Domain Transferable Semantic Frames for Expert Interview Dialogues

Taishi Chika[◇], **Taro Okahisa**[*], **Takashi Kodama**[†]
Yin Jou Huang[†], **Yugo Murawaki**[†], **Sadao Kurohashi**[†♣]
[◇]Kansai University [*]Shizuoka University [†]Kyoto University [♣]National Institute of Informatics
nanou7614@gmail.com, okahisa-taro@inf.shizuoka.ac.jp,
{kodama, huang, murawaki, kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

Interviews are an effective method to elicit critical skills to perform particular actions in various domains. In order to understand the knowledge structure of these domain-specific actions, we consider semantic role and predicate annotation based on Frame Semantics. We introduce a dataset of interview dialogues with experts in the culinary and gardening domains, each annotated with semantic frames. This dataset consists of (1) 308 interview dialogues related to the culinary domain, originally assembled by Okahisa et al. (2022), and (2) 100 interview dialogues associated with the gardening domain, which we newly acquired. The labeling specifications take into account the domain-transferability by adopting domain-agnostic labels for frame elements. In addition, we conducted domain transfer experiments from the culinary domain to the gardening domain to examine the domain transferability with our dataset. The experimental results showed the effectiveness of our domain-agnostic labeling scheme.

Keywords: Semantic frame, Domain Transfer, Interview, Dialogue

1. Introduction

In an expert interview, an interviewer interacts with a domain expert (e.g., skilled mold polisher) to elicit the expert's knowledge of the field. The conversation with the interviewer serves as an important stimulus for the expert to explore aspects they may not have previously considered. Therefore, an interview is a valuable tool, not only for eliciting explicit knowledge, but also for unearthing the implicit or tacit knowledge unconsciously possessed by the domain experts. The elicited domain knowledge can help the transmission and preservation of skills from domain experts to apprentices, thereby offering substantial benefits to numerous industrial fields. Nonetheless, the conversational nature of interviews often results in extended and less succinct forms of knowledge representation. Therefore, it is beneficial to condense the domain knowledge elicited in the interview dialogues to facilitate further use.

In this work, we utilize semantic frame analysis as a method for structuring domain-specific knowledge in expert interviews (Figure 1). Semantic frame analysis is a predicate-centered approach. First, the frame-evoking predicates that signal the occurrence of a specific event are identified, such as "*rest*" and "*fermenting*" in the example. Further, the frame elements of the event are also identified. Frame elements are the participants and attributes of the event that provide a detailed understanding of the event, such as "*tart dough*", "*refrigerator*", "*open-air*", and "*lightly fluffy in texture*".

Semantic frame analysis for expert interviews

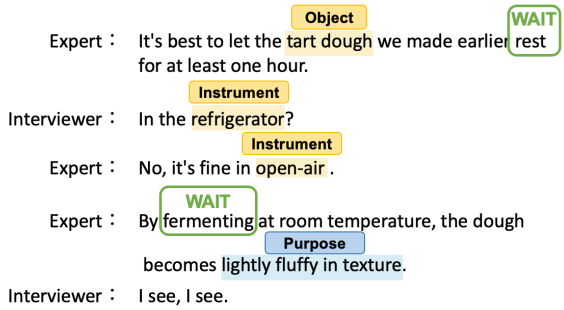


Figure 1: Semantic frame in dialogues. Triggers are surrounded by green boxes, the arguments and specifiers are highlighted in yellow and blue, respectively.

presents two distinct challenges. The first challenge is about the domain-specificity of expert interviews. Traditional semantic frames are tailored to the unique requirements of each specific domain, and adapting the semantic frame structure from one domain to another is frequently a challenging task (Matsubayashi et al., 2009; Ruppenhofer et al., 2010). This means that the semantic frame annotation data collected from one domain is often useless for another domain. In highly specialized domains such as mold polishing, obtaining a substantial number of interviews can be infeasible, making a data-driven approach to semantic frame analysis challenging.

The second challenge is tackling the characteristics of dialogues, which makes the analysis more difficult when compared to monologue data. In a dialogue, speakers frequently revisit the same

LREC-COLING 2024, pages 5299–5308
20-25 May, 2024.

events, repeating and paraphrasing them while incorporating additional details about the event (Shibata et al., 2014). For example, the two triggers "*rest*" and "*fermenting*" in Figure 1 represent the same event. We can also see that the information about an event often scatters across multiple utterances of multiple speakers, adding to the difficulty of the task.

