Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6819–6826
Marseille, 11–16 May 2020

Discourse Component to Sentence (DC2S): An Efficient Human-Aided Construction of Paraphrase and Sentence Similarity Dataset

Won Ik Cho[1], Jong In Kim[2], Young Ki Moon[3,4], Nam Soo Kim[1]
Department of Electrical and Computer Engineering and INMC, Seoul National University[1],
Interdisciplinary Program in Cognitive Science, Seoul National University[2],
Department of Computer Engineering, Inha University[3], Voithru Co., Ltd[4]
wicho@hi.snu.ac.kr, prows12@gmail.com, ykmoon0814@gmail.com, nkim@snu.ac.kr

Abstract

Assessing the similarity of sentences and detecting paraphrases is an essential task both in theory and practice, but achieving a reliable dataset requires high resource. In this paper, we propose a discourse component-based paraphrase generation for the directive utterances, which is efficient in terms of human-aided construction and content preservation. All discourse components are expressed in natural language phrases, and the phrases are created considering both speech act and topic so that the controlled construction of the sentence similarity dataset is available. Here, we investigate the validity of our scheme using the Korean language, a language with diverse paraphrasing due to frequent subject drop and scramblings. With 1,000 intent argument phrases and thus generated 10,000 utterances, we make up a sentence similarity dataset of practically sufficient size. It contains five sentence pair types, including paraphrase, and displays a total volume of about 550K. To emphasize the utility of the scheme and dataset, we measure the similarity matching performance via conventional natural language inference models, also suggesting the multi-lingual extensibility.

Keywords: discourse component, paraphrasing, generation, sentence similarity test

1. Introduction

How should the paraphrase be defined for the utterances with a relatively clear act? *“Could you pass me over the salt?”* is a request, but in some cultures, it might have a polite sibling such as *“I can see salt on your side, so I wonder if it does not matter...”*, though some unexpected nuances are inserted in. One may claim they are not precisely the same, but in real-life usage, paraphrasing does not necessarily aim to convey the identical possible world to the addressee. How should the condition be defined for the above statement? How about other tricky sentences in real life?

Checking sentence similarity and generating paraphrases has been studied a lot recently for their direct relationship with enriching the human language and its understanding (Fernando and Stevenson, 2008; Agirre et al., 2012; Agirre et al., 2013). The closest example is probably in question answering (QA) domain by matching the user query to the questions that are already in the database (Achananuparp et al., 2008), providing substantial information as an output. Similar approaches can be useful for recommendation systems or automatic counselors. In the non-task-oriented dialog between human and machine, sentence similarity datasets can help machines determine if the input is relevant to the conversation history. Also, the sentence relevance test has been of significant importance for some self-supervised pretraining of language models (Devlin et al., 2018).

Theoretically, there has been an extensive discussion on what the paraphrase is, both on logical and semantic viewpoint (Vila et al., 2014). They aimed to make a robust common ground on why and how we should make up the paraphrases. The logical viewpoint defines paraphrases as the sentences conveying the identical possible world as the original sentence indicates (Martin, 1976), while in semantics and computational linguistics, it is permitted to have a shift of a word or phrase (Bhagat and Hovy, 2013), if the sentence meaning or purpose is preserved. In specific, the discussion on the directives has been widely done with its automation (Tomuro, 2003; Prakash et al., 2016), due to the process being closely related to checking the relevance of the sentences. Many ideas were suggested on mechanically achieving the lexical equivalents (Park et al., 2016; Dong et al., 2017), but little was considered on the further meaning of paraphrase and more flexible human generation.

Throughout this paper, we are going to argue the problem of making up paraphrases for the directive utterances, i.e., questions and commands, adopting the concept of discourse component. What are the paraphrases for the questions and commands, and how can we make it without discarding the core information? Why is it required in reality, and how can we relate it to the sentence similarity task? How can the result be utilized? We answer these questions afterward, with the human-aided approach that regards several topics and speech act labels.

2. Background and Proposal

In this section, we mainly discuss how the core content of an utterance is extracted and transformed into a sentence of a new format.

