Abstract

Intention identification and slot filling is a core issue in dialog management. However, due to the non-canonicality of the spoken language, it is difficult to extract the content automatically from the conversation-style utterances. This is much harder for languages like Korean and Japanese since the agglutination between morphemes make it difficult for the machines to parse the sentence and understand the intention. In order to suggest a guideline to this problem, inspired by the neural summarization systems introduced recently, we propose a structured annotation scheme for Korean questions/commands which is widely applicable to the field of argument extraction. For further usage, the corpus is additionally tagged with general-linguistic syntactical informations.

1 Introduction

In a semantic and pragmatic view, questions and commands differ from interrogatives and imperatives, respectively. We can easily observe the particular types of declaratives (1a,b) which explicitly require the addressee to give an answer or to take action. Also, some rhetorical questions (1c) and commands (1d) do not require a response.

(1) a. I want to know why he keeps that hidden.
   b. I think you should go now.
   c. Why should you be surprised?
   d. Imagine what it must have been like for them out there.

In identifying the intention and filling slots for conversational sentences, aforementioned characteristics make it difficult for the spoken language understanding systems to catch what the speaker intends. For these reasons, the concept of dialogue act (Stolcke et al., 2000) was introduced to categorize the sentences regarding the illocutionary act, but the categorization is detailed and does not correspond with the concept of discourse component (Portner, 2004) which is close to what is investigated for slot-filling and dialog management.

In this study, we construct a criteria on materializing arguments from non-rhetorical questions and commands, especially annotating corpus on Seoul Korean. The agglutinative property of Korean is taken into account, in the way of omitting redundant functional particles. For the sentences with overt or covert speech act (SA) layer of question/command, both extractive and abstract paraphrasing are utilized depending on the content.

2 Related Works

The literature on pointing out the important feature of the document include traditional extractive approaches (Chuang and Yang, 2000; Aliguliyev, 2009; Kågebäck et al., 2014) and abstract approaches inspired by deep learning techniques (Rush et al., 2015). In the field of sentence reduction, the trend is heading to data-driven abstractive approaches (Chopra et al., 2016) from traditional statistic-based approaches (Le Nguyen et al., 2004; Shen et al., 2007). For Korean text, a sentence paraphrasing (Park et al., 2016) and news summarization (Jeong et al., 2016) were suggested, but little was done on argument extraction.

3 Corpus annotation

In this section, the annotation scheme regarding the patterns of questions and commands is described. Note that punctuations were removed since the study investigates the transcript of spoken language.
3.1 Questions

For each question, its argument and question type label were annotated. Here, questions include not only the interrogatives but also the declaratives with predicates such as want to know or wonder. In the annotation process, rhetorical questions (Rohde, 2006) were excluded.

Question type label was tagged with three classes, namely yes/no, alternative, and wh-questions (Huddleston, 1994). Yes/no question, also known as polar question, has a possible answer set of yes/no (2a). Alternative question is the question which gives a multiple choice and requires a selection (2b). Wh-question is the type of questions regarding wh-particles, namely who, what, where, when, why, and how (2c-h).

(2a) 너 의료 보상 신청 했어요
ne uulyo pongsa sinchchay hayss-e
you medical service apply did-INT
Did you apply for medical service?

(2b) 버스로 올거야 택시로 올거야
pesu-lo ol-keya thyksi-lo ol-keya
bus-by come-INT taxi-by come-INT
Will you come by bus or taxi?

(2c) 오늘은 누구 왔니
onul-un nwkuku wass-ni
today-TOP who came-INT
Who came today?

(2d) 스톡옵션이 될 줄 아니
suthokopsyen-i mwen cwul a-ni
stock-option-NOM what is/ACC know-INT
Do you know what stock option is?

(2e) 어디 있니 로비야
etti iss-ni Robi-ya
where be-INT Robi-VOC
Where are you, Robi?

(2f) 대구 몇 시에 도착이야
taykwu myech sii ey tochak-iya
Daegu what hour-TIM arrive-INT
When do you arrive in Daegu?

(2g) 이 동네 갓자기 왜 이렇게 막히지
i tongney kapcaki way ileh-key makh-i ci
this town suddenly why this-like jam-INT
Why is this town suddenly jammed like this?

(2h) 해외 송금 어떻게 하는 거야
hayoy songkum ettehkey hanun ke-ya
aboard remittance how doing thing-INT
How can I send money abroad?

