expert, with further participation granted only to those whose annotations demonstrate high agreement ($\geq 80\%$) with the expert-provided answers. Data generated by a worker trained according to the GI protocol, without duplicate data allocations, can significantly improve the classification performance of a relation extraction model relative to the data generated by multiple workers through collaborative effort and majority approval.

V. CONCLUSION

This study introduced a cost-effective framework for knowledge extraction in low-resource environments, exemplified by crowdsourcing data and a Korean knowledge-extraction task. The framework facilitated the evaluation of all knowledge extraction tasks using consistent source data, yielding improved understanding and results. Our experiments determined that only a quarter of the crowdsourcing data were necessary to attain high performance. Furthermore, the implementation of an NA filter in relation extraction significantly reduced errors and enhanced noise detection by 16%.

We also examine the versatility of the framework across different domains and languages, demonstrating that our unified corpus approach and cost-reduction method are universally applicable. Ontology mapping has been suggested as a strategy for developing training data aligned with dynamically changing target knowledge bases. In addition, we proposed employing few-shot learning for domain adaptation when the corpus type shifted within the same language.

Our findings underscore the potential of the proposed framework and dataset to advance future research on knowledge extraction in low-resource settings. The adaptability and cost efficiency of the framework are crucial for researchers and practitioners aiming to develop effective knowledge extraction systems for underrepresented languages.

REFERENCES

[1] S. R. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A nucleus for a web of open data," in *Proc. Int. Semantic Web Conf.*, vol. 4825, 2007, pp. 722–735.

[2] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A core of semantic knowledge," in *Proc. 16th Int. Conf. World Wide Web*, May 2007, pp. 697–706.

[3] D. Vrandečić and M. Krötzsch, "Wikidata: A free collaborative knowledge base," *Commun. ACM*, vol. 57, no. 10, pp. 78–85, Sep. 2014.

[4] I. Augenstein, S. Padó, and S. Rudolph, "LODifier: Generating linked data from unstructured text," in *Proc. Extended Semantic Web Conf.* Berlin, Germany: Springer, 2012, pp. 210–224.

[5] A. Gangemi, F. Draicchio, V. Presutti, A. G. Nuzzolese, and D. Reforgiato, "A machine reader for the semantic web," in *Proc. 12th Int. Semantic Web Conf.*, vol. 1035, 2013, pp. 149–152.

[6] P. Vossen, R. Agerri, I. Aldabe, A. Cybulska, M. van Erp, A. Fokkens, E. Laparra, A.-L. Minard, A. P. Aprosio, G. Rigau, M. Rospocher, and R. Segers, "NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news," *Knowl.-Based Syst.*, vol. 110, pp. 60–85, Oct. 2016.

[7] F. Corcoglioniti, M. Rospocher, and A. P. Aprosio, "Frame-based ontology population with PIKES," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3261–3275, Dec. 2016.

[8] L. Song, A. Wang, X. Pan, H. Zhang, D. Yu, L. Jin, H. Mi, J. Su, Y. Zhang, and D. Yu, "OpenFact: Factuality enhanced open knowledge extraction," *Trans. Assoc. Comput. Linguistics*, vol. 11, pp. 686–702, Jun. 2023.

[9] B. Xu, Y. Xu, J. Liang, C. Xie, B. Liang, W. Cui, and Y. Xiao, "CN-DBpedia: A never-ending Chinese knowledge extraction system," in *Proc. Int. Conf. Ind., Eng. Other Appl. Appl. Intell. Syst.*, 2017, pp. 428–438.

[10] N. Ma, D. Wang, H. Bao, L. He, and S. Zheng, "KEPL: Knowledge enhanced prompt learning for Chinese hypernym-hyponym extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Singapore: Association for Computational Linguistics (ACL), Dec. 2023, pp. 5858–5867.

[11] P. Le and I. Titov, "Improving entity linking by modeling latent relations between mentions," 2018, *arXiv:1804.10637*.

[12] L. B. Soares, N. FitzGerald, J. Ling, and T. Kwiatkowski, "Matching the blanks: Distributional similarity for relation learning," 2019, *arXiv:1906.03158*.

[13] N. Kassner, F. Petroni, M. Plekhanov, S. Riedel, and N. Cancedda, "EDIN: An end-to-end benchmark and pipeline for unknown entity discovery and indexing," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Abu Dhabi, United Arab Emirates: Association for Computational Linguistics (ACL), Dec. 2022, pp. 8659–8673.

[14] K. Clark and C. D. Manning, "Improving coreference resolution by learning entity-level distributed representations," 2016, *arXiv:1606.01323*.

[15] K. Lee, L. He, M. Lewis, and L. Zettlemoyer, "End-to-end neural coreference resolution," 2017, *arXiv:1707.07045*.