2.1. The Core Content

The core content is challenging to define for an arbitrary sentence since there exist many viewpoints regarding a single utterance. One can keep only the emotion and intention, another can preserve the information, and others can change only one or two words to minimize the loss of data. However, at least for directive utterances, it is plausible that one can make up a paraphrase by maintaining the factors of interest (requirement), which are represented by keywords or a keyphrase. For instance, in paraphrasing a colloquial expression *“hey can you tell me when the rain stops in tokyo”*, the keywords are *‘rain’*, *‘stops’*, and *‘tokyo’*, and the keyphrase will be *‘when the rain stops in tokyo’*. Based

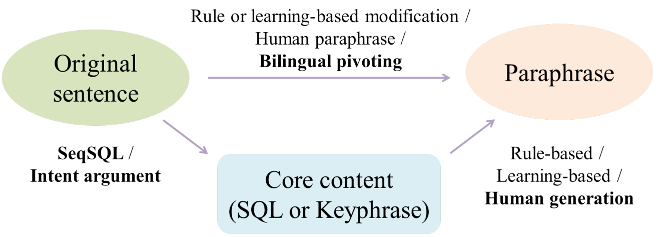


Figure 1: The relationship between the core content and the paraphrase regarding a sentence. Bold phrases are the contents that are covered in this paper.

on this, one may reconstruct sentences such as *"when does the rain stop in tokyo"*, *"i want you to tell me when the rain will stop in tokyo"*, and *"find when the rain stops in tokyo"*. Furthermore, for a more structured and productive generation process, it is meaningful to investigate if the information in the utterance can be delineated in other formats, e.g., in the way of table or structured query language (SQL) (Zhong et al., 2017).

2.2. Literature

The preceding discussion on the core content leads us to the literature of paraphrasing, as can be concisely represented like Figure 1. Identifying and generating the paraphrases, both for human and machine, not only reduces the confusion coming from a small perturbation in the word order (Zhang et al., 2019) but also makes it possible to express the same idea in various and non-typical fashion. We want to take a look at the procedures by which one can make up the collection of paraphrases that are utilized in training identification systems. Beyond direct human paraphrasing or canonical hand-crafted approaches (Bhagat and Hovy, 2013), the related procedures can be categorized regarding mainly three formats, namely multi-lingual, monolingual non-NL[1], and monolingual NL (Figure 1). The first two are introduced as preceding approaches, and the latter is proposed in this study.

Multi-lingual approach is generally known as bilingual pivoting and back translation (Mallinson et al., 2017; Prabhumoye et al., 2018), which are widely used to augment the machine translation data (Resnik et al., 2010). First, one can use the machine translation models in service to project the given sentence into various multi-lingual formats that convey the same information, and then back-translate them to obtain paraphrases. Although the process has little to do with extracting the core content, the acquisition of the sentences can be automated to guarantee an efficient achievement of a large-scale dataset. However, two main points to be considered are: primarily, 1-1 correspondence of the content is not necessarily guaranteed in translation, and second, that the errors or biases in machine translation can be amplified in the forward-backward translation process. Also, the approach may not fit with the recent machine translation algorithms that aim to make an isomorphic relation between the source and target language (Ponti et al., 2018).

Next, **monolingual non-NL approach** can be expressed in various ways. The most general term may be semantic parsing (Berant and Liang, 2014; Su and Yan, 2017), but is referred diversely in the literature. It is genuinely the most efficient way to store the information of an utterance. For example, the sentence above *"hey can you tell me when the rain stops in tokyo"* can be summarized as the following: {**type**: *wh-question*, **domain**: *weather*, **intent**: *get_time*, **argument**: *rain_stop*, **item**: *tokyo*}. The format can vary, but still the information of the utterance is arranged into a structured format, here SQL (Zhong et al., 2017), and sometimes probably table or so. The format makes it easier for the machines to understand the genuine intent of the directive utterances, and there is rarely the chance that the information is omitted.

However, on the other hand, making a paraphrase from such a format is another issue. At a glance, it is straightforward that human participants are not familiar with the above-structured format in case they make up the sentences with it. The organized data conveys the semantic properties of the sentence effectively, but such decomposition may not be helpful for the human language processing if one wants to understand the overall purpose, meaning, or nuance of the utterance.

Though SQL or table-based sentence generation is widely done automatically these days utilizing the pre-trained networks (Liu et al., 2018), human-aided diversification of the sentences is still essential. Thus, we concluded that there is more room to be developed beyond the mentioned schemes, especially considering the flexibility of the content format and human-friendliness of the generation scheme.

2.3. The Proposed Scheme

Type	Denotations	Discourse Component	Force
Declaratives	proposition (p)	Common Ground	Assertion
Interrogatives	set of propositions (q)	Question Set	Asking
Imperatives	property (P)	To-Do List Function	Requiring

Table 1: Clause types and their properties (Portner, 2004).

