Text Implicates Prosodic Ambiguity: A Corpus for Intention Identification of the Korean Spoken Language

WON IK CHO, Seoul National University, Dept. ECE and INMC
NAM SOO KIM, Seoul National University, Dept. ECE and INMC

Phonetic features are indispensable in understanding the spoken language. Especially in Korean, which is wh-in-situ and head-final, the addressee of spoken language sometimes finds it hard to discern the speaker’s original intention if not provided with the sentence prosody. However, acoustic information may not be guaranteed for all spoken language processing, due to the difficulty of managing and computing speech data. This paper suggests a corpus that aims to distinguish utterances with ambiguous intention from clear-cut ones, utilizing the prosodic ambiguity of the text input. In detail, the resulting classification system decides whether the given text input is one of fragment, statement, question, command, rhetorical question/command, or indecisive, taking into account the intonation-dependency of the text. Based on an intuitive understanding of the Korean language engaged in the data annotation, we construct a corpus with seven intention categories, train classification systems, and validate the utility of our dataset with quantitative and qualitative analyses.

CCS Concepts: • Computing methodologies → Natural language processing; Discourse, dialogue and pragmatics; Language resources.

Additional Key Words and Phrases: Korean spoken language, Speech act, Intention identification, Prosodic ambiguity, Directiveness, Rhetoricalness, Intonation-dependency

1 INTRODUCTION

Understanding speech intention includes all aspects of syntax, semantics, and phonetics. For example, even when the transcript of a spoken utterance is given and the syntactic structure is apparent, the speech act may differ depending on dialogue or social context [45]. Besides, phonetic features such as prosody can influence the genuine intention, which can be different from the speech act that is grasped from the textual information [1].

In conventional spoken language processing pipelines, namely automatic speech recognition (ASR) and intention identification, phonetic features tend to be inadvertently removed during speech transcription. For instance, transcripts usually do not contain punctuation which is sometimes essential for proper understanding of spoken utterances. While more accurate understanding of the intention is expected if phonetic features are available, as observed in co-utilization of audio and text [5, 13], such acoustic information is not always guaranteed in real-world spoken language processing since handling speech data requires quite a significant burden compared to lightweight text data. This sometimes incurs ‘prosodic ambiguity’ [12], the possibility of diverse interpretation of the given text regarding the acoustic feature of speech.
Such a phenomenon may not threaten the spoken language understanding (SLU) in many languages, especially if the lexical usage is straightforward given the sentence form and type (as in English, where interrogatives and imperatives are distinguished from declaratives in general). However, it matters if the prosodic ambiguity significantly affects the intention understanding of transcribed utterances. Up-to-date end-to-end models tackle such weaknesses of pipeline approaches, but they are often vulnerable to unexplainable errors. Instead, we deemed it worthwhile to first tell utterances whose intentions are determined solely upon the text, from those that incorporate prosodic ambiguity, preserving the conventional pipeline structure. The further processing of ambiguous texts with context or audio will help the whole system reach the desired decision.

In this study, the language of interest is Korean, a wh-in-situ language with the head-final syntax. Natural language processing in Korean is known to be burdensome, not only because the Korean language is agglutinative and morphologically rich, but also for frequent pro-drop and high context-dependency. Moreover, to make it challenging to understand the utterance meaning only by text, the intention of certain types of sentences is significantly influenced by prosodic information such as the intonation of sentence ender [20]. Consider the following sentence, of which the meaning depends on the sentence-final intonation:

(1) 천천히 가고 있어
   chen-chen-hi ka-ko iss-e
   slowly go-PROG\textsuperscript{1} be-SE\textsuperscript{2}

With a high rise intonation, the sentence becomes a question (Are you/they going slowly?), and given a fall or fall-rise intonation, it becomes a statement (I am going slowly.). Also, given a low rise or level intonation, the sentence becomes a command (Go slowly.). This phenomenon partially originates in particular constituents of Korean utterances, such as multi-functional particle –어 (-e), or other sentence enders determining the sentence type [35]. Although similar tendencies are observed in other languages as well (e.g., declarative questions in English [14]), syntactical and morphological properties of the Korean language strengthen the ambiguity of spoken utterances.

Here, we propose a corpus that can help identify the intention of a spoken Korean utterance, particularly when there are textual utterances that can have diverse meanings depending on the intonation. The system trained upon our corpus classifies an input utterance into seven speech act categories of fragment, statement, question, command, rhetorical question ·command, and intonation-dependent, where the final one suggests that the intention is indecisive and the decision requires further acoustic information. A total of 61,225 lines of text utterances were annotated or generated, including about 20K lines manually tagged with a fair agreement. We claim the following as our contribution:

- A new kind of text annotation scheme that reflects prosodic sensitivity of Korean utterances (effective, but not restricted, to a head-final language)
- A freely available corpus of 61K Korean sentences with seven categories (including human-annotated 20K), validated with conventional and up-to-date pretrained language models

In the following section, we take a look at the literature on intention classification and demonstrate the background of our categorization. In the next section, the corpus construction process is described with a detailed annotation scheme. Afterwards, the trained system is evaluated quantitatively and qualitatively with the test set. Besides, we will briefly explain how our methodology can be effective in real-world applications.

\textsuperscript{1}Denotes a progressive marker.
\textsuperscript{2}Denotes the underspecified sentence enders; final particles whose role vary.
2 BACKGROUND

Our study is most significantly influenced by the existing work on sentence-level semantics. Unlike the syntactic concept presented in Sadock and Zwicky [40], the speech act or intention\(^3\), has been studied in the area of pragmatics, especially along with illocutionary act and dialogue act (DA) [42, 45].

Among all, we concentrate on lessening vague intersections between the classes, which can be observed between e.g., statement and opinion [45], in the sense that some statements can be regarded as an opinion and vice versa. Thus, slightly different from apparent boundaries between the sentence forms declaratives, interrogatives, and imperatives [40], we extend them to syntax-semantic level adopting discourse component (DC) [38]. It involves common ground, question set, and to-do-list: the constituent of sentence types that comprise natural language. We interpret them in terms of speech act, considering the obligation that the sentence impose on the listeners; whether to answer (question), to react (command), or neither (statement).