[16] K. Lee, L. He, and L. Zettlemoyer, "Higher-order coreference resolution with coarse-to-fine inference," 2018, *arXiv:1804.05392*.

[17] M. Joshi, O. Levy, D. S. Weld, and L. Zettlemoyer, "BERT for coreference resolution: Baselines and analysis," 2019, *arXiv:1908.09091*.

[18] J. Lehmann, D. Gerber, M. Morsey, and A.-C. N. Ngomo, "DeFacto-deep fact validation," in *Proc. Int. Semantic Web Conf.* Berlin, Germany: Springer, 2012, pp. 312–327.

[19] S. Ortona, V. V. Meduri, and P. Papotti, "Robust discovery of positive and negative rules in knowledge bases," in *Proc. IEEE 34th Int. Conf. Data Eng. (ICDE)*, Apr. 2018, pp. 1168–1179.

[20] M. Eberts and A. Ulges, "Span-based joint entity and relation extraction with transformer pre-training," 2019, *arXiv:1909.07755*.

[21] A. Liu, S. Soderland, J. Bragg, C. H. Lin, X. Ling, and D. S. Weld, "Effective crowd annotation for relation extraction," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 897–906.

[22] J. Chamberlain, M. Poesio, and U. Kruschwitz, "Phrase Detectives Corpus 1.0 crowdsourced anaphoric coreference," in *Proc. LREC*, 2016, pp. 2039–2046.

[23] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux, "ZenCrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking," in *Proc. 21st Int. Conf. World Wide Web*, Apr. 2012, pp. 469–478.

[24] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proc. Joint Conf. 47th Annu. Meeting ACL 4th Int. Joint Conf. Natural Lang. Process. (AFNLP)*, vol. 2. Singapore: Association for Computational Linguistics (ACL), Aug. 2009, pp. 1003–1011.

[25] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith, "Annotation artifacts in natural language inference data," 2018, *arXiv:1803.02324*.

[26] S. Nisioi, S. Štajner, S. P. Ponzetto, and L. P. Dinu, "Exploring neural text simplification models," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2017, pp. 85–91.

[27] S. Nam, E.-K. Kim, J. Kim, Y. Jung, K. Han, and K.-S. Choi, "A Korean knowledge extraction system for enriching a KBox," in *Proc. 27th Int. Conf. Comput. Linguistics, Syst. Demonstrations*, 2018, pp. 20–24.

[28] C. Park, C. Lee, J. Ryu, and H. Kim, "Contextualized embedding and character embedding-based pointer network for Korean coreference resolution," in *Proc. 30th Annu. Conf. Human Cognit. Lang. Technol.*, 2018, pp. 239–242.

[29] C.-E. Park, K.-H. Choi, and C. Lee, "Korean coreference resolution using the multi-pass sieve," *J. KIISE*, vol. 41, no. 11, pp. 992–1005, Nov. 2014.

[30] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang, "CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes," in *Proc. Joint Conf. EMNLP CoNLL-Shared Task*. Jeju Island, South Korea: ACL, Jul. 2012, pp. 1–40.

[31] C. Park and C. Lee, "Coreference resolution for Korean pronouns using pointer networks," *J. KIISE*, vol. 44, no. 5, pp. 496–502, May 2017.

[32] M. Choi, C. Lee, J. Wang, and M.-G. Jang, "Reference resolution for ontology population," in *Proc. 19th Annu. Conf. Human Cognit. Lang. Technol.*, 2007, pp. 140–144.

[33] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum, "Robust disambiguation of named entities in text," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Edinburgh, U.K.: ACL, 2011, pp. 782–792.

[34] P. Vossen, G. Rigau, L. Serafini, P. Stouten, F. Irving, and W. R. Van Hage, "NewsReader: Recording history from daily news streams," in *Proc. LREC*, 2014, pp. 2000–2007.

[35] N. Gupta, S. Singh, and D. Roth, "Entity linking via joint encoding of types, descriptions, and context," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2681–2690.

[36] J. Kim, S. Nam, and K.-S. Choi, "Universal schemas using shortest dependency paths for free word order languages," in *Proc. Int. Semantic Web Conf. (P&D/Industry/BlueSky)*, 2018.

[37] X. Han, H. Zhu, P. Yu, Z. Wang, Y. Yao, Z. Liu, and M. Sun, "FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation," 2018, *arXiv:1810.10147*.

[38] D. Zeng, K. Liu, Y. Chen, and J. Zhao, "Distant supervision for relation extraction via piecewise convolutional neural networks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1753–1762.

[39] Y. Wu, D. Bamman, and S. Russell, "Adversarial training for relation extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1778–1783.

[40] P. Qin, W. Xu, and W. Yang Wang, "Robust distant supervision relation extraction via deep reinforcement learning," 2018, *arXiv:1805.09927*.

