creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

Ph.D. DISSERTATION

Ambiguity Resolution in Spoken Language Understanding

음성언어 이해에서의 중의성 해소

BY

CHO WON IK

AUGUST 2022

DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Ph.D. DISSERTATION

Ambiguity Resolution in Spoken Language Understanding

음성언어 이해에서의 중의성 해소

BY

CHO WON IK

AUGUST 2022

DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Ambiguity Resolution in Spoken Language Understanding

음성언어 이해에서의 중의성 해소

지도교수 김 남 수
이 논문을 공학박사 학위논문으로 제출함

2022년 7월

서울대학교 대학원

전기 · 정보공학부

조 원 익

조원익의 공학박사 학위 논문을 인준함

2022년 6월

위 원 장: _____ (인)

부위원장: _____ (인)

위 원: _____ (인)

위 원: _____ (인)

위 원: _____ (인)

Abstract

Ambiguity in the language is inevitable. It is because, albeit language is a means of communication, a particular concept that everyone thinks of cannot be conveyed in a perfectly identical manner. As this is an inevitable factor, ambiguity in language understanding often leads to breakdown or failure of communication.

There are various hierarchies of language ambiguity. However, not all ambiguity needs to be resolved. Different aspects of ambiguity exist for each domain and task, and it is crucial to define the boundary after recognizing the ambiguity that can be well-defined and resolved.

In this dissertation, we investigate the types of ambiguity that appear in spoken language processing, especially in intention understanding, and conduct research to define and resolve it. Although this phenomenon occurs in various languages, its degree and aspect depend on the language investigated. The factor we focus on is cases where the ambiguity comes from the gap between the amount of information in the spoken language and the text.

Here, we study the Korean language, which often shows different sentence structures and intentions depending on the prosody. In the Korean language, a text is often read with multiple intentions due to multi-functional sentence enders, frequent pro-drop, *wh*-intervention, etc. We first define this type of ambiguity and construct a corpus that helps detect ambiguous sentences, given that such utterances can be problematic for intention understanding.

In constructing a corpus for intention understanding, we consider the directivity and rhetoricalness of a sentence. They make up a criterion for classifying the intention of spoken language into a statement, question, command, rhetorical question, and rhetorical command. Using the corpus annotated with

sufficiently high agreement on a spoken language corpus, we show that colloquial corpus-based language models are effective in classifying ambiguous text given only textual data, and qualitatively analyze the characteristics of the task.

We do not handle ambiguity only at the text level. To find out whether actual disambiguation is possible given a speech input, we design an artificial spoken language corpus composed only of ambiguous sentences, and resolve ambiguity with various attention-based neural network architectures. In this process, we observe that the ambiguity resolution is most effective when both textual and acoustic input co-attends each feature, especially when the audio processing module conveys attention information to the text module in a multi-hop manner.

Finally, assuming the case that the ambiguity of intention understanding is resolved by proposed strategies, we present a brief roadmap of how the results can be utilized at the industry or research level. By integrating text-based ambiguity detection and speech-based intention understanding module, we can build a system that handles ambiguity efficiently while reducing error propagation. Such a system can be integrated with dialogue managers to make up a task-oriented dialogue system capable of chit-chat, or it can be used for error reduction in multilingual circumstances such as speech translation, beyond merely monolingual conditions.

Throughout the dissertation, we want to show that ambiguity resolution for intention understanding in prosody-sensitive language can be achieved and can be utilized at the industry or research level. We hope that this study helps tackle chronic ambiguity issues in other languages or other domains, linking linguistic science and engineering approaches.

keywords: Spoken language understanding, Natural language processing, Ambiguity, Intention understanding
student number: 2014-22579

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Spoken language processing (SLP) encompasses speech signal processing, which is essential for managing spoken data, and natural language processing, which utilizes transcription that has already been transformed from speech into text. Sometimes, these two processes are integrated to comprise an end-to-end structure. Especially, Spoken language understanding (SLU) refers to the overall process of performing tasks related to language understanding by processing spoken language in a smart device or spoken dialogue system [1, 2, 3, 4]. Downstream tasks performed upon this include intent understanding and slot-filling [5, 3], emotion recognition [6, 7], dialog act classification [8, 9], and intention understanding [10, 11]. The processes are different from text mining, which targets only textual features, in that SLP or SLU concerns information that can be derived from speech that is actually uttered, and it may contain non-verbal information that cannot be encoded in text.

The automatic speech recognition (ASR) and natural language understanding (NLU) pipeline in SLU is a *de facto* structure widely used in industry and daily life. However, since the ASR-NLU pipeline is a cascaded structure, it may inevitably be threatened by the propagation of errors [12]. For example, an error in the ASR process directly affects the validity of the NLU process, or even

if the ASR is correct, information loss in the speech transcription process may make it difficult to perform downstream tasks correctly. To solve these problems, many end-to-end approaches have been proposed [13, 14, 15], and they are considered as ideal direction so far. However, the dataset for the training of end-to-end architectures may not be sufficient from time to time (and also by language), and interpreting the inference procedure of black-box end-to-end models is also challenging.

The ambiguity of spoken language is the main hurdle for reliable and accurate SLU. Though language itself is ambiguous as an artifact, especially spoken language displays ambiguity regardless of whether data is handled in speech or text format. Even with the speech that contains significantly more information than textual language, ambiguity is caused by the absence of contexts such as dialogue history or social relationship, or the presence of linguistic phenomena such as homophony or polysemy [16]. In addition, relying only on text data where many non-verbal features are absent or omitted, more diverse types of ambiguity can be further observed [17]. Such ambiguity comprehensively affects the performance of tasks related to phonetic and phonological characteristics, for instance, ASR and emotion recognition, as well as syntactic tasks such as dependency parsing and semantic tasks like intention understanding.

1.1 Motivation

Enhancing the reliability of the SLU process depends highly on the usage of architectures and the reduction of ambiguity. However, choosing the appropriate architecture for SLU is an open problem since the utility of each approach (pipeline or end-to-end, or hybrid) differs by the downstream task. Also, since the fully end-to-end structure in the current machine learning application is

not realistic yet, it is still meaningful to develop the problem in a data-centric way and handle the issue in a generalizable framework.

To tackle the ambiguity present in spoken language understanding, we should first understand and define the ambiguity in spoken language caused by the lack of prosody and which problems can be followed. Such ambiguity is defined as one of the various layers of linguistic ambiguity, and it refers to a situation in which multiple interpretations are possible depending on the prosody change even if the same text is given [16]. There are also lexical ambiguity and syntactic ambiguity that may not be resolved even if prosody is given, but the reasons of these ambiguities are not as apparent as the absence of acoustic data, and this means that such ambiguities would not be resolved merely by providing sufficient contextual information. Therefore, defining the ambiguity that can be resolved if provided with prosody and discussing the way to disambiguate it can be an essential step towards ambiguity resolution in SLP.

Let us apply the above issue again to spoken language understanding. SLU is often aligned with concepts such as intent understanding and slot-filling, but in this dissertation, we want to note that good intention understanding is the cornerstone of a reliable SLU process. Unlike the detailed, item/argument-based approach dealt within general SLU, intention understanding concentrates on identifying speech act and specifying the illocutionary act of an utterance [18, 19, 20, 21]. Though defining ambiguity for all SLU domains is broad and challenging, it might be meaningful if such a concept is defined to detect the syntax-semantics properties such as directiveness or rhetoricalness of the sentence input.

In usual intention understanding, an utterance is classified into a specific group of sentences, and such classification is decisive given spoken language or textual data. In contrast, we assume a scenario that a spoken language

can belong to multiple classes of intention if they are projected to a textual form, given a specific categorization for intention understanding. In many SLU pipelines, transcribing the text from speech input results in a certain amount of information loss, such as intonation, duration, or nuance. Even though the utterance is decisive in a spoken manner, it may bring ambiguity in the interpretation if such acoustic features are omitted in the transformation. However, in reverse, this suggests that we can recognize whether a textual utterance incorporates potential ambiguity and check if the ambiguity can be resolved with prosodic information, provided only with textual features. It is more like detecting the possibility of multiple arguments in a written sentence. Such a text-based ambiguity detection module can be developed and applied to a pipeline or hybrid SLU system to help reduce the ambiguity of overall spoken language processing.

1.2 Research Goal

Our research goal is to define ambiguity in the speech intention understanding that can be resolved if provided with prosody and to suggest a step towards its disambiguation, here for a prosody-sensitive language, Korean. The former regards creating a new speech act category that considers prosody and intonation of the spoken utterance only given the textual data, while the latter concerns identifying the genuine intention of ambiguous utterances accompanying additional features such as acoustic information.

With these findings, we finally aim to make up an integrated system that helps SLU systems efficiently deal with ambiguous utterances that may cause malfunction or a bad user experience. In order to reduce the burden of computation for all utterances in spoken language understanding, a system can first detect utterances where the text input encompasses ambiguity that can be

resolved with prosody. With the intention disambiguated given acoustic information, the full SLU process would be more robust and the transcribed text will be adopted in a much more reliable manner in the further application.

1.3 Outline of the Dissertation

In Chapter 2, we briefly skim the literature on spoken language understanding and speech acts. This is extended as well to the Korean language, which is highly contextual and relies largely on prosodic information.

In chapter 3, we show how we define ambiguity in Korean speech intention understanding. We create a Korean spoken language corpus that consists of various directives, non-directives, and intonation-dependent utterances. The system trained upon it aims at distinguishing utterances that require prosodic information for disambiguation of their intention.

In chapter 4, we newly create an artificial language corpus that consists only of prosody-sensitive utterances, and search for the appropriate architecture for the classification. We find out that the SLP model which is aligned with the attention from the text processing module performs best, claiming that text matters but speech influences in prosody-sensitive speech intention understanding.

In chapter 5, we show the utility of our methodology by suggesting that SLP systems aided by the proposed approach can provide more efficient and reliable information for various downstream tasks, such as spoken dialogue management and spoken language translation.

We conclude the dissertation with a summary and further remarks.

Chapter 2

Related Work

2.1 Spoken Language Understanding

Spoken language understanding (SLU) is a subdiscipline of spoken language processing (SLP). SLP includes analyses of speech input, the process of converting speech that is an analog signal, into language, a symbolic data, encoding language into text, and representing various para-linguistic features in the numeric and abstractive way [22]. Language understanding refers to multiple downstream tasks that can be performed by comprehensively using such high-level features or by using semantics of utterance from speech input itself.

SLU is mainly regarded as intent classification and slot-filling in the literature of SLP [4]. This concerns that SLU is widely used in industry or daily life, and it can also be said that SLU is exploited most usefully in that way considering how domain-specific items and arguments are selected in tasks such as ATIS [5] or fluent speech command [3]. However, given that natural language understanding (NLU) does not simply look at the practical aspects of text mining but includes various other subtasks such as linguistic acceptability [23], natural language inference [24], and semantic textual similarity [25], we think it necessary to expand the meaning of SLU considering general

language understanding. Accordingly, with the broader viewpoint of SLP, we could interpret the downstream tasks performed using sentence semantics as SLU, which is beyond merely transforming speech into text (automatic speech recognition, ASR) or verifying the speaker identity (speaker verification).

SLU is an essential issue for both humans and machines. Human language understanding is generally achieved by a combination of auditory sensory and understanding system, which includes a process in which analog signals are symbolized into information in the brain [26]. A similar process occurs in the spoken language understanding of the machine, and followingly, machine understanding refers to various studies on the human understanding process.

As the human understanding process is not fully known, opinions on the effectiveness of various methodologies are not unified in machine understanding as well. For example, approaches on processing spoken language are primarily divided into the conventional ASR-NLU pipeline [1] and end-to-end architectures [3]. The difference between the two lies in whether speech is intermediately converted into text or not, which is an explicit symbolic format that is more appropriate for information retrieval but lacks para-linguistic features. Despite this difference, the above methodologies can be utilized in both a conventionally defined SLU [5, 3] or extended/simplified SLU tasks [6, 7, 8, 9, 10, 11]. Considering it as a problem of categorization or classification, intent classification/slot-filling is a little more narrowed topic because the domain is mainly restricted and accordingly the number of labels or the label of item/argument is limited [2, 4]. However, intention understanding of general utterances should consider many more edge cases.

Among them, we concentrate on spoken intention understanding, or speech act classification. Intention understanding is a downstream task that investigates sentence properties that are a slightly more fundamental than intent classification and slot-filling [10, 27, 11]. This is also related to distinguishing

situation entity types or tweet acts in NLU [28, 29], and classifying dialog act in conversation analysis [8, 9]. In the literature of NLU, intention and act are used interchangeably across various domains, situations, or target tasks, and we aim to adopt similar terminology and standard in our spoken language understanding. Also, for a theoretical background on this, we would like to look at the basis of the illocutionary and speech acts.

2.2 Speech Act and Intention

The speech act has its origin in the performative hypothesis of Austin and the illocutionary act of Searle.

2.2.1 Performatives and statements

Austin (1962) [18], the beginning of speech act theory, suggests the concept of performative and statement as follows.

- There are numerous utterances that do not belong between true and false categorization (command and question, etc.).
- Sentences that cannot be analyzed as truth condition, that is, *performatives* exist, and are interpreted more than it's spoken, and are effective more than it is described.

Behind this background lies the performative hypothesis to explain implicit performative, which is that "The latent main clause of every sentence incorporates a specific structure of the performative verb." Unfortunately, the theory fails to get much support due to the rebuttal that "There are performatives that cannot be transformed into the structure of performative verb".

However, beginning with the ideal assumption of the above conditions, research on speech act develops to include the following.

- **Expansion of the concept of performatives**: A performative sentence is a particular class of sentences/utterances syntactically and semantically, and includes not only explicit performance sentences but also implicit performative sentences.
- **General theory of dialog act**: Beyond the early dichotomy of performative sentences and statements, the typology expands to general dialog act theory encompassing various performative sentences and statements.

2.2.2 Illocutionary act and speech act

The typology of utterance, in which Austin initially distinguished performatives and statements, is developed to a more complex level involving the speaker and the addressee. When someone utters a sentence, there is also an uttering action, but there is another possible world that the speaker intends, and it is accompanied by phenomena that may actually appear to the addressee who hears it. Austin describes each of these as follows [18, 19]:

- **Locutionary act**: Uttering action
- **Illocutionary act**: What the speaker intends through uttering
- **Perlocutionary act**: An action that is likely to appear as a result to the addressee due to the uttering

Among the acts related to speech, the illocutionary act that we are interested in is the intention of the speaker and is determined by the illocutionary force. Unlike the locutionary act in which the action is determined only by the existence of an utterance or the perlocutionary act in which the action of an addressee is assumed, several illocutionary acts can be accompanied in a single utterance. Searle (1976) [20] categorizes illocutionary acts into five major categories.

- **Representatives**: Corresponds to Austin's statement. It conveys the speaker's beliefs and expresses a proposition with the truth value. Includes conclusions, reports, statements, etc.
- **Directives**: Actions in which the speaker intends the addressee to do something, including advice, orders, questions, prohibitions, etc.
- **Commissives**: The speaker promises to do something by her/himself, including proposals, oaths, promises, refusals, etc.
- **Expressives**: Expressing the speaker's psychological attitude or state, including accusations, congratulations, thanks, praise, etc.
- **Declaration**: A kind of institutionalized performative, including bidding, declaration of war, expulsion, appointment, etc.

Searle's categorization is quite comprehensive, but since there is no definite answer to this type of classification, it is challenging to create a well-defined boundary between sentence types without overlap. For example, a single utterance may simultaneously command something and express the speaker's psychological status, and sometimes declaration and representatives do not seem completely separable. Since these issues are interpreted differently depending on the factors such as dialog history, nuance, the relationship between participants and their social status, and cultural context, the process of finding a suboptimal typology is often a research progress itself.

However, one thing to be clear is that though the speech act is related to the syntactic concept of 'sentence form' [30], they are not the same concept at all. This point is often overlooked when constructing an intention understanding corpus. For example, there are several ways to make a request 'Wake the speaker up on the phone tomorrow morning at 8 o'clock'.

- *I'd like you to call me tomorrow at eight o'clock.*

- *Can you give me a wake-up call tomorrow morning around 8?*
- *Wake me up on the phone tomorrow morning at eight o'clock.*

Each sentence type is declarative, interrogative, and imperative, but ultimately what the speaker wants is to make the addressee wake the speaker up on the phone tomorrow morning. In the third case, the sentence type of imperative corresponds to a direct speech act that matches the dialogue act 'request', but other two cases correspond to an indirect speech act that borrows the form of a declarative sentence or a question. To be detailed, the sentence form determines the illocutionary force in the direct speech act [20], but not in indirect ones.

Comrie and Sadock (1976) [31] and Levinson [32, 33] claimed the literal force hypothesis from these kinds of observations. The key point is that the sentence form is determined according to the speech act, and furthermore, one illocutionary act corresponds to one clause type. However, in general, the sentence form is considered independent of the speech act. Based on these ideas, Gazdar (1981) [34] rebutted the literal force hypothesis with two arguments: 'A single utterance can encompass several dialogue acts' and 'The decision of a dialogue act is influenced by the addressee's interpretation'.

2.2.3 Formal semantic approaches

Portner (2004) [35], grounded on formal semantics, abstracts the problem from situations that can induce blurry boundaries as above. It helps define the clause type of a sentence in terms of speech act. For this, we must first look at how clause types can be categorized. A clause type is a grammatically decisive, straightforward categorization that checks whether a given clause or sentence is an interrogative including *wh-* particles, an imperative with a covert subject, or a declarative. This has been dealt with syntactically in Sadock and Zwicky

(1985) [30], but Portner claims that each can be defined formally semantic as follows.

- **Declarative**: The process of adding the set of Proposition (p) to the common ground (CG) of the addressee and the speaker. In this case, the sentence force is defined as an assertion.
- **Interrogative**: Since a question can be generally defined as 'a set of corresponding answers', it is expressed as a set of propositions (q). The process is adding a question to a set of questions (question set, QS) to be answered by the addressee. At this time, the conventional force of asking is the process of adding an interrogative to QS.
- **Imperative**: The process of adding a property (P) that can be expressed as an imperative to the addressee's to-do list (TDL). More specifically, it means that the conventional force of requiring must be imposed on the addressee, which means that requiring A means adding P, expressed as imperative through the TDL function, to the TDL about A. can. Imperative is divided into three subtypes: order based on social authority, request that is not based on such authority (where speaker and listener both have benefit), and permission to update the speaker's own TDL.

In addition, Portner further proposes the following hypotheses in the paper.

- Discourse context, universally, contains at least a common ground, a set of questions, and a to-do list.
- The generalized update function F = "take a set of x's and another x, and add the new x to the set" is universal.
- Update functions other than F are not universal and F is preferred, since if F can be used to determine the force of a sentence, it must be used.

In a similar perspective to Portner, Allwood (1995) [36] explains this topic more qualitatively, considering the concept of communicative function. However, there is an additional consideration; in this process, both the expressive factor and the evocative factor are considered in the uttering process. Expressive refers to the attitude of the speaker, and evocative refers to the addressee's reaction. In statements and exclamations, expressive is usually emphasized, but in questions and requests, the opposite phenomenon occurs.

- **Statement** - Expressive: Belief / Evocative: (that listener shares) Belief judgment
- **Question** - Expressive: Desire for information / Evocative: (that listener provides) the desired information
- **Request** - Expressive: Desire for X / Evocative: (that listener provides) X
- **Exclamation** - Expressive: Any attitude / Evocative: (that listener attends to attitude)

Here too, multiple acts can be interpreted from one utterance as well. However, in Portner or Allwood's approach, thinking of these four types of utterances as default syntactic-semantic or pragmatic type lessens the concern about the existence of other verbal expressions that do not fall into these categories, though some utterances require more detailed categorization of speech act. This is also confirmed in Beyssade and Mandarin (2006) [37], which considers both speaker-oriented impact and addressee-oriented impact. Also, exclamation appears as a unique sentence type without a particular addressee-oriented impact.

Based on the above categorization studies on speech act, we explore when the problem of ambiguity arises in the intention understanding process.

2.3 Ambiguity of Intention Understanding in Korean

2.3.1 Ambiguities in language

The ambiguity that affects correct intention understanding can be analyzed at multiple levels. In fact, the word 'ambiguity' is itself ambiguous. Ambiguity can be viewed from a linguistic or multimodal perspective, where various types and levels exist within linguistic ambiguity, and there are solvable ambiguities and non-solvable ones.

Here, we want to limit the ambiguity in intention understanding to the linguistic ambiguity. In other words, among the ambiguities caused by linguistic phenomena, we focus on ones that are caused by the diversity of the prosody, and especially, we investigate the ambiguity that can be resolved by the presence of the prosody.

There are various levels of ambiguity in language and speech [38], ranging from word to morpheme, compound, derivation, phrase, and whole sentence level [16]. Occupying the largest portion is word-level ambiguity, which is caused by linguistic phenomena such as homography, homophony, homonymy, and polysemy. This word-level ambiguity expands to the compound level or derivation level, introducing a new kind of ambiguity (e.g., *clam prod* vs. *clamp rod*, *undressable* as 1. able to be undressed vs. 2. not able to be dressed). When this is extended to the phrasal level, it brings syntactic ambiguity (e.g., the old *men and women* vs. the *old men* and women).

Some of the above types of ambiguity can be resolved with prosodic information. For example, homography related to word prosody uses distinguished usage of phonemes and durations, and in the case of compound or phrasal-level ambiguity, disambiguation is feasible using the placement of pauses within compound or phrase. However, in general, prosodic ambiguity leads to a semantic shift of the whole sentence level [16]. For instance, assuming a sen-

tence "*We didn't come because we were tired*", the presence of speakers changes depending on whether or not a continuation tone comes after *come*. Also, in the case of a declarative question such as "*know what he means*", depending on whether the intonation of 'means' is rising or falling, it may differ whether the sentence is interpreted as a question or a statement [39].