Building on the concept of discourse component revisited, we take into account the rhetoricalness of question and command, which yields additional categories of rhetorical question (RQ) [39] and rhetorical command (RC) [19]. We claim that a single utterance falls into one of the five categories, or be classified as a fragment, if given enough acoustic information\(^4\). Our categorization is much simpler than the conventional DA tagging schemes [3, 45] and is rather close to tweet act [49] or situation entity types [11], but less relies on the dialogue history and is more suitable with short command conditions. We will show that our scheme can be helpful in spoken language processing by discerning the prosodic ambiguity from text, which is achieved by introducing a new class of intonation-dependent utterances.

**Intention and speech act.** At this point, it may be beneficial to point out that the term intention or speech act is to be used as a domain non-specific indicator of the utterance type. We argue that two terms are different from intent, which is used as a specific action in the literature [15, 29, 44], along with the concept of item, object and argument, generally for domain-specific tasks. Also, unlike dialogue management, where a proper response is created upon the dialogue history [27], the proposed system aims to find the genuine intention of a single input utterance and guide the following direction.

**Korean sentence types.** For corpus annotation, research on Korean sentence types is essential. Although the annotation process partly depends on the intuition of annotators, we refer to the work related to syntax-semantics and speech act of Korean [16, 34, 43], to handle some unclear cases regarding optatives, permissives, promisives, requests, suggestions, and rhetorical questions. We expect this research to be cross-lingually extended, but first, we concentrate on resolving the aforementioned language-specific problems.

3 ANNOTATION PROTOCOL

We clarify the annotation concept of the proposed corpus, which can be adopted to train a text-level classifier for spoken language\(^5\). As its motivation is briefly introduced in Section 1, we primarily aim to discern the existence of prosodic ambiguity that is determined upon the lexical features.

\(^3\)In this paper, intention and act are often used interchangeably. In principle, the intention of an utterance is the object of grasping, and the act of speech is a property of the utterance itself. However, we denote determining the act of a speech, such as question and demand, as inferring the intention.

\(^4\)More elaborate definition of fragments and intonation-dependency will be discussed in the next section.

\(^5\)Throughout this paper, text refers to the sequence of symbols (or letters) with the punctuation marks removed, which is a frequent output format of speech recognition. Also, sentence and utterance are interchangeably used to denote an input, where usually the latter implies an object with intention while the former does not necessarily.
We assume that the Korean sentences in our study can be assigned with one of five intention categories, namely statement, question, command, rhetorical question, and rhetorical command. However, only given a single sentence (or a sentence-like text), one might not be able to determine the exact category. First, the annotator should check if the given sentence is a fragment or not, where fragments (FR) denote a single word or a chunk whose intention is underspecified under our criteria. Next, if the sentence is not necessarily determined as a fragment, the annotator may check if the sentence connotes some intentions among the five candidates, and whether the intention can be decided as a unique one. If the decision is not feasible due to the prosodic ambiguity, the sentence is labeled as an intonation-dependent utterance (IU). If the sentence is uniquely determined as one of the pre-defined categories, we call such utterance a clear-cut case (CC), and it includes the above five utterance types.

A brief illustration of the annotation process is depicted in Fig. 1. For a detailed description, we describe each sentence type in the order of FR, CCs and IU, with example sentences.

### 3.1 Fragments (FR)

From a linguistic viewpoint, fragments often refer to a single noun-verb phrase where ellipsis occurred [31]. However, colloquial expressions often show omission, replacement, and/or scrambling, hindering us from applying the same definition as the written language. Thus, in this study, we also count some sentence segments whose intention is underspecified. If the input sentence is not a fragment, it is assumed to belong to clear-cut cases or be an intonation-dependent utterance.

Some might argue that fragments can be interpreted as command or question under some circumstances. For instance, simply uttering a noun in a rising intonation can be interpreted as an echo question, and loudly uttering some objects can be considered as a command to bring it on. We observed that a large portion of the intention concerning context is represented in the prosody, which leads us to define prosody-sensitive cases afterwards.

However, for fragments, we found it difficult to assign a specific intention to them even given audio, since they highly rely on the dialogue or situational context. Interpreting a single noun as an echo question requires the existence of the original question, and uttering some objects as a command requires

---

ACM Trans. Asian Low-Resour. Lang. Inf. Process.
the circumstance that the speaker urgently demands the addressee. That is, discerning such implication is not usually feasible, especially in a short command context. Thus, we decided to leave the intention of fragments underspecified, and let them be combated with the help of the context in real-world usage. Here are some examples of fragments:

(2a) 마우스
mawusu
mouse

(2b) 키보드와 마우스
kipodu-wa mawusu
keyboard-AND mouse

(2c) 마우스로
mawusu-lo
mouse-WITH

with mouse

Not only a single word (2a) is a fragment, but a noun phrase (2b) or a postposition phrase (2c) can also be the case. We concluded that determining the intention of such phrases requires the dialogue history even if the prosody is given.

3.2 Clear-cut cases (CCs)

Clear-cut cases include utterances of five categories: statement, question, command, rhetorical question, and rhetorical command, as described detailed in the annotation guideline with examples. Questions are utterances that require the addressee to answer (3a,b), and commands are ones that require the addressee to act physically or psychologically (3c,d). Even if the sentence form is declarative, words such as wonder or should can let the sentence be a question or command. Statements are descriptive and expressive sentences that do not apply to both cases (3e).

(3a) 너 집에 갈거니
ne cip-ey kal-ke-ni
you home-to go-PRT-INT
Will you go home?

(3b) 내일 날씨 좀 알려줘
nayil nalssi com ally.e-cwu.e
tomorrow weather POL inform.PRT-give.SE
Please tell me tomorrow’s weather.

(3c) 세시 반에 나 좀 깨워
sey si pan-ey na com kkaywu.e
three hour half-at I POL wake.SE

\(^6\)Currently uploaded online in Korean. https://docs.google.com/document/d/1-dPL5MfsxLbWs7vfwczTKqBq_1DX9u1wOqOGPltOQs
\(^7\)Denotes a functional particle.
\(^8\)Denotes an interrogative ender.
\(^9\)Denotes a polarity item for the politeness in asking something.

ACM Trans. Asian Low-Resour. Lang. Inf. Process.
Please wake me up at three thirty.

(3d) 목소리 좀 낮추는 게 어때
moksooli com nacchwu-nun key ettay
voice POL lower-PRT thing.NOM how
How about lowering your voice?

