applied sciences

MDPI

*Article*

# A System for Interviewing and Collecting Statements Based on Intent Classification and Named Entity Recognition Using Augmentation

**Junho Shin [1], Eunkyung Jo [2], Yeohoon Yoon [3] and Jaehee Jung [1,*]**

[1] Department of Information and Communication Engineering, Myongji University, Yongin 17058, Republic of Korea; entus0123@naver.com
[2] College of Police and Criminal Justice, Dongguk University, Seoul 04620, Republic of Korea; ekjo@dongguk.edu
[3] Forensic Science Division, Supreme Prosecutor's Office, Seoul 06590, Republic of Korea; yhyoon@spo.go.kr
* Correspondence: jhjung@mju.ac.kr

**Abstract:** In cases of child sexual abuse, interviewing and obtaining trustworthy statements from victims and witnesses is essential because their statements are the only evidence. It is crucial to ascertain objectively the credibility of the victim's statements, which may vary based on the nature of the questions posed by the forensic interviewer. Therefore, interview skills that eliminate subjective opinions require a high level of training for forensic interviewers. To reduce high-risk subjective interviews, objectively analyzing statements is essential. Understanding the victim's intent and named entity recognition (NER) in the statements is necessary to give the victim open-ended questions and memory recall. Therefore, the system provides an intent classification and NER method that follows the National Institute of Child Health and Human Development Investigative Interview Protocol, which outlines the collection of objective statements. Large language models such as BERT and KoBERT, along with data augmentation techniques, were proposed using a restricted training dataset of limited size to achieve effective intent classification and NER performance. Additionally, a system that can collect objective statements with the proposed model was developed and it was confirmed that it could assist statement analysts. The verification results showed that the model achieved average F1-scores of 95.5% and 97.8% for intent classification and NER, respectively, which improved the results of the limited data by 3.4% and 3.7%, respectively.

**Keywords:** child sexual abuse interview; intent classification; named entity recognition; data augmentation; NICHD

check for updates

**Citation:** Shin, J.; Jo, E.; Yoon, Y.; Jung, J. A System for Interviewing and Collecting Statements Based on Intent Classification and Named Entity Recognition Using Augmentation. *Appl. Sci.* **2023**, *13*, 11545. https://doi.org/10.3390/app132011545

Academic Editor: Douglas O'Shaughnessy

Received: 15 August 2023
Revised: 14 October 2023
Accepted: 18 October 2023
Published: 21 October 2023

**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

In most child sexual abuse cases, the victims' statements are the only evidence for a court ruling. Thus, it is crucial to obtain trustworthy victims' statements. In Korea, the Supreme Prosecutor's Office and the National Police Agency are adopting the interview guidelines developed by the National Institute of Child Health and Development (NICHD), which is world-renowned for child investigative interviewing. The NICHD protocol is a set of structured forensic interview guidelines for children based on input from a wide range of practitioners, including psychologists, child interview experts, police officers, and lawyers [1]. It is designed to maximize the amount of information retrieved from interviewees' free recall memory. It also considers the age-specific developmental characteristics of children, such as memory recall, communicative skills, and suggestibility. In addition, this protocol recommends open-ended prompts and minimizes the effects of investigators' individual abilities or subjective interpretations of a case, in terms of the interview process and quality [2,3].

However, even if investigators are trained to follow the NICHD protocol, most field studies tend to rely heavily on closed-ended questions influenced by subjective opinions. This means that, after listening to statements, investigators ask leading questions based on their suppositions or ask questions about intent based on psychological factor. These problems have been described in many studies in the literature [4,5] and emphasize the need for highly trained investigators to conduct interviews with open-ended questions [6,7].

Furthermore, investigators often conduct repeated interviews with children when they believe that insufficient information was obtained in the initial interview. In cases where victim statements are considered critical evidence (e.g., child sexual abuse and sexual harassment), investigators are encouraged to minimize the number of interviews. Repeated interviews might cause secondary victimization as the interviewees are forced to recall negative memories repeatedly. Research on child sexual abuse victims indicates that children have difficulty recounting the full details of encounters, owing to trauma, excessive self-criticism, and guilt, because they often blame themselves for what happened. Thus, a face-to-face interview with a stranger could burden child victims.

Therefore, obtaining reliable statements from child sexual abuse victims requires a system that (1) follows the principles of the NICHD interview protocol, (2) minimizes any physical or psychological adverse effects, and (3) provides a child-friendly interview environment. Mihas et al. [8] said that artificial intelligence and cognitive interviewing could be essential in eliciting quality evidentiary statements from victims and witnesses. This paper proposes a system that has the function of assisting in obtaining trustworthy statements based on NICHD. This requires accurate intent classification and NER from statements. This study investigated and validated methods for intent classification and NER to establish a foundation for a system that adheres to the NICHD interview rules for child sexual assault. This study demonstrates that colloquial language can be recognized in a particular environment by augmenting limited data and applying intent classification and NER. Our primary contributions are summarized as follows.

- The model is proposed for intent classification and NER for objective interviews using augmentation due to a limited statement dataset.
- The system produces objective questioning in accordance with both the proposed model and the NICHD protocol, which is able to elicit quality statements based on proven interview guidelines.
- The approach and results of this study show that it can be used for objective intent classification and NER even in special environments such as child sexual abuse cases, indicating that it can be used in various domains through transfer learning.

## 2. Materials and Methods

Computer systems can communicate and interact with humans. From the traditional chatbot first proposed in 1966 by ELIZA [9] to help patients with psychotherapy, ChatGPT AI systems utilizing data-driven artificial intelligence (AI) technology are becoming more widespread in our lives. The fields where chatbots are most often utilized are education [10], medical [11], and healthcare [12], and many of them have been developed to provide guidance or answer questions.

Among the recent studies analyzing chatbots, Amon et al. [13] analyzed the types of text-based chatbots, and Li et al. [14] analyzed the types of conversations and analyzed the cases where conversations in chatbots do not go well. Amon et al. [13] studied and analyzed 83 papers that focused on how to interact with text-based chatbots. It analyzed the domains and types of recent chatbot systems and the satisfaction of expectations from chatbots. Based on the result of this paper, the most common research trend of chatbots is “task-oriented”, which needs to perform something such as giving some information. “Conversation-oriented” occupies about 25% of the total collected papers and only a few studies focus on “task-and-conversation-oriented" aspects. Regarding chatbot domains, commonly and widely researched domains in chatbots are healthcare, education, and customer services. On the other hand, only three studies were conducted on interviews.

Li et al. [14] analyzed why chatbots do not communicate well through conversation analysis. They called this conversational "non-progress" (NP). They concluded that the reasons for NP conversations with chatbots are the inability to accurately recognize the user's conversational intentions, the inability to engage in new conversations by staying on the previous topic, and the termination of conversations with chatbots due to difficulties in identifying intentions. Based on the above literature review, we know that using chatbots for interviews is a challenging task and identifying the user's intentions is an important part of successful interviewing. Therefore, a problem that can arise in chatbot forensic interviews is determining the intent of the speaker or respondent.

The second problem with chatbot systems for forensic interviews is the reliability of the conversations obtained by the chatbot. The interaction with the computer was expected to be weaker than the emotional and human interaction. Therefore, the interview response to the chat was seen to be trustworthy. Since they were talking to a machine, researchers believe that their conversations with chatbots were not trustworthy and thought their responses were disingenuous. However, Sidaoui [15] found primary qualitative data generated via chatbot interviews to be sentimentally meaningful. The impact of emotional disclosure was consistent, regardless of whether it was with a chatbot or person [16,17]. Moreover, the study found a high willingness to use AI-based healthcare chatbots [12]. The reliability of chatbots in forensic interviews has also been studied psychologically. In forensic investigations, witness or victim statements are important. The use of artificial intelligence and a cognitive interview (AICI) proved to be much more effective than other tools, such as free recall or questionnaires [8]. From this research, it was experimentally shown that the response from the chatbot achieved the same effect as the face-to-face interview.

A third possible problem with chatbot systems for interviews is data bias. Ji et al. [18] mentioned a hallucination arising from the divergence between the source and the reference. This means that artifact heuristic data can be used as the training set. A chatbot that is trained with too much one-sided content has the disadvantage of providing information or conversing with biased content. However, in the proposed system, the questions can be asked following the established guidelines for forensic interviewing of statements given the correct analysis of the intent and NER.

To summarize, the problems that may arise in interviews for statement elicitation using chatbots are difficulty in identifying intentions and data bias. To solve these problems, we recognize that identifying the responses' intention is an important issue as well as named entity recognition. Therefore, in this section, we review language-related methodologies that can support these issues and show examples of their application. We also discuss the NICHD protocol's promised rules, which are part of the interview guidelines, and describe how our system utilizes them.

### *2.1. NICHD Protocol*

The NICHD protocol comprises structured forensic interview guidelines used to enhance the quality and quantity of statements provided by child victims or witnesses. Studies conducted in various countries, including Korea, have shown that the quality of statements elicited from children improved when investigators followed the NICHD protocol [1,3,19,20]. The NICHD protocol comprises three phases: preliminary investigation, incident-related investigation, and termination.