Among levels of ambiguity above, we would like to focus on the types of ambiguities in which prosodic ambiguity can influence sentence-level semantics, especially those that change the sentence type. Homophony, homonymy, and polysemy, which are difficult to resolve with prosody, require the intervention of cultural or dialog context and therefore, much more information encoded in the first place. Also, prosody-resolvable word/compound, or derivation or phrasal-level ambiguity is less likely to change sentence type by their disambiguation (though not impossible). In other words, a local word meaning change or content modification may have a subtle effect on whether a question is rhetorical or a statement is modal, but except for such tricky cases, it seldom happens that a statement becomes a question or a request is read rhetorically. In addition, even if the whole sentence-level semantics changes, the frequency where speakers' presence affects the sentence type will not be significant. Contrarily, intonation assigned to particles that initiate or terminate sentences, prosody around vocatives or polarity items, etc. are likely to influence the sentence type itself, and this phenomenon is commonly observed in languages [40, 41]. We will define this ambiguity as a prosodic ambiguity related to speech act and intention, and conduct a language-specific case study on this kind of issue.

To prevent the term 'prosodic ambiguity' from being understood as an ambiguity that comes from the acoustic cue, we note that 'ambiguity' henceforth denotes the ambiguity that can be resolved if provided with prosodic information.

2.3.2 Speech act and intention understanding in Korean

Previously, we discussed SLU and intention understanding place as a subfield of SLP and how the concept of speech intention can be interpreted through concepts such as illocutionary act, speech act, and discourse component. In addition, we investigated the ambiguity issue that is defined in terms of intention or sentence type.

These have been comprehensively studied in Indo-European languages, including English, but more minor in non-Latin alphabet-based and non-IE languages such as Korean. It was studied within theoretical Korean linguistics, but a quantitative study on speech act or intention classification in the Korean language currently lacks, especially for single sentences. Not only building a new intention understanding dataset, but we also want to tackle ambiguity issues that arise in the corpus construction process.

Korean language processing and its challenges

Though studies on intention understanding and dialog act which are mainly focused on English, deal with the linguistic concept that is also observable in other languages, the aspect in which intention or act is represented in English and other languages (especially non-Indo-European) may not be the same. It also requires different problem formulations and new approaches, and may lead to new findings. In particular, investigating less studied and low-resourced languages will be a stepping stone for extending English-centered methodologies to various other language groups.

In this dissertation, we try to capture Korean, which belongs to the Koreanic language family, as the language of interest. Korean is a language spoken by 80 million people mainly in East Asia, and has varieties spoken in South and North Korea, Middle Asia, and so on. Korean is a head-final and agglutinative, *wh-in-situ* language, scrambling and highly contextual. In addition,

Korean is written in a featural writing system called Hangul [42], and the characters of Hangul are equal to syllabic blocks of spoken Korean. We will conduct research focusing on Seoul Korean among wide Korean varieties, and expand the scheme in a direction applicable to as many dialects and languages as possible.

The above characteristics of the Korean language make Korean distinguishable from other language families, but they have often made up a hurdle for natural language processing using Korean. For example, Korean being a head-final language makes it difficult to apply techniques that have been effectively used in head-first languages such as English [43]. In addition, due to the characteristics of the agglutinativeness that morphemes comprise a word-like unit, it is not reliable to directly apply word-level approaches in other languages available upon whitespace segmentation. Also, the possibility of multiple tokenizations reduces the stability of language processing [44].

Previous intention studies in Korean SLU

Speech act in Korean was mainly studied in discourse analysis [45, 46], but there is no precedent on scalable corpus construction in a form available in computational linguistics. The main reference is a dataset that tags dialog act in reservation scenario, which is an approach that has a limited domain and requires dialogue context [47]. It is difficult to use them to classify sentence types for usual situations or single utterances. Another recent study on probably single utterances also utilizes a similar categorization, which seems to have blurry boundaries between some classes of acts and fits with limited scenarios [48].

The tasks mainly used for single sentence classification in Korean are sentiment analysis, topic classification, and relation extraction. For sentiment anal-

ysis, the dataset is typically built for movie review [1], and news or Wikipedia articles are used for topic classification or relation extraction [49]. The above datasets mainly consider the semantic properties of sentences, where the keyword or content is considered more important than sentence structure in solving the task. However, in the case of speech act and intention, not only sentence content but also syntactic components such as sentence enders [45, 50] or polarity items [51] play an important role.

We need to pay attention to that, especially in Korean, studies on speech acts have been mainly studied only at the dialog level. This is highly likely due to common situations in which the intention of a sentence becomes apparent only when the context is given, probably due to linguistic phenomena that make analyses of Korean more tricky, e.g., pro-drop or anaphora [52, 53]. That is, when only a single sentence is given, a machine or even a human can find it difficult to categorize it correctly if there is some lack of information. Problem formulation here slightly differs from multi-label binary classification tasks such as emotion analysis where it is inferred that multiple answers exist. For instance, when guessing the emotion of an expression *'Why on earth did you leave me'*, we can talk about both 'sad' and 'angry' in terms of emotion. However, it is difficult to say that it contains both the characteristics of the question and the rhetorical question; the sentence is highly likely to be a rhetorical question, and in other words, there is just a little chance that the utterance is a pure question. Thus, in the categorization of these kinds of sentence types, unlike judgment on sentiment or emotion, the boundary between types is relatively clear, and utterances hardly belong to several classes at the same time. This formulation becomes more visible when there is a clear boundary defined, such as when sentences are categorized into a statement, a question, or a command, introduced by the concept of communicative act or discourse component as

[1] `https://github.com/e9t/nsmc`

earlier [36, 35].

In Korean, the problem of ambiguity which comes from the gap of information between the original utterance and its text format appears more significant. For instance, the sentence ender–which is the head of the sentence, is often underspecified [54], so even if the same text input is given, a conflict arises when assigning the sentence to a specific sentence type. In addition, due to the *wh-in-situ* language characteristics, *wh*-particles are placed in the same position when they are in a question or other sentence types, playing the role of an existential quantifier [55]. This phenomenon is combined with the underspecified sentence ender above to obscure the sentence meaning [41]. This tendency is particularly frequent in chat or web text where punctuation marks or other prosodic features cannot be accurately known, and it is also often caused by an error in the SLU pipeline where the ASR process cannot correctly infer such symbolic features. That is, we face cases in which it is difficult to grasp the genuine sentence meaning or intention with only the text of spoken utterance.

Chapter 3

Ambiguity in Intention Understanding of Spoken Language

In this study, we define ambiguity in Korean speech act and intention understanding that is caused by the absence of prosody or intonation[1], and construct a corpus for the detection of such ambiguous sentences. We conduct quantitative and qualitative analysis via training conventional neural network architectures and pretrained language models, and verify the utility and consistency of our approach. Most passages of this chapter are directly or indirectly quoted verbatim from the published versions [57, 58], and the figures and tables are reprinted under fair use.

3.1 Intention Understanding and Ambiguity

Before we define the ambiguity of intention understanding, we first introduce the typology of intention categorization.

We basically follow the scheme of Allwood (1995) [36] and Portner (2004) [35], where utterances (which are not exclamation) are classified into three cat-

[1]Though two concepts differ, we view intonation as a melodic facet of prosody [56] and use them interchangeably in this dissertation.

egories of statement, question, and command. Allwood regards this as a problem of communicative action, while Portner claims the concept of discourse component (DC).

Though our study is not on formal semantics, and is rather close to pragmatics, the main reason we apply this categorization is to make the obscure boundary between speech acts clear. For instance, vague intersections between the classes can be observed between e.g., *statement* and *opinion* [9], 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* [30], we extend them to syntax-semantic level adopting DC. It involves *common ground* (CG), *question set* (QS), and *to-do-list* (TDL): the constituent of sentence types that comprise natural language [35]. 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*).

Another reason for this kind of categorization is that the directiveness and rhetoricalness of the utterance are key factors to be disambiguated in usual text-based conversation. It is frequently observable that misunderstandings between two participants of messenger chat take place because of the omission of commas, periods, or other punctuation marks. This incurs some circumstances that a declarative sentence that looked like a monologue was a question from the speaker, or an interrogative sentence that was interpreted as a question was genuinely a rhetorical one, that asks nothing to the addressee. Also, not limited to human communication, it is not unusual for intelligent agents or AI speakers to misunderstand the user's chit-chat as a question or command. 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 in-

stance, transcripts usually do not contain punctuation which is sometimes essential for proper understanding of spoken utterances. Such a phenomenon may not threaten the spoken language understanding (SLU) in many languages, especially if the lexical usage is straightforward depending on the sentence form and type (as in English, where interrogatives and imperatives are distinguished from declaratives in general), but it matters if the ambiguity significantly affects the intention understanding of transcribed utterances.

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 [59]. Consider the following sentence, of which the meaning depends on the sentence-final intonation:

(3-1) 너 가고 있어
ne ka-ko iss-e
you go-PROG[2] be-SE[3]

With a high rise intonation, the sentence becomes a question (*Are you in the way?*), and given a fall-rise or level intonation, becomes a command (*You go first and I will follow later.*). 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 [60]. Although similar tendencies are observed in other languages as well (e.g., declarative questions in English [39]), syntactical and morphological properties of the Korean lan-

[2]Denotes a progressive marker.

[3]Denotes the underspecified sentence enders; final particles whose role vary.

guage 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.

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 [27, 1, 2], 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 [61], the proposed system aims to find the genuine intention of a single input utterance and guide the following direction.

3.2 Annotation Protocol

We clarify the annotation concept of the corpus to be constructed, which can be adopted to train a text-level classifier for spoken language[4]. As its motivation is briefly introduced in Chapter 1, we primarily aim to discern the existence of ambiguity that is determined by the lexical features.

[4]Throughout this chapter, *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 this 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 absence of prosodic information, 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 3.1. For a detailed description, we describe each sentence type in the order or FR, CCs and IU, with example sentences.

3.2.1 Fragments

From a linguistic viewpoint, fragments often refer to single noun·verb phrase where ellipsis occurred [62]. 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

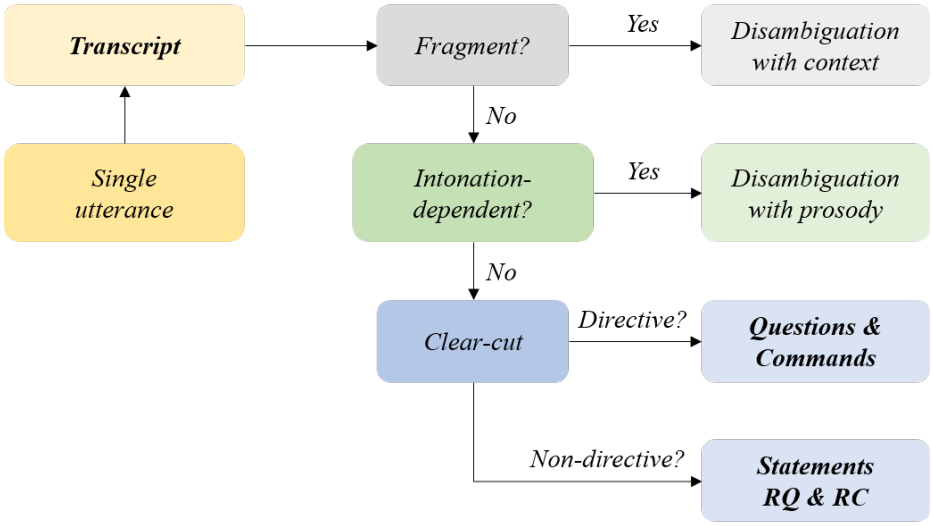

Figure 3.1: A brief illustration on the proposed annotation protocol

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 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 for fragments:

(3-2a) 마우스

mawusu

mouse

mouse

(3-2b) 키보드와 마우스

khipodu-wa mawusu

keyboard-AND mouse

keyboard and mouse

(3-2c) 마우스로

mawusu-lo

mouse-WITH

with mouse

Not only a single word (3-2a) is a fragment, but a noun phrase (3-2b) or a postposition phrase (3-2c) can also be the case. We concluded that determining the intention of such phrases requires the dialogue history even if the prosody is given.

3.2.2 Clear-cut cases

Clear-cut cases include utterances of five categories: *statement, question, command, rhetorical question*, and *rhetorical command*, as described detailed in the annotation guideline[5] with examples. Questions are utterances that require the addressee to answer (3-3a,b), and commands are ones that require the addressee to act physically or psychologically (3-3c,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 (3-3e).

(3-3a) 너 집에 갈거니

[5]Currently uploaded online in Korean. `https://docs.google.com/document/d/1-dPL5MfsxLbWs7vfwczTKgBq_1DX9u1wxOgOPn1tOss`

ne cip-ey kal-ke-ni

you home-to go-PRT[6]-INT[7]

Will you go home?

(3-3b) 내일 날씨 좀 알려줘

nayil nalssi com ally.e-cwu.e

tomorrow weather POL[8] inform.PRT-give.SE

Please tell me tomorrow's weather.

(3-3c) 세 시 반에 나 좀 깨워

sey si pan-ey na com kkaywu.e

three hour half-at I POL wake.SE

Please wake me up at three thirty.

(3-3d) 목소리 좀 낮추는 게 어때

moksoli com nacchwu-nun key ettay

voice POL lower-PRT thing.NOM[9] how

How about lowering your voice?

(3-3e) 아무래도 내일 나스닥 떨어질 것 같아

amwulayto nayil nasudak tteleci.l kes kath-a

anyway tomorrow NASDAQ drop.FUT[10] 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 (3-4a) [63]. 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*, (3-4b)) [64, 65]. Sentences in

[6]Denotes a functional particle.

[7]Denotes an interrogative ender.

[8]Denotes a polarity item for the politeness in asking something.

[9]Denotes a nominative case.

[10]Denotes a future tense.

these categories are functionally similar to statements but are categorized as separate classes since they usually show a non-neutral tone.

(3-4a) 너 돈 벌기 싫니

ne ton pel-ki silh-ni

you money earn-PRT dislike-INT

Don't you want to make money? (= It seems that you are not interested in making money.)

(3-4b) 쏠 테면 쏴 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 [35] to borrow the concept of DC and extend the formal semantic property to the level of pragmatics. This indicates that we search for a QS or TDL which makes an utterance directive in terms of speech act [20], 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 DC of CG. We provide a simplified criterion in Table 3.1, 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.

3.2.3 Intonation-dependent utterances

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

Table 3.1: A simplified annotation scheme using sentence form and discourse component of each sentence type

Sentence form / DC	**Common Ground**	**Question Set**	**To-do List**
Declaratives	*Statements, RQ, RC*	*Question*	*Command*
Interrogatives	*RQ*	*Question*	*Command*
Imperatives	*RC*	*Question*	*Command*

case it is projected to a textual form, when even the punctuation is omitted or not adequately transcribed with an ASR system. Sentence (3-1) in Section 3.1, which is not accompanied with punctuation but is ambiguous, describes such cases.

Although there have been studies on Korean sentences which handle final particles and adverbs [66, 17], to the best of our knowledge, there has been no explicit guideline on a text-based identification of utterances that are ambiguous without prosody. 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 with the maxims of conversation [33], 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,

(3-5a) 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 the 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*, a deixis for an older sister, used mainly by male speakers)' is augmented at the start of the sentence (3-5b), it is much more plausible to interpret the sentence as:

(3-5b) **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]?

(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 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 (3-3b,c). Again, in (3-5c,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).

(3-5c) 오늘 뭐 **좀** 먹고 싶어
onul mwe **com** mek-ko siph-e
today what *slightly* eat-PRT want-SE
I think I want to eat something today.

(3-5d) 오늘 뭐 **하나** 먹고 싶어
onul mwe **hana** mek-ko siph-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 (3-5a), then it becomes less felicitous to interpret it as a question, like:

(3-5e) 오늘 **뜨끈한 국밥** 먹고 싶어

onul **ttukkun-han kwukpap** mek-ko siph-e

today *warmy Kwuk-pap* eat-PRT want-SE

I want to eat a warmy Kwuk-pap today.

Here, *mwe* is replaced with *ttukkun-han kwukpap*, a heart-warming 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.

3.3 Data Construction

3.3.1 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[11], (ii) a set of frequently used lexicons, released by the National Institute of Korean Language[12], 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.

3.3.2 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 3.2, Corpus 20K). Annotators were well informed on the guideline and had enough debate on the conflicts during

[11] http://slp.snu.ac.kr/

[12] https://www.korean.go.kr/

the annotating process. The resulting inter-annotator agreement (IAA) was κ = 0.85 [67] and the final decision was made by majority voting and adjudication.

3.3.3 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 belongs 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 3.2.

3.3.4 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.

Table 3.2: Composition of the constructed corpus

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**

3.4 Experiments and Results

3.4.1 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, [68]) or bidirectional long short-term memory (BiLSTM, [69]) along with fastText [70] word vectors, and up-to-date pretrained language models (PLMs) such as bidirectional encoder representations from Transformers (BERT, [71]) and ELECTRA [72].

Conventional architectures

Conventional architectures include CNN [73, 68], BiLSTM [69], and self-attentive BiLSTM (BiLSTM-Att [74]). For CNN, two convolution layers were stacked with max-pooling layers in between, summarizing the distributional informa-

tion 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 [70] 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 [75], publicly available in a Github repository[13].

Pretrained language models

For BERT-like PLMs, we adopted multilingual BERT (mBERT), KoBERT [76], KcBERT [77], KoELECTRA [78], KcELECTRA [79], and KLUE-BERT [49], which are all currently available in Hugging face Transformers library[14] [80]. mBERT, KoBERT, KcBERT, and KLUE-BERT follow BERT [71] that builds a bidirectional encoding upon Transformer [81], where the pretraining aims optimizing the model to two subtasks of masked language model (MLM) and next

[13] `https://github.com/warnikchow/raws`

[14] `https://github.com/huggingface/transformers`

sentence prediction (NSP). KoELECTRA and KcELECTRA utilizes replaced token prediction (RTD) of ELECTRA [72], which strengthens the model in 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 small amount of spoken texts from message and web data [82]. 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.

3.4.2 Implementation

All conventional architectures were implemented with Keras Python library [83]. 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 same size as the LSTM hidden layers (64). The optimization was done with Adam (5e-4) [84], 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 mBERT[15], KoBERT[16], KcBERT[17], KoELECTRA[18], KcELECTRA[19] and KLUE-

[15] https://huggingface.co/bert-base-multilingual-cased

[16] Originally provided in https://github.com/SKTBrain/KoBERT, and the served version for Hugging face Transformers was available in https://huggingface.co/monologg/kobert

[17] https://huggingface.co/beomi/kcbert-base

[18] https://huggingface.co/monologg/koelectra-base-v3-discriminator

[19] https://huggingface.co/beomi/KcELECTRA-base

BERT[20] 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) vocabs 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)[21] and linear scheduler with 100 warm-up steps[22]. The training with batch size 32 ran for 3 epochs which is sufficient for the fine-tuning of the created data, and the final trained model was directly adopted for the test.

3.4.3 Results

Table 3.3 shows the performance of conventional architectures and up-to-date PLMs, where all results were obtained by inferring the test set. Dictionary size for the conventional architectures indicate 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.

Quantitative analysis

Among all the conventional architectures and up-to-date PLMs, KoELECTRA, which is pretrained upon both colloquial and written texts with adequate size of vocabulary, exhibited the highest accuracy. This proves that strategies for

[20] `https://huggingface.co/klue/bert-base`

[21] We additionally set weight decay 0.01, Adam beta1= 0.9, Adam beta2= 0.95, and Adam epsilon 1e-8.

[22] The optimization scheme for PLMs was more delicately set due to the sensitivity of models.

Table 3.3: Test result (accuracy) with conventional architectures and PLMs

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

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 outperforming conventional architectures.

Pretraining corpus influences the result We analyze that the result is also influence 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 [75].

Less sensitive to OOV and follows scaling laws It is also noteworthy that mBERT, trained upon multilingual vocab and corpora, outperform KoBERT which bases on similar monolingual corpora. This suggests that our dataset is less vulnerable to out-of-vocabulary issues which 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 [85], 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.

Further investigation using PLMs

As using PLMs is *de facto* in recent literature, we conducted further investigation to help understand how the constructed dataset can be utilized in analyses and practice. In Table 3.4, we compare the size and domain of the pretraining corpora of PLMs, referring to Hur et al. (2021) [87] and Yang (2021) [88], and

Table 3.4: Comparison of pretraining corpora and performance of each PLM

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 [82], CC-100-Kor [86], Namu Wiki, Newspaper, Petition dataset, etc.	91.72 (72.18)	0.20	96.13 (65.09)

how they perform in various classification scenarios. Note that the size and domain of mBERT denote the pretraining corpora regarding all the languages that are relevant, that size cannot be specified to a specific language.

For all the PLMs, pretrained weight was fixed 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)[23]. 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.

First, as discussed in the previous section, the size of pretraining corpus seems to influence the performance, considering that mBERT outperforms KoBERT and so as for KcBERT, KcELECTRA, and KoELECTRA. However, given that KcELECTRA shows almost the same performance as KoELECTRA in 7-fold

[23]The performance for 7-fold scenario slightly differs from Table 3.3, which recorded the best score, since the score is averaged after five repetitions with different initialization.

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 in overall and IU performance compared to BERT-based ones. This result suggests that the training scheme of ELECTRA which bases on RTD fits with the current downstream task compared to masked language model of BERT, considering that RTD had conventionally been more suitable with logical or factoid problems such as natural language inference [72], which requires slightly different aspect of language understanding in contrast with indecisive tasks such as sentiment analysis. Finding the presence of ambiguity from 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 approach can be further utilized to promote the model development. Unfortunately, we found that 3-fold classification is not yet effective for the 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 up 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.

Qualitative analysis

We made up a confusion matrix with the result of the fine-tuned KoELECTRA module, which shows the most reliable performance (Table 3.5). Fragments, statements, questions, and commands show high accuracy ($> 92\%$) while others show lower ($< 80\%$).

Table 3.5: Confusion matrix for the validation of the fine-tuned KoELECTRA

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 the 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 버린 거예요' (*Nan pelyn keyeyyo*, 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 kinds of 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 shed light on the further utilization of the trained system in the 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 that are 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 let 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.

3.5 Findings and Summary

3.5.1 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 SLU or natural language understanding (NLU) management and at the same time as a Korean sentence classification benchmark.

One of our concerns is that the 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.