(3e) 아무래도 내일 날씨가 떨어질 것 같아
amwulayto nayil nasudak tteleci.l kes kath-a
anyway tomorrow NASDAQ drop.FUT11 thing seem-SE
I have a feeling that NASDAQ may drop tomorrow.

Rhetorical questions (RQ) are questions that do not require an answer because it is already in the speaker’s mind (4a) [39]. Similarly, rhetorical commands (RC) are idiomatic expressions in which imperative structure does not convey a to-do-list that is mandatory (e.g., Have a nice day, (4b)) [16, 19]. Sentences in these categories are functionally similar to statements but are categorized as separate classes since they usually show a non-neutral tone.

(4a) 너 돈 벌기 싫니
ne ton pel-ki silih-ni
you money earn-PRT dislike-INT
Don’t you want to make money? (= It seems that you are not interested in making money.)

(4b) 쏴 태면 써 봐
sso.l tey-myen sso.a po.a
shoot.FUT thing.NOM-if shoot.PRT see.SE
Shoot me if you can. (= You won’t be able to shoot me.)

In making up the guideline, we carefully looked into the dataset so that the annotation can cover ambiguous cases. As stated in the previous section, we refer to Portner [38] to borrow the concept of discourse component and extend the formal semantic property to the level of pragmatics. This indicates that we search for a question set (QS) or to-do-list (TDL) which makes an utterance directive in terms of speech act [42], taking into account non-canonical and conversation-style sentences which contain idiomatic expressions and jargon. If we cannot find such components (QS for asking a question, TDL for asking an action), the utterance is determined to display a discourse component of common ground (CG). We provide a simplified criterion in Table 1, where the discourse components (CG, QS, and TDL) imply the core concept of the sentence and sentence forms denote the syntactical property of the sentence ender.

3.3 Intonation-dependent utterances (IU)

Given the decision criteria for clear-cut cases, we further investigate whether the intention of a given sentence can be determined without information on prosody or intonation. That is, we consider the potential interpretation of an utterance in case it is projected to a textual form, when even the punctuation is omitted or not adequately transcribed with an ASR system. Sentence (1) in Section 1, which is not accompanied by punctuation but is ambiguous, describes such cases.

Although there have been studies on Korean sentences which handle final particles and adverbs [5, 32], to the best of our knowledge, there has been no explicit guideline on a text-based identification of

10 Denotes a nominative case.
11 Denotes a future tense.

ACM Trans. Asian Low-Resour. Lang. Inf. Process.
Table 1. A simplified annotation scheme regarding discourse component and sentence form: Discourse component (DC) in the table implies the concept that extends the original formal semantic property [38] to speech act level.

| Discourse component / Sentence form | Common Ground | Question Set | To-do List |
|-------------------------------------|---------------|--------------|------------|
| Declaratives                        | Statements, RQ, RC | Question     | Command    |
| Interrogatives                      | RQ            | Question     | Command    |
| Imperatives                         | RC            | Question     | Command    |

utterances that incorporates prosodic ambiguity. On top of this, we set up some principles, or rules of thumb, concerned with the empirical result of our data analysis. Note that the last two (e,f) are closely related to the maxims of conversation [26], e.g., “Do not say more than is required.” or “What is generally said is stereotypically and specifically exemplified.”.

(a) Take into account possible prosody/intonation of a text input, given no non-lexical information such as emojis and punctuation. Remember that the sentence-final part mainly concerns the intonation-dependency of the intention.

(b) A wh-particle is interpreted as an existential quantifier in the case of wh-intervention due to Korean being wh-in-situ, changing the wh-questions to another type of question or a statement.

(c) Since the subject is dropped in many Korean spoken utterances, one may have to assign all the agents (1st to 3rd) in investigating the sentence type, which depends on the intention. In this process, an awkward combination can be ignored. For instance,

| (5a) 오늘 뭐 먹고 싶어 onul mwe mek-ko siph-e today what eat-PRT want-SE |

- can be interpreted as either “I wanna eat something today.” or “What do you want to eat today?”
  - Depending on the prosody around mwe (what or something), making the sentence either a statement or wh-question. Refer to (b).
  - In this process, for the former case, the sentence-final intonation falls and the reverse holds for the latter. Refer to (a).
  - At the same time, it can be inferred without awkwardness that for the statement, the covert subject turns out to be the speaker (I), and for the question, it becomes the addressee (you).

(d) The presence of vocatives can sometimes restrict the role of the utterance. For instance, in the preceding example, if a vocative ‘누나’ (nwuna, deixis for an older sister, used mainly by male speakers) is augmented at the start of the sentence (5b), it is much more plausible to interpret the sentence as:

| (5b) 누나 오늘 뭐 먹고 싶어 nwuna onul mwe mek-ko siph-e nwuna today what eat-PRT want-SE |

What do you want to eat today, [the name of the older sister]?

ACM Trans. Asian Low-Resour. Lang. Inf. Process.
(e) Adding adverbs or numeric polarity items may not always preserve the intention of the sentence. Therefore, one should be aware of the loss of felicity in the interpretation (as to a specific speech act) that is induced by introducing such components. For instance, in Korean, 좀 (com, slightly) or 하나 (hana, one) are respectively an adverb and numeric polarity item that induce politeness, as seen in (3b,c). Again, in (5c,d), com and hana can come right after mwe to cautiously convey that the speaker wants to eat something today (and the addressee may feel an obligation to eat something together with the speaker).

(5c) 오늘 뭐 좀 먹고 싶어
   onul mwe com mek-kok sih-e
   today what slightly eat-PRT want-SE
   I think I want to eat something today.

(5d) 오늘 뭐 하나 먹고 싶어
   onul mwe hana mek-kok sih-e
   today what one eat-PRT want-SE
   I think I should eat something today.

(f) Some sentences can have both an underspecified sentence ender (that can let the sentence be either a question or statement) and excessively specific information. Although the sentence form is not a direct link to the intention, in that case, the sentence is more likely to be determined as a statement rather than a declarative question. This matches with the intuition that it is not felicitous to ask too specific information as a question, except for some affirmative questions. For instance, if a specific cuisine comes in place of mwe (what) in (5a), then it becomes less felicitous to interpret it as a question, like:

(5c) 오늘 뜨끈한 국밥 먹고 싶어
   onul ttukkun-han kwukpap mek-kok sih-e
   today warmy gukbap eat-PRT want-SE
   I want to eat a warmy gukbap today.