The preliminary investigation phase involves an introduction, explanation of ground rules, rapport-building, and pre-interview training. In the introduction and rule explanation stages, the interviewer introduces themselves to the child and explains the conversation rules that the child should follow during the interview. Next, in the rapport-building stage, the interviewer and child have a conversation on a neutral topic unrelated to the case to establish trust, or rapport, between them and create a comfortable environment for the child [21]. After establishing rapport and before entering the actual statement-taking phase, pre-interview training is conducted, during which the child can practice memory

recall. Then, various open-ended questions are used to train the child to provide detailed answers based on memory recall. Rapport-building and pre-interview memory retrieval training have been shown to increase the number of statements made by children during the interview phase [22].

The incident-related investigation phase is central to the forensic interview, which involves incident-related interviewing, breaks, and follow-up questions. In an incident-related interview, the interviewer is encouraged to maximize open-ended questions and non-suggestively facilitate statements. Answers to open-ended questions generally elicit more detail, and are more informative and accurate than answers to closed or suggestive questions [23,24]. Using facilitators encourages the interviewee to keep talking, thus eliciting important additional information [25]. The types of questions in the forensic interview could be classified into (1) open-ended questions, such as an invitation, time-segment invitation, cued invitation, and follow-up invitation; (2) directive questions for extracting detailed information in a focused manner; (3) option-posing questions for focusing on details not mentioned; (4) suggestive questions that include leading questions about information not mentioned; and (5) facilitator questions [26], as presented in Table 1. The last termination phase is the step of finalizing the interview.

**Table 1.** Types of questions proposed in the NICHD protocol.

| Type | Description |
| --- | --- |
| Invitation | Open-ended questions to help the child recall information about the incident |
| | "Could you tell me everything that happened that day?" |
| Facilitator | Non-suggestive prompt to elicit a continuous response |
| | "I see.", "You must have had a hard time." |
| Cued-invitation | Refocusing on the information already mentioned by the child to prompt free recall of information |
| | "You mentioned 'that guy.", "Could you tell more about this person?" |
| Directive | Refocusing on information already mentioned to extract more detailed information |
| | "When did this happen?" |
| Option-posing | Prompting the child to focus on aspects or details not mentioned. Confirmation, negation, or selection of interviewer's words |
| | "Did it hurt?", "Did he touch you over or under your clothes?" |

### 2.2. Data

Table 1 presents the types of interview techniques adopted in the NICHD protocol. "Invitation" in NICHD protocol is a type of interview that collects information about an incident using open-ended questions, such as "Could you tell me everything about what happened that day?" to help the interviewee recall as much information about the incident as possible. A "clued-invitation" question enables the interviewee to freely remember events by refocusing on the mentioned part, such as "You mentioned 'that guy'. Could you tell me more about this person?" "Directive" type questions are used to refocus the interviewee's attention on previously mentioned information to elicit more detailed information. In addition, other types of inquiries related to "facilitation" and "option-posing" exist to elicit responses from the interviewee efficiently.

For such interview processes, identifying the intent of a statement is crucial. To elicit a continued response, subjective psychological description (SPD), interaction description (ID), and acknowledging lack of memory (ALM) were assigned with the intent to refocus on the mentioned parts and obtain details. These intents were among the most frequently analyzed intents of Criteria-Based Content Analysis [27], which was the protocol used to judge the credibility of a statement. ALM means not remembering information about an

incident and corresponds to sentences like "I'm not sure" or "I don't remember". SPD refers to the emotions the interviewee is feeling and is expressed in sentences such as "I was scared". or "I was angry". ID refers to an act committed directly by the perpetrator on the victim, and is expressed in sentences such as "He hit me'. or "He touched me". To extract words related to incidents in the sentences of statements provided by interviewees, entities (i.e., WHO, WHEN, WHERE, ACTION, and NO) were formed. WHO, WHEN, and WHERE refer to the person, time, and place information mentioned, respectively, ACTION refers to the action taken by the interviewee or by the perpetrator. NO is an entity that means the victim can't remember key information about the incident, such as "I don't know". Tags of entities were established according to the beginning, inside, outside (BIO) schema, which was a tagging format used for tagging sentence tokens for NER. The B- and I-prefixes refer to the beginning and internal parts of a named entity, respectively. O refers to an insignificant entity within a sentence that was not used for the NER at the training step.

### 2.3. Data Augmentation

Data augmentation is a method proposed to increase the accuracy of deep learning model training by increasing the volume of data required for such training. Wei et al. [28] demonstrated that models applying simple text augmentation techniques practically exhibited substantial performance in sentence classification tasks. They proposed data augmentation methods using internal data instead of language model and external data. These methods included synonym replacement (SR), which replaced a certain word in a sentence with a synonym; random insertion (RI), which inserted a random word; random swap (RS), which swapped the location of two random words in a sentence; and random deletion (RD), which deleted a random word. In these easy data augmentation (EDA) methods, to convert the number ($n$) of words to be adjusted according to the technique applied, the length ($l$) of the original sentence was multiplied with a parameter ($\alpha$) that had a value ranging from "0" to "1" ($n = \alpha l$). Accordingly, studies have been conducted to increase the performance of classification models by applying these EDA methods. Dhiman et al. [29] proposed an opinion classification model according to government health systems, which applied an EDA method to solve the problem related to an insufficient amount of label data for model training. Dai et al. [30] adjusted an NER task and EDA method according to projects in the biomedical and materials science domains, and compared the derived datasets based on the adjusted methods with existing datasets. Based on the comparison results, they verified that the adjusted methods exhibited better performance levels than conventional methods.

The intent classifier and entity recognizer were trained using face-to-face survey statements as data. A total of 102 statements were obtained from the Supreme Prosecutor's Office, encompassing crimes such as rape, attempted rape, indecent acts by compulsion, pseudo-sexual acts, and other offenses falling under the "sexual assault act", "act on the protection of children", "juveniles from sexual abuse", and "criminal act". The dataset was developed based on statements that were deemed reliable by the testimony analyst, who synthesized the investigation atmosphere, tone of the victim, expression method, psychological state, and internal–external information to determine the intent to be classified and entity to be recognized. Among the intent classifications, ALM, ID, and SPD comprised 19%, 64%, and 17% of the total premise intent classifications, respectively. However, this distribution was imbalanced and required balancing for more accurate intent classification. Similarly, the recognition of entity names exhibited significant imbalances across categories, ranging from 1% to 40%.

Given the limited amount of data obtained from the statement, there was a concern about overfitting during model training, potentially leading to reduced classification and recognition performance in practical applications. Recognizing the critical importance of these models in legal proceedings, data augmentation techniques were employed to mitigate these issues, and enhance the performance of intent classification and named

entity recognition. These efforts were essential to ensure the reliability and effectiveness of the suggested models in court decisions.

We searched for and applied various techniques to solve the low resource data problem and improve classification and recognition performance by diversifying the expressions of sentences and words in the learning data. Hence, few-shot learning was adopted to improve classification performance in a low-resource environment and vocabulary replacement using a pre-trained language model. However, the vocabulary used by children during the face-to-face investigation of child sex crimes was used only in special situations. Few-shot learning had the disadvantage that adaptation was difficult when the difference between domains of datasets was large, and the Korean language pre-trained models currently provided as open source did not have sufficient training on vocabulary related to sexual harassment and sexual assault. As a result, we used the expression method of words in the training data that have Korean characteristics; that is, the meaning of words changes in various ways depending on the postposition attached to the word. The EDA method actively used internal data and the FastText [31] embedding model showed high performance in Korean, where postposition was frequently used. Therefore, the entire dataset was divided into training and test data at a ratio of 8:2. Subsequently, the training data were augmented using the EDA method and FastText.

Algorithm 1 presents the text augmentation algorithm. The EDA method conducts data augmentation by applying the RS and RD techniques, which swap or delete the locations of words, and the SR and RI techniques, which replace target words with synonyms or insert synonyms for target words. Therefore, building a word dictionary consisting of words similar to the words to be converted became a priority. This study used Mecab [32], an open-source Korean morpheme analyzer, to establish a word dictionary by extracting words that belonged to five entity sets for NER based on the statement transcripts. To augment the sentences, including various expressions belonging to the entities and avoiding the loss of meaning of the intent, the RI, SR, RS, and RD parameters were set with ratios of 0.7, 0.7, 0.3, and 0.1, respectively. We set the number of augments per sentence to 10, 20, and 30. Table 2 lists the number of augmentations per sentence using the EDA method in different ratios.

