Machines Getting with the Program: Understanding Intent Arguments of Non-Canonical Directives

Won Ik Cho[1], Young Ki Moon[2,3]*, Sangwhan Moon[4,5]*, Seok Min Kim[1], Nam Soo Kim[1]
Department of Electrical and Computer Engineering and INMC, Seoul National University[1],
Department of Computer Engineering, Inha University[2], Voithru Co., Ltd.[3],
Department of Computer Science, Tokyo Institute of Technology[4], Odd Concepts Inc.[5]
wicho@hi.snu.ac.kr, ykmoon0814@gmail.com, sangwhan@iki.fi
smkim@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract

Modern dialog managers face the challenge of having to fulfill human-level conversational skills as part of common user expectations, including but not limited to discourse with no clear objective. Along with these requirements, agents are expected to extrapolate intent from the user's dialogue even when subjected to non-canonical forms of speech. This depends on the agent's comprehension of paraphrased forms of such utterances. Especially in low-resource languages, the lack of data is a bottleneck that prevents advancements of the comprehension performance for these types of agents. In this regard, here we demonstrate the necessity of extracting the intent argument of non-canonical directives in a natural language format, which may yield more accurate parsing, and suggest guidelines for building a parallel corpus for this purpose. Following the guidelines, we construct a Korean corpus of 50K instances of question/command-intent pairs, including the labels for classification of the utterance type. We also propose a method for mitigating class imbalance, demonstrating the potential applications of the corpus generation method and its multilingual extensibility.

1 Introduction

The advent of smart agents such as Amazon Echo and Google Home has shown relatively wide market adoption. Users have been familiarized with formulating questions and orders in a way that these agents can easily comprehend and take actions. Given this trend, particularly for cases where questions have various forms such as *yes/no*, *alternative*, *wh-*, *echo* and *embedded* (Huddleston, 1994), a number of analysis techniques have been studied in the domain of semantic role labeling (Shen et al., 2007) and entity recognition (Molla et al., 2006). Nowadays, various question answering tasks have been proposed (Yang et al., 2015; Rajpurkar et al., 2016) and have yielded systems that demonstrate significant advances in performance. Studies on the parsing of canonical imperatives (Matuszek et al., 2013) have also been done for many intelligent agents.

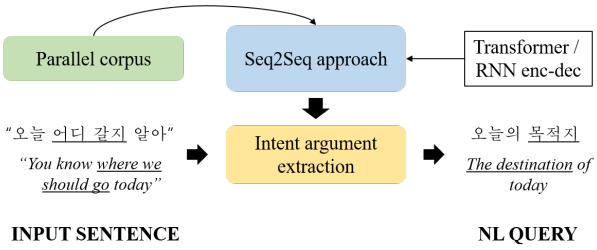

Figure 1: A diagram of the proposed extraction scheme. Unlike in the Korean example that is to be investigated, in English translation, the *wh*-related noun (here, *destination*) is placed at the head part of the output.

However, discerning the intention from a conversational and non-canonical directive (question or command) and correctly extracting its intent argument is still a challenge. It usually matters when the user is not familiar with the *canonical* commands, namely where the direct speech act meets the genuine intention. That is, sometimes, the speech act can be hard to guess merely from the sentence form, as in inferring (1),

(1) *why don't you just call the police*

as a representation of the to-do list '*to call the police*'. Although advanced dialog managing systems may generate a plausible reaction regarding the input utterance, it is different from extracting the exact intent argument (a *question set* or a *to-do-list*) that should be investigated for actual operation.

Additional complexity is introduced when the target text is in a speech recognition context, as the input text may not contain punctuation. For example, given an unclear declarative question (Gunlog-

*Both authors contributed equally to this manuscript.

son, 2002) such as (2),

(2) *you know where we should go today*

a human listener can interpret the subject of inquiry as '*the destination of today*', while this can be challenging for a machine. The basis of our work is that if a system is trained to extract a structured natural language (NL) query from directive sentences, it may help the language understanding systems be more robust at understanding non-canonical expressions in executing the command.

Some may argue that the structured information retrieval we aim to support may benefit from the data augmentation technologies that are concurrent with the studies on paraphrasing (Xie et al., 2019; Kumar et al., 2020). However, complexities as in the examples above have not seen much exploration outside of English, especially in the context of languages with a distinguished syntax or cases which do not use Latin-like alphabets. Also, it is not guaranteed that such technologies fit with less explored languages, where sufficient pre-training resources may not be readily available.