In this work, we propose a semantic frame labeling scheme for expert interview dialogues. In this labeling scheme, owing to the impact of the second challenge, frame elements frequently extend across multiple utterances, and various mentions of the same event are connected through event coreference links. To address the first challenge, we facilitate domain transfer by introducing domain-agnostic labels for frame elements. To explore domain transferability, we conducted annotations on Japanese expert interviews from two distinct domains: (1) interview dialogues related to the culinary domain, originally assembled by Okahisa et al. (2022), and (2) interview dialogues associated with the gardening domain, which we newly acquired. While we have our ultimate goal set on industrial applications, we explored these two domains due to the necessity of acquiring a substantial amount of data.

Utilizing these annotations, we performed domain transfer experiments from the culinary domain to the gardening domain. The experimental results show that the annotation collected in the culinary domain helped the semantic frame identification in the gardening domain, showing the feasibility of our domain-transferable semantic frame labeling scheme.

In conclusion, our contributions are three-fold:

- We proposed a labeling scheme for domain-transferable semantic frames that incorporates domain-agnostic frame elements.
- We constructed a multi-domain expert interview dataset based on an existing interview dataset of the culinary domain.
- We verified the feasibility of our proposed annotation scheme by performing domain transfer experiments with the collected data.

The dataset is made available under the Creative Commons Attribution 4.0 International License and can be accessed via `https://nlp.ist.i.kyoto-u.ac.jp/?EIDC`.

2. Related Work

Dialogue datasets are commonly used for developing task-oriented dialogue systems, such as medical diagnosis (Zeng et al., 2020) or travel planning (McLeod et al., 2019). In this field of research, the standard practice involves annotating texts with named entities or discourse acts. For our purpose, however, we posit the necessity of capturing intricate, structurally indeterminate information, and we regard semantic frames as a suitable approach.

Our annotation of knowledge structures is part of a larger research program aimed at extracting the art of relevant skills mentioned by experts in the form of dialogue. For this purpose, it is necessary to know what processes each utterance represents in a particular domain and what elements (objects of action, instruments) are involved in those actions. While we do not align with a specific linguistic theory for the annotation of expert interviews, our approach to semantic role labeling draws significant influence from Frame Semantics (Fillmore, 1986) and FrameNet (Ruppenhofer et al., 2016).

Frame semantics emphasizes our subjective understanding of what kind of situation a word invokes and what role the lexical item plays, instead of taking the objective approach of setting necessary and sufficient conditions for the meaning of words based on primitive semantic features. For instance, one cannot understand the meaning of the verb "*sell*" without conceiving the entire situation involving selling, such as the commercial transaction scenario, and the participants (Frame Elements; FEs), such as the Seller, Buyer, and Goods.

As for a dialogue dataset based on Frame Semantics, Skachkova and Kruijff-Korbayova (2021) annotated frames and FEs to extract the knowledge of team communication in disaster response.

3. Domain-Transferable Semantic Frame Annotation

3.1. Semantic Frame Structure

In an expert interview dialogue, the interviewer and the expert typically talk about how to fulfill a domain-specific task, such as making a dish or doing pest control. In the process, they talk about the critical actions and events towards the completion of the task, and discuss the details (methodologies, tools, reasons, etc.). In this work, we aim to use semantic frames to capture the knowledge structure of the events depicted in the dialogues.

Figure 2 shows an example of this annotation. Given the transcription of an interview dialogue, we use four types of annotation labels — **trigger**, **argument**, **specifier**, and **attribute** — to represent the knowledge structure of an event of interest. In our annotation, only the trigger labels are domain-specific, while other labels are domain-agnostic. The details for the labels used in our an-

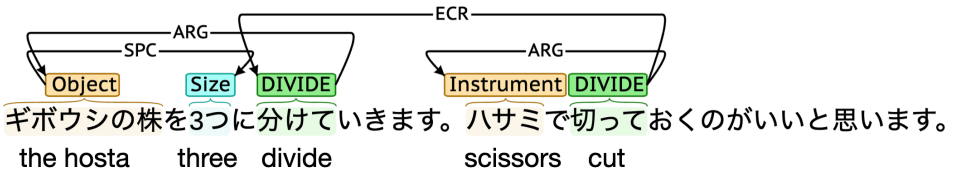


Figure 2: An example of the gardening domain annotation (*Divide the hosta into three pieces. You'd better cut with scissors.*). ARG stands for argument relation between the trigger (enclosed in green) and the argument (colored yellow), SPC for specifier relation between the specifier (colored blue) and other elements, and ECR forms the coreferent relation between two triggers, respectively.

notation are described below:

Trigger A trigger is a frame-evoking predicate that best signals the occurrences of the event. Each trigger is labeled with the frame type of the event. In this work, we use domain-specific labels for triggers. These labels are designed manually by considering the frequent-occurring predicates in the domain:

- 11 frame types for the culinary domain: BAKE-FRY, DIVIDE, CHANGE, SIMMER, HEAT, MIX, PUT-ON, PLACE, WAIT, COMPOUND, and REMOVE.
- 12 frame types for the gardening domain: MIX, CHANGE, PLACE, SUPPLY, SOW, REMOVE, DIVIDE, TRANSPLANT, COVER, ELIMINATE, ARRANGE, and HARVEST.