Here, mwe is replaced with ttukkun-han kwukpap, a warmy stew with rice, which makes the sentence more plausible to be interpreted as a statement, or a declaration that the speaker wants to eat a specific cuisine, rather than a question.

4 CORPUS BUILDING
4.1 Source scripts

To cover a variety of topics, utterances used for the annotation were collected from (i) a corpus provided by Seoul National University Speech Language Processing Lab\(^{12}\), (ii) a set of frequently used lexicons, released by the National Institute of Korean Language\(^{13}\), and (iii) manually created questions/commands. In specific, (i) contains short utterances with topics covering e-mail, housework, weather, transportation, stock, etc. (ii) is an official Korean word dictionary organized in lexicographical order, and (iii) was created by Seoul Korean speakers based on the annotation scheme of question and command.

\(^{12}\)http://slp.snu.ac.kr/
\(^{13}\)https://www.korean.go.kr/
| Categories (total 7 classes) | Intention   | Instances | Corpus 20K | Whole |
|-----------------------------|-------------|-----------|------------|-------|
| Fragment                    | -           | 384       | 6,009      |       |
| Clear-cut cases             |             |           |            |       |
| Statement                   | 8,032       | 18,300    |            |       |
| Question                    | 3,563       | 17,869    |            |       |
| Command                     | 4,571       | 12,968    |            |       |
| Rhetorical Q.               | 613         | 1,745     |            |       |
| Rhetorical C.               | 572         | 1,087     |            |       |
| Intonation-dependent utterance | Unknown (among 5 candidates) | 1,583 | 3,277 |
| Total                       | 19,318      | 61,255    |            |       |

4.2 Agreement

From (i), 20K lines were randomly selected, and three Seoul Korean L1 speakers classified them into seven categories of fragments, intonation-dependent utterances, and five clear-cut cases (Table 2, Corpus 20K). Annotators were well informed on the guideline and had enough debate on the conflicts during the annotating process. The resulting inter-annotator agreement (IAA) was $\kappa = 0.85$ [10] and the final decision was made by majority voting and adjudication.

4.3 Augmentation

Considering the shortage of certain types of utterances in Corpus 20K, (i)-(iii) were utilized in the data supplementation. First, we trained a simple classifier with Corpus 20K. Then, we extracted rhetorical questions and commands, and statements, from the rest of (i). We checked and relabeled the outcome to supplement each category. Next, in (ii), about 6,000 Korean words were investigated, and only single nouns were collected and augmented to fragments. Finally, for (iii), paid participants made up question and command given topics of e-mail, housework, weather, and schedule, which are frequent categories appearing in Corpus 20K. With a total of 20,000 sentences created, where most of the portion belonged to questions or commands, the authors manually checked the outcome and relabeled some of them as statements or IU. The composition of the final dataset is stated in Table 2.

4.4 Train split

The Whole corpus was split into train, validation, and test set, for model-based experiments. Seven classes of utterances were distributed with balance in each set. The size of sets reach 49,620, 5,514, and 6,121, respectively. The dataset is available at https://huggingface.co/datasets/kor_3i4k and the validation set is obtained by splitting the last 10% of the train set, in the currently uploaded version.
5 EXPERIMENT

5.1 Models

To check how our annotation scheme works with the machine learning-based classification algorithms, we investigate the training and validation process with conventional architectures such as convolutional neural network (CNN, [21]) or bidirectional long short-term memory (BiLSTM, [41]) along with fastText [2] word vectors, and up-to-date pretrained language models (PLMs) such as bidirectional encoder representations from Transformers (BERT, [9]) and ELECTRA [8].

5.1.1 Conventional architectures. Conventional architectures include CNN [21, 23], BiLSTM [41], and self-attentive BiLSTM (BiLSTM-Att [28]). For CNN, two convolution layers were stacked with max-pooling layers in between, summarizing the distributional information lying in an input vector sequence. For BiLSTM, the hidden layer of a specific timestep was fed together with the input of the next timestep, to infer the subsequent hidden layer in an autoregressive manner. For a self-attentive embedding, the context vector whose length equals that of the hidden layer of BiLSTM, was jointly trained along with the network to provide the weight assigned to each hidden layer. The input format of BiLSTM equals that of CNN except for the channel number which was set to 1 (single channel) in the CNN model.

For the input featurization of conventional architectures, we tokenized sentences into character-level and adopted 100-dim fastText dense vectors [2] that correspond to each character. Although the featurization of conventional architectures may not fully match the data-driven representation of BERT-like models, we aimed to accommodate language model pretraining that may make models compatible with up-to-date PLMs. Thus, we exploited the word vector pretrained with 200M lines of drama scripts instead of one-hot vectors or TF-IDF, which was reported to display a satisfactory result with spoken language processing tasks such as word segmentation [4], publicly available in a Github repository.

5.1.2 Pretrained language models. For BERT-like pretrained language models (PLMs), we adopted multilingual BERT (mBERT), KoBERT [47], KcBERT [24], KoELECTRA [36], KcELECTRA [25], and KLUE-BERT [37], which are all currently available in Hugging face Transformers library [51]. mBERT, KoBERT, KcBERT, and KLUE-BERT follow Devlin et al. [9] that builds a bidirectional encoding upon Transformer [48], where the pretraining aims to optimize the model to two subtasks of masked language model (MLM) and next sentence prediction (NSP). KoELECTRA and KcELECTRA utilize replaced token prediction (RTD) of ELECTRA [8], which strengthens the model from the perspective of logical reasoning. mBERT and KoBERT are pretrained with written-style texts such as Wikipedia. KoELECTRA and KLUE-BERT utilize a large amount of texts available online, including a small number of spoken texts from messages and web data [33]. KcBERT and KcELECTRA are pretrained with online news comments that are much more colloquial and informal than written text. Note that input features of PLMs are all customized tokens, where the token set differs by the model utilized.