3.5.2 Summary

In this chapter, 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 SLU, which is scarce in the literature that is dominant of tasks with written texts.

Though we could not investigate the case using speech signal input in this chapter, 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 [3, 13], 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 benefit 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 ensembled 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 study 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 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 are to be handled in our future research that addresses the social convention.

Chapter 4

Disambiguation of Speech Intention

In this study, we investigate how speech intention in Korean can be disambiguated for utterances that incorporate ambiguity in their text format. We construct a corpus that consists only of text scripts that have ambiguity, record them for all corresponding intention categories, and conduct a model-based study to see if and how such utterances can be disambiguated. Most passages of this chapter are directly or indirectly quoted verbatim from the published versions [89, 17], and the figures and tables are reprinted under fair use.

4.1 Ambiguity Resolution

Resolving syntactic ambiguity is a core task in spoken language analysis, since identifying the sentence type and understanding the intention of a text form utterance is challenging for some prosody-sensitive cases. Notably, in some *wh-in-situ* languages like Korean and Japanese, some uttered word sequences incorporate syntactic ambiguity, which leads to difficulties discerning directive speech from constative or rhetorical ones. For example, the following sentence in Seoul Korean can be interpreted differently depending on the prosody [89]:

(4-1) 몇 개 가져가 **myech kay kacye-ka**

how quantity bring-USE[1]

(a) *How many shall I take?*

(LHLLH%; ***wh*-Q**)

(b) *Shall I take some?*

(LMLLH%; ***yes/no* Q**)

(c) *Take some.*

(LMLML%; **command**)

where L, M, and H denote relative pitch of each syllabic block and USE denotes an underspecified sentence ender. Unlike English translations, if given only the text with periods or question marks removed (usually provided as an output of automatic speech recognition (ASR)), the language understanding modules may not be able to determine if it is a statement or a question. Even with such marks, it is vague whether the question is *yes/no* or *wh-*.

As such, for an utterance that contains components whose roles are decided by prosody, it requires both the acoustic and textual data for spoken language understanding (SLU) modules (and even humans) to correctly infer the speech intention. In this process, the pitch sequence, the duration between words, and the overall tone all together decide the intention of an utterance. Thus, we concluded that introducing prosodic information is indispensable for resolving syntactic ambiguity, as depicted in Figure 4.1.

4.1.1 Prosody and syntax

The interaction of prosody and syntax has long been investigated regarding sentence types, especially for the question types including *wh*-intervention [90] and declarative forms [39]. Moreover, for some head-final languages, sentence-final intonation can play a significant role in clarifying the sentences. For in-

[1]Denotes an underspecified sentence ender.

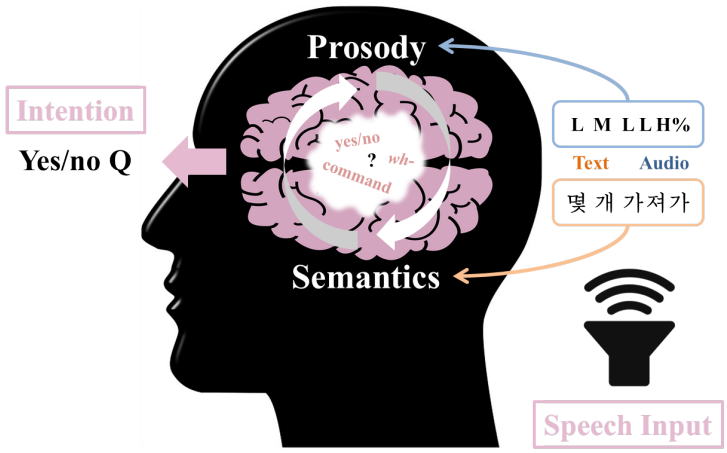

Figure 4.1: Prosody-syntax-semantics interface in Korean

stance, the prosody assigned to the final particle or word of non-scrambled Korean sentences usually decides the sentence form, e.g., declaratives or interrogatives [41].

In a broader perspective, syntactic ambiguity has been dealt with not only in studies on sentences but also phrases. In Korean, datives [91] and comparatives [92] have been mainly investigated. Also, linking syntax with phonetics, Baek (2018) [93] demonstrated that syntactic ambiguity is resolved via prosodic information that can elaborate the lowness/highness of the attachment. They handled several cases in Korean where the syntax differentiates upon phonetic properties, especially among long phrases (e.g., *diligent boy's sister*). This approach is mainly concerned with contiguity theory [94], which claims that syntax can make reference to phonological structure, and that movement operations can be triggered by the need to produce phonologically acceptable objects.

4.1.2 Disambiguation with prosody

The aforementioned studies on the prosody/phonetics-syntax-semantics interface deal with various types of disambiguation, which incorporates the variance of the topic, agent, experiencer, and object of syntactically ambiguous sentences. Among them, a few concerns questions, commands, and their directivity (see [41]). We suggested the seven-class categorization (namely statement, *yes/no* question, *wh*-question, rhetorical question, command, request, and rhetorical command) in Cho et al. (2019) [89], based on (i) sentence-middle intonation that affects topic and *wh*-intervention, (ii) sentence-final intonation that changes the sentence form, and (iii) the overall tone that influences rhetoricalness. The limitation is that, the analysis beyond the categorization has yet been performed. For now, the prosodic activeness-based disambiguation [94, 93] well formulates the phonetic segments that clarify syntax. However, we deemed it necessary to resolve the ambiguity within the wider range of sentence types, promoting possibly automatic management. In this regard, we consider computational approaches that autonomously discover the latent and non-codified criteria.

4.1.3 Approaches in SLU

Early studies on spoken language analysis adopt a simple concatenation of acoustic and textual features [11], where parallel convolutional or recurrent neural networks (CNN/RNNs) were used to summarize each feature. A recent study includes hierarchical attention networks (HAN) [95] that point out the components essential for inferring the answer. In the related area of speech emotion recognition, multi-hop attention (MHA) [96] was introduced to encourage comprehensive information exchange between textual and acoustic features. Nonetheless, since the experiments in literature generally utilize speech utterances with less confusing intention or emotion (e.g., syntactically non-

ambiguous sentences or emotion utterances without semantic cue), there has been little study concentrating on the resolution of ambiguous sentences as in (4-1).

In terms of the prosody-syntax-semantics interface, we concluded that the interaction between acoustic and textual information is required for such cases. We aim to materialize this approach in our co-attentional architectures in the form of MHA [96] and cross-attention (CA) [97], which have shown their power in the area of speech emotion recognition and image-text matching respectively.

4.2 Dataset Construction

The dataset for the analysis of ambiguous speech, which requires disambiguation with prosody, is a corpus that contains about 1.3K sentences with two to four different types of prosody (and the corresponding intention) [89]. Specifically, each sentence (i) starts with a *wh*-particle, (ii) incorporates a predicate made up of general verbs and pronouns, and (iii) ends with underspecified sentence enders so that the overall prosody varies according to intention (and sometimes with politeness suffix).

All the sentences received the consensus of three native Korean speakers, and the total number of speech utterances reached 3,552. Male and female speakers recorded each utterance with appropriate prosody for each intention, to obtain a dataset of size 7,104. The number of intentions is seven, namely statement (S), *yes/no* question (YN), *wh*-question (WH), rhetorical question (RQ), command (C), request (R), and rhetorical command (RC). The categorization is slightly modified from the one used in Section 3.2, to reflect *wh*-intervention as illustrated in (4-1). The specification of the corpus is given in Table 4.1. The construction process is to be explained in detail.

Table 4.1: Frequency matrix on *wh-* particles and the intention types

	S	***YN***	***WH***	***RQ***	***C***	***R***	***RC***
Who	547	544	446	202	112	26	18
What	294	283	186	64	32	14	4
Where	64	64	49	6	11	4	1
When	37	54	40	22	0	4	15
How	59	62	28	8	6	0	0
How much	84	40	100	0	14	8	0

4.2.1 Script generation

In generating the corpus script, namely five factors were considered: ***wh*-particles** that initiate an utterance, **predicates** that convey the content, **reportative particles** that give the utterance evidentiality, **sentence enders** that possess potential to represent various intentions, and **politeness suffixes** which come just after the sentence ender to assign honorific mood to the sentence.

wh-particles

Among the six *wh*-particles, namely '누구 (*nwukwu*, who)', '뭐 (*mwe*, what)', '어디 (*eti*, where)', '언제 (*encey*, when)', '어떻게 (*ettehkey*, how)', and '왜 (*way*, why)', only the first five were utilized in constructing the corpus. This is because way is rarely used as a quantifier, except for some cases in child language. Instead of *way*, we used '몇 (*meych*, the number of)', which is widely used as a quantifier for counting. For the purpose of variation, in some cases, nominative (NOM) or accusative cases (ACC) were attached to the *wh*-particles.

Predicates

Predicates largely depend on the *wh*-particle they are aligned with. For instance, *nwukwu* (who) harmonizes with the verbs that are related to interaction, such as 'give' and 'receive'. In contrast, *eti* (where) matches with the verbs concerning location, such as 'come' and 'go'. In selecting the verbs, we referred to the set of 5,800 frequently used lexicons, released by the National Institute of Korean Language[2]. Depending on the verbs, appropriate particles were agglutinated and the phrases that contain object/complement were inserted. In some circumstances, polarity items such as '좀 (*com*, bit)' or '하나 (*hana*, a piece)' were augmented to modify or restrict the implicature.

Reportative particles

The reportative particles (RPT) provide utterances with evidential mood. Usually '-대 (*tay*)', '-래 (*lay*)', and '-쟤 (*cyay*)' are used for statements, commands, and hortatives [60]. The particles were selectively added considering the content.

Sentence enders

The sentence enders (SEs) with various roles are components that influence the sentence type and intention of the utterance. There are mainly two types of SEs; the first type is SEs with a fixed role, e.g., '-다 (*-ta*)' for declaratives and '-니 (*-ni*)' for interrogatives [60]. For these, the sentence type is fixed but the intention can vary regarding *wh*-intervention and rhetoricalness. The second type is the underspecified SEs whose feature is not fixed (e.g., '-어 (*-e*)', '-지 (*-ci*)'). They have the potential to display various intention types depending on the prosody. Both types of SEs were utilized in the generation.

[2] https://www.korean.go.kr/

Politeness suffix

The politeness suffix (POL), '-요 (*-yo*)', can be agglutinated to SEs and in most cases does not affect the functional variability of the sentence, except for rhetoricalness. For some SEs such as '-지 (*-ci*)' or '-야지 (*-yaci*)', the augmented form is modified to '-죠 (*-cyo*)'. On the other hand, the utterances with SEs to which the politeness suffix is not attachable, such as '-냐 (*-nya*)', were left without the politeness suffix. An example sentence incorporating the aforementioned concepts is as follows:

(4-2) 뭐 좀 먹었대요 — mwe com mek-ess-tay-yo

what bit eat-PST-RPT[3]-POL

Statement: *S/he told me that s/he ate something.*

Yes/no question: *Did s/he tell that s/he ate something?*

4.2.2 Label tagging

The labels used for the tagging of intention are statement, *yes/no* question, *wh*-question, rhetorical question, command, request, and rhetorical command, a modified version of the categorization suggested in the previous chapter.

- **Statement (S)** indicates an utterance that conveys information or the speaker's thought.
- ***Yes/no* question (YN)** indicates a question where the answer set is limited to yes or no.
- ***Wh*-question (WH)** indicates a question where the answer set is open and variable.
- **Rhetorical question (RQ)** indicates a question whose answer set is in the speaker's mind, usually being adopted to express the thought.

[3]Reportative particle.

- **Command (C)** incorporates an order that corresponds to imperatives in English with a covert subject, hortative that indicates an order with a politeness particle (e.g., please), and modal that indicates a statement with particles which correspond with should or must.
- **Request (R)** indicates a command expressed in an interrogative form.
- **Rhetorical command (RC)** indicates a command where the to-do-list is not mandatory, usually used as an idiomatic expression.

We list some examples regarding several *wh*-particles, incorporating more than three intention (and prosody) types. The case for *mwe* (what) is explained in the previous section, and the case for *ettehkey* (how) is omitted in this study since the intention variability is small (two cases at most). Q denotes question and C denotes command. L, M, H and '=' denote the relative pitches.

(4-3) **누가 보러 간대** **nwu-ka po-le kan-tay**

who-NOM see-to go-RPT

(a) *Who will go see it?*

(LHL==H%; ***wh*-Q**)

(b) *Will sbd go see it?*

(LML==H%; ***yes/no* Q**)

(c) *Does anyone say I'm gonna go see it?*

(LMLMLH%; **rhetorical Q**)

(d) *I heard sbd will go see it.*

(L==HL=%; **statement**)

(4-4) **어디 가고 싶어** **e-ti ka-ko siph-e**

where go-to want-USE

(a) *Where do you want to go?*

(LHL==H%; ***wh*-Q**)

(b) *Do you want to go somewhere?*

(L==MLH%; ***yes/no* Q**)

(c) *I want to go somewhere.*

(L==HL=%; **statement**)

(4-5) **언제 다시 봐** **en-cey ta-si pwa**

when again meet-USE

(a) *When will we meet again?*

(LHL=H%; ***wh*-Q**)

(b) *Shall we meet again someday?*

(LML=H%; ***yes/no* Q**)

(c) *I want to go somewhere.*

(LMLML%; **rhetorical C**)

(4-6) **몇 개 가져가** **myech kay ka-cye-ka**

how quantity bring-USE

(a) *How many shall I take?*

(LHL=H%; ***wh*-Q**)

(b) *Shall I take some?*

(LML=H%; ***yes/no* Q**)

(c) *Take some.*

(LMLML%; **command**)

4.2.3 Recording

The first version of the sentence list was generated by the methodology explained above, and only the sentences that received the consensus of three native speakers of the Seoul Korean dialect were taken into account. In total, the corpus contains 3,552 utterances that fall into one of the seven classes of intention. All the utterances were recorded by two native Koreans, a male and

a female. The speech corpus containing a total of 7,104 (= 3,552 * 2) utterances are available on-line[4] as with the corpus.

4.3 Experiments and Results

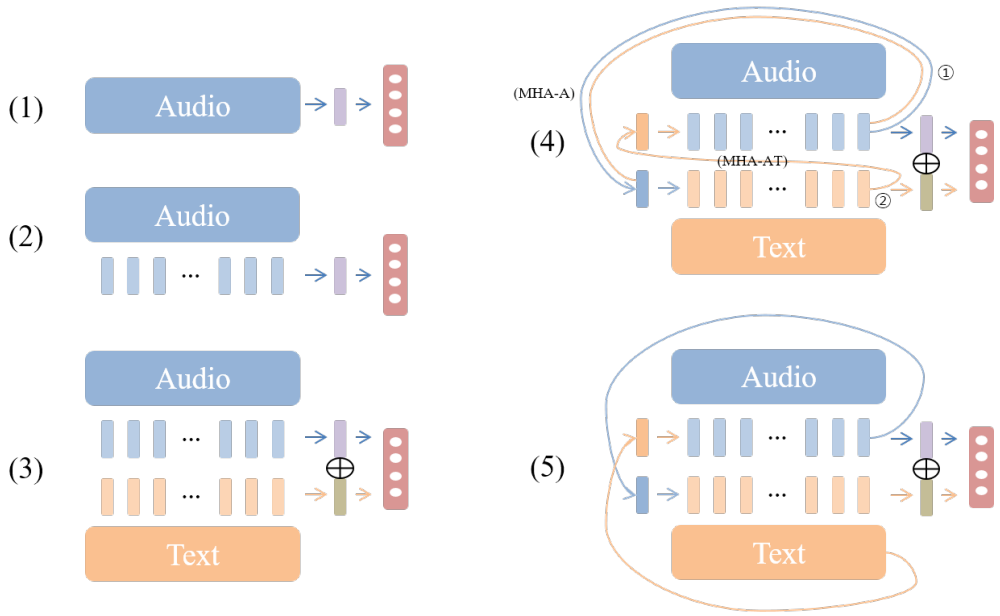

Figure 4.2: Block diagrams of the implemented models

4.3.1 Models

Here, we describe how the co-attention frameworks are constructed in terms of speech processing, self-attentive embedding, text-aided analysis, multi-hop attention, and cross-attention, as shown visually in Figure 4.2. In all models, the input is either audio-only (1-2) or audio-text pair (3-5). The text-only model is not taken into account since the text alone does not help resolve the syntactic ambiguity.

[4] https://www.github.com/warnikchow/prosem

Audio-only model (Audio-BRE)

The baseline model utilizes only audio input. Frame-level audio features are fed as an input to bidirectional long short-term memory (BiLSTM) [69], for which the expression BRE (bidirectional recurrent encoder) is assigned following [96]. The final hidden state is fully connected to a multi-layer perceptron (MLP) to yield a correct answer as a maximum probability output in the final softmax layer. Refer to (1) in Figure 4.2 for an illustration of the model's architecture.

Audio-only model with self-attentive embedding (Audio-BRE-Att)

Since the audio-only BiLSTM[5] model lacks information regarding the identification of the core parts in analyzing an utterance, we augmented a self-attentive embedding layer as utilized in the sentence representation [74]. In brief, a context vector, which has the same width as the hidden layers of the BiLSTM, is jointly trained to assign weight vector to the hidden layer sequence thereof. The whole process, as in (2) of Figure 4.2, implies that the weight is decided upon the overall distribution of acoustic features. Since the acoustic feature reflects the lexicon and the syntactic property, the weight ends up playing a crucial role in predicting the intention of the speech.

Parallel utilization of audio and text (Para-BRE-Att)

Unlike emotion analysis, where either textual or acoustic features do not necessarily dominate, in intention analysis, obtaining textual information can bring a significant advantage [11], even when a period or a question mark is omitted as in our experiment. Here, text input for the ambiguous sentences are identical (without punctuation marks) for two to four different versions of the

[5]In this chapter, we interchangeably use BRE and BiLSTM.

speech, but feeding them as an input of separately constructed *Audio-BRE* may provide supplementary information. The final hidden layer of *Audio-BRE-Att* is concatenated with that of the *Text-BRE-Att* (BRE-Att that exploits textual features) to make up a new feature layer, as illustrated in (3) of Figure 4.2.

Multi-hop attention (MHA)

In multi-hop attention [96], which is proposed for speech emotion recognition, textual and acoustic features interact by sequentially transmitting information to each other. This is the background of the expression 'multi-hop', where the hopping is performed by adopting the final representation of each feature as a context vector of the other as in (4) of Figure 4.2. The final output of the former and the latter are eventually concatenated. Here, we first implement hopping only *from audio to text* (4a) (MHA-A), and then augment *from text to audio* to make up (4b) (MHA-AT). They showed better performance than the further hopped model (i.e., MHA-ATA) in the original study [96]. Also, it is empirically more acceptable than the reverse case (e.g., MHA-T/TA) since auditory sensory first faces acoustic data than semantic information.

Cross-attention (CA)

From the perspective of another co-attention framework, we adopt cross-attention that fully utilizes the information flow exchanged simultaneously by both acoustic and textual features, as depicted in (5) of Figure 4.2. In the preceding study on image-text matching [97], image segments are utilized in determining the attention vector for the text, and similarly in reverse. Thus, not limited to using the representation regarding one feature as a context vector of the other's attention weight, we assumed it also plausible to utilize the final representation of *Audio-BRE-Att* in making up a weight vector for *Text-BRE* and vice versa. In this case, self-attentive embedding was not applied to the textual features, in

order to reflect the auditory-first nature[6].

Table 4.2: Experimental results on the 10% test set

	Accuracy (F1)		Param.s	Comp.
	Sparse	Dense		
(1) Audio-BRE	83.9 (0.652)		116K	65s
(2) Audio-BRE-Att	89.3 (0.759)		190K	67s
(3) Para-BRE-Att	**93.2 (0.919)**	92.8 (0.919)	260K	70s
(4a) MHA-A	**93.8 (0.928)**	**93.5 (0.922)**	266K	67s
(4b) MHA-AT	92.8 (0.909)	91.8 (0.904)	270K	67s
(5) CA	91.8 (0.884)	**93.5 (0.919)**	326K	65s
(3′) Para-ASR	90.0 (0.822)	-	-	-
(4a′) MHA-ASR	90.2 (0.799)	-	-	-

4.3.2 Results

Table 4.2 shows the comparison result utilizing the corpus dataset. Both train and test sets in (1-5) incorporate the scripts of ground truth, and for the others, the test set scripts were ASR results. Input materials are either sole audio or audio-text combined, both in the training and test phase.

Attention matters

First, in (1) and (2), we observed that audio itself incorporates substantial information regarding speech intention, and physical features such as duration,

[6]Since auditory sensory meets the speech before the audio is encoded to lexical components, we considered it fair to assign different levels of representation and weight regarding both modalities. Here, we implement it in the way of giving self-attentive embedding only to the audio features. In fact, providing attention to the textual features as well, resultingly degraded the performance; we also tried to avoid that case.

pitch, tone, and magnitude can help yield semantic understanding via attention mechanism. This seems to be related to the phenomenon where people often catch the underlying intention of a speech even when they fail to understand the whole words [98]. Also, it was shown that attaching the attention layer guarantees stable convergence of the learning curve in the training phase.

Text matters

Next, as expected, the text-aided models (3-5) far outperform the audio-only ones (1, 2), notwithstanding bigger trainable parameter set size and the computation time. Although the character-level features we utilized do not necessarily represent semantic information (which is held at least at morpheme-level), this result can be interpreted as implying that utilizing textual features can help recognize the prosodic prominence within the audio features [99]. It was beyond our expectation that the sparse vectors outperform the dense ones in general. The exception was in CA, which implies that CA takes more advantage from the distributional semantics within the text embedding. We infer that CA may exhibit significance if the utterances become more cumbersome, where pre-trained language models (PLMs) prevalent these days might be helpful.

Co-attention framework helps

In (3-5), we noticed that co-utilizing both audio and text in making up the attention vectors as in (4) *MHA* or (5) *CA* shows better performance than a simple concatenation in *Para-BRE-Att*. Since the studies on speech emotion analysis [100, 101] claim that prosody and semantic cues cooperatively affect inferring the ground truth, we suspect that a similar phenomenon takes place in the case of speech intention. That is, *acoustic and textual processing signifi-*

cantly benefit from a consequent or simultaneous interaction with each other.