**Algorithm 1** Text augmentation with these parameters: SR = 0.7, RI = 0.7, RS = 0.3, RD = 0.1 $D_{EDA}$ = Korean dictionary of similar words, $D_{FastText}$ = FastText embedding model

**Input:** The trained dataset
**Output:** A set of augmented sentences, $Aug$

**for** sentence in the train dataset **do**
  $L \leftarrow$ length of sentence
  $N \leftarrow L \times$ SR (or RI, RS, RD)
  $E_{word} \leftarrow$ Extract N words randomly from a sentence
  $S_{word} \leftarrow$ Replace with words similar to the $E_{word}$ using $D_{EDA}, D_{FastText}$

  **for** SR augmentation number **do**
    $Aug \leftarrow$ Add by replacing $E_{word}$ with $S_{word}$ in the sentence
  **end for**
  **for** RI augmentation number **do**
    $Aug \leftarrow$ Add by inserting $S_{word}$ after $E_{word}$ in the sentence
  **end for**
  **for** RS augmentation number **do**
    $Aug \leftarrow$ Add by randomly changing the position of $E_{word}$ in the sentence
  **end for**
  **for** RD augmentation number **do**
    $Aug \leftarrow$ Add by deleting $E_{word}$ from the sentence
  **end for**
  $Aug \leftarrow$ Add the original sentence.
**end for**
**return** $Aug$

**Table 2.** Number of total sentences with different augmentation ratios.

| Original Sentence | Augmentation Techniques | | | | Total Augmented Sentences |
|---|---|---|---|---|---|
| | SR | RI | RS | RD | |
| 1 | 5 | 1 | 2 | 1 | 10 |
| 1 | 10 | 3 | 4 | 2 | 20 |
| 1 | 15 | 5 | 6 | 3 | 30 |

The Skip-gram model of FastText was employed for data augmentation. The EDA method performed word replacement or conversion based on the word dictionary related to the entity set, which was developed based on the extraction results of the transcripts. It adopted this process because the forms of entities used in the sentences of statements were frequently observed in the thesaurus. However, this method was limited by its range of expression. Accordingly, the Skip-gram model of FastText was used to replace target expressions with various entity expressions that were not included in the internal data based on external data. Hence, in this experiment, the embedding model of FastText was trained on 700,000 datasets of Korean conversations, news, and translations provided by AI-Hub [33]. Based on the trained embedding model, the target words were replaced with words that had similar vectors to the target words. We solely used words that exhibited 95% (or higher) similarity with the corresponding target words and data augmentation was applied in the same way as the EDA method. Tables 3 and 4 present the number of augmented data points and examples of tokenizing augmented sentences based on the EDA method and FastText.

**Table 3.** Total number of training data using different augmented ratios.

| Augmentation Method | Augmented Ratio | INTENT | ENTITY | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | ALM | ID | SPD | WHO | WHEN | WHERE | ACTION | NO |
| Original Data | | 557 | 1849 | 497 | 523 | 305 | 256 | 1059 | 445 |
| EDA | 10 | 5570 | 18,490 | 4970 | 4737 | 2855 | 2403 | 9793 | 4205 |
| | 20 | 11,140 | 36,980 | 9940 | 9106 | 5472 | 4553 | 18,716 | 8012 |
| | 30 | 16,710 | 55,470 | 14,910 | 13,455 | 8117 | 6739 | 27,541 | 11,860 |
| FastText | 10 | 5570 | 18,490 | 4970 | 3307 | 2123 | 1639 | 6494 | 2849 |
| | 20 | 11,140 | 36,980 | 9940 | 6189 | 4037 | 3031 | 12,215 | 5312 |
| | 30 | 16,710 | 55,470 | 14,910 | 9111 | 5828 | 4468 | 17,913 | 7895 |

**Table 4.** ((1) is an augmented sentence and (2) means tokenized (1) sentence; (3) means tokens of (2) tagged by BIO tagging; (4) is a sentence that translates (1) into English. In the example below, the KoBERT tokenizer is used as the tokenizer). ($^{*}$ NO entity, √ WHERE entity, $^{\diamond}$ WHO entity, $^{\triangle}$ ACTION entity, $^{\heartsuit}$ ACTION entity).

| Intent | Example |
|---|---|
| ALM | (1) 몰라요$^{*}$<br>(2) _몰, 라, 요$^{*}$<br>(3) (B-NO$^{*}$, I-NO$^{*}$, I-NO$^{*}$)<br>(4) (I don't know$^{*}$.) |
| ALM | (1) 뭘로 밀었는지 그게 기억이 안나네요$^{*}$<br>(2) _, 뭘, 로, _밀, 었, 는, 지, _그, 게, _기억, 이, _안, 나, 네요$^{*}$<br>(3) (O O O O O O O O O O O O B-NO$^{*}$ I-NO$^{*}$ I-NO$^{*}$)<br>(4) (I don't remember$^{*}$ what he shoved me with.) |

**Table 4.** *Cont.*

| Intent | Example |
|---|---|
| SPD | (1) 되게 기분 나빴어요.<br>(2) _되, 게, _기분, _나, 빴, 어요<br>(3) (O O O O O O)<br>(4) (I felt really bad.) |
| SPD | (1) 그때 그래서 정말 화가 났어요.<br>(2) _그때, _그래서, _정말, _화, 가, _, 났, 어요<br>(3) (O O O O O O O O)<br>(4) (So, I got really angry then.) |
| ID | (1) 저를 학교$^{\surd}$에서 오빠가$^{\diamond}$ 때렸어요$^{\triangle}$<br>(2) _저, 를, _학교$^{\surd}$, 에서, _오빠, 가,$^{\diamond}$ _때, 렸, 어요$^{\triangle}$<br>(3) (O O B-WHERE$^{\surd}$ I-WHERE$^{\surd}$ B-WHO$^{\diamond}$ I-WHO$^{\diamond}$ B-ACTION$^{\triangle}$ I-ACTION$^{\triangle}$ I-ACTION$^{\triangle}$)<br>(4) (My brother$^{\diamond}$ hit$^{\triangle}$ me at$^{\surd}$ school$^{\surd}$.) |
| ID | (1) 쌤이$^{\diamond}$ 어제$^{\heartsuit}$ 괴롭혔어요$^{\triangle}$ 아침에$^{\heartsuit}$<br>(2) _, 쌤, 이,$^{\diamond}$ _어제,$^{\heartsuit}$ _괴, 롭, 혔, 어요,$^{\triangle}$ _아침, 에$^{\heartsuit}$<br>(3) (O B-WHO$^{\diamond}$ I-WHO$^{\diamond}$ B-WHEN $^{\heartsuit}$ B-ACTION$^{\triangle}$ I-ACTION$^{\triangle}$ I-ACTION$^{\triangle}$ I-ACTION$^{\triangle}$ B-WHEN$^{\heartsuit}$ I-WHEN$^{\heartsuit}$<br>(4) (The teacher$^{\diamond}$ bullied$^{\triangle}$ me at yesterday$^{\heartsuit}$ morning.$^{\heartsuit}$) |

### *2.4. Word Embedding*

An embedding process is required to convert natural languages into machine-readable languages. In other words, this process presents the results of changing natural languages to vectors, in the form of numbers that machines can understand, or a series of these processes. FastTextis an embedding method that is an enhanced version of the Word2Vec [34] model and an example of the distribution hypothesis. The distribution hypothesis assumes that the distribution of each word differed depending on its location in a sentence and the other words that appeared around it. It assumes that a pair of certain words might have similar meanings if they appeared in sentences that had similar meanings. FastText represents each word based on an *n*-gram, the unit of a word. The term "*n*-gram" refers to a sequence of *n* word(s). In other words, a certain word is divided based on *n*. For example, when the word "teach" is divided based on an *n*-gram ($n = 3$), it could be expressed as [te, tea, eac, ach, ch], and [teach]. FastText uses the symbols "[" and "]" to indicate the boundaries of a word, including its beginning and end. Consequently, the sum of the vectors divided based on the unit of *n*-gram is adopted to represent an embedding of the term "teach". FastText uses a skip-gram model to train a pair of target words and context words. If the context word was located near the target word, then this model increased the dot product of the vectors of both words to increase the cosine similarity and correlation between both words. If the context word was not located near the target word, then this model reduced the dot product of the vectors of both words to decrease the cosine similarity. Via this process, the distribution information, which was the peripheral context of the target word, was included in the embedding. The "[te, tea, eac, ach, ch]" vector, which was obtained via the representation of the word "teach", based on the unit of *n*-gram, underwent embedding. Accordingly, if the word "teacher" did not exist in a dataset, FastText could infer that these words were similar to each other because both words had the same *n*-gram vectors and the context word "study" was located nearby. This method was advantageous because it could respond flexibly to changes in verbs or nouns, and it performed well when dealing with spacing, spelling, and out-of-vocabulary words. It was especially effective in the Korean language, which had a wide range of postpositions and endings [35].

### *2.5. Language Model*

Bidirectional Encoder Representations from Transformers (BERT), introduced by Devlin et al. (2018) [36], is a significant language model publicly released by Google. Its

architecture is built upon a modified version of the transformer encoder proposed by Vaswani et al. (2017) [37]. BERT is a domain-specific transfer learning method based on a pre-trained model represented as a vector by learning contextual relationships and word relationships in both directions from large amounts of data. In BERT, specific words within a sentence are masked using the [MASK] token and the model is trained in both directions to predict these masked words. One of the core mechanisms in BERT is the scaled dot-product attention, which calculates the relationship of a word with other words in the sentence. This attention mechanism is expressed in Equation (1). BERT's ability to capture bidirectional dependencies has made it a powerful tool in various natural language processing tasks.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

To generate dot-product attention $Q$, $K$, and $V$ vectors are used. $W_i^Q$, $W_i^K$, and $W_i^V$, matrices corresponds to $Q$, $K$, and $V$, respectively. Dot-product attention based on a query and key is performed to identify a correlation between them, and the result of the dot-product attention is scaled based on $\sqrt{d_k}$ ($k$ = the number of dimensions in a key vector). Subsequently, the softmax function is applied. When a certain query and key exhibited an important relationship in the context, their dot product is increased to raise the attention weight. In the transfer learning and fine-tuning stages, the BERT embedding, including the semantic and grammatical relationships of a training corpus, is adopted to re-adjust the BERT model according to a downstream task to be established. In this method, an embedding model contains abundant language expressions pre-trained via unsupervised learning based on many corpora. As a result, it could train with relatively few resources and data to be optimized for various domains and tasks. In addition, it could reduce the amount of time required for training. BERT models that are language-specific have also been proposed [38,39]. BERT has been used in many languages for intent classification purposes [40–42]. In South Korea, Korean BERT pre-trained cased (KoBERT) [43], a model specialized in the Korean language, has been presented. The pre-trained BERT-based Multilingual Cased [44] and KoBERT models are compared in Table 5.