5.2 Implementation

All conventional architectures were implemented with Keras Python library [7]. CNN includes two convolutional layers of window size 3, with one max pooling layer in between, and BiLSTM is made up of two 64-dim forward and backward LSTM layers. For both architectures, the maximum length was set to 50, and empty areas were padded with zeros. The word vector size was fixed to 100, while CNN had a single channel with 32 filters. For self-attentive BiLSTM, the context vector was set up with the

14https://github.com/warnikchow/raws
15https://github.com/huggingface/transformers
Table 3. Test result (accuracy) with conventional architectures and PLMs. Dictionary size for the conventional architectures indicates the number of character vectors. Epochs denote the training from scratch for conventional models and fine-tuning for large-scale PLMs. In pretraining, ‘Mono’ denotes that the model pretraining was done with monolingual data, while ‘Multi’ denotes the multilingual case. ‘Mono (Emb)’ means that the pretraining was done only for the embedding vectors (with fastText), not the weight for the whole architecture.

| Model       | Feature (dimension - length)     | Performance | Pretraining  | Dictionary size | Epochs |
|-------------|----------------------------------|-------------|--------------|-----------------|--------|
| CNN         | Dense fastText vector (100 - 50) | 87.06       | Mono (Emb)   | ~2,500          | 50     |
| BiLSTM      | Dense fastText vector (100 - 50) | 88.07       | Mono (Emb)   | ~2,500          | 50     |
| BiLSTM-Att  | Dense fastText vector (100 - 50) | 88.69       | Mono (Emb)   | ~2,500          | 50     |
| mBERT       | Tokenized raw text (768 - 512)   | 89.56       | Multi        | ~120,000        | 3      |
| KoBERT      | Tokenized raw text (768 - 512)   | 61.73       | Mono         | 8,000           | 3      |
| KcBERT      | Tokenized raw text (768 - 512)   | 91.08       | Mono         | 30,000          | 3      |
| KoELECTRA   | Tokenized raw text (768 - 512)   | 92.47       | Mono         | 35,000          | 3      |
| KcELECTRA   | Tokenized raw text (768 - 512)   | 92.08       | Mono         | 50,000          | 3      |
| KLUE-BERT   | Tokenized raw text (768 - 512)   | 91.95       | Mono         | 32,000          | 3      |

same size as the LSTM hidden layers (64). The optimization was done with Adam (5e-4) [22], with batch size 16, and a dropout rate of 0.3. The model for the test was chosen as the best performing one with the validation set, after training for 50 epochs.

Up-to-date PLMs were adopted from Hugging face model hub; namely mBERT16, KoBERT17, KcBERT18, KoELECTRA19, KcELECTRA20 and KLUE-BERT21 and follows the default setting. mBERT is a multilingual model for around 100 languages and utilizes 119,547 tokens for the dictionary, while other five are monolingual models and utilize 8,002 (KoBERT), 30,000 (KcBERT), 35,000 (KoELECTRA), 50,135 (KcELECTRA), and 32,000 (KLUE-BERT) vocab for dictionary, respectively. All the tokens were projected to 768 dimension output layers, while the length was set to 512 following the original Transformers setting. The dropout rate was set to 0.1, with Adam optimizer (1e-4)22 and linear scheduler with 100 warm-up steps23. The training with batch size 32 ran for three epochs which is sufficient for the fine-tuning of the created data, and the final trained model was directly adopted for the test.

5.3 Result

Table 3 shows the performance of conventional architectures and up-to-date PLMs, where all results were obtained by inferring the test set.

16https://huggingface.co/bert-base-multilingual-cased
17Originally provided in https://github.com/SKTBrain/KoBERT, and the served version for Hugging face Transformers was available in https://huggingface.co/monologg/kobert
18https://huggingface.co/beomi/kcbert-base
19https://huggingface.co/monologg/koelectra-base-v3-discriminator
20https://huggingface.co/beomi/KcELECTRA-base
21https://huggingface.co/klue/bert-base
22We additionally set weight decay 0.01, Adam beta1= 0.9, Adam beta2= 0.95, and Adam epsilon 1e-8.
23The optimization scheme for PLMs was more delicately set due to the sensitivity of models.
5.3.1 Quantitative analysis. Among all the conventional architectures and up-to-date PLMs, KoELECTRA, which is pretrained upon both colloquial and written texts with an adequate size of vocabulary, exhibited the highest accuracy. This proves that strategies for language model pretraining and the property of source corpora both benefit the classification performance for our dataset.

PLMs outperform conventional architectures in general, but not all. It is notable that not all the fine-tuned PLMs outperform conventional architectures, which differs from recent reports that PLMs leveraging information from massive corpora have an advantage over models trained solely upon the target task. In our experiment, CNN and BiLSTM(-Att) modules showed competitive performance with some BERT modules, and KoBERT, with the smallest dictionary size among PLMs, seems to fail to outperform conventional architectures.

Pretraining corpus influences the result. We analyze that the result is also influenced by the type of source corpora utilized in pretraining of fastText word vectors or PLMs. Different from other PLMs of which the source corpus of pretraining includes colloquial texts, training corpora for mBERT and KoBERT are more concentrated on written texts such as Wikipedia, which may not fit with the processing of the spoken language. In the proposed task, some utterances are more challenging to categorize due to prosodic cues that are not explicit in the textual form. Such property may have made it difficult for mBERT or KoBERT to meet the desired standard, at the same time guaranteeing the competitive performance of conventional modules where the fastText-based word vectors are trained upon colloquial and non-normalized drama scripts [4].

Less sensitive to OOV and follows scaling laws. It is also noteworthy that mBERT, trained upon multilingual vocab and corpora, outperforms KoBERT, which bases on similar monolingual corpora. This suggests that our dataset is less vulnerable to out-of-vocabulary issues that lie in mBERT with shortened Korean Hangul vocabs (about 3.3K). Instead, it can be inferred that models follow the scaling laws for neural language models [18], as can be observed similarly in KcBERT and KcELECTRA, or KLUE-BERT and KoELECTRA (though weakly significant).

Data fits with models. Despite some results beyond expectation, it is still encouraging that PLMs show adequate performance only with simple fine-tuning of three epochs. In future, the updated PLMs pretrained with more various spoken language corpora and advanced strategies may show higher performance with lightweight architectures, which can be helpful for the real-world application of the trained module.

5.3.2 Further investigation using PLMs. As using PLMs is de facto in recent literature, we conducted a further investigation to help understand how the constructed dataset can be utilized in analyses and practice. In Table 4, we compare the size and domain of the pretraining corpora of PLMs, referring to Hur et al. [17], Yang [52], and how they perform in various classification scenarios.