Over-stack may bring a collapse

We first hypothesized that (4b) or (5) would show better performance compared to (4a) due to a broader or deeper exchange of information between both sources. Instead, we observed performance degeneration, leading to the conclusion that the inference becomes unstable if too much information is stacked. It is assumed that speech intention analysis is *affected dominantly by the combination of speech analysis and a speech-aided text analysis* (4a, 5), preferably with the smaller contribution of text-aided speech analysis (4b), though the performance of the models may not be directly linked to actual human processing mechanism. This shows that *text matters but speech influences*, as will be discussed further below.

In-depth analyses

For a practicability of the systems, model parameter size and training time per epoch were recorded (Table 4.2). Taking into account that audio processing itself incorporates huge computation, co-utilizing the textual information seems to bring significant improvement.

Then, we performed an additional experiment on ASR result[7] (3', 4a'), especially for the test utterances, where (3) and (4a) were chosen to observe how the degeneration differs in concatenation and co-attention frameworks. The training was performed with the ground truth, and the models for the sparse textual features were chosen upon the result with it as well (3, 4a). It is notable that both perform competitively with the case of perfect transcription, but the

[7] ASR was performed with a freely available API:
`https://aibril-stt-demo-korean.sk.kr.mybluemix.net/`

degeneration was more significant in the co-attention framework. This implies that the framework utilizing textual information more aggressively is ironically more vulnerable to errors. Thus, precise ASR and error-compensating text processing are both required for the improvement and application of the system.

Lastly, we observed that (i) the co-utilization of acoustic and textual features shows strength in identifying the intention classes that are highly influenced by prosody itself, e.g., distinguishing RQs from pure questions. Some cases deeply concerned the lexicon, e.g., distinguishing commands from statements or requests from *yes/no* questions. (ii) On the other hand, figuring out *wh*-intervention between *yes/no* and *wh*-questions, depended more on the interaction of audio and text processing, shown by a superior performance of *MHA* than *Para-BRE-Att*. These two observations can be explained as follows: in Korean spoken language, identifying rhetoricalness often accompanies non-neutral emotion (as suggested for a syntactically similar language [102]), whereas *wh*-intervention mostly involves phonological properties. Thus, we assume that (i) the emotion-related identification concerns a comprehensive understanding of the utterance as in *Para-BRE-Att*, while (ii) the elaborate processing of verbal data requires an analysis that pays more attention to the details of audio and text.

4.4 Summary

In this chapter, we constructed a speech intention recognition system using co-attentional frameworks inspired by psycholinguistics and the prosody-semantics interface of human language understanding. Multi-hop attention and cross-attention outperformed the conventional speech/attention-based and text-aided models, as shown by the evaluation using the audio-text pair recorded with

manually created scripts. An additional experiment with ASR output was also conducted to guarantee real-world usage and placed the room for improvement in text processing.

As stated previously, ambiguous utterances are disturbing factors for speech intention understanding, which can mislead a computational model to provide a wrong intent or item. However, aggregating both audio and text actively in analyzing such utterances can help more precisely predict the intention, if given a transcription with high accuracy. We are optimistic that our approach will prove meaningful for solving intriguing problems. In real life, co-attention frameworks can help machines or aphasia patients understand speech. Followingly, the system users or social chatbots may be able to provide proper responses/reactions in free-style or goal-oriented conversations with others.

From a slightly different viewpoint, the proposed strategy can also be utilized by patients who find it difficult to understand the emotion or intent conveyed by voice tone and prosody. The model may recognize the emotion and intention of the speaker and report it to the users so they can make a proper reaction/response. Beyond the intention-related syntactic ambiguity, the implemented structures can be utilized in other kinds of natural language processing systems that incorporate multi-modal inputs that are expected to be interactive with each other. For example, the proposed network can be utilized to provide a proper translation in a multi-modal context. Not just focusing on a text-to-text transformation, the system might capacitate abstracting and utilizing the source speech or image as an auxiliary input for the conventional machine translation process.

Chapter 5

System Integration and Application

In this chapter, we draw a brief sketch of how we can integrate previous studies and make up a useful spoken language understanding (SLU) module for real-world application. We provide a theoretic view and a simple experiment as a preliminary study. By this, we try to shed light on how our approach resolves ambiguity, improves system reliability, and can be utilized in making up a free-running spoken dialog system that helps intelligent agents flexibly execute user instructions without false alarm. Also, we see how our viewpoint can alleviate ambiguity issues chronic in multilingual language processing such as speech translation. Some passages of this chapter are directly or indirectly quoted verbatim from from the published versions [57, 58], and the figures and tables are reprinted under fair use.

5.1 System Integration for Intention Identification

5.1.1 Proof of concept

First, we define an intention identification system S, which has speech x as input and intention y as an output. Here, x is an audio data which is represented as a sequential data, and y is included in the set of intention categories I_{speech},

which consists of $|I_{speech}|$ components. The goal of intention identification is to assign each speech x a proper intention label y, and we represent this as $S(x) = y$. In our study, $|I_{speech}| = 6$, namely `fragment`, `statement`, `question`, `command`, `rhetorical question` (RQ), and `rhetorical command` (RC), and we assume that these categories does not have an overlap between each other if given a speech input.

One intuitive approach is to train an end-to-end system S_{e2e}, where all the parameters are jointly trained based only on x and y, with a given data. However, it is not usually affordable to construct a spoken language corpus of satisfying size due to excessive requirements of budget and time. It is the main background of our study in Chapter 3, which proposes a new category `intonation-dependent utterance` (IU) that incorporates the potential to be any of six pre-defined categories but is indecisive only with the transcript. Regardingly, some may leverage S_{asr}, an automatic speech recognition (ASR) module that transforms speech data to text, for a hybrid inference. This equals to the text-audio co-utilization suggested in Chapter 4, and may be much more accurate since symbolic information is added. Nonetheless, processing acoustic data twice for all the input may not be computationally efficient.

Different from S_{e2e}, a pipeline system S_{pipe} consists of S_{asr}, the ASR module, and S_{nlu}, which infers the intention with given transcription. Let the speech input x be mapped to transcript z. If it were a conventional natural language understanding (NLU) process, S_{nlu} would infer one of six intention categories from the text input. However, for the proposed system $S_{proposed}$, we split S_{nlu} into two parts, namely ambiguity detection module S_{ambi} suggested in Chapter 3 and intention decision module S_{deci} handled in Chapter 4. In this process, we add `IU` to our data, to filter out the ambiguous sentences and make the overall classification reliable.

To let the problem be more clear, we newly define I_{text} as a set of intentions where the input is in textual format, which is distinguished from I_{speech} with no indecisive category. I_{text} for conventional S_{nlu} originally shares the components with I_{speech}. Our approach for S_{ambi} augments a new category to I_{text} to make up $I^{+}_{text} := I_{text} \cup \{\texttt{IU}\}$, where `IU` is not necessarily independent with but is distinguished from the rest of categories. In total, $|I_{speech}| = |I_{text}| = 6$ and $|I^{+}_{text}| = 7$ in our study. Texts that are not intonation-dependent are assigned with pre-defined six categories as well in S_{ambi}. Main difference of the proposed system with S_{nlu} is that texts inferred as `IU` are fed as an input of S_{deci} to yield the final categorization, provided along with the original speech input x.

In brief, the list of systems can be simplified as following:

- $S_{e2e} : x \rightarrow y$, where $x \in \texttt{speech}, y \in I_{speech}$
- $S_{asr} : x \rightarrow z$, where $x \in \texttt{speech}, z \in \texttt{text}$
- $S_{nlu} : z \rightarrow y$, where $z \in \texttt{text}, y \in I_{text}$
- $S_{e2e-hybrid} : (x, z) \rightarrow y$, where $(x, z) \in (\texttt{speech}, \texttt{text}), y \in I_{speech}$
- $S_{ambi} : z \rightarrow y'$, where $z \in \texttt{text}, y' \in I^{+}_{text}$
- $S_{deci} : (x, z') \rightarrow y$, where $(x, z') \in (\texttt{speech}, \texttt{IU}), y \in I_{speech}$

The integrated systems can be rewritten again as:

- $S_{hybrid} = S_{asr} + S_{e2e-hybrid}$
- $S_{pipe} = S_{asr} + S_{nlu}$
- $S_{proposed} = S_{asr} + S_{ambi} + S_{deci}$

In this regard, we compare four systems in total, namely S_{e2e}, $S_{e2e-hybrid}$, S_{pipe}, and $S_{proposed}$. Let us assume a spoken language corpus C which consists of tuples (x,z,y) where x is a speech, z is the transcript and y is the target intention.

(Case a) For S_{e2e}, we only need to infer the intention y from given speech x, which means that the available corpus without further recording is only C at this point.

(Case b) However, in the case of S_{pipe}, the overall accuracy is expected to be lowered compared to the case of S_{e2e}, since it is inevitable that the usual ASR process drops some critical information about the intention of the utterance. That is, some portion of C would show an ambiguous label only with z, not (x,z), which can incur a degrade of S_{pipe}. Thus, we can define again C^{-}, which is a corpus with `IU` omitted from C, where `IU` are labeled based on our criteria in Chapter 3. Training with this corpus may prevent the model from training with samples of gray areas. However, this still does not help the correct inference of samples with ambiguous transcripts.

(Case c) Thus, we can use `IU` again in the text classification to make the NLU process two-stage. Here, we use all texts of C, but it is defined as a combination of C^{-} and $\{\texttt{IU}\}$. This is the training corpus for S_{ambi}. Fragments and clear-cut cases are filtered out as a final category, and `IU` texts are handled with S_{deci} module which is built upon speech and text hybrid input. Note that S_{deci} equals to S_{e2e} since we do not augment any speech data.

(**Case d)** Or, as a combined methodology of (Case a) and (Case b), we can utilize the textual data and audio in a hybrid manner ($S_{e2e-hybrid}$), upon proper ASR process. However, this makes the input (x, z), which implies that the overall computation cost may boost up.

5.1.2 Preliminary study

We designed a simple experiment to see how the strategies discussed above (a-c) affect the reliability of SLU process. We first make up challenging test cases and validate them with a fixed training corpus C, and implement (a-c) with lightweight neural networks trained from scratch.

Dataset

For train and validation, namely C in our formulation, we adopted a spoken corpus of size 7,000 which consists of (x,z) tuples. It is originally a set of recorded scripts created for Korean speech synthesis dataset construction [103], and consists of i) drama lines with the punctuation marks removed, and ii) the recorded audio. We annotated each tuple with intention y, using two kinds of target intention set, I_{speech} and I^{+}_{text}. Annotating C with I_{speech} denotes assigning six intention categories (including fragments) to speech utterances, and annotating with I^{+}_{text} indicates tagging each text utterance as six intention categories or an intonation-dependent utterance. Both annotation processes follow the guideline introduced in Section 3.2, while the speech was also referred to in the annotation with I_{speech}. Also, if we omit IUs from I^{+}_{text}, we get a text corpus C^{-} whose size is smaller than C but without ambiguous sentences. In total, we have three kinds of corpora for training set, namely (C, I_{speech}), (C, I^{+}_{text}), and (C^{-}, I_{text}).

For the test, we separately constructed an evaluation set of size 2,000. Half of the set contains 1,000 challenging utterances also randomly sampled from the same source corpora that above C was excerpted from. For another half, 1,000 question/command sentences in the corpus i) of Section 3.3 (not necessarily overlapping with the ones randomly chosen for the proposed corpus) were recorded and manually tagged. The former incorporates highly scripted lines, while the latter encompasses the utterances in real-life situations such as

calling or asking intelligent agents. By binding them, we assign balancedness to the test of the models trained with both types of data.

Implementation

For the validation of the overall mechanism, we implement the models with vanilla convolutional neural network (CNN) [73, 68] or self-attentive bidirectional long short-term memory (BiLSTM-Att) [69, 74], not using the pre-trained language models used in Section 3.4 or other cross-attention models implemented in Section 4.3, to guarantee that the overall performance change comes from the property and composition of the dataset used for training. For CNN, five convolution layers were stacked with the max-pooling layers in between, summarizing the distributional information lying in the input of a spectrogram (acoustic features) or a character vector sequence (textual features, although used for CNN only in the pilot study). 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 layers of BiLSTM, was jointly trained along with the network so that it can 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 models.

The architecture specification is provided in Table 5.1. L_f (for the audio frames) was set to 300 and L_{max} (for the character sequence) was set to 50, considering the utterances' length. Taking into account the syntactic property of the Korean language, sentence-final frames/syllables were utilized. The batch normalization [104] and dropout [105] were utilized only for the CNN (audio) and the multi-layer perceptrons (MLPs).

First, S_{ambi} is constructed using a character BiLSTM-Att [74] alone, which

Table 5.1: Specification of the implemented architectures

	Specification	
CNN (audio)	Input size (single channel)	(L_f, 129, 1)
	# Conv layer	5
	Window size (# filters) ($\rightarrow$ Batch normalization) $\rightarrow$ Max pooling size ($\rightarrow$ Dropout)	5 by 5 (32) $\rightarrow$ 2 by 2 5 by 5 (64) $\rightarrow$ 2 by 2 3 by 3 (128) $\rightarrow$ 2 by 2 3 by 3 (32) $\rightarrow$ 2 by 1 3 by 3 (32 $\rightarrow$ 2 by 1
CNN (text)	Input size (single channel)	(L_{max}, 100, 1)
	# Conv layer	2
	Window size (# filters) $\rightarrow$ Max pooling size	3 by 100 (32) $\rightarrow$ 2 by 1 3 by 1 (no max-pooling)
BiLSTM -Att	Input size	(L_f, 129) (audio) (L_{max}, 100, 1) (text)
	Hidden layer nodes	128 (64 x 2)
	Context vector size	64
MLP	Hidden layer nodes	64 or 128
Others	Optimizer	Adam (0.0005) [84]
	Batch size	16
	Dropout	0.3 (for CNN/MLP)
	Activation	ReLU (CNN/MLP) Softmax (attention, output)

shows the best performance among the implemented models (Table 5.2)[1]. CNN is excellent at recognizing a syntactic distinction that comes from the length of utterances or the presence of specific sentence enders, but not appropriate for handling the scrambling of the Korean language, worsening the performance in the concatenated network.

Table 5.2: Validation performance for S_{ambi} architectures

Models	F1 score	accuracy
CNN	0.7691	0.8706
BiLSTM	0.7811	0.8807
CNN + BiLSTM	0.7700	0.8745
BiLSTM-Att	**0.7977**	**0.8869**
CNN + BiLSTM-Att	0.7822	0.8746

Next, for S_{deci}, especially in abstracting the acoustic features, the concatenation of CNN and BiLSTM-Att was utilized, in the sense that prosody concerns both shape-related properties (e.g., mel spectrogram) and sequential information. Also, as expected, the models which use root mean square energy (RMSE) sequence seem to emphasize the syllable onsets that mainly affect the pitch contour in Korean. For the textual features, a character BiLSTM-Att is adopted as in S_{ambi}. Eventually, the output layer of the acoustic feature is concatenated with the output layer of the character BiLSTM-Att, making up a thought vector that concerns both audio and text. The concatenated vector is fed as an input of an MLP to infer the final intention. The structure of the S_{deci} equals to that of $S_{e2e-hybrid}$, concerning the end-to-end input and output format.

In brief, S_{ambi} adopts a self attentive BiLSTM (acc: 0.88, F1: 0.79). For S_{deci},

[1]Note that all the models are in character-level. Also, the usage of large-scale pre-trained language models in Section 3.4 was prevented here to i) compensate for the smaller size of the training set and ii) to lessen the influence of pretraining corpora.

the networks each utilizing audio (CNN and BiLSTM-Att merged) and text (BiLSTM-Att) were jointly trained via simple concatenation, to make up a multi-modal network (acc: 0.75, F1: 0.61). For all the modules, the dataset was split into train and validation set with the ratio of 9:1. The class weight was taken into account in the training session concerning the imbalance of the volume for each utterance type. The implementation of the whole system was done with Librosa[2] [106], fastText[3] [70, 75] and Keras [83], which were used for extracting acoustic features, embedding character vectors, and making neural network models, respectively.

Experiment

Let (a-d) denote **(Case a-d)** of above.

For (a) S_{e2e}, we adopt the speech corpus (C, I_{speech}) annotated in the previous section. Here, all the IUs were tagged with their genuine intention regarding audio. For (b) $S_{pipe} = S_{asr} + S_{nlu}$, we adopt ($C^{-}$, I_{text}) and perform 6-class classification. For (c) $S_{proposed} = S_{asr} + S_{ambi} + S_{deci}$, we utilize both script and speech, but in a cascading manner. That is, S_{ambi} adopts (C, I^{+}_{text}) and S_{deci} uses (C, I_{speech}). For (d) $S_{hybrid} = S_{asr} + S_{e2e-hybrid}$, ($C$, I_{speech}) is used and the property of $S_{e2e-hybrid}$ equals to S_{deci} of (c).

The overall dataset size, neural network architecture, computation time, and evaluation results are described in Table 5.3. In architecture, RNN denotes BiLSTM-Att and the computation denotes the time spent in the inference of 1,000 utterances. Also, to prevent the influence that comes from ASR performance, we used the ground truth script of each utterance as ASR output in all experiments.

Notably, $S_{proposed}$ yields a comparable result with $S_{e2e-hybrid}$ while reducing

[2] https://github.com/librosa/librosa

[3] https://pypi.org/project/fasttext/

Model	Speech corpus	Text corpus	Training scheme	Architecture	Computation	Accuracy (F1 score)
S_{e2e}	(C, I_{speech})	-	end2end	CNN + RNN	3m 20s	50.00% (0.1972)
S_{pipe}	x	(C^{-}, I_{text})	end2end	RNN	8s	57.20% (0.3474)
$S_{proposed}$	(C, I_{speech})	(C, I^{+}_{text})	3-stage	(b) → (c)	10s	58.50% (0.3814)
S_{hybrid}	(C, I_{speech})	-	2-stage	(a) + (b)	3m 30s	58.65% (0.3706)
$S^{+}_{proposed}$	(C, I_{speech})	(C, I^{+}_{text})	3-stage	(b^{+}) → (c)	10s	75.55% (0.5227)

Table 5.3: Specification of the models compared in the evaluation

the computation time to about 1/20. The utility of the text-based sieve is also observed in the performance of $S_{pipe} = S_{asr} + S_{nlu}$, which is much reliable than S_{e2e} and is close to $S_{proposed}$ and $S_{e2e-hybrid}$.

With the large-scale corpus constructed in Chapter 3, the models which show much higher performance were obtained ($S^{+}_{proposed}$). The models were enhanced with both accuracy and F1 score by a large portion compared to the models trained with the small corpus while preserving the short inference time. This kind of advance seems to be quite tolerable, considering that many recent breakthroughs in NLU tasks accompanied pretrained language inference systems that benefit from out-of-data information.

Analysis

We want to clarify some points about the head-final language and our work's scalability. Head-final syntax regards languages such as Japanese/Korean/Tamil (considering only the rigid head-final ones). We claim that the scheme can be expanded to the other languages that display underspecified sentence enders or *wh-* particles *in-situ*. Moreover, we expect the scheme to be adopted in non-head-final languages that incorporate the type of utterances whose intention is ambiguous without prosody/punctuation (e.g., declarative questions in English).

Referring again to the literature, the result can be compared to the case

of utilizing a fully multi-modal system as suggested for English [11], where the accuracy of 0.83 was obtained with the test set split from the English corpus. Such kinds of systems are uncomplicated to construct and might be more reliable in the sense that fewer human factors are engaged in the implementation. Nevertheless, our approach is meaningful for the cases where there is a lack of resources in labeled speech data. The whole system can be partially improved by augmenting additional text or speech. Also, the efficiency of the proposed system lies in utilizing the acoustic data only for the text that requires additional prosodic information. Resultingly, it lets us avoid redundant computation and prevent confusion from unexpected prosody of users.

We do not claim that our approach is the optimal for intention identification. However, we believe that the proposed scheme might be helpful in the analysis of some low-resource languages since text data is easier to acquire and annotate. We mainly target the utilization of our approach in goal-oriented/spoken-language-based artificial intelligent (AI) systems, where the computation issue is challenging to apply acoustic analysis for all the input speech.

5.2 Application to Spoken Dialogue System

Applying our corpus and the trained system to the real world is an essential consideration for the broader impact of our research. Our protocol makes it possible for conventional SLU 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 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 the feedback, while simply gen-

erating chit-chat for other non-directive utterances (*Omakase dialogue system*).

5.2.1 What is 'Free-running'?

In Table 3.1, 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.

5.2.2 Omakase chatbot

Omakase dialogue system is a coined term for a dialogue manager that adopts the module trained based on our dataset. Figure 5.1 depicts a simplified architecture for the system.

For the transcript of a single utterance in a 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 man-

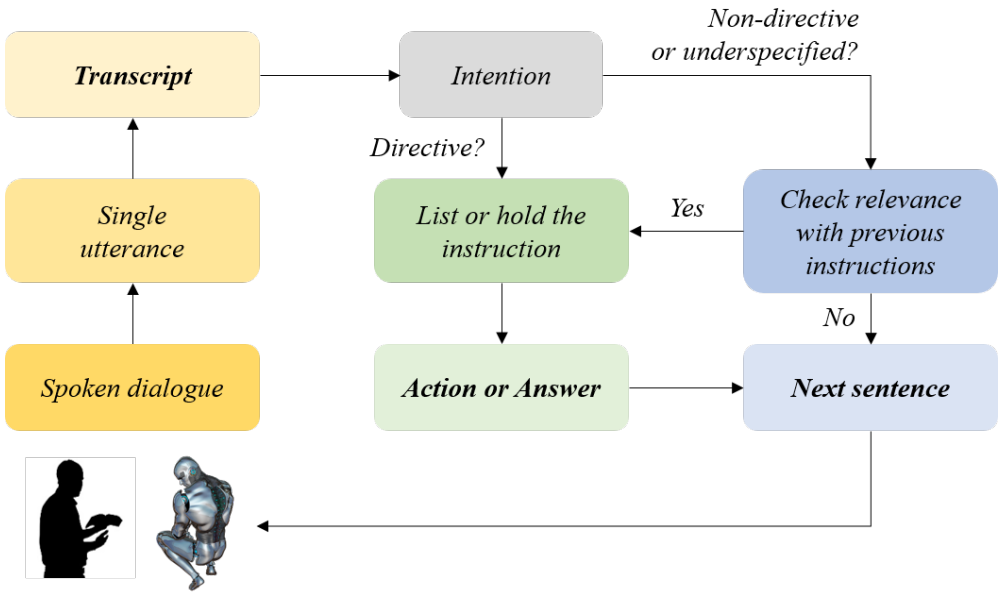

Figure 5.1: A brief illustration on the Omakase dialogue system

ager lists it to the array of instructions so that the module that follows 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 generates the next sentence, for instance, a superficial 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.