As a more concrete example, in the Korean language, the morphology is agglutinative, the syntax is head-final, and scrambling (non-deterministic permutations of word/phrase ordering) is a common practice between native speakers. Primarily, the agglutinative property of Korean requires additional morphological analysis, which makes it challenging to identify the component of the sentence that has the most substantial connection to core intent. Moreover, the head-finality characteristic introduces an additional layer of complexity, where an under-specified sentence ender incorporates a prosodic cue which requires disambiguation to comprehend the original intent (Yun, 2019; Cho et al., 2020a). Finally, considering the scrambling aspect, which frequently happens in spoken utterances, further analysis is required on top of recognizing the entities and extracting the relevant phrases. These make it difficult for dialog managers to directly apply conventional analysis methods that have been used in Germanic or other Indo-European languages.

In this paper, based on such aspects of the conversation-style utterances of Korean, we propose a structured NL query[1] extraction scheme, which can help enrich the human-like conversation with artificial intelligence (AI). For automation, we construct a corpus of sentence-phrase pairs via annotation and then augment the dataset to mitigate class imbalance, demonstrating the flexibility, practicality, and extensibility of the proposed methods. To further prove that the scheme is not limited to a specific language, we demonstrate the methodology using English examples and supplement specific cases with Korean. We describe the followings as our contribution to the field:

- We propose the scheme for building the parallel corpora of non-canonical Korean directives and their intent arguments, along with speech act type labeled, and release it publicly.
- We suggest a visible result on the content extraction scheme with conventional Seq2Seq systems, probing the application potential.

2 Concept and Related Work

The theoretical background of this proposal builds on literature from speech act (Searle, 1976) and formal semantics (Portner, 2004). Although many task-oriented systems identify the intents as a specific action that the agent should take (Liu and Lane, 2016; Li et al., 2018), to make our approach generic in the aspect of the domain and sentence structures, we hypothesized that it would be beneficial for the natural language understanding (NLU) modules first to recognize the directiveness and represent the core content in a structured format.

Once an utterance is identified to be directive, conventional systems rely on slot-filling to extract the item and argument (Li et al., 2018; Haghani et al., 2018), where the number of the categories is generally restricted. Instead, for non-task-oriented dialogues, we hypothesized that the arguments should be attained in NL format rather than structured data, by, e.g., rewriting the utterances into some nominalized or simplified terms which correspond to the source text. There have been studies on paraphrasing of questions concerning the core content (Dong et al., 2017), but little has been done on NL formalization. Also, our study targets the extraction of commands, which is equivalently essential but has not been widely explored outside of the robotics domain (Matuszek et al., 2010, 2013).

The closest problem to this task is probably semantic parsing (Berant and Liang, 2014; Su and Yan, 2017) and structured query language (SQL) generation, Zhong et al. (2017) which propose

[1] Hereafter, we interchangeably use *NL query* and *(intent) argument* to indicate the structured core content, depending on the context.

Seq2Seq (Sutskever et al., 2014)-like architectures to transform NL input into a structured format. These approaches provide the core content of the directive utterances as a sequence of queries, both utilizing it in paraphrasing (Berant and Liang, 2014) or code generation (Zhong et al., 2017). However, the proposed source sentence formats are usually *canonical* and mostly information-seeking, rather than being in a colloquial context.

Our motivation builds on the basis that real-world utterances as input (e.g., smart home commands from the less tech-familiar audience), in particular for Korean, can diverge from the expected input form, to the *non-canonical* utterances that require actual comprehension for classifying as a question or command. Moreover, as we discuss in the latter part of our work, we intend the extracted NL queries to be reusable as building blocks for efficient paraphrasing, following the approach in Berant and Liang (2014).

Recently, in a related view, or stronger linguistic context emphasis, guidelines for identifying non-canonical natural language questions or commands have been suggested for Korean (Cho et al., 2018a). We build on top of this corpus for the initial dataset creation, and extend the dataset with additional human-generated sentences.

3 Proposed Scheme

In this section, we describe the corpus construction scheme along with the motivation of this work. As discussed in the first section, our goal is to propose a guideline for discerning the intent argument for conversational and non-canonical questions and commands. These appear a lot in everyday life, but unlike cases where the input is in a canonical form, algorithmically extracting the core intent is not straightforward. We suggest that a data-driven methodology should be introduced for this task, which can be done by creating a corpus annotated with the core content of the utterances. While our work in this paper is for Korean, the example sentences and the proposed structured scheme are provided in English, for demonstrative purposes.

3.1 Identifying Directives

Identifying directive utterances is a fundamental part of this work, though our main content is not just classification. Thus, at this moment, we briefly demonstrate the Korean corpus whose guideline is for distinguishing such utterances from the non-directives such as fragments and statements (Cho et al., 2018a).