Finally, as claimed above, one can think of **monolingual NL approach**, which can be replaced with terms such as summarization, sentence rewriting, keyphrase extraction, etc. This was tackled previously as rewriting canonical form of the statement and is still valid, as shown in the recent studies (Dong et al., 2017). The different point is, we adopt the sentence segments, to be specific a *phrase*, to represent the core content of a question or command. The phrase is called either a question set (QS) or a to-do list (TDL) as the terms arranged in Portner (2004) (Table 1)[2]. In detail, QS is defined as a set of curious components that the speaker wants to exploit from the addressee, and TDL implies a counterpart of QS regarding real action.

[1] Here, NL denotes natural language, especially sentences or sentence segments.

[2] Basically QS and TDL are syntactic representatives regarding interrogatives and imperatives, but here interpreted as speech act-level components that denote (possibly nominalized) wishlist, each carrying the force of asking or requiring. Also, in this paper, we use the terms *discourse component*, *intent argument*, and *keyphrase* interchangeably.

Although there are some points to be clarified regarding rhetorical question/commands, suggestions or *would* questions in the actual process[3], instead, here we want to note that the monolingual NL approach displays its extensibility, reproducibility, and human-friendliness.

Extensibility: The less competitive point of the previous SQL-based methods is that the notation of the core content accompanies the limitation in the number of possible intents or arguments. In many cases, the parsing aims to obtain a structured output that is determined upon proper value mapping or classification for the determined slots, that some information can be inadvertently removed. The NL-based approach prevents such phenomenon by preserving the contents that are relevant to the topic or intention.

Reproducibility: For SQL-based information extraction, a complicated annotation system is required, and also the quality checking is not straightforward. Instead, the NL-based approach provides an easily reproducible scheme, especially for the directive utterances that are of interest. The consistency of representation can happen to be lower than the table or SQL-based approaches, but especially in the extraction process that aims a structuring of non-canonical utterances, NL-based approach may be helpful.

Human-friendliness: The most competitive point of the proposed scheme regarding paraphrase generation is that the approach is both monolingual and NL-based. It is straightforward that making up the paraphrase usually accompanies either human generation (Cohn et al., 2008) or human quality check (Park et al., 2019; Zeng et al., 2019). For either process, it is beneficial for the participants to be aware of the core content of every utterance they generate or check. Moreover, especially for directives, the NL-based representation can be more accessible to the participants, since it is usually considered difficult to insert the act-related features (e.g., if the question is polar or alternative) to non-NL-based formats.

3. Corpus Construction

In this section, we aim to show how the canonical or non-canonical directives can be paraphrased via structured monolingual simplification. Here, we first define the types of directives that are of interest and then introduce how they are simplified to a structured format. Next, we explain the procedure of the human-aided generation in detail, along with the feedback and modification process.

3.1. Utterances of Interest

Our utterances of interest are some subtypes of question and command, which were categorized for sentence-keyphrase pairs in Cho et al. (2018b) (Table 2). Thus, we aimed to perform data augmentation and make up the sentence similarity dataset at the same time, utilizing the human resource.

The original corpus, which was constructed regarding the Korean language, lacked three kinds of utterance types, and especially the need for more *wh*-questions was significant.

	Types		**Correspondings**
Questions	*Yes/no*		whether or not *-(in)ci, yepwu*
	Alternative		what is/to do between *-lang -cwung -han/hal kes*
	Wh-questions	*Who*	person, identity *sa-lam, ceng-chey*
		What	meaning *uy-mi*
		Where	location, place *wi-chi, cang-so*
		When	time, period, hour *si-kan, ki-kan, si-kak*
		Why	reason *i-yu*
		How	method, measure *pang-pep, tay-chayk*
Commands	*Prohibitions*		Prohibition: not to - *-ci anh-ki*
	Requirements		Requirement: to - *-(ha)-ki*
	Strong Requirements		Requirement: to - *-(ha)-ki*

Table 2: Structured annotation scheme for the Korean language; more details available in Cho et al. (2018b). The corresponding terms are specific for the Korean language, but can be interpreted into other languages if given proper replacements (e.g., *riyuu* for 'reason' in Japanese).

We first made up 400 intent arguments for alternative questions (Alt. Q), prohibitions (PH)[4], and strong requirements (Str. REQ)[5], and consequently, 800 more phrases were created for *wh-* questions (*Wh-* Q), to reflect the above motivations. Considering the content, we have balanced the number of keyphrases for the widely used topics such as email, smart home, appointment scheduling, and weather.