Dialog management

Though just conceptional at this point, we named this system *Omakase* since it aims at a well-serving and intelligent 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 idea recently suggested in Sun et al. (2021) [107]. However, our approach intends a more heuristic and less data-driven, but assistive and attachable module. Also, by incorporating sentences with various syntax and sentence forms that are labeled with their intention, the resulting classifier may fit with a wide range of users who are not familiar with talking to intelligent agents in a commanding manner. In other words, our approach heads more human-familiar and inclusive usage of SLU modules.

Since our dialog system considers both task-oriented service and non-task-oriented conversation, it contains two modules, an intention identifier that yields an instruction flag, and a dialog manager that is operated by a spontaneous update of history indices and status flag. Briefly on the whole process, as an input utterance goes in, the system decides if it should take action depending on the instruction flag which is yielded by the intention identification module. In case the utterance is instructional (question/command), the system turns on the status flag (to Q or C) and keeps the instruction for a further obligation or an error correction. Afterward, in case the input utterance is non-instructional or incomplete, the system checks the relevance of the input regarding the preceding instruction and appends the sentence to the instruction list. In all the other non-instruction-related circumstances, a response is generated and the history index is accumulated, resetting the attention if the index reaches a specific value (here three).

Intention identification In the intention identification module, the system categorizes an input text utterance into instructions, non-instructions, and underspecified utterances containing fragments. Here, instructions are specified as questions and commands, which are similar in the sense that they both call for the attention of the agents. For this stage, either one of S_{e2e}, S_{pipe}, $S_{proposed}$,

or S_{hybrid} suggested in Section 5.1 can be exploited.

Instruction flag In each turn, the instruction flag (F_{ins}) immediately displays the result of intention identification. If the input utterance is a question or command, then F_{ins} which is originally `NULL` is turned on. For further usage, F_{ins} is specified into a question flag (F_{ins} = Q) or a command flag (F_{ins} = C). In brief, the instruction flag informs the system whether the current input is the type of utterance that requires the agent's response or not.

Status flag Next, we adopt the status flag (F_{stat}) that indicates whether the agent is paying attention to a specific type of instruction, here, either question or command. For instance, if the question status flag is turned on (F_{stat} = Q), it is recommended that the dialog system treat the forthcoming utterance as one that is relevant to the given question set. Similarly, the forthcoming utterance is considered relevant to the to-do list if the command status flag is turned on (F_{stat} = C). Here, the status flag mainly depends on the flow of conversation, at the same time influenced by the history index.

History index The history index (I_{hist}) is the number that indicates how many turn-takings have taken place without an instructional voice. I_{hist}) is 0 at the starting point, where F_{ins} and F_{stat} are `NULL` by default. As conversation proceeds and once F_{ins} is turned on, I_{hist} is prepared to be updated.

For each turn, if the input utterance is instructional (that F_{ins} = Q or C), or if given the relevant utterance in case of a non-trivial status flag (that F_{stat} = Q or C), consequently I_{hist} is reset to `NULL`. This implies that the agent becomes attentive to a specific instruction. Otherwise, I_{hist} is updated in a positive direction, to count the number of turns that the user goes through irrelevant and non-instructional conversation. As I_{hist} reaches I_{MAX}, F_{stat} is turned to `NULL`. This indicates the moment when the agent gives up waiting for the further re-

quirement or correction of the user regarding the preceding instruction. In an implementation, we heuristically set the maximum value I_{MAX} to 3, since in a human conversation, it seems unlikely and inefficient to mention something about the original instruction after more than two turns of out-of-topic talk. It is supported by the empirical studies that the adjacency pairs generally have three turn-takings [108], and also, the subsequent utterance pairs usually share the relevance topic unless there exists a deep chain which might be required in cumbersome circumstances [109] that may go beyond the conversation with the social robots.

Nevertheless, we set the value to 3, to compensate for the risk of true negative cases regarding the relevance check. If given enough scenario data, our tentative module can be replaced with the one that probabilistically decides whether the attention should be further maintained. In such a case, the preferred approach may be supervised learning that has proven its performance with various natural language processing tasks. Else, it might be possible in the way of reinforcement learning, that adopts the given instruction as an input utterance, and F_{ins}, F_{stat} and I_{hist} as arguments. Though beyond the scope of this study, the latter issue is to be tackled since it also fits well with the data shortage circumstances we are concerned with.

Checking relevance For a pair of input utterances, namely an utterance incorporating instruction and one of the forthcoming utterances, it is not a simple problem to determine if the two share a relevant topic. One possible approach is to identify the topic of each sentence with a specific categorization and check if the classification results are the same. This fits well with task-oriented services since the intent (in domain-specific tasks) may be classified into topics regarding weather, music, schedule, etc. However, for such an approach, an additional annotation on the topic-specified corpus is inevitable.

In this regard, both to guarantee reproducibility and reliability, it would be much more beneficial to adopt publicly released pre-trained language models, especially fine-tuned with semantic textual similarity data.

Dialog generator (optional) Although we have demonstrated managing the instruction utterances so far, for many social robots and companion AIs, making a proper response and continuing dialog is indispensable in the interaction with the human (or possibly non-human) subjects. Thus, a dialog generator can be augmented to the system to provide the users with a communication experience.

Albeit here, we do not adopt a specific conversation model, since the purpose and persona of social chatbots are so diverse these days. However, there is a simple guideline. For each input utterance, a generated answer may be provided; but except for the questions and commands, and also for the potentially instructional sentences with the relevant topic, a more obedient voice would be preferred. It can be prepared in a manner that is easily observable in conventional rule-based chatbots (e.g., *wait, I will find you the answer*). We considered that this resembles a fluent chef chatting with the guest, at the same time taking orders and making the cuisine. Some users may want their social robots to simultaneously provide a reply for their utterance while the instruction is being undertaken [107].

Implementation For the whole dialog system, we designed a framework that manages text-based chatting in colloquial task-oriented circumstances. We provide a pseudo-code in Algorithm 1 for a concise understanding of the whole process. L_{inst} denotes the list of instructions and $s \sim z$ means that s and z regards similar instruction.

Algorithm 1 Persona-switching Dialog System

procedure DIALOG
Initialization:
    $I_{hist} \leftarrow 0, F_{stat} \leftarrow$ `NULL`, $F_{inst} \leftarrow$ `NULL`
    $L_{inst} \leftarrow [\ ]$
For every turn:
    $F_{inst} \leftarrow$ `NULL`
    $s \leftarrow$ `input`, $i_s \leftarrow$ `intention of the input`
    if $i_s \in \{Q, C\}$ **then**
        $F_{inst} \leftarrow i_s, F_{stat} \leftarrow i_s$
        $L_{inst} \leftarrow L_{inst} + [s]$
    else if $i_s \notin \{Q, C\}$ and $F_{stat} \in \{Q, C\}$ **then**
        if $I_{hist} \leq 3$ and $\exists z \in L_{inst} : s \sim z$ **then**
            $F_{inst} \leftarrow i_s$
            $L_{inst} \leftarrow L_{inst} + [s]$
        end if
    end if
Update indices (and generate answer):
    if $i_s \in \{Q, C\}$ **then**
        $I_{hist} \leftarrow 0$
    else if $i_s \notin \{Q, C\}$ **then**
        if $F_{stat} \in \{Q, C\}$ and $I_{hist} \leq 3$ **then**
            $I_{hist} \leftarrow I_{hist} + 1$
        else
            $I_{hist} \leftarrow 0$
        end if
    end if
end procedure

Demonstration To help the readers' comprehension, we display here a scenario that best explains the characteristics of the proposed system. The flags and indices are notated, after being inferred by the system for each turn.

Structured reply implies the case when the system encounters an instruction for which the answer can be ready-made, and simple chit-chat denotes a freestyle response, which is possibly generated. Note that the switching of personas takes place in the sentences (2), (4), and (7), flexibly managing both dialog and to-do-list.

USER: (1) hey come here

AGENT: (Structured reply)

[***intention***: *command*, ***instruction list***: *[1]*, ***relevance***: *False*]

[***ins flag***: *C*, ***stat flag***: *C*, ***hist ind***: *0*]

USER: (2) i was just boring

AGENT: (Simple chit-chat)

[***intention***: *non-instruction*, ***instruction list***: *[1]*, ***relevance***: *False*]

[***ins flag***: *False*, ***stat flag***: *C*, ***hist ind***: *1*]

USER: (3) isn't it a good day for a short travel

AGENT: (Simple chit-chat)

[***intention***: *non-instruction*, ***instruction list***: *[1]*, ***relevance***: *False*]

[***ins flag***: *False*, ***stat flag***: *C*, ***hist ind***: *2*]

USER: (4) how's the weather in tokyo now

AGENT: (Structured reply)

[***intention***: *question*, ***instruction list***: *[4]*, ***relevance***: *False*]

[***ins flag***: *Q*, ***stat flag***: *C*, ***hist ind***: *0*]

USER: (5) oh i mean kyoto

AGENT: (Structured reply)

[***intention***: *non-instruction*, ***instruction list***: *[4,5]*, ***relevance***: *True*]

[***ins flag***: *Q*, ***stat flag***: *Q*, ***hist ind***: *0*]

USER: (6) and also arashiyama

AGENT: (Structured reply)

[***intention***: *non-instruction*, ***instruction list***: *[4,5,6]*, ***relevance***: *True*]

[***ins flag***: *Q*, ***stat flag***: *Q*, ***hist ind***: *0*]

USER: (7) okay i got it i'll be get dressed

AGENT: (Simple chit-chat)

[***intention***: *non-instruction*, ***instruction list***: *[4,5,6]*, ***relevance***: *False*]

[***ins flag***: *False*, ***stat flag***: *Q*, ***hist ind***: *1*]

USER: (8) let's depart as soon as i'm ready

AGENT: (Structured reply)

[***intention***: *command*, ***instruction list***: *[8]*, ***relevance***: *False*]

[***ins flag***: *C*, ***stat flag***: *C*, ***hist ind***: *0*]

5.3 Beyond Monolingual Approaches

So far, we have concentrated on the monolingual application of the proposed concept and systems. However, another circumstance that the ambiguity matters is when ambiguous sentences should be translated into another language. Appropriate translation from a language to another may reduce the ambiguity that is inevitable in a single language, especially in textual format. However, if such ambiguity is not resolved due to the deficiency of acoustic or contextual information, semantic errors may propagate from the transcription to the machine translation (MT) stage.

Recent spoken language translation (SLT) research has been expanded to focus on the prevention of such error propagation and utilization of small resources [110, 111]. Along with the advent of knowledge distillation [112] and model compression techniques [113], up-to-date schemes are proposed to compensate for the standard training schemes that require a large amount of data

and parameters. However, in those studies, the translation of para-linguistic and functional components has rarely been considered compared to its importance.

Comprehension of this diversity can be challenging in translation through the ASR-MT pipeline. First, para-linguistic features such as intonation, overall tone, and prosody are mostly simplified in the speech to text process. This can, of course, appear indirectly through punctuation marks, silence notation, spacing, etc., in text [114]. However, in Korean, it is often not viable to describe the acoustic features delicately only with textual representation. In addition, some functional components that may not have cognates are often inserted in, to make the conventional pipeline structure difficult to catch the nuance just via intermediate text output.

There are quite a few works which deal with the information that speech incorporates more than text [115, 116, 117, 118, 3, 112]. However, they proceeded mainly on increasing the accuracy of intent understanding or translation, and as previously mentioned, it was difficult to find a discussion on the resolution of ambiguous utterances and the reduction of the linguistic information loss. Here, we skim the literature on spoken language translation and show how we created a resource for an end-to-end speech translation that may help resolve the chronic issue of spoken language translation.

5.3.1 Spoken language translation

Early SLT approaches had actively utilized the ASR-MT (+ speech synthesis) pipeline [119]. Literature focused on the distinct improvement of ASR and MT, and research on end-to-end SLT to overcome the problems in this process was recently boosted by B´erard et al. (2016) [110]. It intensively explored the SLT of English to French, and a subsequent study [120] suggested an efficient corpus construction scheme through the dataset proposal. Quite a few

benchmarks such as the mentioned Augmented LibriSpeech, IWSLT shared tasks [121], and MuST-C [122] were proposed. On top of these, the feasibility was explored by analyzing baseline performances [123, 111, 117, 118].

Despite these advances, while investigating the languages studied so far, we observed that the researchers concentrate on Romance or Germanic languages and Chinese. Though this regards the efficiency of low-resource translation [117] and probably the industrial necessity, specific languages being set as a benchmark is one of the factors that can deter the typological development of SLT technology. First, Romance and Germanic languages share many lexical similarities with English [124, 125], mainly the target or source language. Such languages may receive translation advantages over other languages like Altaic and Austronesian ones, in terms of lexicon, sentence order, and phonetics.

Next, though the characteristics of Chinese are more close to the languages that use Chinese characters (e.g., Japanese or Korean) rather than to English, since the tonality is significant from the perspective of phonetics, much of the acoustic information is exploited to determine the lexicon itself correctly. Ironically, and followingly, if the outcome of the ASR is exact, it is assumed that the translation process can be more straightforward than expected, concerning that Chinese incorporate tens of thousands of isolated tokens. In contrast, in Korean, the language of interest in this study, many properties of the sentences are determined according to the prosody, while the lexicon does not necessarily change [57, 93, 41]. That is, even if the output of the ASR displays a low word error rate (WER), in a text-to-text translation, it is difficult to fully comprehend whether the sentence is a question or a statement, whether a question is *wh-* or *yes/no*, whether the sentence is rhetorical or not, and so on. We also observed that some functional features such as sentence enders exhibit their role dominated by the acoustic property [126, 60, 127].

Without a doubt, this tendency, para-linguistic features determining sen-

tence meaning, is not a distinct characteristic of Korean. One can observe a similar phenomenon concerning sentence-end particles in Cantonese to English [128]. In English as well (that is, in from-English translation), for the declarative questions [39], though grammatically unacceptable in some sense, the sentence-end intonation gives the utterance directiveness. Besides, in many languages, nuance (such as a question being rhetorical) is determined by the overall tone of the sentence. Such factors include a variance in pitch and magnitude, negative polarity items, and some functional components that determine the mood [129, 130]. Nevertheless, we scrutinize the Ko-En case in this study because the various linguistic characteristics previously described are simultaneously observable. This implies that when creating an SLT dataset for some language pair, one may need to consider a little more factors than for other pairs. We want to materialize those in this section.

5.3.2 Dataset

For this study, we augmented the English translation to a speech corpus of Korean (syntactically) ambiguous sentences constructed in Section 4.2. In the original corpus, 1,292 ambiguous scripts are provided, and the number of speech utterances with distinct speech acts derived from the scripts is 3,552. The dataset was constructed in such a way as to infer the intention labels given audio and script, with the number of labels being 7; namely statement (S), *yes/no* question (YN), *wh*-question (WH), rhetorical question (RQ), command (C), request (R), and rhetorical command (RC), slightly modified from [58]. To this end, *wh*-particles were first decided, and then predicates (verbs), pronouns, particles and suffixes were subsequently augmented. This dataset deals only with sentences that have multiple interpretations, and aims to disambiguate them with prosody from a syntax-semantics point of view. In addition, the sentences were created using lexicons that can be used in as many

colloquial circumstances as possible, referring to Korean word dictionaries.

Augmenting English translation concerned both audio and script data. The English sentences were translated and edited by three people, namely Korean natives with intermediate English fluency. The sentences that failed to reach a full consensus of suitability in English translation were deleted in the verification process. Also, for each sentence, if present, properties such as reportativeness, affirmativeness, and politeness were separately indicated, as to be described in Section 4.5. Alternative interpretations were prepared for rhetorical questions/commands.

Directiveness

Directiveness is a factor that decides whether an utterance is a statement or a question/command [19]. Often, the directiveness of rhetorical questions [63] or rhetorical commands [65] is also discussed, but to make our argument clear, we will only consider pure questions or commands as directive utterances here. In the Korean language, where directiveness is displayed as a prosodic segment [93], it is mainly represented by sentence-final intonation [57]. For example, the following two translation results can be compared:

(5-1) 천천히 가고 있어 chenchen-hi ka-ko iss-e
slow-ADV go-PRT be-SE

(a) *I am going in a slow phase.* (statement)

(b) *Are you going slowly now?* (question)

The way to correctly translate this sentence in the pipeline structure may be manually augmenting the punctuation in text [114], or providing additional information that the sentence is interrogative [60]. If so, the ambiguity regarding directiveness can be partially resolved. However, in the process, the loss of such phonetic information, which is a chronic drawback of the pipeline struc-

ture, can be problematic. This phenomenon also bothers in an in-service situation where the user is unfamiliar with the target or the source language.

Wh-intervention

While the issue of directiveness originates to a certain extent that Korean is a head-final language [60], another cumbersome characteristic of Korean is that it is a wh-in-situ language [55]. In other words, a wh-particle can be interpreted as a component of wh-question depending on prosody, or as an existential quantifier. (5-2a) and (5-2b) are representative examples.

(5-2) 누가 먹고 싶대 nwu-ka mek-ko siph-t.ay
who-NOM eat-to want-DEC.RPT

(a) *Who wants to eat?* (*wh*-Q, reportative)

(b) *Does anyone want to eat?* (*yes/no* Q, reportative)

Detecting the occurrence of *wh*-intervention is more cumbersome than the decision of directiveness in SLT, because it is challenging to denote the property as a separate component in the sentence after the ASR process. Even though the sentence structure is the same, the question becomes *yes/no* when the *wh*-expression is an existential quantifier, and the *wh*-question is yielded when the *wh*-particle is utilized as it is. In other words, unlike in English, where *do*-support and *wh*-movement distinguish the sentence types, the textual shape of the two utterances is identical in Korean. Therefore, in this case, unless one records the *wh*- or *yes/no* attribute separately, a direct translation from the speech would be advantageous.

Rhetoricalness

Rhetoricalness mainly decides if the question or command is a requirement for an answer or action. In English, for rhetorical questions, the tone assigned

to the wh-particle, related verbs, or polarity item comprehensively determines the nuance [131, 130], as well as in German [129]. The Korean language incorporates correspondings, but the aspect is slightly different. Slightly different from English, where do-support or wh-movement can have a significant influence on the overall prosody and tone as in (5-3a-c), in ambiguous Korean utterances where the rhetoricalness is mainly expressed through the modification of prosody in a textually fixed sentence (5-3), there is a strong tendency that people insert more accent around the wh-particles, along with its corresponding cases or verbs. For instance, for the sentence (5-3) to be read rhetorical in Korean, there comes a short pause just before *nwu* (who) and the corresponding verb *wa* (come) has a dramatic rising accent.

(5-3) 오늘 회사에 누구 와 onul hoysa-ey nwu-ka wa

today office-LOC[4] who-NOM come

(a) *Does anyone come to office today?* (*yes/no* Q)

(b) *Who will come to office today?* (wh- Q)

(c) *Who (the hell) comes to office today?* (RQ)

Besides, albeit dependent on acoustic features, rhetoricalness is a highly cultural and pragmatic property that cannot be easily detected [131, 63]. Thus, the corresponding agreement between language users is relatively lower than other syntactic properties [57]. Followingly, identifying such a tone only via text in a syntactically ambiguous sentence is almost impossible without a dramatic prosodic segment [130] or a lexical feature such as a polarity item [132]. From this, in Korean, it can be inferred that SLT corpora might better be augmented with some scripts containing polarity items or speeches with sufficient dynamic or pitch range, to boost the performance of understanding rhetoricalness.

[4]Locative case.

Subject and object drop

A core factor for the ambiguity regarding above three phenomena is the frequent drop of subjects and objects in Korean [133]. For example, revisiting (5-2), na (나, I) is omitted in reading (5-1) as (5-1a) (check 5-4a), and ne (너, you) is omitted in reading as (5-1b) (check 5-4b).

(5-4a) (na) chenchen-hi ka-ko iss-e

I am going in a slow phase.

(5-4b) (ne) chenchen-hi ka-ko iss-e

Are you going slowly now?

This originates in that the Korean language is used in high-context culture [134, 135], and followingly, Korean speakers assume a subject through context and thus variate prosody in the process of reflecting it covertly. This differs from English, which requires pronouns such as *I, you, s/he*, etc., except for sentences in particular forms such as imperative. Consequently, correctly grasping the above attributes plays an essential role in translation. This is because the reconstruction of the subject or object is crucial in determining the first, second, and third person in the target language, if exists. Especially in the case of rhetorical questions or commands, if such information is lost, nuance can be transferred incorrectly. For instance, a single sentence (5-5) can be translated into either sentence whose tone and meaning are different (5-5a,b).

(5-5) 누가 갖다 달래 nwu-ka kac-ta tal-l.ay

who-NOM bring-PRT give-IMP.RPT

(a) *Somebody told me to bring it here.* (statement)

(b) *Who on earth told you to bring it here?* (RQ)

Multi-functional particles and politeness suffix

So far, we have investigated the factors that precede the text-level representation. On the other hand, in terms of morpho-syntax, Korean is an agglutinative language that incorporates various particles, or functional morphemes, that make up the words. They determine the grammatical role of each *eojeol* (word), which sometimes dominates the act of the whole sentence or influences the nuance. What we observed in the corpus are reportative and affirmative sentences yielded by specific sentence-final particles. Such components convey nuances that are difficult to translate into the target languages.