For questions, interrogatives which might be represented by *do* support or *wh-* movement in English, were primarily considered[2]. The ones in an embedded form were also counted, possibly with the predicates such as *wonder*. Also, a large number of the declarative questions (Gunlogson, 2002) were taken into account. Since the corpus utilized in both Cho et al. (2018a) and this annotation process does not contain punctuation marks, the final work was carried out for clear-cut questions that were selected upon the majority voting of the annotators, at the same time removing the utterances that depend on acoustic features. For all the types of questions, the ones in rhetorical tone were removed since their argument usually does not perform as an effective question set (Rohde, 2006).

For commands, the imperatives in a covert subject and with the modal verbs such as *should* were primarily counted. The requests in question form were also taken into account. All the types incorporate the prohibition. Conditionalized imperatives were considered as command only if the conditional junction does not negate the to-do-list. Same as the former case, the ones in rhetorical tone or usage were removed despite it has an imperative structure (Han, 2000; Kaufmann, 2016). All the other types of utterances except questions and commands were considered non-directive[3].

3.2 Annotating Intent Arguments

The following section exhibits example annotation of intent arguments for non-canonical directives, as shown in Figure 2. We want to note again that while we describe the procedure based on simplified English sentence examples, the actual data and process were significantly more complicated.

3.2.1 Questions

For the three major question types, which are defined as *yes/no*, *alternative* and *wh-*[4], we applied different extraction rules. For *yes/no* questions

[2]Note that this does not always hold for the Korean language, which is *wh-in-situ*. A more complicated and audio-aided identification is required in those cases, as in Cho et al. (2020a)

[3]We aim to explain the type of utterances which are also counted as non-directive in other languages, even if a 1:1 mapping might not be possible through translation. We plan to publish an expansion of this work, which is specific to English sentences accompanied by sample corpora as separate work.

[4]Note that here, these are not the syntactic properties, but the level of speech act.

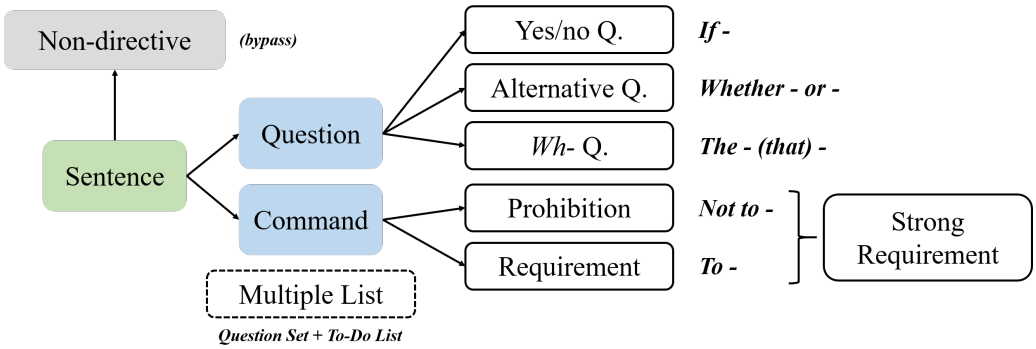

Figure 2: A simple description on the categorization and annotation. The sentence is either a given text utterance or a transcript. The lexicons on the right side denote the head of the arguments (which goes to the tail of a phrase in Korean). Multiple list denotes the rare cases where question and command co-exist, but was not detected in the construction phase. The strong requirement, which is a serial placement of PH and REQ, is to be explained afterward since it originates in an empirical study and may not be a universal phenomenon.

(*yes/no* Q), we employ an *if-* clause which constraints the candidate answers to yes or no (3a). For *alternative* questions (Alt. Q), we employ a *whether - or -* clause accompanied by the list of possible answers (3b). For *wh-* questions (*wh-* Q), the extraction process starts with a lexicon which corresponds with the *wh-* particle that is displayed (3c-d). It is notable that some alternative questions also show the format that is close to the *wh-*questions, with possibly *between* that corresponds with *whether - or -* (3e).

(3) *a. did I ever tell you about how*

→ **if** the speaker told the addressee about how

b. you hungry or thirsty or both

→ **whether** the addressee is hungry **or** thirsty

c. how many points you got

→ **the number** of points that the addressee got

d. i want to know about treadstone

→ **the information** about treadstone

e. you know which is hotter in hawaii or guam

→ **the place** hotter **between** hawaii *and* guam

3.2.2 Commands

Since the main intent of the commands is analogous to a to-do-list (Portner, 2004), we annotated an action which the addressee may take, in a structured format. All of these lists start with *to* indeterminate (4a, REQ, requirement), with possibly *not to* for the prohibitions (4b, PH). During this process, non-content-related lexicons such as politeness strategies (e.g., *please*) were not considered in the extraction (4c).