Argument extraction from the questions was done depending on the question type. For yes/no questions, the content was appended with the term ‘(으)니가지 or 여부’ ([-(in)c] or [yepwu], both meaning whether or not), to make up a nominalized term for the query (3a). For alternative questions (3b), all the items were sequentially arranged in the form of ‘(A B 중 ’-한/할 것’ (([A B cwung) - han/hal kes), what is to do - between A and B). For various types of wh-questions we tried to avoid repeating the wh-particles in the extraction and instead used the wh-related terms such as ‘사람’ ([sa-lam], person), ‘의미’ ([uy-ni], meaning), ‘위치’ ([wi-chi], place), ‘시간’ ([si-kan], time), ‘이유’ ([i-yu], reason), ‘방법’ ([pang-pep], method) to guarantee the structuredness of the extraction and the utility for further usages such as web searching (3c-h). The result below correspond with the sentences (2a-h).

(3a) 의료 보상 신청 여부
uulyo pongsa sinchchay yepwu
medical service apply presence
Whether or not applied to medical service

(3b) 버스 택시 중 타고 1 올 것
pesu thyksi cwung tha-ko ol kes
bus taxi between ride-PRG come thing
What to ride between bus and taxi

(3c) 오늘은 사람
onul on salam
today came person
The person who came today

(3d) 스톡옵션 의미
suthokopsyen uymi
stock-option meaning
The meaning of stock option

(3e) 지금 있는 위치
ckum iss-nun wichi
now be-PRG place
The place currently belong to

(3f) 대구 도착 시간
taykwu tochak sikan
Daegu arrival time
Arrival time for Daegu

(3g) 막히는 이유
makh-i nun iyu
jam-PRG reason
The reason for jam

(3h) 해외 송금 방법
hayoy songkum pangpep
aboard remittance method
The way to send money abroad

1 타-/ride is usually accompanied with the transportation.
3.2 Commands

For each command, its argument and the positivity label were annotated. Here, commands include not only the imperative forms with covert subject and the requests in the interrogative form (different from the categorization in Portner (2004)), but also the wishes and exhortatives that induce the addressee’s response. Imperatives used as exclamation or evocation are not included since they are considered rhetorical. The optatives that are used idiomatically, such as *Have a nice day!* (Han, 2000), are also not included since the feasibility of the to-do-lists is beyond the addressee’s capacity.

**Positivity** label was tagged with three classes, namely prohibitions, requirements, and strong requirements. **Prohibition (PH)** is the type of command that stops or prohibits an action. It possibly contains negations (4a1) or the predicates/modifiers that induce the prohibition (4a2).

**Requirement (REQ)** is the type of command that are positive, with no terms that induce the restriction (4b1,2), and corresponds with various sentence forms aforementioned. **Strong requirement (SR)** is the type of command where the prohibition and requirement are concatenated sequentially, appearing in spoken Korean as an emphasis (4c), due to its head-final property.

(4a1) 태풍 오니가 밖에 나가지 마
thayphwung o-nikka **pakk-ey naka-ci ma**
typhoon come-because **outside-to go-ci NEG**
Don’t go outside, typhoon comes.

(4a2) 안전히 안 해 볼까요
ancentti **an-may-myen khunil-na**
seatbelt no-take-if danger-occur.DEC
It’s dangerous if you don’t take a seatbelt.

(4b1) 간접사항 확인 바랄니다
inceksahang hwakin palap-nita
personal-info check want-HON.DEC
I want you to check the personal info.

(4b2) 이번 주 일정을 모두 말해
ipen cwu **ilceng-ul motwu** mal-hay
this week schedule-ACC all tell-IMP
Tell me all the schedules this week.

(4c) 욕심부리지 말고 지금 팔아
yoksim-pwuli-ci malko **cikum phal-a**
greedy-be-ci not-and **now sell-IMP**
Don’t be greedy, just sell it now!

**Argument extraction from the commands** was done depending on the positivity. For PH, the action that is prohibited is annotated (5a1). For REQ, the requirement is annotated (5b1). For SR we only annotated the action that is required (5c), for a disambiguation and an effective representation of a to-do-list. Most of the arguments ended with a nominalized predicate ‘-(하)기’ (**-(ha)ki**), *doing/to do something*, for consistency and a flexible application. (5a1-c) correspond with (4a1-c).

| Types          | Correspondings       |
|----------------|----------------------|
| **Yes/no**     | whether or not       |
| **Alternative**| what is/to do between |

| Questions                  | Types  | Correspondings       |
|---------------------------|--------|----------------------|
| Who                       | Person  | identity             |
| What                      | Meaning | 이리                  |
| Where                     | Location, place | 위치, 장소            |
| When                      | Time, period, hour | 시간, 기간, 시각      |
| Why                       | Reason  | 이유                  |

| Commands                  | Types  | Correspondings       |
|---------------------------|--------|----------------------|
| **Prohibitions**          | PH     | -(기) (금지)          |
| **Requirements**          | REQ    | -(기) (요구)          |
| **Strong Requirements**   | SR     | -(기) (요구)          |

Table 1: Structured annotation scheme.
Table 2: Dataset specification, denoted with the number of instances for each category and the portion. For strong requirements, NEG imply the phrase or word regarding a negation. PH-REQ indicates the sentences like (4c), and REQ-PH implies the scrambled order. NEG-PH implies a double negation as in (4a2).