**Table 5.** Comparison of pre-trained BERT and KoBERT.

| | **BERT-Multilingual-Cased** | **KoBERT** |
|---|---|---|
| Trained languages | 104 languages | Korean |
| Number of layers | 12 | 12 |
| Hidden size | 768 | 768 |
| Maximum length | 512 | 512 |
| Attention heads | 12 | 12 |
| Vocabulary size | 30,522 | 8002 |
| Parameter size | 110 M | 92 M |
| Tokenizer method | Wordpiece [45] | Sentencepiece [46] |

The BERT-multilingual-cased model is pre-trained on Wikipedia data from 104 languages, including English and Korean. It employs the data of languages (e.g., English and French) that comprise a comparatively abundant amount of label data to increase the performance related to languages that have a small amount of label data. However, unlike English, Korean is an agglutinative language, where the meanings of words change depending on roots and affixes. The BERT-multilingual-cased model showed limited capability in optimizing sentence expressions in Korean. Conversely, KoBERT trained its tokenizer with only the Korean corpora, exhibiting exceptional performance for Korean grammar processing. However, because this model was solely trained in Korean, it had a smaller scale of vocabulary and training parameters than the BERT-multilingual-cased model trained in

numerous languages. These limitations affect KoBERT's ability to handle words outside its vocabulary, ultimately impacting its classification performance. By comparing two different LLMs, we want to analyze the performance evaluation of intent classification and NER to find the best way to optimize the system.

The pre-trained models, BERT-multilingual-cased and KoBERT, were adopted to fine-tune models for intent classification. In the previous section, the method of augmentation is explained and a large dataset for making a training model is described. However, the training target data used are different from the sources of BERT or KoBERT, which are literary and intellectual data. The used data are colloquial and oral sentences delivered in interviews. The data also come from the limited domain related to sexual violence. Therefore, the similarity of the data to the source of the large language model is low; therefore, the internal parameters were fine-tuned in the process of retraining the entire model. Specifically, the model trained labeled data that had been augmented 10, 20, and 30 times by FastText and the EDA method.

As Figure 1 illustrates, two pre-trained models were fine-tuned and adopted as the models of intent classification and NER, respectively. The data used are augmented statement analysis data, and the entire pre-trained model was applied to intent classification and object name recognition. For the intent classification model, a classifier comprising a single dropout and fully connected layer was added at the end of the pre-trained model. To fit the model training for intent classification, the [CLS] output token, which included the meanings of the entire sentences, was used as input data in the classifier. The same classifier adopted for the intent classification model was added to the entity classification model. Subsequently, this model trained the output tokens located between the [CLS] token, indicating the beginning of a sentence, and the [SEP] token, indicating the end of a sentence. It is important to have a fine-tuning learning rate in order to feed the entire data back into the pre-training model. So, the author of BERT suggested an appropriate batch size and learning rate for good fine-tuning. We applied the training process to achieve the optimal performance for various parameters, for which the batch sizes assessed were 16 and 32, and the learning rates were $5 \times 10^{-5}$, $3 \times 10^{-5}$, and $2 \times 10^{-5}$ In addition, the Adam optimizer [47] was used, the epoch was set to 10, and the model with the lowest validation loss was selected.

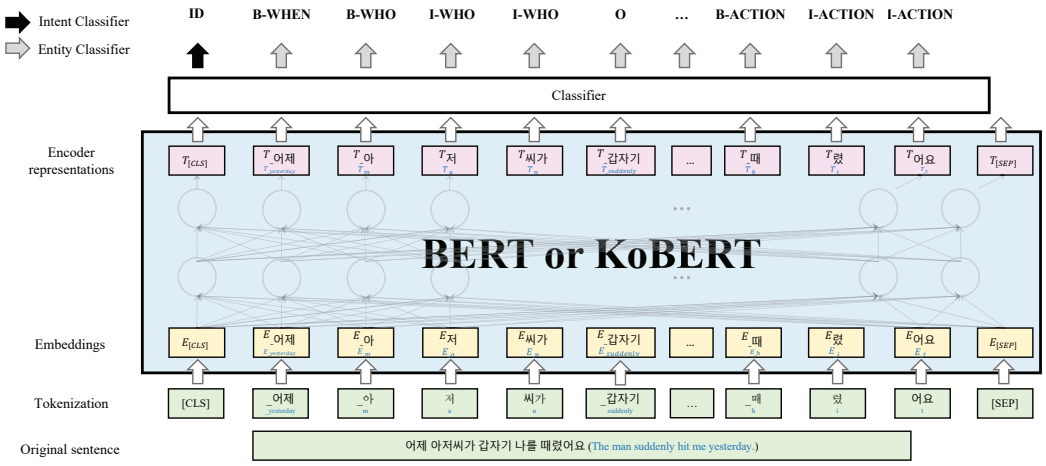



**Figure 1.** Architecture of the fine-tuning learning model for intent classification and NER by the pre-trained models. A fine-tuning of the entire model as performed using the given statement data and the augmented data.

### *2.6. Configuring and Implementing the System*

Currently, many cloud service platforms provide several free APIs related to chatbots, such as Google DialogFlow and Microsoft Azure Bot. However, among them, no chatbot was developed to investigate statements from victims of child sex crimes. Therefore, we

developed our chatbot framework specializing in statement investigation. The chatbot developed here comprises the preliminary research, incident-related investigation, and termination phases according to the NICHD protocol. The initial investigation phase consists of a stage in which the chatbot explains the rules it should know from the child before the incident-related investigation phase begins and a stage in which the child freely describes their hobbies. In the incident-related investigation phase, the child makes a statement about the incident they experienced. In the termination phase, when the chatbot determined that sufficient critical information about the incident had been collected, it expressed gratitude for the child's statements, informed about the possibility of future interviews, and ended the investigation. Figure 2 presents the system configuration related to the incident-related investigation phase implemented in full-scale investigative interviews. The incident-related investigation phase was divided into three stages: (1) training the classifier and recognizer models using the transcript (yellow box in Figure 2); (2) identifying intent and entities in sentences based on trained models (blue box in Figure 2); and (3) the question generation part based on NICHD protocol to collect objective statements using classified intents and recognized entities (green box in Figure 2). The objective of our chatbot was to impartially gather extensive information about the event from the child's statements. Hence, the primary focus of this study was to assess the efficacy of intent classification, NER, and the construction of the training model within the chatbot system.

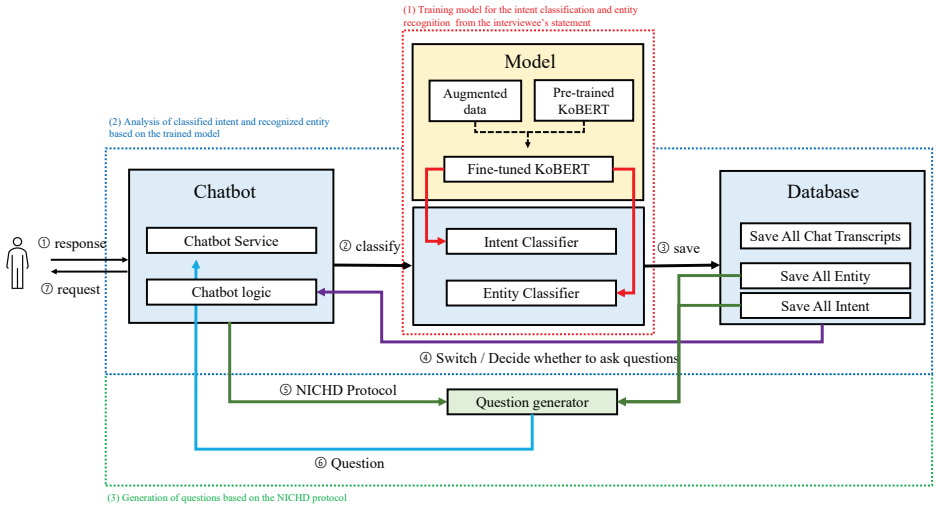



**Figure 2.** System configuration.

#### 2.6.1. Training Model for Intent Classification and Entity Recognition from the Interviewee's Statements

The chatbot was meticulously designed with a slot-filling structure, enabling it to generate questions by identifying and analyzing crucial incident keywords while also classifying the intent of interviewees based on their statements. A model capable of continuous intent classification and entity recognition was necessary to converse with interviewees continuously. Consequently, a high-performance BERT-based model from the NLP domain was employed and interview transcripts supplied by the Supreme Prosecutor's Office of the Republic of Korea were utilized for model training. Furthermore, to enhance classification and recognition performance, a well-established data augmentation method was implemented.

#### 2.6.2. Analysis of the Classified Intent and Recognized Entity Based on the Trained Model

The basic framework of the system was adopted from the open-source chatbot and Kochat [48]. The interviewee first connects to the web server through a browser and begins an interview with the chatbot. The chatbot analyzes the intent and entities in the sentences obtained from the interviewee using open-ended questions. When the interviewee's statement was set as input data in the chatbot system, the intent of the sentence is classified into

the category with the largest softmax predicted value among the three predefined intent categories. Entity analysis is conducted using the entity recognizer that trained entities that could be derived as detailed clues. Then, the final classified intent and recognized entities are stored in the database, along with the interviewee's sentences. The blue dotted line in Figure 2 depicts these processes.