- {**Topic:** Weather, **Type:** *Wh-* Q}
  - **Argument:** The time that rain stops today
  - **Example**: *do you know when the rain stops today*

- {**Topic:** Email, **Type:** PH}
  - **Argument:** Not to clean the spam mail box
  - **Example**: *please don't remove the spam mails*

- {**Topic:** Smart home, **Type:** Alt. Q}
  - **Argument:** The place that the light is on between living room and terrace
  - **Example**: *between living room and terrace you know where the light is on*

- {**Topic:** Appointment scheduling, **Type:** Str. REQ}
  - **Argument:** To make an appointment in the afternoon on the weekend
  - **Example**: *on the weekend you should make an appointment in the afternoon, not in the evening*

[3] Specification on the disambiguation can be found in (Cho et al., 2018a).

[4] For clarification, PH does not necessarily denote negated statements. All the directives that prevent the addressee from doing some action can be regarded as PH. Though some might overlap with REQ, there is clearly the set of utterances that the only PH can represent; e.g., "*Could you refrain from touching my hands?*" for *not touching the hand of the speaker*, which cannot be replaced with other REQ utterances.

[5] Str. REQ is the command type that contains both a requirement and its negative counterpart, as in the example above.

Act/Topic	Email	Scheduling	S. Home	Weather
Alt. Q	100	100	100	100
***Wh-* Q**	200	200	200	200
PH	100	100	100	100
Str. REQ	100	100	100	100

Table 3: Composition of the intent arguments (keyphrases). For each intent argument 10 different paraphrases are generated.

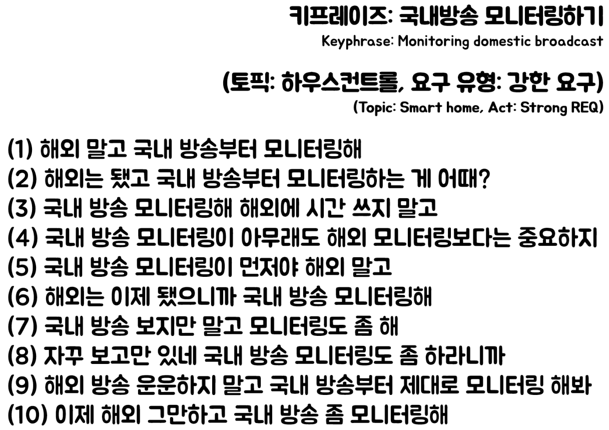
All the phrases were checked their felicity by at least three native Korean speakers. We display the balancedness of the intent argument set regarding the topic, in Table 3.

3.2. Sentence Generation and Verification

Next, we performed a human-aided generation with 4 participants. Each created the utterances regarding four topics; the request was to make up ten utterances for each intent argument, as diversely as possible. The final output sentences were checked by two more Korean natives and have undergone the adjudication and modification. All the mutual feedbacks related to making up the sentences were done actively as the generation went on. The followings were mainly informed to the participants before and during the process:

- Ten sentences should be written in different styles as much as possible. At this time, the style incorporates all of politeness, honorific, and nuance.
- The participant does not have to repeat the terms that are in the argument, and can put in different words/phrases/idioms depending on the circumstance. The expression should be suitable for spoken language.
- It is also recommended to pursue the diversity of sentence structure through scrambling.
- In the case of *wh*-question, *wh*-particles are essential, and alternative questions may be inserted in some cases. Both utterance types need not be written interrogatively.
- In the case of prohibition, the participant should generate an utterance that prevents any action that the addressee may do, at the same time conveying substantial force compared to *permission for not doing* (i.e., *you don't have to ##*). If prohibiting the action is equivalent to requiring another action, replacing the expression is allowed (e.t., *stay* for *do not leave*).
- Both prohibition and strong requirements need not be imperative but should have the purpose of preventing or forcing the action of the addressee.
- For arguments that include the notation of speaker/addressee, the participant should use a corresponding pronoun expression in the sentence generation.

To display the diversity of the generated sentences and how the paraphrases are not obvious nor repetitive, we present a sample case on an intent argument regarding smart home (Figure 2). In specific, the phrase is *monitoring domestic broadcast* and the act is *strong REQ*. The exact translation for all the ten utterances is omitted here due to the difficulty in conveying the nuance, but we notate a raw text so that the detail can be useful for some audience.