For all the PLMs, pretrained weight was freezed and we conducted additional training for a single fully-connected network added on the highest 768-dim layer regarding [CLS] token of the input. For the statistical validation, we had several trials for each scenario and defined ‘error’ as a standard deviation of results divided by the average (normalized standard deviation)24. Also, to see how the dataset can be used in multi-stage scenarios such as first distinguishing IUs from fragments and clear-cut cases, we added experimental results on 3-fold scenarios (FR, CCs, IU). For both 7-fold and 3-fold classification, we accompany the accuracy on IUs.

24The performance for the 7-fold scenario slightly differs from Table 3, which recorded the best score, since the score is averaged after five repetitions with different initialization.
Table 4. Comparison table of pretraining corpora and performance of each PLM module. Note that the size and domain of mBERT denote the pretraining corpora regarding all the languages which are relevant, that size cannot be specified to a specific language.

| Model       | Pretraining corpora                                      | Performance (Average)                |
|-------------|----------------------------------------------------------|--------------------------------------|
|             | Size | Domain                                   | 7-fold (IU) | Error (%) | 3-fold (IU) |
| mBERT       | 2.5B (words) | Wikipedia (of 104 languages) | 89.57 (66.65) | 0.19 | 93.12 (17.23) |
| KoBERT      | 5.4M (words) | Korean Wikipedia                         | 52.87 (22.23) | 20.19 | 92.40 (0) |
| KcBERT      | 12GB | Korean online news comments               | 90.93 (69.92) | 0.11 | 94.76 (41.63) |
| KoELECTRA   | 34GB | Korean Wikipedia, Namu Wiki, Newspaper, Messages, Web, etc. | 91.98 (72.86) | 0.36 | 96.37 (68.13) |
| KcELECTRA   | 17GB | Korean online news comments               | 91.95 (72.16) | 0.11 | 96.72 (70.81) |
| KLUE-BERT   | 63GB | Modu Corpus [33], CC-100-Kor [50], Namu Wiki, Newspaper, Petition dataset, etc. | 91.72 (72.18) | 0.20 | 96.13 (65.09) |

First, as discussed in the previous section, the size of pretraining corpus seems to influence the performance, considering that mBERT outperforms KoBERT and so for KcBERT, KcELECTRA, and KoELECTRA. However, given that KcELECTRA shows almost the same performance as KoELECTRA in 7-fold and even outperforms in 3-fold despite its half-sized pretraining corpus, it seems that how the model is familiar with colloquial text is crucial to the practical utilization of the proposed dataset. In other words, effective fine-tuning using the dataset requires more domain-specific (especially prosodic and phonetic) linguistic knowledge, such as sentence structure for spoken language that helps disambiguate the role of polarity items or sentence enders that can completely change or diversify the meaning of utterances. Also, it seems that concentrating on domain-specific dictionary seems to lessen the statistical uncertainty of the training and inference, given relatively stable results shown by KcBERT and KcELECTRA compared to other written text-based or general-domain models.

Next, ELECTRA models (KoELECTRA, KcELECTRA) show higher performance for overall and IU, compared to BERT-based ones. This result suggests that the training scheme of ELECTRA that bases on RTD, fits with the current downstream task compared to the masked language model of BERT, considering that RTD had conventionally been more suitable with logical or factoid problems such as natural language inference [8], which requires a slightly different aspect of language understanding in contrast with indecisive tasks such as sentiment analysis. Finding the presence of ambiguity in a given text is more close to detecting some attribute rather than deciding the intensity of it. In contrast, detecting rhetoricalness (as in RQ and RC) being less clear problem and depending more on context or other non-verbal terms, may have yielded the lower accuracy in those classes.

Last, we see how each module distinguishes intonation-dependent utterances from fragments or clear-cut cases, and how such an approach can be further utilized to promote the model development. Unfortunately, we found that 3-fold classification is not yet practical for performance enhancement, since integrating CCs to one class yields a severe imbalance between FR & IU and CCs. However, concerning that detecting IU is promising in both scenarios using ELECTRA models, we expect that making the dataset balanced (beyond merely integrating classes) can boost the performance and help multi-stage
classification, which would benefit the detection of IUs and the classification of CCs. We leave the adequate sampling strategy and dataset reformulation as our future direction.

5.3.3 Qualitative analysis. We made up a confusion matrix with the result of the fine-tuned KoELECTRA module, which shows the most reliable performance (Table 5). Fragments, statements, questions, and commands show high accuracy (> 92%) while others show lower (< 80%).

| Pred \ Ans | FR | S   | Q   | C   | RQ  | RC  | IU |
|------------|----|-----|-----|-----|-----|-----|----|
| Fragment (FR) | 586 | 4   | 3   | 2   | 0   | 0   | 5  |
| Statement (S) | 6   | 1,676 | 7   | 61  | 15  | 12  | 53 |
| Question (Q) | 0   | 8   | 1,737 | 19  | 12  | 0   | 10 |
| Command (C) | 1   | 34  | 23  | 1,223 | 3   | 7   | 5  |
| Rhetorical Q (RQ) | 0   | 25  | 25  | 3   | 118 | 0   | 3  |
| Rhetorical C (RC) | 3   | 9   | 4   | 9   | 0   | 83  | 0  |
| Into-dep. U (IU) | 0   | 56  | 16  | 14  | 4   | 0   | 237 |

Challenges. RQs show the lowest accuracy (73%), and a large portion of wrong answers were related to utterances that are even difficult for a human to disambiguate since nuance is involved. Such cases include questions without tags or wh-particles, for example, ‘난 빠린 거예요’ (Nan pelyn keeyeyyo, Did you dump me?). The sentence can be interpreted as interrogative and declarative in Korean, at a glance, since there is no subject nor polarity item that determines the rhetoricalness of the sentence. However, people may not ask ‘Did you dump me?’ to the addressee because they are curious about it. The model found it hard to tell these rhetorical sentences from declarative statements.

RCs and IUs also showed low accuracy. Nevertheless, it is encouraging that the frequency of false alarms regarding RCs and IUs is generally low (except for statements predicted as IU). For RCs, the false alarms might induce an excessive movement of the addressee (e.g., AI agents), in the case that involves the optatives (‘Have a nice day!’) or greetings (‘See you later!’). For IUs, an unnecessary analysis of the speech data could have been performed if clear-cut cases were classified incorrectly as IU. The low false alarm rate of both categories sheds light on the further utilization of the trained system in circumstances with single short commands.