#### 2.6.3. Generating Questions Based on the NICHD Protocol

Whenever the chatbot obtains an answer from the interviewee, it checks the internal database to determine whether sufficient statements about the incident were obtained. The recognized entities are keywords for the main information on the incident and are used to determine the sufficiency of the statement: WHO, WHEN, WHERE, action of the perpetrator (ACTION), inability to remember (NO), and others (O). The chatbot verifies whether at least two or more entities and additional information about each entity were stored for each of the four important entities (who, when, where, and action) within the database. If the chatbot decides that the necessary entities were not obtained via open-ended questions, then it generates additional questions based on the NICHD protocol, and identifies the intent and entities in the sentence. It repeatedly asks invitation-based questions until all information about the necessary entities is obtained. The chatbot classifies the interviewee's sentences into one of three intents (i.e., subjective psychological description (SPD), interaction description (ID), and acknowledging lack of memory (ALM)), and the question is formulated differently depending on the intent. After identifying major pieces of information about the incident using invitation questions, the chatbot asks cued-invitation, directive, and option-posing questions about the recognized entities to request additional information on the entities. If the interviewee acknowledges a lack of memory, then the chatbot generates facilitators and open-ended questions, such as "Could you tell me everything you can remember?" If the interviewee described a subjective psychological experience, then the chatbot adopted facilitators and generated other questions about the collected entity. This process is repeated as the interview continues and, once sufficient statements about the incident have been obtained, the interview is terminated.

## 3. Results

Figures 3 and 4 show the accuracy of the performance evaluation for each batch size and learning rate. KoBERT-EDA20 and KoBERT-FastText30, which showed high-performance distributions for each parameter, were used as the classification and recognition measurement models for each intent and entity. During the extraction of the main information of the case from the victim's statement, it was important to recognize the entire entity, not only the tokens divided through BIO tagging. In Korean, nouns and verbs are mainly classified as "B-tags" and suffixes are classified as "I-tags". The personality of an entity might change depending on the combination of the B- or I-tag. Therefore, the accuracy measurement method in entity recognition was applied to the case in which all B- and I-tagged entities constituting one entity were predicted. Both data augmentation methods enhanced the accuracy of the intent classification and entity recognition. As Figure 3 shows in the intent classification, the EDA method achieved a higher accuracy than the FastText method. Moreover, based on an analysis, the EDA method, which applied a data augmentation based on internal data, facilitated excellent intent classification performance because the victims' statements exhibited similar patterns according to the types of intent. The FastText method had the highest accuracy in entity recognition, as shown in Figure 4. This result was obtained because the models trained various changes in words that were not included in the internal data by employing external data instead of internal data, considering the characteristics of the Korean language, where words easily change. The highest performance was achieved at a batch size of 16 for intent classification and entity recognition, and the optimal performance was achieved for learning rates of $3 \times 10^{-5}$ and $5 \times 10^{-5}$, respectively. For both intent classification and named entity recognition, the accuracy was higher when using different batch sizes and augmentation than using only

original data. The lower performance on the original was due to the smaller data size in each classification for both intent and person names, which resulted in better training for classes with more data.