Reportatives and affirmatives Reportative ender is a component that reflects evidentiality [136]. In the case of a statement, it implies that the speaker has heard the information from somebody. In the case of a question, the addressee is asked for information that s/he has heard from somewhere. In the case of a command, it implies that someone other than the speaker has assigned a task to the addressee (or speaker her/himself). In Korean, this case is expressed as a single morpheme such as *-ay* (hearsay marker) augmented after the canonical sentence enders (e.g., declarative, interrogative, and imperative).

Unlike the prosody-sensitive cases, the evidentiality is preserved at the sentence level in the pipeline translation process. However, we assumed it could be more concise if given as an additional label, in case of question and command. In other words, e.g., for (5-6), which can be interpreted as either statement or question, it is natural to express the statement as (5-6a), but (5-6b), which demonstrates elaborately, was considered awkward. Instead, we wrote it as (5-6c) and augmented the reportative (RPT) label. The same holds for the commands, as like 'Bring some water' rather than 'I heard that you should bring some water'.

(5-6) 뭐 좀 먹고 싶대 mwe com mek-ko siph-t.ay

what little[5] eat-PRT want-DEC.RPT

(a) *I have heard that s/he wants to eat something.*

(b) *Is it what you heard that s/he wants to eat something?*

(c) *Does s/he wants to eat something?* (reportative)

On the other hand, the affirmative case is mainly observed in the questions. Rather than asking the addressee for the purpose of information-seeking, it aims to confirm that what the speaker already knows is correct. The notation of such purpose is more challenging compared to the case of reportatives. The affirmativeness of the questions is usually conveyed by ending the question in *-ci* as in (5-7a). If it is not interpreted as an ender for questions, it just acts as a declarative ender (5-7b), which implicates a kind of self-confidence. The similar holds in English, but in Korean, either case is fully grammatical.

(5-7) 그건 어떻게 잘 끝냈지 ku-ken ettehkey cal kkuthnay-ss-ci

it-NOM somehow well[6] finish-PST-AFM

(a) *Did you finish it somehow?* (affirmative)

(b) *I finished it somehow.*

Politeness The last is the politeness suffix, which adds a new property to the sentence, similarly to the two cases above. This mainly appears in the form of augmenting a particle such as *-yo* at the end of the sentence, and in fact, no expression can replace it grammatically in English, notwithstanding it is possible to represent the manner indirectly through would- or may- questions. Therefore, it was considered that a situation in which the spoken language had such a factor that could be separately notated.

The main difference between politeness and reportativeness/affirmativeness is that the role of the preceding content does not change in the former case. Due

[5] A polarity item that means 'a little'.

[6] A passing word, and does not necessarily mean the goodness.

to the sentence enders utilized in the dataset being mostly underspecified, we found that the politeness suffix can come after almost all the candidates, only adding a functional attribute.

We note that the necessity of notification owes more to the cultural factor than in the case of reportatives or affirmatives. That is, in some target languages, such particles can be inserted in the sentences as well. However, in Ko-En, we had to add it mechanically, albeit the usage is possibly restricted.

5.3.3 Analysis

In the corpus, primarily to clarify where the directiveness matters, we sought the sentences that can be interpreted as a statement and either of *yes/no* or *wh*-question. The percentage of such sentences reached 1,023 of all 1,292, which shows that the directiveness is observed as a core factor of the ambiguity (Table 5.4).

Next, the appearance of *wh*-intervention was measured by the co-existence of interpretation as *wh*-question and other statements or directives. Directives here include *yes/no* questions, commands and requests, all in which the *wh*-particles are regarded as an existential quantifier. We found 848 cases among all 1,292, implying that the wh-intervention happens for all the cases in the corpus that concerns *wh*-questions. Since yes/no and *wh*-questions co-appear very frequently (675 among all), their distinction is crucial to the detailed understanding of directives.

The rhetorical utterances occupy 298 among all 1,292 sentences, which might seem insignificant. However, taking into account the low portion of rhetorical questions in colloquial environment[7], the percentage is not to be ignored. It instead suggests that the rhetoricalness is relatively frequent among the ambiguous utterances, and the interpretation should be sensitive to prosody around

[7] [58] suggests 1,185 over 19,318 utterances (6.13%) for a spoken language corpus.

	Instances	Portion (%)
Directiveness	1,023 / 1,292	79.17
Wh- Intervention	848 / 1,292	65.63
Rhetoricalness	298 / 1,292	23.06
Reportativeness	353 / 3,552	9.93
Affirmativeness	55 / 3,552	1.54

Table 5.4: Statistics on the frequency of appearance for three acoustics-related attributes and two functional properties

specific terms, to be aware of the nuance of the speech.

As for the functional attributes, we observed quite a few cases where the two kinds of attitudes were represented through multifunctional particles in the corpus. We categorized them by some labels for the nuances to be conveyed separately. In specific, 353 utterances showed reportativeness and 55 were affirmative, among the total of 3,552. There was no overlap between those two types. Politeness information, in numerics, does not incorporate a specific statistical meaning since almost all the conversation-style utterances can be converted via just adding a suffix. Thus, the portion of the utterances that show politeness is about half. Some exceptions occurred in the cases when it is unable to augment a suffix, such as the sentences terminating with *-nya* (an informal interrogative ender).

5.3.4 Discussion

Considering the factors mentioned in the analysis, Ko-En translation through the ASR-MT pipeline often faces challenges. These problems can be solved depending on the amount of acoustic information or context.

It is promising that the first three factors are resolvable via various modeling using acoustic features, if given little context for the rhetoricalness. The fourth factor, covert subjects or objects, can consequently be derived if the

first three are found to be definite. Among the rest, namely the functional attributes, politeness has little to do with the existence of speech or context, while the other two are influenced by the sentence type that the intonation decides. Below we suggest a simple scheme to consider these while making up the SLT corpus.

What should be checked?

Factors to be disambiguated From the perspective of Ko-En translation, to solve the first three quests, a separate script annotation is required for the utterances that are interpreted distinctly depending on intonation. In other words, this can be organized as identification of intonation-dependent sentences and insertion of additional information. In the Korean language itself, the methodology has been discussed as a corpus linguistic approach [57, 58]. Adopting this will not only aggregate the additional data into the SLT corpus, but will also yield a more delicate outcome. More specifically, it is necessary to investigate the SLT corpus currently being distributed or utilized, and check out the followings:

1. Does the sentence ender specify the role of the utterance as in Pak (2008) [60]? If not, what can the candidates be? It should be checked if the sentence incorporates excessively specific information to be regarded as a question, in view of conversation maxims [33].

2. In case of *wh*-questions, can *wh*-particle be interpreted as an existential quantifier [55]? If so, interpreting it as a *yes/no* question acceptable?

3. Is acknowledging the pure question as a rhetorical question or a statement still acceptable?

4. If the subject or object is dropped, what can it be? What will the resulting

Verb	Wh-	Portion (%)	Type	Translation	Politeness	Miscellaneous
가다 (go)	Who	누가 간대	s	I heard sbd will go	n	
	Who	누가 간대	yn	Is anyone going?	n	rep
	Who	누가 간대	wh	Who is going?	n	rep
	Who	누가 간대	rq	Who says I'm going?	n	(= I won't go)
	Who	누가 간대요	s	I heard sbd will go	p	
	Who	누가 간대요	yn	Is anyone going?	p	rep

Table 5.5: Excerpt of the augmented dataset

sentence type be? What is considered awkward among assigning all the agents (1^{st} to 3^{rd} person)?

5. In case of the presence of vocatives, is there a chance that the utterance can be differently interpreted?

6. In case of the presence of the polarity items, usually in the form of adverbs or numerics, is there a possibility that the utterance can be interpreted as a pure question or command?

Some are adopted directly from [57] and [58]. In addition to the lexicons, we infer that the length may matter in deciding the acceptability of the ambiguity [137]. Overall, we note that detecting the prosody-sensitivity of a sentence has a positive influence on SLT as well as a correct understanding of intention (Table 5.5).

Though the fourth issue, the subject and object drop, is expected to be resolved upon the above factors, one of the difficulties here is that the distinction between I, we, or you is challenging, if the corresponding components are dropped. Fortunately, this issue is not a part that can have a great impact on the assertion, question set, or to-do list that is conveyed, due to Korean being not sensitive to number agreement. However, it might give the translation far more possibilities of interpretation. Thus, we added all possible candidates as

a note, to supplement the information that could have been missed. This phase also went through the adjudication of the translators regarding acceptability.

Factors to be identified The other grammatical factors, ones related to multi-functional particles and politeness, might not be the component of explicit translation output as aforesaid, unless stated as in (5-6a). However, they often need to be preserved in from-Korean translation, though they do not have corresponding lexicons. In this regard, the reportatives, affirmatives, and politeness, can be materialized in a structured format (Table 5.5). The first two columns contain information on the predicates and *wh*-particles that are utilized, and the following two columns state the sentence and its intention type. The augmented are translation, politeness, and miscellaneous, where the last one contains the information regarding reportativeness, affirmativeness, and an alternative translation on rhetorical utterances.

The data augmentation is easily achievable since the factors are grammatical. In other words, the nuance of multi-functional particles and politeness are to be resolved by generating more input data mechanically or just by simple morphological analysis. Also, it might make the whole annotation process much concise, compared to the scheme that reflects all the functional features in the translation, as in (5-6b).

The labeled attributes can be learned jointly in the translation training phase (as multi-task learning), or be provided by a separately trained network. Of these, the latter is straightforward, since it is consistent with the conventional classification scheme. On the other hand, the former's approach is assumed to let the machine learn sentences and properties simultaneously. It may give attention to the sentence components which help grasp the particular nuance successfully.

Factors not covered in this study Though not covered in this section, in Korean, there are a few more features to be taken into account when introduced with speech-based disambiguation. For datives, constituent length affects NP ambiguity [91]. For comparatives, the duration or pause between the words can clarify the subjects that are indicated [92]. Similarly, for the phrases containing genitive terms and modifiers, the duration in between determines the attachment and the following syntactic structure [93]. We assume these are to be considered as possible in constructing the larger dataset.

Beyond these para-linguistic features, more subtle semantics is driven by the non-canonical usage of the standard sentence enders such as *-ta* (declaratives) and *-ni* (interrogatives). For *-ta*, which generally comes with fall intonation, if used with rise, a boasting is implicated [127]. Instead, for *-ni*, which is usually exploited with the rise ending, the falling intonation makes it more likely to be a rhetorical question, frequently used as rejection, refutation, or reproach [126]. As a morphologically rich language, Korean encompasses a much more variety of functional particles that add a specific mood to the question. We leave these as a further study.

Application

The dataset we proposed can mainly be adopted as a scheme to augment a new corpus, as the format in Table 5.5. Simplifying the procedure is as follows:

- Detecting prosody-sensitive utterances in the script
- Augmenting the supplementary utterances after deciding how the scripts can be read and translated in various ways
- Adding the information regarding multifunctional particles and politeness to the translation output of the corpus, using rule and learning hybrid method

After creating an augmented speech corpus in this way, how can it be used? First, the augmented dataset can be utilized directly in training, as the most intuitive and promising method. However, the number of ambiguous sentences may not be significant[8], which may lessen the performance of inferences in the training process.

Thus, another approach is to intensively utilize ambiguous utterances in the process of fine-tuning or distillation of the pre-trained SLT module, possibly adopting the scheme suggested in Bansal et al. (2019) [117]. In this process, the scheduling of the training procedure [138] may be critical for the learning not to be biased to one side, between the inference in the pre-trained MT module and the new challenging data regarding ambiguous utterances.

Both approaches are recommendable ways to translate spoken utterances concerning the syntax-semantic ambiguity and functional features of the sentences. Implementing such a procedure for the pre-trained models in service is our next research direction.

5.4 Summary

In the first section, we proposed a multi-stage system for the identification of speech intention. The system first checks if the speech is a fragment or has determinable intention, and if neither, it conducts an intonation-aided decision, associating the underspecified utterance with the genuine intention. For a data-driven training of the modules, 7K speeches were additionally collected or manually tagged, yielding an accuracy of 58.65% with the built corpus and 75.55% with the aid of pre-trained intention classification model, utilizing an additionally constructed challenging test set. The possible application of the

[8]Though they frequently appear in Korean, the size of spoken language corpus is not sufficiently large usually, and we cannot guarantee that each sentence type comes up with a balanced portion.

proposed system is the SLU modules of intelligent agents, especially those targeting a free-style conversation with humans. Our future work aims to enhance a multimodal system for the disambiguation module, which can be reliable only by making up a large-scale and accurately tagged speech database.

In the second section, we suggested a natural language-based instruction understanding system that flexibly handles the conversation style and error correction within the user dialog. The proposed system can be usefully adopted in the social robots and companion AIs, whose dialog flow should be managed in a persona-switching manner between task-oriented service and non-task-oriented conversation. Although we did not cover all the dialog situations that take place between the user and the agent, it is expected that our approach can be utilized as a primary and basic module for conversational AI, especially where the simple adaption and implementation schemes are preferred or required. Many parts of the system have the potential to be enhanced via introducing an augmented training set, pre-trained sentence encoding models, and deep reinforcement learning assisted by annotated scenario data, which are achievable if the temporal and economic burdens are alleviated. As a next step, we aim to attack those obstacles and enhance the naturalness of persona-switching in human-AI conversation.

In the last section, we checked the points to be considered in Korean-to-English SLT, based on the dataset concerning the ambiguous sentences. There are a total of six categories, namely directiveness, *wh*-intervention, rhetoricalness, subject and object drop, multi-functional particles including reportativeness and affirmativeness, and politeness. The first four items, which mainly come from the Korean language being high context and wh-in-situ, can hopefully be resolved in the SLT. The others, which require additional tagging on the corpus, are recommended to be considered in the future dataset construction phase. In further research, we will investigate whether the constructed

corpus is practical regarding the above observations, and whether the expected performance is achievable by fine-tuning with only a small amount of data constructed in this research. The dataset is to be freely available online[9].

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

Chapter 6

Conclusion and Future Work

Through this dissertation, we dealt with the ambiguity of spoken language understanding that can be resolved with prosody.

We studied how such ambiguity affects intention understanding within the prosody-sensitive language Korean, and constructed a corpus to scrutinize this phenomenon for the Korean spoken language, conducting quantitative and qualitative analysis using pre-trained language models. Also, to see whether such ambiguity can be resolved by using acoustic information in spoken language, we created an artificial speech corpus that consists of ambiguous sentences and verified the feasibility with attention models.

On the application side, we discussed the utilization of proposed methodologies in the real-world dialogue system, assuming intention identification technology that accompanies ambiguity resolution. In addition, considering the points discussed above, we can also tackle ambiguity that occurs in speech translation, suggesting that the consideration of ambiguity can be applied beyond monolingual approaches.

In this dissertation, we narrowly defined the problem of ambiguity and presented methodologies to solve it. However, ambiguity may exist in many other ways for various semantic/syntactic structures, as well as for intention

understanding. Not all of these ambiguities need to be resolved, since it is the nature of the human language. However, we saw that it is crucial to distinguish the types of ambiguity that cannot be or do not need to be resolved from those that need resolution, and further improve natural language understanding systems by understanding/reflecting the resolvable ambiguity.

In future work, we want to define and resolve ambiguity in a way that applies to multiple tasks and languages. Also, for ambiguity in spoken language understanding that can be resolved with acoustic factors, we want to proceed by defining a prosodic segment in human spoken language, beyond merely conducting post-mortem corpus analysis. It may help tackle the limitations of text-oriented language analysis, and is also suitable for leveraging pre-trained spoken language models currently in progress.

We hope that this research will be utilized theoretically and industrially. In addition, it is expected that the corpus, construction methodology, and analysis results built in this process will play a significant role in handling the ambiguity in spoken language understanding, especially in Korean language processing.

Bibliography

[1] B. Liu and I. Lane, "Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling," in *Proc. Interspeech 2016*, 2016, pp. 685–689.

[2] P. Haghani, A. Narayanan, M. Bacchiani, G. Chuang, N. Gaur, P. Moreno, R. Prabhavalkar, Z. Qu, and A. Waters, "From audio to semantics: Approaches to end-to-end spoken language understanding," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 720–726.

[3] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech Model Pre-Training for End-to-End Spoken Language Understanding," in *Proc. Interspeech 2019*, 2019, pp. 814–818.

[4] L. Qin, T. Xie, W. Che, and T. Liu, "A survey on spoken language understanding: Recent advances and new frontiers," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Z.-H. Zhou, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2021, pp. 4577–4584, survey Track.

[5] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "The ATIS spoken language systems pilot corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*, 1990. [Online]. Available: https://aclanthology.org/H90-1021

[6] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[7] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[8] J. Allen and M. Core, "Draft of DAMSL: Dialog act markup in several layers," 1997.

[9] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–374, 2000. [Online]. Available: https://aclanthology.org/J00-3003

[10] S. Matsubara, S. Kimura, N. Kawaguchi, Y. Yamaguchi, and Y. Inagaki, "Example-based speech intention understanding and its application to in-car spoken dialogue system," in *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. [Online]. Available: https://aclanthology.org/C02-1107

[11] Y. Gu, X. Li, S. Chen, J. Zhang, and I. Marsic, "Speech intention classification with multimodal deep learning," in *Canadian Conference on Artificial Intelligence*. Springer, 2017, pp. 260–271.

[12] Y. Liu, J. Zhang, H. Xiong, L. Zhou, Z. He, H. Wu, H. Wang, and C. Zong, "Synchronous speech recognition and speech-to-text translation with in-

teractive decoding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8417–8424.

[13] W. I. Cho, D. Kwak, J. W. Yoon, and N. S. Kim, "Speech to Text Adaptation: Towards an Efficient Cross-Modal Distillation," in *Proc. Interspeech 2020*, 2020, pp. 896–900.

[14] M. Kim, G. Kim, S.-W. Lee, and J.-W. Ha, "St-bert: Cross-modal language model pre-training for end-to-end spoken language understanding," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7478–7482.

[15] S. Kim, G. Kim, S. Shin, and S. Lee, "Two-stage textual knowledge distillation for end-to-end spoken language understanding," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7463–7467.

[16] D. Gibbon, "The ambiguity of 'ambiguity': beauty, power, and understanding," *Jodłowiec Maria and Leśniewska Justyna*, 2010.

[17] W. I. Cho, J. Cho, W. H. Kang, and N. S. Kim, "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*, S. Denison, M. Mack, Y. Xu, and B. C. Armstrong, Eds. cognitivesciencesociety.org, 2020. [Online]. Available: https://cogsci.mindmodeling.org/2020/papers/0448/index.html

[18] J. L. Austin, *How to Do Things with Words: The William James Lectures Delivered at Harvard University in 1955*. Oxford, England: Oxford University Press, 1962.

[19] J. R. Searle, "Austin on locutionary and illocutionary acts," *The Philosophical Review*, vol. 77, no. 4, pp. 405–424, 1968.

[20] ——, "A classification of illocutionary acts," *Language in Society*, vol. 5, no. 1, pp. 1–23, 1976.

[21] J. F. Allen and C. R. Perrault, "Analyzing intention in utterances," *Artificial intelligence*, vol. 15, no. 3, pp. 143–178, 1980.

[22] R. De Mori, "Spoken language understanding: a survey," in *2007 IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, 2007, pp. 365–376.

[23] A. Warstadt, A. Singh, and S. R. Bowman, "Neural network acceptability judgments," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 625–641, 2019.

[24] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1112–1122. [Online]. Available: https://www.aclweb.org/anthology/N18-1101

[25] E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, R. Mihalcea, G. Rigau, L. Uria, and J. Wiebe, "SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado: Association for Computational Linguistics, Jun. 2015, pp. 252–263. [Online]. Available: https://www.aclweb.org/anthology/S15-2045

[26] J. Obleser and F. Eisner, "Pre-lexical abstraction of speech in the auditory cortex," *Trends in Cognitive Sciences*, vol. 13, no. 1, pp. 14–19, 2009.

[27] K. Shimada, K. Iwashita, and T. Endo, "A case study of comparison of several methods for corpus-based speech intention identification," in *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING2007)*, 2007, pp. 255–262.

[28] A. Friedrich, A. Palmer, and M. Pinkal, "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)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1757–1768. [Online]. Available: https://aclanthology.org/P16-1166

[29] S. Vosoughi and D. Roy, "Tweet acts: A speech act classifier for Twitter," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 10, no. 1, pp. 711–714, Aug. 2021. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14821

[30] J. M. Sadock and A. M. Zwicky, "Speech act distinctions in syntax," *Language Typology and Syntactic Description*, vol. 1, pp. 155–196, 1985.

[31] B. Comrie and J. Sadock, "Toward a linguistic theory of speech acts," *Philosophical Quarterly*, vol. 26, no. 104, p. 285, 1976.

[32] S. C. Levinson, *Pragmatics*, ser. Cambridge Textbooks in Linguistics. Cambridge University Press, 1983.

[33] S. C. Levinson, C. Stephen, and S. C. Levinson, *Presumptive meanings: The theory of generalized conversational implicature*. MIT press, 2000.

[34] G. Gazdar, "Speech act assignment," in *Elements of Discourse Understanding*, A. Joshi, B. H. Weber, and I. A. Sag, Eds. Cambridge University Press, 1981, pp. 64–83.

[35] P. Portner, "The semantics of imperatives within a theory of clause types," in *Semantics and Linguistic Theory*, vol. 14, 2004, pp. 235–252.

[36] J. Allwood, "An activity based approach to pragmatics," *Gothenburg Papers in Theoretical Linguistics*, no. 76, pp. 1–38, 1995.

[37] C. Beyssade and J.-M. Marandin, "The speech act assignment problem revisited: Disentangling speaker's commitment from speaker's call on addressee," *Empirical Issues in Syntax and Semantics*, vol. 6, no. 37-68, 2006.

[38] L. N. Kennette, *On the disambiguation of meaning: the effects of perceptual focus and cognitive load*. Wayne State University, 2012.

[39] C. Gunlogson, "Declarative questions," in *Semantics and Linguistic Theory*, vol. 12, 2002, pp. 124–143.