False alarms. Though the significance is lower compared to the above challenging cases, we observed a tendency within wrong answers regarding the prediction as statements. We found that most of them have a long sentence length that can confuse the system as the descriptive expression, especially those originally a question, command, or RQ. For example, some of the misclassified commands contained a modal phrase (e.g., -야 한다 (-ya hanta, should)) that is frequently used in prohibition or requirement. This lets the utterance be recognized as a descriptive one. Also, we could find some errors incurred by the morphological ambiguity of Korean. For example, ‘베란다 (peylanta, a terrace)’ was classified as a statement due to the existence of ‘란다 (lanta, declarative sentence ender)’, albeit the word (a single noun) has nothing to do with descriptiveness.
6 DISCUSSION

6.1 Findings

In the experiment, we found that the proposed corpus, constructed with a satisfactory agreement (0.85), shows the accuracy that fits the industrial needs (around 0.9) with conventional architectures and up-to-date PLMs. Since we publicly open the corpus and training schemes to facilitate future research, we expect that our dataset can serve as a source of efficient spoken language understanding (SLU) or natural language understanding (NLU) management and, at the same time as a Korean sentence classification benchmark.

One of our concerns is that adequate classification performance or agreement does not necessarily guarantee the optimality of our sentence categorization scheme. For instance, if we merely categorize the sentences based on their sentence form (declaratives, interrogatives, and imperatives), the scheme would be clearer and the classification performance may be far higher. However, it does not resolve the problem of ambiguity that is frequently observed in SLU environments.

To attack this, we adopted the concept of discourse component, assuming that the genuine intention of the sentence can be categorized into one of CG, QS, and TDL, regardless of the sentence form. Also, we took into account SLU environments where only transcripts are available, even with no punctuation marks. This is the background we set a broader categorization, including fragments and intonation-dependent utterances, where the former is considered underspecified and the latter is indecisive without prosodic information. Although experimental results do not guarantee that our categorization comprises the whole Korean sentence types, a well-defined annotation guideline with examples and the resulting corpus may benefit the application of the trained modules.

6.2 Application

Applying our corpus and the trained system to the real world is an essential consideration for the broader impact of our research. We claim that our protocol makes it possible for conventional spoken language understanding systems that utilize an ASR-NLU pipeline to perform more efficiently at handling transcribed utterances. First, the corpus can make the system function without the requirement of wake-up words such as ‘Siri’ or ‘Bixby’, with the proper aid of ASR and speaker verification technologies (Free-running environment). Besides, the corpus can be exploited in making the system react only to utterances that require feedback, while simply generating chit-chat for other non-directive utterances (Omakase dialogue system).

6.2.1 Free-running environment. In Table 1, sentence types with the discourse component of common ground, namely statements, RQs and RCs, are non-directive utterances. Such utterances may require the addressee’s reaction (answering or acting) if there is a specific context, but usually not in the case they are used to start a dialogue. In usual SLU environments where the user’s command starts the conversation between human and agent, it is essential to discern the directive intention from a single input utterance.

In this regard, given that an acoustic channel is open for the device, the system trained upon our corpus may suggest which input utterance to accept as a command or not, instead of requiring wake-up words from the user. This simple detection system prevents unnecessary wake-up of agents caused by false alarms (e.g., wake-up caused by non-directive sentences that contain words pronounced similar to ‘Siri’), and in the case of IU, the device may provide acoustic information for further processing. At the same time, the system induces the agents’ reaction without starting with the wake-up words. Eventually, agents may not interrupt users’ non-directive utterances in usual conversation.
6.2.2 Omakase dialogue system. Omakase dialogue system is a coined term for a dialogue manager that adopts the module trained based on our dataset. Fig. 2 depicts a simplified architecture for the system.

For the transcript of a single utterance in spoken dialogue, first, the trained module categorizes the intention into one of seven sentence types. If the intention is discerned as a directive one, namely a question or command, the manager lists it to the array of instructions so that the following module can understand the instruction and take action (to commands) or give an answer (to questions). If the intention of the utterance is underspecified or non-directive, the manager checks if the topic of the utterance is shared with any listed instruction, and holds the instruction if relevant. If the topic is not relevant to any of the instructions listed in the array, the manager merely generates the following sentence, for instance, a simple chit-chat for the user’s fun. Even if the utterance is directive and instructional, such chit-chat is inserted to accommodate a smooth continuation of the dialogue.

Though just conceptional at this point, we named this system Omakase since it aims at a well-serving and smart task-oriented agent, which is also fluent at chit-chatting with the user. This is quite similar to an Omakase chef who is a guru in making sushi and at the same time fluent at talking with the customers. The spirit is aligned with the recently suggested Sun et al. [46]. However, our approach intends a more heuristic and less data-driven, but assistive and attachable module. Also, since our approach incorporates Korean sentences with various syntax and sentence forms that are labeled with their intention, the resulting classifier may fit a wide range of users who are not familiar with talking to AI agents in a commanding manner. In other words, our approach heads more human-familiar and inclusive usage of SLU modules.

7 CONCLUSION
In this paper, we proposed a textual classification scheme for the spoken Korean language, which considers the intonation-dependency of the given sentence. The corpus was created based on the annotation principle that first detects fragments and categorizes the sentence into one of five intention types, considering if such categorization is available depending on the presence of prosodic information.
For a data-driven training of deep learning models, 61K sentences were collected, with a fairly high inter-annotator agreement using 20K manually tagged samples. The neural network model-based classification yielded adequate accuracy, proving the validity of our approach. Also, we found that the PLMs trained upon colloquial texts more fit with our task, suggesting that our corpus can be a new benchmark for Korean spoken language understanding, which lacks in the literature that is dominant of tasks with written texts.

Though we could not investigate the case using speech signal input in this paper, direct usage of trained systems might enhance the accuracy of spoken language processing. Particularly, there are emerging needs and studies on end-to-end SLU systems [6, 30], which are conducted to reduce the error propagation and computation issues of conventional ASR-NLU pipelines. In this regard, up-to-date SLU modules are being used along with or replacing conventional pipelines. However, we believe that our scheme can be beneficial for both pipeline and end-to-end modules in weighing the importance of each approach. For instance, the probability of predicting the input as IU can be aligned with the output distribution of the end-to-end module, to tell how the output distribution should be taken into account in the final decision. This kind of application does not harm the power of the ensemble guess, at the same time allowing an efficient computation if the pipeline and end-to-end modules are calculated subsequently.