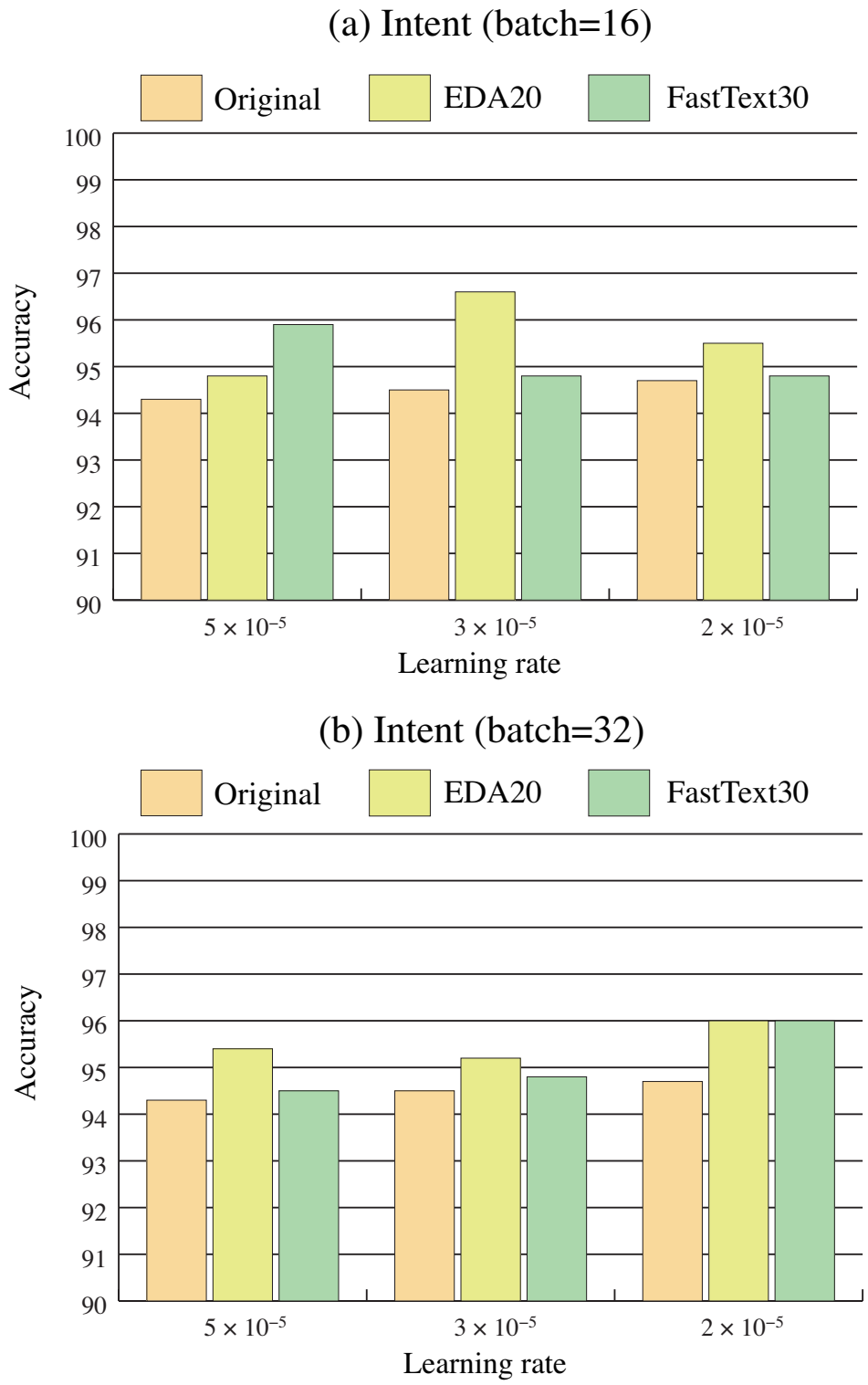



**Figure 3.** Intent classification performance by the training parameters.

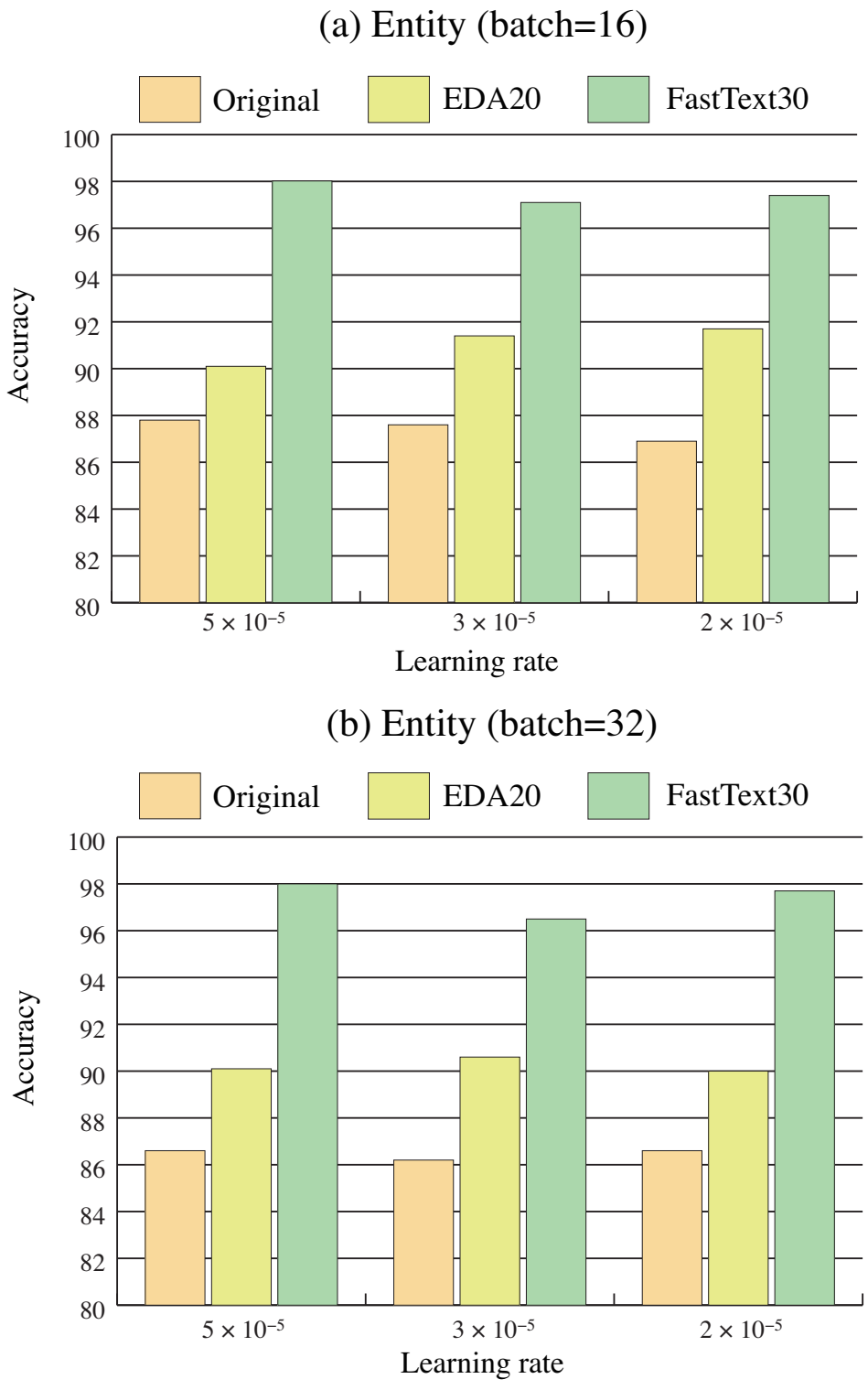



**Figure 4.** Entity recognition performance by the training parameters.

## 4. Discussion

Table 6 contains an excerpt from an incident-related investigation using the suggested system. The scenario was assessed by adapting the sentences stated by the child in the actual face-to-face investigation transcript. Figure 5 shows a few screens of the corresponding content. A full-scale statement investigation began with a sentence in which the system recalled the entire memory of the incident, according to the invitation question type of the NICHD protocol. First, the system recognized the entity of the incident from the victim's

statement and, if sufficient entities were not collected, it induced recall once again. This process is shown in lines 1 and 3. When the intent of the statement was classified as an ID, the system stored the recognized entity in the database because it contained the main information of the event. In lines 2 and 4, the system recognized yesterday (WHEN), old man (WHO), touched, and grabbed (ACTION) entities and confirmed that it did not collect information about the WHERE entity. In lines 5–16, the system additionally asked questions about the WHERE entity not mentioned and the WHO, WHEN, ACTION entities mentioned. Based on the type of invitation question, if the intent of the statement was classified as ALM and there was no recognized entity, it was set up to make a statement as much as possible if the child did not know it well. In lines 10–12, the system generated a question to collect additional information about the old man (WHO), but the child's statement was classified as ALM. However, the WHO entity in the sentence was recognized again, and the system judged that additional information about the WHO entity was collected, although the intent was classified as ALM. The question continued about other entities stored in the database. In line 13, the child's statement was classified as SPD. In line 16, as SPD was classified from the previous victim's statement, additional questions were asked about the feelings of the child. Therefore, the proposed system is capable of efficient statement investigation based on a total of three items: (1) From the victim's response sentences collected through NICHD questions that induce free recall of the incident, the main information of the incident can be collected using intent classification and NER. (2) The system can continuously request additional information about the recognized entities and can cope with various unexpected situations such as admitting lack of memory. (3) Scenarios for statement investigations can be set according to the answers of sex crime victims and sentences that deviate from the flow of statement investigations can be controlled.

Table 7 shows the performance of the intent classifier with the optimal parameters. The F1-score for the SPD was lower than that for the other intent types. This result was derived from the existence of data that had similar structures, such as an ID, that is, "He hit me yesterday", and an SPD, that is, "I was scared because he hit me yesterday". As a result, SPD data necessitated a thorough examination. The results verified that KoBERT exhibited a better performance than BERT in classifying sentences with similar structures but different types of intent. Although KoBERT has fewer training parameters and fewer words than BERT, our belief is that advanced tokenizer techniques, such as SentencePiece, enhance its contextual comprehension significantly. Furthermore, in comparing data imbalance levels, SPD shows more minor data than other intent classifiers. This means that the degree of imbalance for each classification class can significantly influence accuracy evaluations.

Table 8 shows the performance of the entity recognizer with the optimal parameters. In entity recognition, the KoBERT model showed a higher accuracy than the BERT model under the same augmentation method as the intent classification. However, in contrast to the intent classification result, a high F1-score was derived when the FastText augmentation method was applied. The accuracy of the EDA method tended to decrease as the number of augmentations increased but the FastText method improved accuracy.

The named entity ACTION can change the word "때리다 (hit)" to a variety of forms, including "때리다가 (while hitting)", "때리려 다가 (to be about to hit)", and "때리려고 (to hit)". This phenomenon can be attributed to the nature of the Korean language, which is morphological and composed of prefixes, stems, and suffixes. As a result, specific movement words can be expressed in diverse ways. In the case of the FastText method, one action target word (ex: "때리다 (hit)") can have various expressions based on external data; this enabled models to train the meanings of the target entity (ex: "때리다 (hit)") in a wide range.

**Table 6.** Conversation table excerpted from a part of the incident-related investigation with a chatbot. The gray rows are the chatbot's responses or questions, and the white rows are the user's responses.

| Line | Speaker | Text | Classified Intent, Recognized Entity | NHCID Adopted Question |
|---|---|---|---|---|
| 1 | Chatbot | As far as I know, something happened to you, can you tell me everything about what happened to you, from beginning to end? | | Invitation |
| 2 | Child | Yesterday, I was going home with my friend after school, and an old man came to me to teach me how to use my cell phone and touched my body. | ID, Yesterday(WHEN), old man(WHO), touched(ACTION) | |
| 3 | Chatbot | Please elaborate a little more on what you just told me. | | Invitation |
| 4 | Child | I was with a friend, and he came to me. That's why he asked me to teach him how to capture it. So, he asked me to help him, so I helped him, but he suddenly grabbed my waist. | ID, old man(WHO), grabbed(ACTION) | |
| 5 | Chatbot | Then I'd like to ask you about the things that you told me in detail. | | Invitation |
| 6 | Chatbot | Please tell me more about that time. | | Cued-invitation |
| 7 | Child | It is probably 3 o'clock because it is time to go home after school. | 3 o'clock (WHEN) | |
| 8 | Chatbot | Please tell me more about the position you were in at the time. | | Option-posing |
| 9 | Child | At that time, I was turning my back on the apartment, and the old man came to us after sitting on a bench. | ID, At that time(WHEN) | |
| 10 | Chatbot | Please tell me everything you know about him. | | Cued-invitation |
| 11 | Child | He was wearing a hat, so I can't remember his face in detail... He was wearing something like a coat, and he was lame. | ALM, he(WHO) | |
| 12 | Chatbot | Could you tell me what action you took in detail when that person "touched" you? | | Cued-invitation |
| 13 | Child | I felt that he was trying to touch my body with force. So, I was so embarrassed that I shook off his hand and ran away with my friend because I was scared. | SPD | |
| 14 | Chatbot | You told me that he "grabbed" it, but could you tell me in more detail about it? | | Cued-invitation |
| 15 | Child | He pushed his face closer to me asking for instructions on how to capture a cell phone and then I pretended to be natural and put my hands around my waist. | ID | |
| 16 | Chatbot | Could you tell me how you felt when it happened? | | Directive |

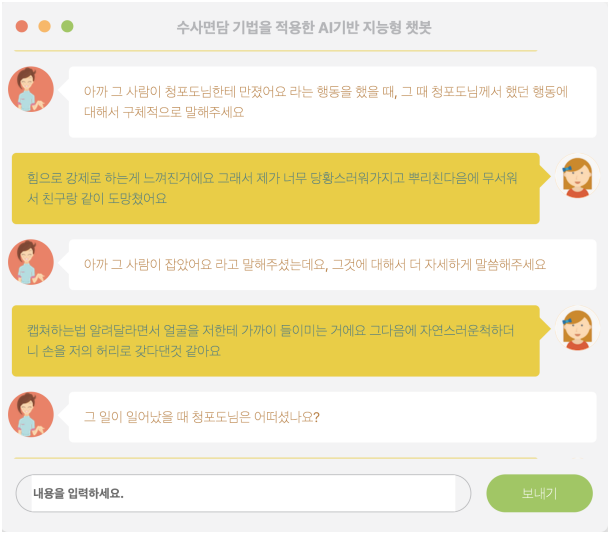



**Figure 5.** A The screen capture is a translation of line numbers 1 through 5 from Table 6.

**Table 7.** Results of the intent classification. Batch size of 16 and learning rate of $3 \times 10^{-5}$ were used as training parameters. The bold font has the best accuracy in this comparison.