[40] J. Neitsch, B. Braun, and N. Dehé, "The role of prosody for the interpretation of rhetorical questions in German," in *9th International Conference on Speech Prosody 2018*, 2018, pp. 192–196.

[41] J. Yun, "Meaning and prosody of wh-indeterminates in Korean," *Linguistic Inquiry*, vol. 50, no. 3, pp. 630–647, 2019.

[42] P. T. Daniels and W. Bright, *The world's writing systems*. Oxford University Press on Demand, 1996.

[43] H. Yamashita, Y. Hirose, and J. L. Packard, *Processing and producing head-final structures*. Springer Science & Business Media, 2010, vol. 38.

[44] K. Park, J. Lee, S. Jang, and D. Jung, "An empirical study of tokenization strategies for various Korean NLP tasks," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 133–142. [Online]. Available: https://aclanthology.org/2020.aacl-main.17

[45] S.-J. Chang, *Korean*. John Benjamins Publishing, 1996, vol. 4.

[46] C. LEE, "Speech act terms and mood indicators (in Korean)," *Acta Linguistica Hungarica*, vol. 38, no. 1/4, pp. 127–141, 1988. [Online]. Available: http://www.jstor.org/stable/44362607

[47] S.-J. Kim, Y.-H. Lee, and J.-H. Lee, "Korean speech act tagging using previous sentence features and following candidate speech acts," *Journal of KIISE: Software and Applications*, vol. 35, no. 6, pp. 374–385, 2008.

[48] H.-S. Kim, C.-N. Seon, and J.-Y. Seo, "Review of Korean speech act classification: machine learning methods," *Journal of Computing Science and Engineering*, vol. 5, no. 4, pp. 288–293, 2011.

[49] S. Park, J. Moon, S. Kim, W. I. Cho, J. Y. Han, J. Park, C. Song, J. Kim, Y. Song, T. Oh *et al.*, "KLUE: Korean language understanding evaluation," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[50] A. S. Byon, "The role of linguistic indirectness and honorifics in achieving linguistic politeness in Korean requests," vol. 2, no. 2, pp. 247–276, 2006. [Online]. Available: https://doi.org/10.1515/PR.2006.013

[51] S. Cho and H.-g. Lee, "NPIs and rhetorical question in Korean," *The Linguistic Association of Korea Journal International Issue*, pp. 145–166, 2001.

[52] N.-R. Han, *Korean zero pronouns: analysis and resolution*. University of Pennsylvania, 2006.

[53] A. Park, S. Lim, and M. Hong, "Zero object resolution in Korean," in *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, Shanghai, China, Oct. 2015, pp. 439–448. [Online]. Available: https://aclanthology.org/Y15-1050

[54] H. H. Ceong, "The morphosyntax of clause typing: Single, double, periphrastic, and multifunctional complementizers in Korean," Ph.D. dissertation, 2019.

[55] Y. Jang, "Two types of question and existential quantification," vol. 37, no. 5, pp. 847–869, 1999. [Online]. Available: https://doi.org/10.1515/ling.37.5.847

[56] Y. Xu, "Prosody, tone and intonation," *The Routledge handbook of phonetics*, pp. 314–356, 2019.

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

[58] W. I. Cho and N. S. Kim, "Text implicates prosodic ambiguity: A corpus for intention identification of the Korean spoken language," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, mar 2022, just Accepted. [Online]. Available: https://doi.org/10.1145/3529648

[59] M. S. Kim, "Evidentiality in achieving entitlement, objectivity, and detachment in Korean conversation," *Discourse Studies*, vol. 7, no. 1, pp. 87–108, 2005.

[60] M. D. Pak, "Types of clauses and sentence end particles in Korean," *Korean Linguistics*, vol. 14, no. 1, pp. 113–156, 2008.

[61] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2017, pp. 986–995.

[62] J. Merchant, "Fragments and ellipsis," *Linguistics and Philosophy*, vol. 27, no. 6, pp. 661–738, 2005.

[63] H. Rohde, "Rhetorical questions as redundant interrogatives," 2006.

[64] C.-h. Han, *The structure and interpretation of imperatives: mood and force in Universal Grammar*. Psychology Press, 2000.

[65] M. Kaufmann, "Fine-tuning natural language imperatives," *Journal of Logic and Computation*, vol. 29, no. 3, pp. 321–348, 2019.

[66] J. Nam, "A novel dichotomy of the Korean adverb nemwu in opinion classification," *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, vol. 38, no. 1, pp. 171–209, 2014.

[67] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[68] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. [Online]. Available: https://aclanthology.org/D14-1181

[69] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[70] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017. [Online]. Available: https://aclanthology.org/Q17-1010

[71] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[72] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in *International Conference on Learning Representations*, 2019.

[73] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[74] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=BJC_jUqxe

[75] W. I. Cho, S. J. Cheon, W. H. Kang, J. W. Kim, and N. S. Kim, "Giving space to your message: Assistive word segmentation for the electronic typing of digital minorities," in *Designing Interactive*

Systems Conference 2021, ser. DIS '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 1739–1747. [Online]. Available: https://doi.org/10.1145/3461778.3462078

[76] S. TBrain, "Korean BERT pre-trained cased (KoBERT)," https://github.com/SKTBrain/KoBERT, 2019.

[77] J. Lee, "KcBERT: Korean comments BERT," in *Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology*, 2020, pp. 437–440.

[78] J. Park, "KoELECTRA: Pretrained ELECTRA model for Korean," https://github.com/monologg/KoELECTRA, 2020.

[79] J. Lee, "KcELECTRA: Korean comments ELECTRA," https://github.com/Beomi/KcELECTRA, 2021.

[80] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.

[81] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[82] N. National Institute of Korean Languages, "NIKL CORPORA 2020 (v.1.0)," 2020. [Online]. Available: https://corpus.korean.go.kr

[83] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[84] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[85] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[86] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave, "CCNet: Extracting high quality monolingual datasets from web crawl data," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4003–4012. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.494

[87] Y. Hur, S. Son, M. Shim, J. Lim, and H. Lim, "K-EPIC: Entity-perceived context representation in Korean relation extraction," *Applied Sciences*, vol. 11, no. 23, p. 11472, 2021.

[88] K. Yang, "Transformer-based Korean pretrained language models: A survey on three years of progress," *arXiv preprint arXiv:2112.03014*, 2021.

[89] W. I. Cho, J. Cho, J. Kang, and N. S. Kim, "Prosody-semantics interface in seoul Korean: Corpus for a disambiguation of wh-intervention," in *Proceedings of the 19th International Congress of the Phonetic Sciences (ICPhS 2019)*, 2019, pp. 3902–3906.

[90] A. Pires and H. Taylor, "The syntax of wh-in-situ and common ground," in *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, vol. 43, no. 2. Chicago Linguistic Society, 2007, pp. 201–215.

[91] H. Hwang and A. J. Schafer, "Constituent length affects prosody and processing for a dative NP ambiguity in Korean," *Journal of Psycholinguistic Research*, vol. 38, no. 2, pp. 151–175, 2009.

[92] J.-B. Kim and P. Sells, "A phrasal analysis of Korean comparatives," *Studies in Generative Grammar*, vol. 20, pp. 179–205, 2010.

[93] H. Baek and J. Yun, "Prosodic disambiguation of syntactically ambiguous phrases in Korean," *MIT Working Papers in Linguistics*, vol. 88, pp. 89–100, 2018.

[94] N. Richards, *Contiguity theory*. MIT Press, 2016, vol. 73.

[95] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2225–2235.

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

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

[98] N. Hellbernd and D. Sammler, "Prosody conveys speaker's intentions: Acoustic cues for speech act perception," *Journal of Memory and Language*, vol. 88, pp. 70–86, 2016.

[99] A. Talman, A. Suni, H. Celikkanat, S. Kakouros, J. Tiedemann, and M. Vainio, "Predicting prosodic prominence from text with pre-trained contextualized word representations," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 2019, pp. 281–290.

[100] M. D. Pell, A. Jaywant, L. Monetta, and S. A. Kotz, "Emotional speech processing: Disentangling the effects of prosody and semantic cues," *Cognition & Emotion*, vol. 25, no. 5, pp. 834–853, 2011.

[101] B. M. Ben-David, N. Multani, V. Shakuf, F. Rudzicz, and P. H. van Lieshout, "Prosody and semantics are separate but not separable channels in the perception of emotional speech: Test for rating of emotions in speech," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 1, pp. 72–89, 2016.

[102] I. Miura and N. Hara, "Production and perception of rhetorical questions in Osaka Japanese," *Journal of Phonetics*, vol. 23, no. 3, pp. 291–303, 1995. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447095801626

[103] J. Y. Lee, S. J. Cheon, B. J. Choi, N. S. Kim, and E. Song, "Acoustic Modeling Using Adversarially Trained Variational Recurrent Neural Network for Speech Synthesis," in *Proc. Interspeech 2018*, 2018, pp. 917–921.

[104] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*. PMLR, 2015, pp. 448–456.

[105] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[106] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, vol. 8, 2015, pp. 18–25.

[107] K. Sun, S. Moon, P. A. Crook, S. Roller, B. Silvert, B. Liu, Z. Wang, H. Liu, E. Cho, and C. Cardie, "Adding chit-chat to enhance task-oriented dialogues," in *Proceedings of the 2021 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 1570–1583.

[108] A. B. Tsui, "Beyond the adjacency pair," *Language in Society*, vol. 18, no. 4, pp. 545–564, 1989.

[109] E. Goffman, "Replies and responses," *Language in Society*, vol. 5, no. 3, pp. 257–313, 1976.

[110] A. Bérard, O. Pietquin, L. Besacier, and C. Servan, "Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation," in *NIPS Workshop on End-to-End Learning for Speech and Audio Processing*, Barcelona, Spain, Dec. 2016. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01408086

[111] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Low-resource speech-to-text translation," in *Proc. Interspeech 2018*, 2018, pp. 1298–1302.

[112] Y. Liu, H. Xiong, J. Zhang, Z. He, H. Wu, H. Wang, and C. Zong, "End-to-End Speech Translation with Knowledge Distillation," in *Proc. Interspeech 2019*, 2019, pp. 1128–1132.

[113] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=S1XolQbRW

[114] F. N. Akinnaso, "On the differences between spoken and written language," *Language and Speech*, vol. 25, no. 2, pp. 97–125, 1982.

[115] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," in

Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 949–959. [Online]. Available: https://aclanthology.org/N16-1109

[116] Y. Qian, R. Ubale, V. Ramanaryanan, P. Lange, D. Suendermann-Oeft, K. Evanini, and E. Tsuprun, "Exploring asr-free end-to-end modeling to improve spoken language understanding in a cloud-based dialog system," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 569–576.

[117] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 58–68. [Online]. Available: https://aclanthology.org/N19-1006

[118] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model," in *Proc. Interspeech 2019*, 2019, pp. 1123–1127. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1951

[119] B. Zhou, D. Déchelotte, and Y. Gao, "Two-way speech-to-speech translation on handheld devices," in *Proc. Interspeech 2004*, 2004.

[120] A. C. Kocabiyikoglu, L. Besacier, and O. Kraif, "Augmenting librispeech with French translations: A multimodal corpus for direct speech

translation evaluation," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://aclanthology.org/L18-1001

[121] J. Niehues, R. Cattoni, S. Stüker, M. Cettolo, M. Turchi, and M. Federico, "The IWSLT 2018 evaluation campaign," in *Proceedings of the 15th International Conference on Spoken Language Translation*. Brussels: International Conference on Spoken Language Translation, Oct. 29-30 2018, pp. 2–6. [Online]. Available: https://aclanthology.org/2018.iwslt-1.1

[122] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: a Multilingual Speech Translation Corpus," 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2012–2017. [Online]. Available: https://www.aclweb.org/anthology/N19-1202

[123] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6224–6228.

[124] O. W. Robinson, *Old English and its closest relatives: a survey of the earliest Germanic languages*. Stanford University Press, 1992.

[125] T. McArthur and T. Macarthur, *The English Languages*, ser. Canto (Cambridge University Press). Cambridge University Press, 1998. [Online]. Available: https://books.google.co.kr/books?id=m0XVCSfvfPkC

[126] H.-J. Min and J. C. Park, "Analysis of indirect uses of interrogative sentences carrying anger," in *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation*. Seoul National University, Seoul, Korea: The Korean Society for Language and Information (KSLI), Nov. 2007, pp. 311–320. [Online]. Available: https://aclanthology.org/Y07-1032

[127] C. Lee, G. B. Simpson, Y. Kim, and P. Li, *Handbook of East Asian Psycholinguistics*. Cambridge University Press, 2015, vol. 3.

[128] K. C. Chiu, "Exploring the role of utterance-final particle lō in turn allocation in Cantonese conversation," in *Proceedings of the 29th North American Conference on Chinese Linguistics (NACCL-29)*, vol. 1, 2017.

[129] D. Wochner, J. Schlegel, N. Dehé, and B. Braun, "The prosodic marking of rhetorical questions in German," in *Interspeech 2015 : 16th Annual Conference of the International Speech Communication Association*. ISCA Archive, 2015, pp. 987–991. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2015/i15_0987.html

[130] N. Dehé and B. Braun, "The prosody of rhetorical questions in English," *English Language & Linguistics*, vol. 24, no. 4, pp. 607–635, 2020.

[131] C.-H. Han, "Deriving the interpretation of rhetorical questions," in *Proceedings of West Coast Conference in Formal Linguistics*, vol. 16. Citeseer, 1998, pp. 237–253.

[132] S. Oraby, V. Harrison, A. Misra, E. Riloff, and M. Walker, "Are you serious?: Rhetorical questions and sarcasm in social media dialog," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Saarbrücken, Germany: Association for Computational Linguistics,

Aug. 2017, pp. 310–319. [Online]. Available: https://www.aclweb.org/anthology/W17-5537

[133] N. Kwon, M. Polinsky, and R. Kluender, "Subject preference in Korean," in *Proceedings of the 25th West Coast Conference on Formal Linguistics*, 2006, pp. 1–14.

[134] D. Kim, Y. Pan, and H. S. Park, "High-versus low-context culture: A comparison of Chinese, Korean, and American cultures," *Psychology & Marketing*, vol. 15, no. 6, pp. 507–521, 1998.

[135] S. Nishimura, A. Nevgi, and S. Tella, "Communication style and cultural features in high/low context communication cultures: A case study of Finland, Japan and India," *Teoksessa A. Kallioniemi (toim.), Uudistuva ja kehittyvä ainedidaktiikka. Ainedidaktinen symposiumi*, vol. 8, no. 2008, pp. 783–796, 2008.

[136] J. Song, "Evidentiality in Korean," in *Evidentials and Modals*. Brill, 2020, pp. 412–444.

[137] B. Hemforth, S. Fernandez, C. Clifton Jr, L. Frazier, L. Konieczny, and M. Walter, "Relative clause attachment in German, English, Spanish and French: Effects of position and length," *Lingua*, vol. 166, pp. 43–64, 2015.

[138] "Towards making the most of BERT in neural machine translation," vol. 34. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6479

초 록

언어의 중의성은 필연적이다. 그것은 언어가 의사 소통의 수단이지만, 모든 사람이 생각하는 어떤 개념이 완벽히 동일하게 전달될 수 없는 것에 기인한다. 이는 필연적인 요소이기도 하지만, 언어 이해에서 중의성은 종종 의사 소통의 단절이나 실패를 가져오기도 한다.

언어의 중의성에는 다양한 층위가 존재한다. 하지만, 모든 상황에서 중의성이 해소될 필요는 없다. 태스크마다, 도메인마다 다른 양상의 중의성이 존재하며, 이를 잘 정의하고 해소될 수 있는 중의성임을 파악한 후 중의적인 부분 간의 경계를 잘 정하는 것이 중요하다.

본고에서는 음성 언어 처리, 특히 의도 이해에 있어 어떤 양상의 중의성이 발생할 수 있는지 알아보고, 이를 해소하기 위한 연구를 진행한다. 이러한 현상은 다양한 언어에서 발생하지만, 그 정도 및 양상은 언어에 따라서 다르게 나타나는 경우가 많다. 우리의 연구에서 주목하는 부분은, 음성 언어에 담긴 정보량과 문자 언어의 정보량 차이로 인해 중의성이 발생하는 경우들이다.

본 연구는 운율(prosody)에 따라 문장 형식 및 의도가 다르게 표현되는 경우가 많은 한국어를 대상으로 진행된다. 한국어에서는 다양한 기능이 있는(multi-functional한) 종결어미(sentence ender), 빈번한 탈락 현상(pro-drop), 의문사 간섭(*wh*-intervention) 등으로 인해, 같은 텍스트가 여러 의도로 읽히는 현상이 발생하곤 한다. 이것이 의도 이해에 혼선을 가져올 수 있다는 데에 착안하여, 본 연구에서는 이러한 중의성을 먼저 정의하고, 중의적인 문장들을 감지할 수 있도록 말뭉치를 구축한다. 의도 이해를 위한 말뭉치를 구축하는 과정에서 문장의 지향성(directiv-

ity)과 수사성(rhetoricalness)이 고려된다. 이것은 음성 언어의 의도를 서술, 질문, 명령, 수사의문문, 그리고 수사명령문으로 구분하게 하는 기준이 된다. 본 연구에서는 기록된 음성 언어(spoken language)를 충분히 높은 일치도($\kappa = 0.85$)로 주석한 말뭉치를 이용해, 음성이 주어지지 않은 상황에서 중의적인 텍스트를 감지하는 데에 어떤 전략 혹은 언어 모델이 효과적인가를 보이고, 해당 태스크의 특징을 정성적으로 분석한다.

또한, 우리는 텍스트 층위에서만 중의성에 접근하지 않고, 실제로 음성이 주어진 상황에서 중의성 해소(disambiguation)가 가능한지를 알아보기 위해, 텍스트가 중의적인 발화들만으로 구성된 인공적인 음성 말뭉치를 설계하고 다양한 집중(attention) 기반 신경망(neural network) 모델들을 이용해 중의성을 해소한다. 이 과정에서 모델 기반 통사적/의미적 중의성 해소가 어떠한 경우에 가장 효과적인지 관찰하고, 인간의 언어 처리와 어떤 연관이 있는지에 대한 관점을 제시한다.

본 연구에서는 마지막으로, 위와 같은 절차로 의도 이해 과정에서의 중의성이 해소되었을 경우, 이를 어떻게 산업계 혹은 연구 단에서 활용할 수 있는가에 대한 간략한 로드맵을 제시한다. 텍스트에 기반한 중의성 파악과 음성 기반의 의도 이해 모듈을 통합한다면, 오류의 전파를 줄이면서도 효율적으로 중의성을 다룰 수 있는 시스템을 만들 수 있을 것이다. 이러한 시스템은 대화 매니저(dialogue manager)와 통합되어 간단한 대화(chit-chat)가 가능한 목적 지향 대화 시스템(task-oriented dialogue system)을 구축할 수도 있고, 단일 언어 조건(monolingual condition)을 넘어 음성 번역에서의 에러를 줄이는 데에 활용될 수도 있다.

우리는 본고를 통해, 운율에 민감한(prosody-sensitive) 언어에서 의도 이해를 위한 중의성 해소가 가능하며, 이를 산업 및 연구 단에서 활용할 수 있음을 보이고자 한다. 본 연구가 다른 언어 및 도메인에서도 고질적인 중의성 문제를 해소하는 데에 도움이 되길 바라며, 이를 위해 연구를 진행하는 데에 활용된 리소스, 결과물 및 코드들을 공유함으로써 학계의 발전에 이바지하고자 한다.

주요어: 음성언어 이해, 자연어 처리, 중의성, 의도 파악
학번: 2014-22579

ACKNOWLEDGMENT

대학원에 들어온 지도 어느덧 팔 년, 관악에 몸담은 지는 거의 십삼 년이 지났습니다. 물론 훨씬 오래 계신 분들도 계시겠지만, 서른이 좀 넘게 살아온 세월 동안 이렇게 오래 한 곳에 머무는 것은 처음인 것 같아, 이 끝이 시원섭섭하기도 합니다. 철없이 기타치고 하고 싶은 공부만 하던 학부 시절을 거쳐 대학원에 처음 들어오게 된 것은 음악을 연구하고 싶어서였던 것 같은데, 졸업할 때 보니 사람의 언어를 연구하고 있는 제 자신을 볼 때마다 묘한 기분입니다.

제가 휴먼인터페이스 연구실에 들어와 음악 연구에서 방향을 돌려 자연어를 연구하게 되고, 그 과정에서 여러 가지 언어 데이터셋을 구축할 기회를 갖게 된 것, 그리고 시간이 지나 이렇게 무사히 졸업하고 학위를 받게 된 것에 대해 모두, 옆에서 걱정스런 마음으로 지도 및 조언을 아끼지 않으셨던 김남수 교수님께 감사의 인사를 드리고 싶습니다. 아직도 기억나는 장면은, 언어도 음악과 비슷한 것이니 한번 언어를 연구해 보면 어떻겠냐고 하시던 순간, 그리고 하고 싶은 연구들을 열심히 보고하는 저에게 연구하는 것이 재밌느냐고 물어보셨던 순간입니다. 관련 전공 수강 내역 없이 무작정 문을 두드린 때부터, 음대와 뇌인지과학과의 수업을 듣겠다고 220동을 기웃거리던 석사 시절, 음성명령 수행 과제를 배정받고 정작 언어학과 가서 수업을 듣고 있었던 박사 초입, 연구실에서 그간 제출하지 않았던 학회 및 아카이브에 논문을 내고 리뷰에 깨지던 박사 시절까지, 사실 저는 많은 경우 제가 하고 싶었던 방향대로 문제를 풀어나갔던 것으로 기억합니다. 그 과정에서 종종 교수님의 생각과 다른 부분들에 열띤 토론을 하기도 했지만, 항상 합리적으로 생각하고 발전적인 방향으로 지도해 주시는 모습, 그리고 인격적인 대우들에 감사와 존경의 마음을 다

시금 표하고 싶습니다.