키프레이즈: 국내방송 모니터링하기
Keyphrase: Monitoring domestic broadcast

(토픽: 하우스컨트롤, 요구 유형: 강한 요구)
(Topic: Smart home, Act: Strong REQ)

(1) 해외 말고 국내 방송부터 모니터링해
(2) 해외는 됐고 국내 방송부터 모니터링하는 게 어때?
(3) 국내 방송 모니터링해 해외에 시간 쓰지 말고
(4) 국내 방송 모니터링이 아무래도 해외 모니터링보다는 중요하지
(5) 국내 방송 모니터링이 먼저야 해외 말고
(6) 해외는 이제 됐으니까 국내 방송 모니터링해
(7) 국내 방송 보지만 말고 모니터링도 좀 해
(8) 자꾸 보고만 있네 국내 방송 모니터링도 좀 하라니까
(9) 해외 방송 운운하지 말고 국내 방송부터 제대로 모니터링 해봐
(10) 이제 해외 그만하고 국내 방송 좀 모니터링해

Figure 2: Ten sentences generated from a single *strong REQ* argument: *monitoring domestic broadcast*.

Nonetheless, one may observe that there is sufficient non-triviality. For instance, imperatives such as (1) (commands) and interrogatives such as (2) (request), indirective speech acts such as (4) (meaning "*Monitoring domestic broadcast is more important than overseas, by all means*") or (5) (meaning "*Monitoring domestic broadcast is first dude, not the overseas*") all share the same core content. Moreover, due to the Korean language being scrambling, (1, 2, 6, 9) and (3, 5, 7) indicate the same intent argument despite their order of the clauses reversed[6].

4. Sentence Similarity Dataset

As a next step, we aimed to address the relevance between the collected utterances and make a dataset on non-trivial paraphrases. In specific, with the sentences that are created, we have undertaken a further arrangement to make a labeled corpus that contains the sentence pairs. Every sentence pair is labeled in five categories:

- **Class 0:** No relevant speech act nor topic
- **Class 1:** Relevant act but no relevant topic
- **Class 2:** No relevant act but the relevant topic
- **Class 3:** Relevant act and topic
- **Class 4:** Paraphrase

The classification of the sentence pairs was performed in a straightforward manner. First, for all the generated sentences regarding each intent argument, we labeled the intent argument index (IAI), topic (TOP), and speech act (ACT) of the utterances. Two hypotheses were presumed that (1) the utterances with the same intent argument are the paraphrases of each other, and (2) act and topic are two independent components that decide the similarity of the utterances, at least in our formulation.

[6] For the former, the negation on monitoring overseas comes first, while it comes last for the latter.

Property / Label (Volume)	**SAME_ARG**	**SAME_ACT**	**SAME_TOP**
Class **0** (270,000): **No relevance**	0	0	0
Class **1** (105,000): **The act matches**	0	1	0
Class **2** (90,000): **The topic matches**	0	0	1
Class **3** (34,500): **Topic & act match**	0	1	1
Class **4** (45,000): **Paraphrase**	1	1	1

Table 4: The property of the sentence pair classes and the corresponding number of pairs.

Among the sentence pairs with the same intent argument index (SAME_ARG, $\text{IAI}_1 = \text{IAI}_2$), 45 sentence pairs (in other words, $n(n-1)/2$ for $n = 10$) were extracted. Thus, for total 1,000 intent arguments, we obtained 45,000 pairs of paraphrases.

Next, we adopted some heuristics to arrange the sentence pairs regarding the relations defined above. First, we assumed all the combinations where the intent argument differs ($\text{IAI}_1 \neq \text{IAI}_2$), where the number of all the cases turned out to be $499,500 = 1000 \times 999/2$. Consequently, we defined two relations: SAME_TOP (if $\text{TOP}_1 = \text{TOP}_2$) and SAME_ACT (if $\text{ACT}_1 = \text{ACT}_2$). We randomly chose a sentence among each group of sentences and defined the sentence pair relations as: 0 (if not SAME_TOP and not SAME_ACT), 1 (if not SAME_TOP and SAME_ACT), 2 (if SAME_TOP and not SAME_ACT), and 3 (if SAME_TOP and SAME_ACT). The whole decision process is organized in Table 4. In total, we achieved a total dataset of size 544,500, with the specifications as in the table.

There are two main characteristics of this categorization process. First, since the process involves topic and act, which are two main factors that can influence the core content of the directive utterance, the distinction between non-relevant utterances (label 0) and the relevant utterances (labels 1-4) could be available. Secondly, by labeling the utterances in the order of degree of similarity, the result can be utilized in both classification and regression tasks[7].