(4) *a. i suggest that you ask your wife*

→ **to** ask one's wife

b. yeah but don't pick me up

→ **not to** pick the speaker up

c. please don't tell my daddy

→ **not to** tell the speaker's daddy

3.2.3 Phrase Structure

As discussed above, the argument of the questions are transformed into *if-* clause, *whether-* clause or *the-* phrase. Following this logic, the commands are rewritten to either a *to*-clause or *not to*-clause. Except for the *wh-* questions and some alternative questions, all the rewritten sentences contain at least one predicate (verb). Here, note that unlike the English examples displayed above, in the Korean samples, the components that decide the phrase structure (e.g., *if-*, *whether-*, *(not) to-*) are all placed at the end of the sentence, mainly due to head-finality. This is to be further described.

3.2.4 Coreference

Coreference is a critical issue when extracting the information from the text. It appears a lot in conversational utterances, in the form of pronouns or anaphora. In the annotation process, we decided to preserve such lexicons except for *I*/*we* and *you* since they are participants in the dialog. The concepts which correspond with the two were replaced with either *the speaker(s)* or *the addressee* as shown in (3a-c) and (4b-c); and in some cases with *one(self)* to make it sound more natural (4a).

3.2.5 Spatial-Temporal and Subjective Factors

Unlike other question or command corpora, the proposed scheme includes content which requires an understanding of spatial (5a) and temporal (5b) dependencies, namely deixis. These factors are related to the coreference in the previous section, in particular, involving lexicons such as *there* and *then*. Also, the dialog being non-task-oriented results in the content unintentionally incorporating the subjective information, such as current thoughts of the speaker or the addressee. The proposed scheme tries not to ignore such factors in the intent argument (5c-d), to ensure that the core content is preserved.

(5) *a. put your right foot there*
→ to put the right foot **there**
b. i i don't want to see you tomorrow
→ not to meet **tomorrow**
c. any ideas about the colour
→ **the idea** about the colour
d. you ought to know what our chances are
→ **to be aware** about the speaker's chances

4 Dataset Construction

4.1 Corpus Annotation

For the argument annotation process, we adopted the corpus constructed in Cho et al. (2018a), a Korean single utterance corpus for identifying directives/non-directives that contains a wide variety of non-canonical directives. About 30K directive utterances were adopted for the creation of their intent arguments, which are labeled either question or command. The broader categorization on whether the utterance is question or command had been done with moderate agreement $\kappa = 0.85$ (Fleiss, 1971), thus, we only annotated the NL queries, simultaneously tagging the subcategories that directly follow the query. The additional tagging and annotation were done by three Korean natives with a background in computational linguistics, and the cross-checking was done with discussion and modification on the conflicts (improper summarization). In detail, the draft query generation was done by two of the annotators, where they cross-checked the work of each. The last annotator thoroughly checked the validity and appropriateness, so that the consensus can be attained from at least three speakers. The detail on this process with the Korean examples is available in Cho et al. (2018b).

We want to emphasize here that our work is not precisely an *annotation task*, but closer to a *rewriting task* with lax constraints on the expected answer. Although the written NL queries may not be identical for all the same kind of utterances, we hypothesize that there is a plausible semantic boundary for each utterance.

Although our examples are in English, this kind of rewriting supports that the natural language-formatted intent argument can be robust in preserving the purpose of input directives, although the cultural factors such as politeness might influence. We claim that the constraints of our method guarantee this, as we utilize the nominalized and structured terms. Specific considerations when creating a Korean dataset are discussed below.

Head-finality In the Korean language, due to the head-finality, all of the structured expressions which are used to construct the phrase structure (Section 3.2.3.) goes to the end of the intent arguments. However, in a cross-linguistic perspective, this may not necessarily change the role of the intent arguments. For example, in the Korean sentence *SENT* = "*mwe ha-ko siph-ni* (what do you want to do)", which has an intent argument *ARG* = '*cheng-ca-ka ha-ko siph-un kes* (the thing that the addressee wants to do)', the original *SENT* can be rewritten as *SENT** = "ARG-*i mwu-ess-ip-ni-kka*". Here, *SENT** can be interpreted as "*what is ARG*" or "*tell me about ARG*", where the core content *ARG* is not lost in translation.