| Model | Augmentation Method | Augmented Ratio | INTENT (F1-Score) | | | | Accuracy |
|---|---|---|---|---|---|---|---|
| | | | ALM | ID | SPD | Average | |
| BERT | Original | | 95.5 | 94.4 | 79.2 | 89.7 | 91.9 |
| | EDA | 10 | 96.0 | 95.5 | 82.7 | 91.4 | 93.5 |
| | | 20 | 94.2 | 95.5 | 82.8 | 90.8 | 93.1 |
| | | 30 | 92.0 | 95.4 | 81.2 | 89.5 | 92.4 |
| | FastText | 10 | 95.5 | 95.9 | 83.9 | 91.8 | 93.8 |
| | | 20 | 93.8 | 95.1 | 80.6 | 89.8 | 92.4 |
| | | 30 | 94.1 | 96.1 | 84.8 | 91.7 | 93.8 |
| KoBERT | Original | | 95.2 | 96.9 | 85.1 | 92.4 | 94.5 |
| | EDA | 10 | 96.0 | 97.3 | 91.6 | 94.9 | 96.0 |
| | | 20 | 97.3 | 97.7 | 91.5 | 95.5 | **96.6** |
| | | 30 | 94.5 | 96.0 | 84.4 | 91.6 | 93.6 |
| | FastText | 10 | 96.4 | 96.9 | 87.6 | 93.6 | 95.2 |
| | | 20 | 92.9 | 97.1 | 87.6 | 93.0 | 95.0 |
| | | 30 | 96.5 | 96.6 | 86.6 | 93.2 | 94.8 |

In actual statement, words representing the time of the crime and perpetrator were not as diverse as the ACTION entity. For example, most words only denoted a specific range, such as “Monday” to “Sunday”, “01:00” to “24:00”, “morning”, “evening”, “night”, “midnight”, “he”, “father”, “grandfather”, and “maternal grandfather”. However, since the WHEN and WHO entities that recognize the primary information of an incident are significant when gathering testimonies, we set them as entities to be recognized despite their small scope of expression. Subsequently, during the augmentation process, the expressions of the WHEN and WHO entities presented in the sentence were further trained. Unlike other entities, they achieved a relatively high F1-score of 100 compared with the recognition result of the original data. According to the analysis results, the best combination of intent classification and NER for collecting sentences is KoBERT-EDA20 for intent classification and KoBERT-FastText30 model for NER. This result shows that KoBERT, an LLM model based only on Korean, is more appropriate than BERT for recognizing intent and NER. Also, EDA is more suitable as an augmentation considering the relationship between words in a sentence.

**Table 8.** Results of the entity classification. Batch size 16 and learning rate $5 \times 10^{-5}$ were used as training parameters. The bold font has the best accuracy in this comparison.

| Model | Augmentation Method | Augmented Ratio | ENTITY (F1-Score) | | | | | | | | | | | Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | B-WHO | I-WHO | B-WHEN | I-WHEN | B-WHERE | I-WHERE | B-ACTION | I-ACTION | B-NO | I-NO | Average | |
| BERT | Original | | 96.1 | 96.7 | 94.6 | 97.6 | 89.7 | 86.4 | 84.3 | 82.6 | 97.0 | 97.6 | 92.3 | 84.9 |
| | EDA | 10 | 98.6 | 98.4 | 97.9 | 98.8 | 95.6 | 95.2 | 90.6 | 92.8 | 92.7 | 97.6 | 95.8 | 81.7 |
| | | 20 | 97.2 | 97.9 | 96.9 | 98.8 | 94.6 | 94.3 | 91.3 | 93.7 | 93.2 | 96.5 | 95.4 | 80.3 |
| | | 30 | 96.3 | 96.9 | 96.8 | 98.8 | 92.6 | 93.2 | 90.6 | 91.6 | 97.0 | 98.2 | 95.2 | 80.9 |
| | FastText | 10 | 98.6 | 100 | 97.8 | 98.8 | 94.1 | 93.7 | 91.3 | 93.1 | 97.6 | 98.2 | 96.3 | 90.6 |
| | | 20 | 98.1 | 98.9 | 97.8 | 98.8 | 95.2 | 94.9 | 90.9 | 94.0 | 98.2 | 98.8 | 96.6 | 91.0 |
| | | 30 | 99.0 | 99.5 | 97.9 | 97.6 | 94.0 | 94.9 | 92.9 | 94.3 | 98.2 | 98.8 | 96.7 | 93.3 |
| KoBERT | Original | | 96.2 | 97.1 | 96.8 | 97.4 | 89.5 | 87.7 | 92.0 | 90.9 | 97.6 | 97.6 | 94.3 | 87.8 |
| | EDA | 10 | 98.1 | 97.2 | 99.0 | 98.7 | 95.7 | 93.2 | 96.0 | 95.9 | 97.1 | 97.1 | 96.8 | 92.2 |
| | | 20 | 96.8 | 96.7 | 100 | 100 | 94.6 | 93.2 | 95.6 | 95.4 | 97.1 | 97.1 | 96.6 | 90.1 |
| | | 30 | 99.1 | 98.3 | 100 | 97.4 | 94.6 | 91.9 | 93.6 | 92.8 | 97.6 | 97.6 | 96.3 | 89.0 |
| | FastText | 10 | 98.6 | 98.9 | 100 | 100 | 97.7 | 98.5 | 94.2 | 93.7 | 97.6 | 97.6 | 97.7 | 94.4 |
| | | 20 | 99.5 | 100 | 98.9 | 100 | 96.6 | 95.5 | 95.3 | 94.8 | 98.2 | 98.2 | 97.8 | 97.7 |
| | | 30 | 99.0 | 98.9 | 98.9 | 100 | 97.7 | 97.0 | 95.0 | 94.5 | 98.2 | 98.2 | 97.8 | **98.0** |

## 5. Conclusions

To elicit specific descriptions from the victim with memory recall, the interviewer's questions must be structured to minimize the introduction of subjective opinions. Depending on the question, interviewers can also ask leading or opinionated questions, so the interviewer should learn how to ask questions in a more trained approach. The introduction of a computational system can substantially reduce the time and effort required for interviewer training. In addition, recent studies have shown that collecting victims' statements via computational systems such as chatbot and 3D avatar results are comparable to those obtained through face-to-face interviews regarding the richness of free recall memory within the statements.

So, this study introduces an interviewing system for non-face-to-face statement investigation. The system's ability to comprehend the victim's statement's intent and NER results in asking proper and objective questions. So, it is an essential role of intent classification and object recognition. Since the statement data is imbalanced according to intent and named entity due to the limited data, the proposed model employed an augmentation process to improve the intent classification and NER performance. As a result of model verification, the EDA technique achieved an average F1-score of 95.5%, which is 3.4% better than the classification result of limited data in intent classification by replacing words that distinguish the characteristics of each intent with similar words. The FastText embedding model achieved an average F1-score of 97.8%, which is 3.7% better than the result of limited data in NER by learning various changes of Korean words, an agglutinative language. The above advancements exemplify the potential of this interviewing system to significantly improve the effectiveness of non-face-to-face statement investigations, offering a reliable and objective means of obtaining vital information from victims.

We expect the NLP system can collect more described statement data than face-to-face interviews with a trained person. We also expect to be able to collect data on psychological states that children can express. However, there is still work to be done. Depending on their characteristics, individuals may express themselves differently during the interview. There are also concerns about typos in the case of a typing chatting system. Some words used by different regions had the same meaning but appeared in various forms (e.g., dialects), and minors tend to use metaphors and personifications because they are not accustomed to describing their personal experiences. In the future, we plan to increase the number of entities to be recognized and the intentions to be classified by the chatbot, so that it can continue investigating statements while coping with various types of victims' statements. Furthermore, we plan to expand the system to determine the reliability, accuracy, and validity of interviewee statements by comparing and analyzing interview statements saved after the termination of interviews. We believe that guidelines for questions that can be asked in collaboration with statement analysts and psychologists are needed.

**Author Contributions:** Conceptualization, E.J., J.J. and Y.Y.; methodology, J.S. and J.J.; software, J.S.; validation, J.S. and J.J.; formal analysis, J.J. and E.J.; investigation, J.S. and J.J.; resources, Y.Y. and E.J.; data curation, Y.Y. and E.J.; writing—original draft preparation, J.S. and J.J.; writing—review and editing, J.S., J.J. and E.J.; supervision, E.J. and J.J.; project administration, E.J. and J.J.; funding acquisition, E.J. and J.J. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was supported in part by a grant from the National R&D program of the Supreme Prosecutor's Office (SPO) and the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2022R1F1A1061476).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data used in this study are an analysis of victims' statements provided by the Office of the Chief Prosecutor. To comply with data privacy standards due to the sensitive nature of sexual violence and personal information, the Supreme Prosecutor's Office has

authorized the use of the data for research purposes only and not for public disclosure. Therefore, the data are not publicly available but the source codes presented in this study are available on request from the corresponding author.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| NER | Named Entity Recognition |
| NICHD | National Institute of Child Health and Development |
| BERT | Bidirectional Encoder Representations from Transformers |
| KoBERT | Korean BERT pre-trained cased |
| RI | Random Insertion |
| RS | Random Swap |
| RD | Random Deletion |
| SR | Synonym Replacement |
| EDA | Easy Data Augmentation |
| SPD | Subjective Psychological Description |
| ID | Interaction Description |
| ALM | Acknowledging Lack of Memory |
| BIO | Beginning, Inside, Outside |
| CLS Token | Classification Token |