5. Experiment

5.1. Models and Implementation

Here, we design some model architectures that learn and infer how similar the sentences are within the pairs. The two sentences are either concatenated with a separation token in between, or fed as an input of parallel neural networks. In detail, we utilize a bidirectional long short term memory (BiLSTM) (Schuster and Paliwal, 1997), self-attentive BiLSTM (Lin et al., 2017), parallel BiLSTM (Mueller and Thyagarajan, 2016; Yoon et al., 2019), and BiLSTM cross-attention (Lee et al., 2018; Cho et al., 2019a). For the mentioned feature-based approaches, we adopted the multi-hot encoding (Song et al., 2018), which was shown well-performing for the sentence classification (Cho et al., 2019b). All the token embeddings were done at character-level since all the characters in the Korean language can be considered as a sort of subword or word piece (Sennrich et al., 2015). The architecture and hyperparameter specification are to be provided along with an on-line repository that incorporates all the dataset and codes.

	Acc.	**F1 score**	**Param.s**	**Per epoch**
(a) BiLSTM (BRE)	93.84	0.8217	43K	1200s
(b) BRE-Att	98.75	0.9650	163K	1700s
(c) Para-BRE-Att	99.40	0.9833	319K	1200s
(d) Para-BRE-CA	99.44	0.9844	325K	1600s

Table 5: Accuracy (%), F1 score, trainable parameters (param.s), and average training time per epoch for **(a)** BiLSTM; bidirectional recurrent encoder (BRE), **(b)** self-attentive BRE, **(c)** parallel concatenation of self-attentive BREs, and **(d)** parallel BREs with cross-attention. The validation was done on 10% of the total dataset.

5.2. Result

The result is in Table 5. To prevent the choice of the models that have luckily reached a high accuracy or F1 score, we used a new type of selection among the high-performance cases. For each architecture, among the five best accuracy and five best F1 score models, we chose the intersection with the better result, to both guarantee the practicability and generalizability.

It is encouraging that the trained systems show reliable performance considering the quantitative evaluation. The result is convincing, given that we did not utilize any pre-trained dense vectors. The performance is expected to be improved with an additional dictionary, but here we only used the training-free sparse features to emphasize that the created dataset itself has sufficient lexicon and expressions which are useful in paraphrase identification. The phenomenon seems to originate in sufficient size of the dataset, although domain-related overfitting is probable. We propose a model analysis of the performance regarding what each implemented architecture shows.

5.2.1. Attention

With the models (a) and (b), we observed that the self-attentive embedding has a significant effect, enhancing the absolute accuracy by nearly five percent. This implies that pointing out where to look at can be much more beneficial even in our setting of sentence similarity test. Also, it can be interpreted that even though the dataset we constructed contains various non-canonical sentence forms, scrambling, and colloquial expressions, an attention-based neural network training is expected to find a latent factor that determines similarity, beyond the observable features.

5.2.2. Parallel than series

The next issue regards the placement of the features, whether to put them in series in (a, b) or to make a parallel organization. We expected that the latter would prevent the vanishing gradient and also discover the latent relationship

[7] It can be controversial that the topic-related utterances are more similar than the utterances with the same intention, but we decided this upon the relevance; the subsequent dialog utterances happen more to be of same topic rather than same act, as can also be observed in the subtasks such as next sentence prediction (Devlin et al., 2018).

between the components of each sentence. It turned out to be promising, and the convergence was much faster in the parallel case. The main reason is assumed to be the flexibility of attention weight that is more guaranteed in the parallel case due to the shorter hidden layer sequence length.

5.2.3. Cross-attention

The final part of our model analysis is using cross attention (d), not a simple parallelization (c). The architecture was first proposed in an image-text matching task (Lee et al., 2018), and was recently extended into the co-utilization of sequential data (Cho et al., 2019a). Here, we borrowed this implementation[8] for the simultaneous interaction of information between the given input sentences. We found that it makes a slight enhancement in performance, by leveraging the inter-sentence analysis in (b) and flexible attention weight in (c).

6. Discussion

We want to claim that our research has three main contributions as:

- An empirical comparison of the mono/bi-lingual (non-)NL-based approaches on content extraction and paraphrasing
- A detailed scheme for making up a sentence similarity dataset regarding intent argument, topic, and speech act
- A large-scale sentence similarity and paraphrase corpus on Korean, with reasonably high evaluation result utilizing the conventional NLP approaches

6.1. Limitations

The limitation of our work lies in the similar space where the advantage occupies. Thanks to the flexibility of the NL-format core content extraction, we could obtain various expressions that are familiar to human language usage and also easily achievable by people. However, at the same time, it cannot guide us to the structural formalization of the core content. It is not problematic at this point since we aim to make up a large-scale sentence similarity dataset, but when it comes to semantic parsing, the conventional methodologies might be more advantageous. Also, unlike the case of *directive utterances* that structuring a question set or to-do list brings conciseness, for the *statements* that aim to inform something or express the speaker's thoughts, NL-format expression may not give a sufficient benefit.