Strong Requirements The term *strong requirement* is not an official academic term, but was coined and proposed for their existence in the corpus. Simply explained, this can be described as a co-existence of a prohibitive expression (PH) and the canonical requirement (REQ), as we can see in the sentence "*don't go outside, just stay in the house*". Even when the prohibitive expression comes immediately before the requirement, such forbidding expressions are not considered as the core content in the final sentence. That is, in these cases, simply understanding it as "*just stay in the house*" does not harm the process of query extraction that results in the ideal final form: '*to stay in the house*'. In Korean where scrambling is common, both [PH+REQ] and [REQ+PH] can be valid expressions. In our work, we did not encounter

cases where scrambling leads the interpretation of the utterance to be a prohibition.

Speaker/addressee Notation We consider the notation of coreference crucial in this work. A subject drop is a typical pattern that can be observed in casual spoken Korean. This is different from English, where the agent and the experiencer are explicit. In Korean, they can be dropped and are resolved with context or prosody. Thus, to minimize the ambiguity, we created two separate corpora; one with the speaker/addressee notation, and the other with this information removed. In the former, we classify all possible cases into one of the following five categories: only the speaker (*hwa-ca*), only the addressee (*cheng-ca*), both (*hwa-ca-wa cheng-ca*), none, and unknown. We believe this kind of information will be beneficial for both contextual disambiguation and further research. On the other hand, in the latter, while the specification must be inferred from the context, the output will be closer to what one would encounter in everyday life.

4.2 Corpus Augmentation

4.2.1 What Should be Supplemented

Above, we used an existing dataset to annotate intent arguments for questions and command utterances, but encountered an imbalance in the dataset - specifically not having enough data for some utterance types, namely Alt. Q, PH, and Str. REQ. Additionally, we concluded that the amount of parallel data was not large enough for the *wh-* question to be useful in real life, taking into account that the extraction of queries from *wh-* questions involves the abstraction of the *wh*-related concept (e.g., 'destination' from *where-to-*). To address the issues, we expanded the dataset size by obtaining various types of sentences from intent arguments, specifically via human-aided sentence generation.

Data Imbalance First, for Alt. Q, PH, and Str. REQ, we needed to ensure the class balance for each utterance type, or at least a sufficient number for the automation. To this end, we manually wrote 400 intent arguments for each of the three types. Specifically, sentences were created at ratio (1 : 1: 1: 1: 4) for *mail, schedule, smart home, weather*, and *other free topics*[5], which are considered as usual topics of interest in intelligent agents and also follow the original corpus.

[5]Other topics include the ones that are not mentioned previously, e.g., game, politics, commercials.

***Wh-* Questions** To enforce the second goal, the supplement of *wh*-questions, 800 intent arguments were newly written. The topics of each sentence considered in this process are identical to the above. However, the use of *wh*-particles that can hinder the transformations between *wh*-particles and *wh*-related terms was not allowed. This means that the intent arguments were created in the way in which they only expose the nominalized format, and not the *wh*-particles, e.g., *the weather of tomorrow* rather than *what the weather is like tomorrow*. This trend was also applied when constructing additional phrases for some alternative questions above.

4.2.2 Method and Outcome

We recruited eight Seoul Korean natives, with diverse academic backgrounds and sufficient knowledge in Korean grammar, to generate the directive sentences from the queries. In detail, with the 2,000 NL queries suggested above, created by other four Korean native speakers, we requested the participants to write ten utterances per phrase as diversely as possible. The guideline was provided to encourage the use of politeness expressions, scrambling, word replacement, etc., for the diversity of expression, and the process was undergone with free QA hours. The output was cross-checked as in the annotation process and was finally augmented to the corpus. The detailed guideline is demonstrated in Cho et al. (2020b).

The paraphrasing process resulted in a total of 20,000 argument-directive pairs, constructed from 2,000 arguments. Examples of various question and command expressions for phrases obtained in this process include, for example (from Cho et al. (2020b)),

Argument: The most important concept in algebra
Topic: Free, **Type**: *wh-* question

→ *just pick me one important concept in algebra*

→ *what you think the core concept in algebra is*

→ *which concept is the most important in algebra*

→ *what should i remember among various concepts in algebra* · · · (various versions in Korean)

The composition of the entire dataset after augmentation is shown in Table 1. As a result of the above remedies, the class imbalance and practicability, which were problematic at the initial point, have been partially resolved. The details are available online[6].

[6]https://github.com/warnikchow/sae4k

Intention	Types	Original	Augmented	Sum
Question	Yes/no Q	5,715	-	5,715
	Alternative Q	229	4,000	4,229
	Wh- Q	11,988	8,000	19,988
Command	Prohibition	478	4,000	4,478
	Requirement	12,302	-	12,302
	Strong REQ.	125	4,000	4,125
	Total	**30,837**	**20,000**	**50,837**

Table 1: The final composition of the dataset.