## References


1. Orbach, Y.; Hershkowitz, I.; Lamb, M.E.; Sternberg, K.J.; Esplin, P.W.; Horowitz, D. Assessing the value of structured protocols for forensic interviews of alleged child abuse victims. *Child Abus. Negl.* **2000**, *24*, 733–752. [CrossRef] [PubMed]
2. Lamb, M.E.; Orbach, Y.; Sternberg, K.J.; Aldridge, J.; Pearson, S.; Stewart, H.L.; Esplin, P.W.; Bowler, L. Use of a Structured Investigative Protocol Enhances the Quality of Investigative Interviews with Alleged Victims of Child Sexual Abuse in Britain. *Appl. Cogn. Psychol. Off. J. Soc. Appl. Res. Mem. Cogn.* **2009**, *23*, 449–467. [CrossRef]
3. Sternberg, K.J.; Lamb, M.E.; Orbach, Y.; Esplin, P.W.; Mitchell, S. Use of a structured investigative protocol enhances young children's responses to free-recall prompts in the course of forensic interviews. *J. Appl. Psychol.* **2001**, *86*, 997. [CrossRef] [PubMed]
4. Lamb, M.; Brown, D.; Hershkowitz, I.; Orbach, Y.; Esplin, P. *Tell Me What Happened: Questioning Children about Abuse*; John Wiley & Sons: Hoboken, NJ, USA, 2018.
5. Ettinger, T.R. Children's needs during disclosures of abuse. *SN Soc. Sci.* **2022**, *2*, 101. [CrossRef] [PubMed]
6. Fernandes, D.; Gomes, J.P.; Pedro, B. Albuquerque and Marlene Matos. Forensic Interview Techniques in Child Sexual Abuse Cases: A Scoping Review. *Trauma Violence Abus.* **2023**. [CrossRef]
7. Tidmarsh, P.; Sharman, S.; Hamilton, G. The Effect of Specialist Training on Sexual Assault Investigators' Questioning and Use of Relationship Evidence. *J. Police Crim. Psychol.* **2023**, *38*, 318–327. [CrossRef]
8. Minhas, R.; Elphick, C.; Shaw, J. Protecting victim and witness statement: Examining the effectiveness of a chatbot that uses artificial intelligence and a cognitive interview. *AI Soc.* **2022**, *37*, 265–281. [CrossRef]
9. Weizenbaum, J. ELIZA—A computer program for the study of natural language communication between man and machine. *Commun. ACM* **1966**, *9*, 36–45. [CrossRef]
10. Smutny, P.; Schreiberova, P. Chatbots for learning: A review of educational chatbots for the Facebook Messenger. *Comput. Educ.* **2020**, *151*, 103862. [CrossRef]
11. Blanc, C.; Bailly, A.; Francis, É.; Guillotin, T.; Jamal, F.; Wakim, B.; Roy, P. FlauBERT vs. CamemBERT: Understanding patient's answers by a French medical chatbot. *Artif. Intell. Med.* **2022**, *127*, 102264. [CrossRef]
12. Nadarzynski, T.; Miles, O.; Cowie, A.; Ridge, D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digit. Health* **2019**, *5*, 2055207619871808. [CrossRef] [PubMed]
13. Rapp, A.; Curti, L.; Boldi, A. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *Int. J. Hum.-Comput. Stud.* **2021**, *151*, 102630. [CrossRef]
14. Li, C.H.; Yeh, S.F.; Chang, T.J.; Tsai, M.H.; Chen, K.; Chang, Y.J. A Conversation Analysis of Non-Progress and Coping Strategies with a Banking Task-Oriented Chatbot. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI'20), Honolulu, HI, USA, 25–30 April 2020; pp. 1–12. [CrossRef]
15. Sidaoui, K.; Jaakkola, M.; Burton, J. AI feel you: Customer experience assessment via chatbot interviews. *J. Serv. Manag.* **2020**, *31*, 745–766. [CrossRef]
16. Ho, A.; Hancock, J.; Miner, A.S. Psychological, Relational, and Emotional Effects of Self-Disclosure after Conversations with a Chatbot. *J. Commun.* **2018**, *68*, 712–733. [CrossRef] [PubMed]

17. Tsai, W.H.S.; Lun, D.; Carcioppolo, N.; Chuan, C.H. Human versus chatbot: Understanding the role of emotion in health marketing communication for vaccines. *Psychol. Mark.* **2021**, *38*, 2377–2392. [CrossRef]
18. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
19. Hershkowitz, I.; Orbach, Y.; Lamb, M.E.; Sternberg, K.J.; Horowitz, D. Dynamics of Forensic Interviews with Suspected Abuse Victims who do not Disclose Abuse. *Child Abus. Negl.* **2006**, *30*, 753–769. [CrossRef]
20. Yi, M.; Jo, E.; Lamb, M.E. Effects of the NICHD protocol training on child investigative interview quality in Korean police officers. *J. Police Crim. Psychol.* **2016**, *31*, 155–163. [CrossRef]
21. Sternberg, K.J.; Lamb, M.E.; Hershkowitz, I.; Yudilevitch, L.; Orbach, Y.; Esplin, P.W.; Hovav, M. Effects of introductory style on children's abilities to describe experiences of sexual abuse. *Child Abus. Negl.* **1997**, *21*, 1133–1146. [CrossRef]
22. Yi, M.; Jo, E.; Lamb, M.E. Assessing the Effectiveness of NICHD Protocol Training Focused on Episodic Memory Training and Rapport-Building: A Study of Korean Police Officers. *J. Police Crim. Psychol.* **2017**, *32*, 279–288. [CrossRef]
23. Saywitz, K.J.; Camparo, L.B. Contemporary child forensic interviewing: Evolving consensus and innovation over 25 years. In *Children as Victims, Witnesses, and Offenders: Psychological Science and the Law*; Guilford Press: New York, NY, USA, 2009; pp. 102–127.
24. Malloy, L.C.; Brubacher, S.P.; Lamb, M.E. "Because She's One Who Listens" Children Discuss Disclosure Recipients in Forensic Interviews. *Child Maltreat.* **2013**, *18*, 245–251. [CrossRef] [PubMed]
25. Lamb, M.E.; Sternberg, K.J.; Orbach, Y.; Hershkowitz, I.; Horowitz, D.; Esplin, P.W. The Effects of Intensive Training and Ongoing Supervision on the Quality of Investigative Interviews with Alleged Sex Abuse Victims. *Appl. Dev. Sci.* **2002**, *6*, 114–125. [CrossRef]
26. Lamb, M.E.; Orbach, Y.; Hershkowitz, I.; Esplin, P.W.; Horowitz, D. A structured forensic interview protocol improves the quality and informativeness of investigative interviews with children: A review of research using the NICHD Investigative Interview Protocol. *Child Abus. Negl.* **2007**, *11–12*, 1201–1231. [CrossRef] [PubMed]
27. Steller, M.; Köhnken, G. Statement analysis: Credibility assessment of children's testimonies in sexual abuse cases. In *Psychological Methods in Criminal Investigation and Evidence*; Springer: Berlin/Heidelberg, Germany, 1989; pp. 217–245.
28. Wei, J.; Zou, K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. *arXiv* **2019**, arXiv:1901.11196.
29. Dhiman, A.; Toshniwal, D. An Enhanced Text Classification to Explore Health based Indian Government Policy Tweets. *arXiv* **2020**, arXiv:2007.06511.
30. Dai, X.; Adel, H. An Analysis of Simple Data Augmentation for Named Entity Recognition. *arXiv* **2020**, arXiv:2010.11683.
31. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
32. Available online: https://bitbucket.org/eunjeon/mecab-ko-dic/ (accessed on 20 July 2018).
33. AI-Hub. Available online: https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=117 (accessed on 1 October 2023).
34. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
35. Jo, H.; Goo Lee, S. *Korean Word Embedding Using FastText*; The Korean Institute of Information Scientists and Engineers: Busan, Republic of Korea, 2017; pp. 705–707.
36. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
37. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
38. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Wang, S.; Hu, G. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, Online, 16–20 November 2020; pp. 657–668.
39. Kikuta, Y. BERT Pretrained Model Trained on Japanese Wikipedia Articles. 2019. Available online: https://github.com/yoheikikuta/bert-japanese (accessed on 20 October 2019) .
40. Amer, E.; Hazem, A.; Farouk, O.; Louca, A.; Mohamed, Y.; Ashraf, M. A Proposed Chatbot Framework for COVID-19. In Proceedings of the 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), Cairo, Egypt, 26–27 May 2021; pp. 263–268.
41. Lee, J.H.; Wu, E.H.K.; Ou, Y.Y.; Lee, Y.C.; Lee, C.H.; Chung, C.R. Anti-Drugs Chatbot: Chinese BERT-Based Cognitive Intent Analysis. *IEEE Trans. Comput. Soc. Syst.* **2023**, 1–8. [CrossRef]
42. Fernández-Martínez, F.; Luna-Jiménez, C.; Kleinlein, R.; Griol, D.; Callejas, Z.; Montero, J.M. Fine-Tuning BERT Models for Intent Recognition Using a Frequency Cut-Off Strategy for Domain-Specific Vocabulary Extension. *Appl. Sci.* **2022**, *12*, 1610. [CrossRef]
43. SKT-Brain. Korean BERT Pre-Trained Cased (KoBERT). 2021. Available online: https://github.com/SKTBrain/KoBERT (accessed on 20 August 2022).
44. How multilingual is Multilingual BERT? *arXiv* **2019**, arXiv:1906.01502
45. Schuster, M.; Nakajima, K. Japanese and Korean Voice Search. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 5149–5152.

46. Kudo, T.; Richardson, J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv* **2018**, arXiv:1808.06226.
47. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
48. Available online: https://pypi.org/project/kochat/ (accessed on 1 October 2023).

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.