A large portion of this paper concentrates on verifying the validity of our corpus in a computational manner, but our goal in theoretical linguistics lies in making up a new speech act categorization that aggregates potential prosodic cues. It was shown to be successful computationally, but as we discussed in Section 6.1, the promising result does not guarantee theoretical completeness. Still, some challenges exist in handling jussives such as promissives and exhortatives, since utterances that involve social context for disambiguation are not clearly categorized in linguistic viewpoint; such as “It’s so hot here” that asks for the addressee to open the window. In our annotation scheme, such utterances were considered non-directive, and may require the dialogue history or multimodal input to determine it as an instruction. These kinds of disambiguation is to be handled in our future research that addresses the social convention.

A promising application of the proposed system regards spoken language understanding modules for smart agents, especially those targeting human-like conversation with the user. It is the reason our categorization considers the directiveness and rhetoricalness of an utterance. We expect that identifying the rhetoricalness and non-directiveness of an utterance in dialogue management might help people who are not familiar with the way to talk to intelligent agents, at the same time preventing false alarms. It may widen the accessibility of speech-driven AI services and shed light on flexible dialogue management. For a real-life application, we aim to check how the proposed scheme can be extended or distilled to a speech understanding beyond the textual level. We provide the corpora25 and PLM-based recipe26 freely online, to enlighten future research towards human-like Korean spoken language understanding.

ACKNOWLEDGMENT

The annotation guideline and the initial version of the corpus were elaborately constructed with the great help of Ha Eun Park and Dae Ho Kook. Also, the authors appreciate Jong In Kim, Jio Chung, and Kyu Hwan Lee from SNU Spoken Language Processing laboratory (SNU SLP) for providing useful corpora for the analysis. We note that this work was supported by Projects for Research and Development of Police science and Technology under Center for Research and Development of Police

25https://github.com/warnikchow/3i4k
26https://colab.research.google.com/drive/131QcNkPykwxWioby3W0l7pTpJzESsi?#scrollTo=Vbhh6YZZvbb5

ACM Trans. Asian Low-Resour. Lang. Inf. Process.
REFERENCES

[1] Atissa Banuazizi and Cassandre Creswell. 1999. Is that a real question? final rises, final falls, and discourse function in yes-no question intonation. *CLS* 35 (1999), 1–14.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146. https://doi.org/10.1162/tacl_a_00051

[3] Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, et al. 2010. Towards an ISO standard for dialogue act annotation. In *Seventh conference on International Language Resources and Evaluation* (LREC’10).

[4] Won Ik Cho, Sung Jun Cheon, Woo Hyun Kang, Ji Won Kim, and Nam Soo Kim. 2021. Giving Space to Your Message: Assistive Word Segmentation for the Electronic Typing of Digital Minorities. In *Designing Interactive Systems Conference 2021 (Virtual Event, USA)* (DIS ’21). Association for Computing Machinery, New York, NY, USA, 1739–1747. https://doi.org/10.1145/3461778.3462078

[5] Won Ik Cho, Jeonghwa Cho, Woo Hyun Kang, and Nam Soo Kim. 2020. Text Matters but Speech Influences: A Computational Analysis of Syntactic Ambiguity Resolution. In *Proceedings of the 42th Annual Meeting of the Cognitive Science Society - Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020, virtual, July 29 – August 1, 2020*, Stephanie Denison, Michael Mack, Yang Xu, and Blair C. Armstrong (Eds.). cognitivesciencesociety.org. https://cogsci.mindmodeling.org/2020/papers/0448/index.html

[6] Won Ik Cho, Donghyun Kwak, Ji Won Yoon, and Nam Soo Kim. 2020. Speech to Text Adaptation: Towards an Efficient Cross-Modal Distillation. In *Proc. Interspeech 2020*. 896–900. https://doi.org/10.21437/Interspeech.2020-1246

[7] François Chollet et al. 2015. Keras. https://github.com/fchollet/keras.

[8] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[10] Joseph I. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[11] Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1757–1768.

[12] Dafydd Gibbon. 2010. The ambiguity of ‘ambiguity’: beauty, power, and understanding. *Iwdoczce Maria and Lešniewska Justyna* (2010).

[13] Yue Gu, Xinyu Li, Shuhong Chen, Jianyu Zhang, and Ivan Marsic. 2017. Speech intention classification with multimodal deep learning. In *Canadian Conference on Artificial Intelligence*. Springer, 260–271.

[14] Christine Gunlogson. 2002. Declarative questions. In *Semantics and Linguistic Theory*, Vol. 12. 124–143.

[15] Parisa Haghani, Arun Narayanan, Michiel Bacchiani, Galen Chuang, Neeraj Gaur, Pedro Moreno, Rohit Prabhavalkar, Zhongdi Qu, and Austin Waters. 2018. From audio to semantics: Approaches to end-to-end spoken language understanding. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 720–726.

[16] Chung-Hee Han. 2000. *The structure and interpretation of imperatives: mood and force in Universal Grammar*. Psychology Press.

[17] Yuna Hur, Suhyune Son, Midan Shim, Jungwoo Lim, and Heuiseok Lim. 2021. K-EPIC: Entity-Perceived Context Representation in Korean Relation Extraction. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 0448–0453.

[18] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[19] Magdalena Kaufmann. 2019. Fine-tuning natural language imperatives. *Journal of Logic and Computation* 29, 3 (2019), 321–348.

[20] Mary Shin Kim. 2005. Evidentiality in achieving entitlement, objectivity, and detachment in Korean conversation. *Discourse Studies* 7, 1 (2005), 87–108.

[21] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751.
[49] Soroush Vosoughi and Deb Roy. 2016. Tweet acts: A speech act classifier for twitter. In Proceedings of the International AAAI Conference on Web and Social Media, Vol. 10.

[50] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, 4003–4012. https://www.aclweb.org/anthology/2020.lrec-1.494

[51] Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 38–45.

[52] Kichang Yang. 2021. Transformer-based Korean Pretrained Language Models: A Survey on Three Years of Progress. arXiv preprint arXiv:2112.03014 (2021).