5 Experiments

Here, we validate the usefulness of the constructed dataset with multiple sequence-to-sequence (Seq2Seq) (Sutskever et al., 2014) architectures. We would like to note that as we propose both a new dataset accompanied by a new task, there is no baseline or proven evaluation metric as of the time of writing. For these reasons, we used existing evaluation frameworks used by other generation tasks.

5.1 Format

The final format of the corpus is as follows: [***Label / Sentence / Argument***]. Here, the label denotes the six utterance types as in Section 4.1, and the utterance and intent argument are in raw text form. As stated in Section 4.1.2, there are two versions of the corpus: with and without the speaker/addressee notation. The latter is utilized at this phase, to ensure whether the non-functional contents are well captured.

In the automation process, we aimed to infer the intent argument directly, by giving a **sentence** as an input and an **argument** as a target. Here, the correct inference of the intent argument is not independent with the identification of the exact utterance type[7] due to the formats being distinct. Therefore, we separate metrics for different tasks. We discuss this further in the evaluation section.

5.2 Automation

While the total volume is not significant for fluent automation concerning the usual dataset size for machine translation (MT), we proceeded to observe how the proposed scheme works. The implementation was done through a recurrent neural network (RNN)-based encoder-decoder (enc-dec) with attention (Cho et al., 2014; Luong et al., 2015) and a Transformer (Vaswani et al., 2017). For the agglutinative nature of the Korean language, morpheme-level tokenization was done with MeCab[8] tokenizer provided by the KoNLPy (Park and Cho, 2014) library.

For the *RNN enc-dec with attention* that utilizes the morpheme sequence of maximum length 25, hidden layer width and dropout rate Srivastava et al. (2014) was set to 512 and 0.1, respectively. This model was trained for 100,000 epochs.

For the *Transformer*, which adopts a much more concise model compared to the original paper (Vaswani et al., 2017), the maximum length of the morpheme sequence was set to also 25, with hidden layer width 512 and dropout rate 0.5. Additionally, multi-head attention heads were set to 4, and a total of two Transformer layers were stacked, considering the size of the training data. Due to the higher computation budget required, this model was trained for 10,000 epochs.

5.3 Evaluation

The most challenging part of validating a new dataset and task is deciding a fair and robust evaluation framework. This is particularly challenging for generative tasks, such as translation or summarization. For this kind of task, several candidates exist that can be considered felicitous for an input utterance. It means that the same phrase can be expressed in various ways, without harming the core content.

Nonetheless, as it is for paraphrasing or summarization, we believe that there should be a rough boundary regarding our tolerance of the output variance. Specifically, in our task, the answer *has to be* some formalized expression. However, if we utilize only BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004) as a measure, there is a chance that the diversity of possible outputs can result in grammatically incorrect or incomprehensible output (Matsumaru et al., 2020), although it is semantically plausible. Also, in the corpus construction, we have explicitly set the formats for different utterance types, which requires the correct identification of the speech act and thus can largely influence the accurate inference of an argument.

In this regard, we first surveyed a proper metric for the automatic and quantitative analysis of the result, respectively. A part of the conclusion is that the automatic analysis of semantic similarity can be executed utilizing the recent pre-trained language

[7] Nonetheless, we don't consider this task as a classification that identifies the label.

[8] https://bitbucket.org/eunjeon/mecab-ko-dic/src/master/

	RNN S2S + Attention	Transformer	
Test split	9:1	7:3	9:1
Iteration	100,000	10,000	10,000
ROUGE-1	0.5335	0.5383	0.5732
BERTScore	0.7693	0.8601	0.9724
Total	**0.6514**	**0.6992**	**0.7728**

Table 2: Validation result with the test set.

model-based scoring system, namely BERTScore[9] (Zhang* et al., 2020). Such an approach can be adopted regardless of whether the label is correctly inferred and reflects the common sense inherited in the pre-trained language models. Moreover, in case the label is correct and some format-related tokens (e.g., *the method*, *whether*, *not to*) in the output overlap with the ones in the gold data, the lexical similarity can also be taken into account, probably as an extra point. It can be further represented by ROUGE compared to the gold standard.

Considering the different natures, we determined to aggregate both kinds of evaluation values. The final score was obtained by averaging those two results, namely ROUGE-1 and BERTScore. With this, we compensate for the case that the format difference caused by the wrong label leads to the misjudgment on lexical features.

5.4 Result

The validation results are in Table 2. For *Total*, we averaged BERTScore and ROUGE-1.