There are points to be clarified regarding (4a2) and (5a2). Although (4a2) displayed a property of PH induced by ‘큰일이나’, the target action contained a negation ‘안’ that a double negation occurred. Therefore, (5a2) was labeled as SR.

Since the commands did not accompany abstract concept as wh-questions did, the argument was obtained mostly in an extractive way. Also, since the command inevitably includes a detailed to-do-list, the removal of functional particles was done only if they were considered redundant, unlike it was highly recommended for the questions. However, there are some exceptions with the information-seeking commands (4b2) including the terms show, inform, tell, find, check, etc.; despite the clear to-do-lists they show, the intent is close to acquiring information. Thus, the argument extraction for those commands followed the scheme regarding the questions (5b2) as described in Section 3.1, avoiding the nominalizer ‘-(으) 큰일이나’.

4 Dataset Specification

We adopted the spoken Korean dataset of size 800K which was primarily constructed for language modeling and speech recognition of Korean. The sentences are in conversation-style and partly non-canonical, and the content covers the topics such as weather, news, housework, e-mail, and stock. From the corpus we randomly selected 20K sentences and classified them into seven sentence types: fragments, rhetorical questions, rhetorical commands, questions, commands, and statements, with κ = 0.85 (Fleiss, 1971).

Argument extraction was done for the questions and commands which are not rhetorical. The specification of the annotated corpus is displayed in Table 2.

Due to the characteristics of the adopted corpus as a spoken language script targeting smart home agents, the portion of the commands is higher than in the real-life language. We could observe that the alternative questions, PH, and SR (especially the scrambled order and double negation) are relatively scarce, whereas yes/no-wh-questions and REQ dominate in number.

5 Conclusion

In this paper, we proposed a structured annotation scheme for the argument extraction of conversation-style Korean questions and commands, concerning the discourse component and the properties they show. This is the first dataset on question set/to-do-list extraction for spoken Korean, up to our knowledge, and we annotated the syntax-related properties for the potential usage. For interrogatives and imperatives extended to semantic/pragmatic level, this study may provide an appropriate guideline that helps argument extraction of various conversations in real life.

There’s no doubt that the primary application of the dataset is a slot-filling for Korean questions and commands. Although the volume is small, the dataset shows consistency regarding the way it was constructed. In case of need, the utterance-argument pairs can be uncomplicatedly created referring to the examples and flexibly augmented to the original dataset. Also, in the aspect of linguistic characteristics, the annotation scheme can be extended to the languages that is syntactically similar to Korean, such as Japanese. Most importantly, the scheme fits with the spoken language analysis flourishing with the smart agents widely used nowadays. We expect the proposed scheme and dataset can help machines understand the intention of natural language, especially conversation-style directives.

References

Ramiz M Aliguliyev. 2009. A new sentence similarity measure and sentence based extractive technique for
automatic text summarization. *Expert Systems with Applications*, 36(4):7764–7772.

Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98.

Wesley T Chuang and Jihoon Yang. 2000. Extracting sentence segments for text summarization: a machine learning approach. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 152–159. ACM.

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

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

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

Hyoungil Jeong, Youngjoong Ko, and Jungyun Seo. 2016. Efficient keyword extraction and text summarization for reading articles on smart phone. *Computing and Informatics*, 34(4):779–794.

Mikael Kågebäck, Olof Mogren, Nina Tahmasebi, and Devdatt Dubhashi. 2014. Extractive summarization using continuous vector space models. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositional Properties* (CVSC), pages 31–39.

Minh Le Nguyen, Akira Shimazu, Susumu Horiguchi, Bao Tu Ho, and Masaru Fukushi. 2004. Probabilistic sentence reduction using support vector machines. In *Proceedings of the 20th international conference on Computational Linguistics*, page 743. Association for Computational Linguistics.

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

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

Hannah Rohde. 2006. Rhetorical questions as redundant interrogatives.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.