6.2. Applications

Notwithstanding some limitations, our approach and its resulting materials have some advantages in the application viewpoint. First, in the industry where the high-quality similarity dataset is required, the human-aided generation of the paraphrases is valid so far. For such cases, our scheme may provide a structured guideline for the participants to make up the sentences, beyond just suggesting an SQL format or a canonical utterance. Besides, the corpus that is constructed, which shows fairly high classification accuracy, can be of great help for the spoken language understanding (SLU) systems of the personal agents.

Next, the proposed scheme may unexpectedly fit well with the automation, given that there have been approaches that utilize the canonical form of the directives (Dong et al., 2017) in making up a paraphrase. Our recent work (Cho et al., 2019c) includes the automated extraction of the intent arguments from some non-canonical directives, which suggests the probable utility of the proposed scheme and dataset as to make up an automatic paraphrasing system. The detailed implementation is assumed to incorporate the process of integrating the extraction module along with the paraphrasing system constructed based on this dataset.

Moreover, though the categorization of the directive utterances may depend on the target language, the extraction of discourse component and the sentence reproduction are expected to be consistent in reasonably many languages. The supporting details can be found in Huddleston (1994) and Portner (2004), where the syntax-semantic definition on question and command was proposed; questions have mainly three representative types, as adopted in our study. The distinction for the commands is not theoretically complete, but was employed in our previous study (Cho et al., 2018b) based on the frequent appearances in the corpus, and also is semantically straightforward. Note that our work has yet been proved valid typologically. The specification for the intent arguments' format can vary upon language. As future work, we plan to expand our dataset into several more languages, including Japanese and English, by translation and by making up another corpus, respectively.

7. Conclusion

In this paper, we generated the topic-labeled directive utterances based on the concept of discourse component and finally arranged a sentence similarity test set with them, concerning both the topic and act of the utterances. The intent arguments were manually created regarding four topics of email, appointment scheduling, smart home, and weather, with the structured format that corresponds to the four types of act (Alt. Q, *Wh-* Q, PH, and Str. REQ).

The dataset contains five types of sentence pairs, namely paraphrases and four other relations that vary upon the relevance regarding topic and intention. Since the corpus was constructed with a guideline that asks the participants to exploit as many diverse sentence styles as possible, and also the validity of the sentences was reliably checked, it is expected that the resulting corpus fit with various tasks regarding directive utterances including non-canonical expressions. The performance evaluation utilizing the sparse character-level embedding

Our next goal is to find a reliable and efficient corpus construction scheme that can also cover the non-directive utterances, including rhetorical questions, expressives, and statements, possibly in a multilingual manner. The models and generated dataset in this paper are freely available online[9]. The multilingual versions and theoretical investigation will be added as future work.

[8] The implementation that adopts speech-text multimodal dataset is available at https://github.com/warnikchow/coaudiotext where we referred to regarding the other models (a-c) as well for the Keras (Chollet and others, 2015)-based framework.

[9] https://github.com/warnikchow/paraKQC

8. Acknowledgements

This project was funded 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, this work was partly supported by Clova AI Research, NAVER. The corpus construction would not have been possible without the support of total eight great participants, namely Eunah Koh, Kyung Seo Ki, Sang Hyun Kim, Kimin Ryu, Dongho Lee, Yoon Kyung Lee, Minhwa Chung, and Ye Seul Jung. After all, the authors appreciate the helpful advices provided by three anonymous reviewers, Reinald Kim Amplayo, Sanghwan Moon, Jio Chung, and †Kyuwhan Lee.

9. Bibliographical References

Achananuparp, P., Hu, X., Zhou, X., and Zhang, X. (2008). Utilizing sentence similarity and question type similarity to response to similar questions in knowledge-sharing community. In *Proceedings of QAWeb 2008 Workshop, Beijing, China*, volume 214.

Agirre, E., Diab, M., Cer, D., and Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics.

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., and Guo, W. (2013). * sem 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43.

Berant, J. and Liang, P. (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.

Bhagat, R. and Hovy, E. (2013). What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Cho, W. I., Lee, H. S., Yoon, J. W., Kim, S. M., and Kim, N. S. (2018a). Speech intention understanding in a head-final language: A disambiguation utilizing intonation-dependency. *arXiv preprint arXiv:1811.04231*.