The result shows the advantage coming from (a) adopting the self-attention-based (Vaswani et al., 2017) Seq2Seq and (b) setting aside a larger volume of data for the training phase. (a) can be observed in the results, in both ROUGE-1 and BERTScore, where the Transformer model performs better with the same split model, even with the 7:3 split model that has gone through less training. (b) is observed within the two Transformer models. The main reason for the difference is assumed to be the existence of out-of-vocabulary (OOV) terms in the test set, which in our experiments loses information during encoding. As the information has been lost, this in turn affects the performance of the decoder.

Beyond the quantitative analysis that mainly concerns metrics, we checked the model's validity with the output for a test utterance that is fed as a common input. For example, from the original sentence (Str. REQ):

(6) "저번처럼 가지 말고 백화점 세일은 미리 가서 대기하렴" / "*This time, please go to the department store earlier (than its opening time) and wait there for the upcoming sale event*"

the followings are obtained from each model:

(6) a. **RNN Seq2Seq with attention** - 백화점 가 미리 가 서 대기 대기 대기 ... / *department store, go earlier (than its opening time), and wait wait wait* ... (failure of proper phrase ending)

b. **Transformer (split 7:3)** - 백화점 가 서 미리 가 서 도와 주 기 / *to go to the department store earlier (than its opening time) and help (something)*

c. **Transformer (split 9:1)** - 백화점 세일 은 미리 가 서 대기 하 기 / *to go to the depratment store earlier (than its opening time) and wait for the sale event*

Taking into account that the given utterance (6) is a strong requirement, or a series of (less meaningful) PH and (substantial) REQ, it is encouraging that all three models succeeded to place the *department store* (백화점, payk-hwa-cem) at the very first of the sentence, ignoring the PH in the first half clause[10]. However, note that in (6a), the hallucination took place in the RNN model, while the other two Transformer models cope with it and find the right place to finish the inference. Being able to determine when to terminate the sequence is important for matching the sentence type correctly, especially in a head-final language as Korean[11].

Besides, comparing (6b) and (6c), where the tails of the clauses (regarding sentence type) were correctly inferred, the latter fails to choose the lexicon regarding *wait*, instead picking up *help* that may have been trained in a strong correlation with the terms such as *go earlier* in the training phase. Here, it is also assumed that loanwords such as *sale* (세일, seyil), which is expected to be OOV in the test phase, might have caused the failure in (6b), even though it exists in the input sentence. The gold standard for (6) is '백화점 세일은 미리 가서 대기하기, *to go to the department store earlier and wait for the sale event*', which is identical to (6c) if the morphemes are accurately merged.

[9] BERT denotes a bidirectional encoder representation from Transformer (Devlin et al., 2019), a freely available pre-trained LM, and this fine-tuned evaluating toolkit is provided in https://github.com/Tiiiger/bert_score

[10] Though omitted for the fluent translation, '저번처럼 가지 말고' is PH that originally means *not to go as the last time*.

[11] Stably guessing the accurate tail of the phrase is not guaranteed in the auto-regressive inference.

Here are more samples that come from the well-performing Transformer model, especially some tricky input sentences (7) and *wh-* questions (8). We expect such formalizations can be meaningful for the future AIs with personality, aiming at human-friendly interaction. As part of the pre-processing pipeline, all punctuation was removed from the input, and the output phrases were not polished to deliver the original format.

(7) "박사 졸업과 결혼 준비를 비교한다면 어떤게 더 지옥같아" / "*which is more hell if you compare your phd with your wedding preparation*"

→ 박사 졸업 과 결혼 준비 중 더 힘들 었 던 것 / *the tougher process (for the addressee) between getting phd and preparing wedding*

(8) "몇 도 기준으로 열대야라고 해" / "*from what temperature is it called a tropical night*"

→ 열대야 기준 온도 / *the reference temperature of tropical night*

5.5 Discussion

Analysis The results suggest that the self attention-based model architecture can be quite beneficial for stable inference. Moreover, the inference seems to take advantage of grasping the proper interaction between long-distance components of the input sentence. It emphasizes that the intent argument extraction requires the understanding beyond the given lexicons, not merely being a syntactic parsing task. Though we cannot rule out the possibility of overfitting, Seq2Seq-style approaches are validated with a moderate amount of sentence-query pairs (40K - 50K). The overall performance is expected to boost up with the modern noise-robust sentence encoders (Lewis et al., 2020).

Limitation As shown by the dependency on the train set data, domain generalization issues regarding OOVs is critical in coping with the resource shortage and guaranteeing efficiency. However, we assume that our limitation in the topic may not affect much on generalization given the controllable and content-preserving technologies (Logeswaran et al., 2018; Martin et al., 2020), since our transformation rarely changes the domain-specific contents. For instance, "*what will you best recommend memorizing in the algebra textbook*" is transformed to '*the most important concept in the algebra*', where the transformation engages in the general expressions (*best, recommend, most important*). That is, though our baseline experimental results merely attest to the validity of the corpus, we believe that models that have higher robustness to OOV, such as those pre-trained on large corpora, will perform better and leverage our framework.