그리고 항상 생각하는 것은, 저는 과분할 정도로 좋은 동료들만을 옆에 두어 왔다는 것입니다. 제가 입학하자마자 수행해야 했던 매트랩과 C 코딩을 기가 막히게 가르쳐 주었던 연구실의 정신적 지주 신재 형, 워크샵 때마다 형수님과 세나와 함께 오셔서 저희 모두를 삼촌으로 만들어 주셨던 두화 형, 그리고 같은 음향 덕후로써 연구실에서 종종 음악과 장비에 대해 이야기하고, 대학원 초기 저의 천방지축 행동들에도 따뜻한 충고를 해 주셨던 철민이 형에게 감사의 인사를 전합니다. 또, 언제나 냉철한 시선으로 사안을 보고 인생의 조언들을 아끼지 않아 주시는 연구실의 진짜 브레인 태균이 형, 이 년 가까이 윗방에서 함께 지내며 심려도 많이 끼쳐 드렸지만 또 술 마시고 이런저런 이야기하며 인간적인 모습도 많이 공개해 주셨던 연구실 축구왕 기수 형에게도 많이 감사했습니다. 그리고, 동아리의 대선배이기도 하지만 연구실에서 또다른 모습으로 만났던 석재 형에게도 연구 외적으로도 많이 여쭤보고 배웠던 것 같습니다.

아랫방에 함께 내려와 거의 사년을 옆자리에서 함께한 인규 형, 이제는 다시금 더 자주 보게 되지 않을까 싶습니다. 가즈아를 농담으로 받지 못해서 죄송했어요. 인규 형과도 말레이를 함께 다녀왔던 수현이 형, 무슨 이야기를 하든 유쾌하게 잘 받아 주셔서 감사했습니다. 초반에 인식 팀장과 팀원으로 만났던 강현이 형과도 지옥의 패러미터 튜닝 기억도 있고 청춘의 토픽들에 대해 이야기 나누었던 기억들이 스쳐 지나갑니다. 준엽이 형께는 제가 방장일 때 말을 잘 안 들은 것 같아 항상 죄송했는데, 결혼하고 행복하게 지내고 계신 것 같아서 좋아요. 아랫방의 수호자, 항상 어려운 일이 있을때 MIR 팀부터 무향실 녹음, 아랫방 적응, 그리고 워크샵 방콕팸에 함께하였던 정훈이 형도 깊은 감사를 드립니다.

소중한 동기와 후배 분들도 연구실 생활을 계속하는 데에 큰 힘이 되었습니다. 석사 초기에 취미 생활을 함께 했고 결국은 다른 연구분야를 택했지만 항상 따뜻하게 봐 주고 조언 주셨던 성준이 형, 윗방과 아랫방에서 오랫동안 동고동락하며 점심시간에 신세 많이 졌던 형용이 형, 연구 이야기로 함께 시간을 보내며 뉴올리언즈도 함께했던, 지금은 한국에서 보기 어려운 우현이 형, 14학번 동기들 모두 사랑합니다. 또 에스토니아에서 열심히 좋아하는 것을 하며 살고 있을 세영이도 한국에 올

때마다 좋은 선물들 잘 챙겨주어 너무 고마웠고, 방장 하느라 고생 많았던 현승이도 이제 바로 직속 후배로 다음 졸업까지 화이팅입니다. 히오스 너무 못해서 미안했어. 아랫방에서 형용형과 디스커션을 열심히 하던 모습이 기억나는 석완이 형, 그리고 마치 박지성 같은 산소통을 자랑하는 (문)성환이도 연구를 열심히 진행하고 있는 것 같아서 보기 좋습니다.

마곡에 놀러갔을 때 반갑게 반겨 주었던 주현이 형과는 더 자주 보지 못해 아쉬웠고, 병진이 형도 합성 과제로 이것저것 이야기를 많이 주고받았지만 윗방과 아랫방에 나뉘어 있어 많이 보지는 못해 아쉬워요. 현승이와 병진이 형과 셋이 윤리 과제 제안서를 쓰고 훈련소에 들어가던 때는 아직도 잊을 수가 없네요. 먼 옛날, 나와 이런저런 이야기를 하고 연구실에 들어왔던 지환이는 더 챙겨 주지 못해서 미안해. 그래도 회사 가서 즐겁게 연구하고 취미생활하는 모습을 보니 좋다. 주현이 형과 함께 마곡에 있는 형래도 하고 싶었던 음악 연구를 잘 하고 지내는 것 같고. 민현이는 항상 연구실의 듬직한 방장으로, 앞으로도 믿음직스럽게 연구실 운영을 해줄 것이라 기대합니다. 수고가 많습니다. 아랫방 옆자리에서 오랫동안 연구로 소통했던 지원이는 좋은 주제를 잡아서 꾸준히 논문을 내고 있어 대단한 것 같아요. 아랫방의 미식 담당이자 NLP 부사수인 석민이는, 거의 유일한 팀원이었음에도 더 많이 챙겨 주거나 무언가 가르쳐주지는 못한 것 같아서 미안합니다. 워낙 연구실에서 하는 다른 연구와 궤가 달라 적응이 힘들었을 텐데, 맡은 일들을 잘 수행해 주어 고맙습니다.

후배들 중에는 직접 말을 해볼 기회가 많지 않아 아쉬웠던 분들도 많이 있습니다. 이번 과제 덕에 부쩍 이야기를 많이 하게 된 민찬이의 경우도, 이것저것 이야기를 전해 듣기만 했지 직접 이야기를 많이 나누어 보지는 못해 아쉽습니다. 주어진 일을 묵묵히 잘 하는 모습이 분명 연구에 있어서도 좋은 결과를 가져올 것이라 생각합니다. 연구실 생활 동안 짧게 같은 공간에 있었던 병찬이는 나중에 사회에서 다시 반갑게 인사할 수 있었으면 좋겠습니다. 연구실 에이스 형주, 강남역에서 우연히 마주쳐 반가웠는데 이후 연구실 행사들에서도 종종 보고 이야기할 수 있어서 좋았어요. 아랫방 귀염둥이 범준이는 너무 말썽 피우지 않고 전문연도 잘 마무리하고, 맡고 있는 과제로 어깨가 무거울 텐데 무사히 수행해 내길 기원합니다.역시 이번 과제로 인생 이야기를 많이 했던 동준이도, 맘대로 일이 풀리지 않아도 배워두는 것들이 언젠가

의미있는 순간이 있을 테니 너무 상심 말고 연구 계속하며 좋아하는 것을 찾길 바랍니다. 명훈이도 재미있는 주제들로 논문을 내며 점차 연구 역량을 늘려 가는 것 같아서 분명 잘 해낼 것이라 생각합니다. 아랫방을 밤늦게까지 지키는 (안)성환이의 열정과 구현력 등등을 보면서 저도 배우는 것이 많습니다. 좋은 연구결과로도 이어지길 바라 봅니다. 지환이는 어려운 과제 함께 하면서, 팀은 향상이지만 NLP에도 관심 가지면서 든든하게 데이터를 처리해 주어 다시금 고맙다는 말을 전하고 싶습니다. 무사히 전문연도 구했으면 좋겠고. 음악 연구의 물꼬를 다시 한번 터줄 수 있을 것 같은 세민이도 하고 싶은 연구 꿋꿋이 계속 해서 의미있는 결과 거두기를 바랍니다. 연구실에서 연구를 지켜봐주고 응원해 준 모두들 너무 고마웠습니다.

8년 간 공부를 해온 만큼 이러저런 일들이 있었습니다. 무엇보다 저의 박사시절을 함께한, 정민화 교수님 연구실의 로봇과제 세 분께 깊은 감사를 드립니다. 막막한 과제 사 년 동안 함께 헤쳐나간 종인님, 그런 저희를 항상 잘 챙겨주셨던 정 대표님, 그리고 지금은 좋은 곳에 먼저 가 계시는 규환님까지도, 제가 어떤 아이디어를 가져와도 응원하고 격려해 주셔서 너무 고마웠습니다. 그 때 잘 배우고 소통하고 구축했던 내용이 저의 밑바탕이 되고 학위논문이 되었습니다. 좋은 과제에 매닝해 주신 교수님, 그리고 잘 지휘해 주신 장준혁 교수님과 신종원 교수님께도, 커미티를 하며 주신 많은 조언 외에 이러한 점도 있었음을 이 자리를 빌어 감사를 드립니다. 과제에 직접 참여한 멤버는 아니었지만, 졸업 프로젝트로 함께 데이터를 구축했던 하은 학생과 대호 학생, 지금은 미시건에 있는 정화와 학교를 떠난 지민님도, 관련 연구를 함께 발전시킬 수 있어서 너무 즐거운 경험이었습니다. 언젠가 또 사회 혹은 학회에서 만날 일이 있기를 기원합니다. 또, 저만의 생각들로 논문을 쓰던 시절, 온라인으로 거침없이 연락주시고 미숙한 아이디어를 함께 발전시켜 주셨던 영기님과 상환님께도 늘 고맙습니다. 영기님은 항상 지도박사님으로 불러주셔서 참 민망했는데, 이제 정말 박사가 되니 조금 덜 민망할 것 같네요. 상환님도 분명 좋은 아이디어로 논문을 더 출판하여 무사히 학위과정을 마치실 것이라 생각합니다. 함께 밤을 지샜던 용래도 큰 결심하고 진학한 대학원을 결실 있게 마무리짓기를 기원합니다.

또 공부 기간 동안 저를 키워준 것은 그만큼 활발히 활동할 수 있었던 AI 커뮤니티들이었던 것 같습니다. 텐서플로 코리아, 파이토치 코리아, 챗봇 코리아, 그리고

사운들리까지, 온라인/오프라인에서 자유롭게 소통할 분들이 계셨던 것만으로도 저는 성장할 수 있었습니다. 데이터를 공개하여 사람들과 소통하고, 그 과정에서 여러 가지 재미있는 프로젝트들에도 참여하게 되었습니다. 무엇보다 데이터의 의미를 알아봐주시고 코웍을 선뜻 제안해주신 경태님, 그리고 그 과정에서 클로바와 브릿지를 놓아 주셨던 하정우 소장님께 깊은 감사를 드립니다. 연구실과 학교에서만 공부하던 제가 세상으로 나갈 수 있는 길이 되었다고 생각합니다. 대학원 시절 저의 거의 유일한 인더스트리 경험이었던 클로바에서 어려운 부분들을 함께 디스커션해 주신 상우님과 동현님, 그리고 점심을 함께 해 주신 모든 분들께 감사합니다. 이어서 하게 된 파파고와의 과제에서도 잘 내용을 조율해 주셨던 루시와 끝까지 과제를 잘 수행할 수 있게 도와주셨던 현창님 모두 이 자리를 빌어 감사의 말씀을 전합니다. 비록 초반에는 우왕좌왕했지만, 돌이켜 보면 너무 좋았던 프로젝트 경험들이었습니다. 또한, 바벨탑과 싸이그래머의 세계로 저를 인도하고, 다양한 스터디에 저를 초대해 주셨던 무성님과 윤경님께도 깊은 감사를 드립니다. 결국은 함께 의미있는 과제를 수행할 수 있었고, 그 과정에서 대화라는 방대하면서도 신비로운 주제를 본격적으로 다뤄볼 수 있어서 영광이었습니다. 이 과정에서 많이 수고해 주셨던 서연 쌤에게도 충분한 감사 인사를 전하지 못한 것 같습니다. 교감이란 주제 하에 똘똘 뭉쳐 상반기 내내 열심히 대화를 살펴보았던 안녕루다 팀의 수민님, 유정님, 그리고 영훈이와도 너무 좋은 시간 의미있게 함께 보냈습니다. 아직 연구가 마무리되지는 않았지만, 좋은 연구 지원해준 언더스코어 관계자 여러분께도 깊은 감사를 드립니다.

저의 첫 국제학회는 O-COCOSDA였고 처음 가 본 해외 학회는 NAACL이었지만, 본격적으로 제가 사람들과 소통하고 함께 연구할 동력을 얻게 된 것은 아무래도 19년의 ACL이 아니었나 싶습니다. 아무래도 혼자 간 학회이다 보니 누군가와 만나 친해지며 돌아다닐 수밖에 없었고, 그 과정에서 열흘 내내 함께해준 재민이, 파파고 및 네이버 팀과 한층 더 친해지게 해 준 지형이, 그리고 먼 후배이자 초짜 자연어처리 연구자에게 스스럼 없이 다가오고 친해졌던 현중이 형, 모두 고맙습니다. 두오모가 보이는 루프탑에서 먹은 양식은 아직도 기억에 남네요. 그렇게 친해진 지형이와 함께 BEEP! 프로젝트를 진행하게 되고, 파이콘 스피커로만 어렴풋이 알고 있던 준범 님과도 함께 작업하게 되었습니다. 훌륭한 자연어처리 연구자 및 개발자 분들과

여러 방면으로 함께 작업할 수 있었던 것은 영광이었어요. 이러한 연구 내용을 보고 기탄없이 연락주어 APEACH 프로젝트를 함께 진행했던 (양)기창님과 원준님, 그렇게 친해져서 아지트에서, 인덕원에서 함께 술잔을 기울였던 수경님, 기현님과 HK, 모두 학교에서만 있던 저를 잘 받아주셔서 너무 고맙습니다.

주변 사람들에게 다소 생소한 분야를 연구하면서도 계속 하고 싶은 일을 해나갈 힘을 얻은 것은, 한국어 NLP계에서 무언가를 이루어 보고자 하는 사람들이 커뮤니티를 만들어 함께 뜻을 모으는 과정에서 다양한 분들과 이야기하고 울고 웃은 덕분이 아닐까 싶습니다. Korpora라는 귀중한 프로젝트에 선뜻 손을 벌려 주신 (이)기창님과 현중이 형, 그리고 KLUE라는 뜻깊은 프로젝트에 함께했던, 여기에서 모두 언급할 수 없는 많은 한국어 NLP 연구자 분들께 이 자리를 빌어 다시 한번 깊은 감사의 말씀을 드립니다. PM이었던 성준님과 지형이, 또 모델 파트를 총괄해서 맡아 준 성동이가 특히 고생이 많았고, 데이터 구축 파트를 공동으로 리드했던 지윤누나, STS팀을 함께 꾸려나간 명화님과 성원님, 그리고 동준님께 다시 한번 너무 고생 많으셨다는 말씀 드리고 싶습니다. 큰 프로젝트를 통해 누구보다 가까워진 장원이와 태환이, 그리고 치성님도 늘 맛있는 음식과 함께 심심한 대학원생을 잘 챙겨주어서 너무 감사드리고, 그 외에도 자주 보면서 근황을 나누던 주현이와 종원이 형, 승원이와 준성님, 모두 사랑합니다. 또, 이러한 연구를 함께할 뿐 아니라, 랭콘이라는 훌륭한 행사를 통해 항상 사람들의 이야기를 할 기회를 주셨던 영숙님도 너무 수고 많으셨습니다. 그러한 행사를 통해 민주님과 택현님 등 소중한 인연들을 알게 된 것도 감사합니다. 지금은 다들 바빠서 각자의 길을 걷고 있지만, 언젠가 또 다시 모여 의미있는 작당을 해볼 수 있었으면 좋겠습니다.

이외에도, 함께 연구를 하거나 논문을 쓰지는 않았지만 힘들 때마다 서로의 연구 얘기, 혹은 삶 얘기를 하며 애환을 주고받았던 소중한 친구들이 많이 있습니다. 오랜 세월 관악을 함께 지키며 서로의 안부를 물어오던 R반 10학번 동기들, 특히 이번에 같이 졸업하는 명섭이, 다들 잘 자리잡아 본인이 하고싶은 일, 혹은 사회에 의미있는 일을 해나가고 있는 것 같아서 제가 다 뿌듯합니다. 또 저희 주50시간, 15년부터 몇 주에 한 번씩은 저에게 기타를 잡게 했던 대학원생 밴드는 어느새 박사와 교수로 구성된 동호회가 되었네요. 점차 손이 굳어 가던 저에게 한 줄기 오아시스같

은 음악 시간을 선사해 주셔서 감사합니다. 코로나 직전 다녀온 훈련소에서 만났던 초딩때부터 친구 정우와 유난히 잘 맞았던 옆 분대 기범이도, 분야는 다르지만 하는 연구들에 종종 모티베이션을 주어 고마웠습니다. 비록 요 몇 년은 코로나로 클래식 기타 합주를 하지 못했지만, 여러 모로 저에게 대학원 진학의 모티베이션 및 대학원 생활을 해나갈 힘을 주었던 화현회, 그리고 화현회의 모든 선배, 동기, 후배들에게도 이 자리를 빌어 감사의 말을 전합니다. 곡을 고르고, 편곡하고, 함께 연습하고, 피드백을 받아 무대에 올리는 과정은 사실 연구와 다방면으로 맞닿아 있었다고 생각합니다. 그만큼 삶을 대하는 자세에 진지함을 가르쳐 주고, 다른 사람과 소통할 수 있는 능력을 배웠던 것 같습니다. 무엇보다 우리 회장단, 동우랑 지은이, 다들 행복하고 관악 탈출도 하자. 다같이 모여서 카공을 함께 하던 동기와 후배들에게도 모두 고맙다는 이야기를 하고 싶습니다. 저녁이나 주말에 카공을 콜하면 나와서 함께해 준 상문이, 진래, 다혜, 준형이, 횡향 커플, 성훈이와 민이, 또 다른 카공메이트인 수정누나, 보영이, 슬기, 한나, 은영에게도 모두 감사의 말씀을 전합니다.

학교 안팎에서 학술적으로 만났지만 학문 그 이상을 이야기했던 소중한 사람들도 있습니다. 매번 풀잎을 이끌어 주셨던 보섭님은, 저를 항상 박사님으로 불러주셨지만 사실 누구보다 실력자임을 모두 알고 있습니다. 그리고 정말 저에게도, 커뮤니티에도 많은 도움 주시지만 항상 겸손한 성현님, 그렇게 셋이 오버워치 또 하고 싶네요. 오버워치 하면 또 스누워치와 함께했던 박사과정을 빼놓을 수 없을 것 같습니다. 대학원생 비율이 유난히 높은 동호회였지만 또 같은 전공이라고 자주 만났던 달별님, 지크님, 뽀까님, 모두 응원 많이 해주셔서 고마웠습니다. 나중에 또 맛있는 밥과 함께 연구 사주 부탁드리겠습니다. 또, 시작은 공모전이었지만, 케라콘이라는 이름 아래 순수한 열정으로 모여 모티베이션 받고, 주기적으로 만나서 근황을 주고받는 시간들이 굉장히 소중했습니다. 특히나 항상 하는 일을 응원하고 관심가져 주신 우정님과 영진님, 그리고 슬기님께 이 자리를 빌어 다시 감사의 말씀을 드리고 싶습니다. 오프라인에서의 친구들 뿐 아니라, 코로나 시국에 온라인에서 저를 응원해 주고 격려해 주었던 분들도 모두 고맙습니다. 나이 차가 꽤 나는데도 불구하고 박사 말년차의 넋두리를 재밌게 들어주고 응원해주었던 커넥팅 친구들, 석박 4년차부터 8년차가 될 때까지 박사과정의 고락을 함께했던 수과방 사람들,

힘든 시간들 울고 웃으며 함께 보낸 코인코와 김캠방 및 BIT방 사람들 모두 고맙습니다. 특히 먼저 배운 사람으로써 조언을 아끼지 않은 우박사님과, 어려운 일들 있을 때 기탄없이 이야기해 준 빗성과 탄맘성에게 고맙습니다. 학교 떠날 즈음에 녹두 탐방시켜 주신 몇 안남은 대학동 주민 애옹성도 다시금 감사합니다. 무엇보다 언어라는 어려운 주제를 처음 다룰 때부터 이방인이었던 저를 잘 받아주고 같이 놀아주었던 언어학그룹/언어학과 사람들, 특히 매사에 사려깊고 걱정해 주는 윤이와, 각종 언어학 관련 이야기들을 재밌게 풀어주시는 공부메이트 기효님, 언제든 대화에서 언어학적인 특징을 잡아내는 링고지킴이 영주형, 통사론 수업부터 함께했지만 어느새 석사를 마치고 먼저 사회로 나가 있는 유림님, 생명과학 전공의 언어덕후 규환님과 어느새 그 옆을 지키고 있는 오랜 친구 이재현 박사님, 풍부한 언어 지식을 가진 명민한 공저자 재영이와, 어느새 비슷한 길을 걷고 있는 호진님. 만날 때마다 우리 스스로도 너디한 대화라고 생각했지만 너무 재미있는 시간들이었고 모두 다시 모여 즐길 수 있었으면 좋겠어요. 그리고 또, 학교에서 근처의 건물에서 오랜 시간 함께 고생한 고등학교 선배 영하 형. 새내기 때 정말 우연히 만나 아직까지도 인연을 이어오는 존경하는 선배 상화 누나. 먼저 관악을 떠났지만, 남아있는 후배 밥 사주고, 할 수 있다고 언제나 북돋아 주어 정말 감사했습니다. 어릴 적부터 죽마고우인 지훈이랑 승협이도 맨날 나 공부 오래 한다고 밥 사주며 타박했지만 그래도 이제 마치고 나가니까 나도 맛있는 거 살게. 마지막으로 학부생일 때 우연한 계기로 만나 언어와 컴퓨터라는 관심사를 공유하며 항상 서로에게 발전적인 존재가 된 연구메이트 지원에게도 깊은 감사를 표합니다.

그리고 어머니, 아버지, 서른이 넘도록 학생이었던 아들이 이제 비로소 세상으로 나갑니다. 항상 하고 싶은 공부만 하면서 속 많이 썩였는데, 이제 번듯한 사회인으로 효도하고 좋은 곳도 여행 많이 보내드릴게요. 글쓰기 선생님인 어머니의 아낌없는 응원과 기도, 타고난 도전가인 아버지가 주신 자유로운 사고, 어떤 시각에서든 본받을 수 있었던 두 분의 모습 덕에 지금의 제가 있는 것 같습니다. 마지막으로, 대학원이 3년만에 끝나는 것으로 알고 손주가 박사 되는 것을 오매불망 기다려 주셨던 사랑하는 할머니와, 어릴 때부터 줄곧 저를 조 박사라고 불러 주셨던, 지금은 현충원에 계신 할아버지께 이 박사학위의 영광을 전합니다.