[41] S. Riedel, L. Yao, and A. McCallum, "Modeling relations and their mentions without labeled text," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases (ECML/PKDD)*. Barcelona, Spain: Springer, Sep. 2010, pp. 148–163.

[42] X. Zhang, P. Li, W. Jia, and H. Zhao, "Multi-labeled relation extraction with attentive capsule network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 7484–7491.

[43] Y. Yuan, L. Liu, S. Tang, Z. Zhang, Y. Zhuang, S. Pu, F. Wu, and X. Ren, "Cross-relation cross-bag attention for distantly-supervised relation extraction," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 419–426.

[44] I. Beltagy, K. Lo, and W. Ammar, "Combining distant and direct supervision for neural relation extraction," 2018, *arXiv:1810.12956*.

[45] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó. Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz, "SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals," 2019, *arXiv:1911.10422*.

[46] R. Wu, Y. Yao, X. Han, R. Xie, Z. Liu, F. Lin, L. Lin, and M. Sun, "Open relation extraction: Relational knowledge transfer from supervised data to unsupervised data," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 219–228.

[47] R. Bunescu and M. Paşca, "Using encyclopedic knowledge for named entity disambiguation," in *Proc. 11th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2006, pp. 9–16.

[48] N. Kolitsas, O.-E. Ganea, and T. Hofmann, "End-to-end neural entity linking," in *Proc. 22nd Conf. Comput. Natural Lang. Learn.*, 2018, pp. 519–529.

[49] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*.

[50] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2016, pp. 207–212.

[51] X. Han and L. Sun, "Global distant supervision for relation extraction," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 2950–2956.

[52] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld, "Knowledge-based weak supervision for information extraction of overlapping relations," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Human Lang. Technol.*, 2011, pp. 541–550.

[53] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, "Neural relation extraction with selective attention over instances," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2016, pp. 2124–2133.
[54] M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning, "Multi-instance multi-label learning for relation extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.* Jeju Island, South Korea: ACL, 2012, pp. 455–465.
[55] X. Jiang, Q. Wang, P. Li, and B. Wang, "Relation extraction with multi-instance multi-label convolutional neural networks," in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers (COLING)*, 2016, pp. 1471–1480.
[56] J. Feng, M. Huang, L. Zhao, Y. Yang, and X. Zhu, "Reinforcement learning for relation classification from noisy data," in *Proc. AAAI*, 2018, pp. 5779–5786.
[57] G. Angeli, J. Tibshirani, J. Wu, and C. D. Manning, "Combining distant and partial supervision for relation extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1556–1567.
[58] M. Pershina, B. Min, W. Xu, and R. Grishman, "Infusion of labeled data into distant supervision for relation extraction," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2014, pp. 732–738.
[59] C. Zhang, F. Niu, C. Ré, and J. Shavlik, "Big data versus the crowd: Looking for relationships in all the right places," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics, Long Papers*, vol. 1, 2012, pp. 825–834.
[60] O. Inel, K. Khamkham, T. Cristea, A. Dumitrache, A. Rutjes, J. van der Ploeg, L. Romaszko, L. Aroyo, and R.-J. Sips, "CrowdTruth: Machine–human computation framework for harnessing disagreement in gathering annotated data," in *Proc. Int. Semantic Web Conf.*, Riva del Garda, Italy. Cham, Switzerland: Springer, 2014, pp. 486–504.
[61] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—But is it good? Evaluating non-expert annotations for natural language tasks," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2008, pp. 254–263.
[62] K. Bontcheva, L. Derczynski, and I. Roberts, "Crowdsourcing named entity recognition and entity linking corpora," in *Handbook of Linguistic Annotation*. Dordrecht, The Netherlands: Springer, 2017, pp. 875–892.

SANGHA NAM received the B.S. and M.S. degrees in computer science from the University of Kyonggi, Republic of Korea, and the Ph.D. degree from the School of Computing, KAIST, Daejeon, Republic of Korea, in 2021.

He is currently leading the Dialog Model Team with the NC Research, NCSoft, where he is involved in research across several areas of artificial intelligence, including large language models, retrieval-augmented generation (RAG), persona-based dialogue, empathetic dialogue, and the development and application of knowledge graphs.

EUN-KYUNG KIM received the B.S. and M.S. degrees in computer science from Sookmyung Women's University, Republic of Korea, and the Ph.D. degree from the School of Computing, KAIST, Daejeon, Republic of Korea, in 2016.

She is currently an Assistant Professor with the Division of AI Software, Daejeon University. Her research interests include machine learning-based artificial intelligence, such as deep learning, artificial intelligence-based intelligent language processing, and distributed systems for big data analysis.