Application Since the task domain of the proposed approach is not specified, we expect our scheme and output to be worthwhile for a general AI that aims human-friendliness. At the same time, it may prevent users from feeling isolated by talking mechanically. Also, along with the non-task-oriented dialogues, our scheme may be useful for avoiding inadvertent ignorance of the users' will, such as the digitally marginalized.

6 Conclusion

The significance of this research is in proposing the construction and augmentation schemes for rewriting of less explored sentence units, and making it an open, permissive resource for the general public. The sentence set consists of directive utterances in the Korean language, where the morpho-syntactic property often provide difficulties in information retrieval. Additionally, we propose baselines for the constructed dataset using multiple Seq2Seq architectures, exhibiting that our methodology is practically meaningful in real-world applications.

Our next work is to extend this more typologically by showing that the annotation/generation scheme applies to other languages. While the scope of our work is limited to Korean, we hope that the proposed annotation scheme and resources from our work can be reused as a common protocol for intent-argument extraction tasks in other languages.

Acknowledgments

This research was supported by the Technology Innovation Program (10076583, Development of free-running speech recognition technologies for embedded robot system) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea). Also, the authors appreciate Siyeon Natalie Park for suggesting a great idea for the title. After all, the authors are grateful for the invaluable advices and supports provided by Reinald Kim Amplayo, David Mortensen, Jong In Kim, Jio Chung, and †Kyuwhan Lee.

References

Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1415–1425.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Won Ik Cho, Jeonghwa Cho, Woo Hyun Kang, and Nam Soo Kim. 2020a. Text matters but speech influences: A computational analysis of syntactic ambiguity resolution. In *Proceedings of the 42nd Annual Virtual Meeting of the Cognitive Science Society*, pages 1953–1959.

Won Ik Cho, Jong In Kim, Young Ki Moon, and Nam Soo Kim. 2020b. Discourse component to sentence (DC2S): An efficient human-aided construction of paraphrase and sentence similarity dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6819–6826.

Won Ik Cho, Hyeon Seung Lee, Ji Won Yoon, Seok Min Kim, and Nam Soo Kim. 2018a. Speech intention understanding in a head-final language: A disambiguation utilizing intonation-dependency. *arXiv preprint arXiv:1811.04231*.

Won Ik Cho, Young Ki Moon, Woo Hyun Kang, and Nam Soo Kim. 2018b. Extracting arguments from Korean question and command: An annotated corpus for structured paraphrasing. *arXiv preprint arXiv:1810.04631*.

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)*, pages 4171–4186.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Christine Gunlogson. 2002. Declarative questions. In *Semantics and Linguistic Theory*, volume 12, pages 124–143.

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. *arXiv preprint arXiv:1809.09190*.

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

Rodney Huddleston. 1994. The contrast between interrogatives and questions. *Journal of Linguistics*, 30(2):411–439.

Magdalena Kaufmann. 2016. Fine-tuning natural language imperatives. *Journal of Logic and Computation*.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Xiujun Li, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *arXiv preprint arXiv:1807.11125*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *Interspeech 2016*, pages 685–689.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. In *Advances in Neural Information Processing Systems*, pages 5103–5113.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Louis Martin, Éric Villemonte de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4689–4698.

Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. 2020. Improving truthfulness of headline generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Cynthia Matuszek, Dieter Fox, and Karl Koscher. 2010. Following directions using statistical machine translation. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, pages 251–258. IEEE.

Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. 2013. Learning to parse natural language commands to a robot control system. In *Experimental Robotics*, pages 403–415. Springer.

Diego Molla, Menno van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 51–58.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Eunjeong L Park and Sungzoon Cho. 2014. KoNLPy: Korean natural language processing in Python. In *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology*, pages 133–36.

Paul Portner. 2004. The semantics of imperatives within a theory of clause types. In *Semantics and linguistic theory*, volume 14, pages 235–252.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Hannah Rohde. 2006. Rhetorical questions as redundant interrogatives.

John R Searle. 1976. A classification of illocutionary acts. *Language in society*, 5(1):1–23.

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document summarization using conditional random fields. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 2862–2867.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Yu Su and Xifeng Yan. 2017. Cross-domain semantic parsing via paraphrasing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1235–1246.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018.

Jiwon Yun. 2019. Meaning and prosody of wh-indeterminates in Korean. *Linguistic Inquiry*, 50(3):630–647.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.