Cho, W. I., Moon, Y. K., Kang, W. H., and Kim, N. S. (2018b). Extracting arguments from korean question and command: An annotated corpus for structured paraphrasing. *arXiv preprint arXiv:1810.04631*.

Cho, W. I., Cho, J., Kang, W. H., and Kim, N. S. (2019a). Text matters but speech influences: A computational analysis of syntactic ambiguity resolution. *arXiv preprint arXiv:1910.09275*.

Cho, W. I., Kim, S. M., and Kim, N. S. (2019b). Investigating an effective character-level embedding in korean sentence classification. *arXiv preprint arXiv:1905.13656*.

Cho, W. I., Moon, Y. K., Moon, S., Kim, S. M., and Kim, N. S. (2019c). Machines getting with the program: Understanding intent arguments of non-canonical directives. *arXiv preprint arXiv:1912.00342*.

Chollet, F. et al. (2015). Keras. `https://github.com/fchollet/keras`.

Cohn, T., Callison-Burch, C., and Lapata, M. (2008). Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics*, 34(4):597–614.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, L., Mallinson, J., Reddy, S., and Lapata, M. (2017). Learning to paraphrase for question answering. *arXiv preprint arXiv:1708.06022*.

Fernando, S. and Stevenson, M. (2008). A semantic similarity approach to paraphrase detection. In *Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics*, pages 45–52.

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

Lee, K.-H., Chen, X., Hua, G., Hu, H., and He, X. (2018). Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216.

Lin, Z., Feng, M., Santos, C. N. d., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. (2017). A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Liu, T., Wang, K., Sha, L., Chang, B., and Sui, Z. (2018). Table-to-text generation by structure-aware seq2seq learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Mallinson, J., Sennrich, R., and Lapata, M. (2017). Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893.

Martin, R. (1976). Inférence, antonymie et paraphrase éléments pour une théorie sémantique.

Mueller, J. and Thyagarajan, A. (2016). Siamese recurrent architectures for learning sentence similarity. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Park, H., Gweon, G., and Heo, J. (2016). Affix modification-based bilingual pivoting method for paraphrase extraction in agglutinative languages. In *Big Data and Smart Computing (BigComp), 2016 International Conference on*, pages 199–206. IEEE.

Park, S., Hwang, S.-w., Chen, F., Choo, J., Ha, J.-W., Kim, S., and Yim, J. (2019). Paraphrase diversification using counterfactual debiasing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6883–6891.

Ponti, E. M., Reichart, R., Korhonen, A., and Vulić, I. (2018). Isomorphic transfer of syntactic structures in cross-lingual nlp. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1531–1542.

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

Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., and Black, A. W. (2018). Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

Prakash, A., Hasan, S. A., Lee, K., Datla, V., Qadir, A., Liu, J., and Farri, O. (2016). Neural paraphrase generation with stacked residual lstm networks. *arXiv preprint arXiv:1610.03098*.

Resnik, P., Buzek, O., Hu, C., Kronrod, Y., Quinn, A., and Bederson, B. B. (2010). Improving translation via targeted paraphrasing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 127–137. Association for Computational Linguistics.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Song, C., Han, M., Cho, H. Y., and Lee, K.-N. (2018). Sequence-to-sequence autoencoder based korean text error correction using syllable-level multi-hot vector representation. In *Proceedings of HCLT [in Korean]*, pages 661–664.

Su, Y. and Yan, X. (2017). Cross-domain semantic parsing via paraphrasing. *arXiv preprint arXiv:1704.05974*.

Tomuro, N. (2003). Interrogative reformulation patterns and acquisition of question paraphrases. In *Proceedings of the second international workshop on Paraphrasing-Volume 16*, pages 33–40. Association for Computational Linguistics.

Vila, M., Martí, M. A., and Rodríguez, H. (2014). Is this a paraphrase? what kind? paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, 4(01):205.

Yoon, S., Byun, S., Dey, S., and Jung, K. (2019). Speech emotion recognition using multi-hop attention mechanism. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2822–2826. IEEE.

Zeng, D., Zhang, H., Xiang, L., Wang, J., and Ji, G. (2019). User-oriented paraphrase generation with keywords controlled network. *IEEE Access*, 7:80542–80551.

Zhang, Y., Baldridge, J., and He, L. (2019). Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.

Zhong, V., Xiong, C., and Socher, R. (2017). Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.