Note that we cannot rely solely on the surface form of the trigger to decide its frame types. We also need to consider the context of the trigger appropriately to disambiguate the meaning of the trigger. For example, both "種を撒く" (*plant seeds*) and "水を撒く" (*supply water*) are triggered by the predicate "撒く", but represent events of different frame types.

Argument Arguments of a semantic frame are the participants of an event, which play an essential role in representing the conceptual structure of events. Also, each argument is labeled with its argument role with respect to the event. In many works, a distinct set of argument role labels is crafted for each frame type. These frame-dependent argument roles often result in domain-specific argument roles, such as *Seller*, *Cook*, *Plants*. In contrast, our annotation does not utilize frame-dependent argument roles, but covers a more generic set of argument roles that is domain-independent. We define the following 5 types of argument role labels:

- **Object**: An object is the core participant that receives an action and undergoes the effects of an event.
  - 土 に肥料 を混ぜていきます (*mix the compost with the soil*)
- **Instrument**: An instrument is a tool or instrument used to carry out an action.
  - 藁 で土壌を覆っていきます (*use the straw to mulch the soil*)
- **Temperature**: The temperature specified for a certain event (only used in the culinary domain).
  - 弱火 でにんにくを炒めていきます (*fry the garlic over low heat*)
- **Time**: The temporal aspects of the event, giving information about the timing or duration of the action.
  - 1 時間 生地を寝かせます (*leave the dough to rise for one hour*)
- **Manner**: Manner arguments of an event describe how an action is performed, including frequency, method, etc.
  - 朝晩に 1 回 水やりする (*water once in the morning and evening*)
  - バジルの苗を直 植えします (*place the basil directly into the pot*)

Specifier A specifier modifies a trigger, argument, or another specifier. Just like arguments, each specifier is given a domain-agnostic label. In this annotation, we set 5 types of specifiers:

- **Size**: The size specifier provides information about the size of an object.
  - 株を3 センチ に切ります (*cut the plant into 3cm chunks*)
- **State**: The state specifier indicates the condition of an entity caused by a certain event.
  - 霧吹きで苔に水をやれば湿ります (*spray the moss and it'll become moist*)
- **Amount**: The amount specifier gives information about the quantity of an entity.
  - 砂糖は大さじ 2 か 3 くらい 入れる (*add 2 or 3 tablespoons of sugar*)

- **Purpose**: The purpose specifier provides the reason or intended function of an action.
  - 根を傷つけないよう ハサミで切ります (*to avoid damaging the roots, cut with scissors*)
- **Condition**: The condition specifier describes the circumstance or prerequisite of an event.
  - 殺虫剤を持っていない場合 は 酢を使って虫を退治してください (*in case you don't have an insecticide, use a vinegar to repel bugs*)

Attribute For events and entities with specific attributes, we apply additional labels:

- **Prohibition**: Prohibition is an attribute indicating that the speaker refrains from performing the action or applying the entity.
  - 豚肉を焼いていきます。ロース よりもバラ肉がいいですね。 (*Grill the pork. I prefer rib than loin.*)
- **Generalization**: The attribute applied when an event is a general event (e.g. general tips) but not related to the task in question.
  - 時々どのタイミングで株を間引いたらいいのか分からなくなります。 (*I sometimes don't know when to separate the roots.*)

3.2. Relations between Frame Entities

We also annotate the relations between various entities annotated in Section 3.1. We mainly consider three types of relations: **Event Coreference Relation (ECR)**, **Event Narrative Relation**, and **Counterpart (CP) Relations**. Among them, the ECR and narrative relations express event-to-event dependencies, indicating whether multiple events constitute a whole single process or not. On the other hand, the CP relations can occur between any two entities of the same label type.

Event Coreference relation (ECR) When two or more semantic frames refer to the same real-world event, we say that they are coreferent. For example, The events represented by "*rest*" and "*fermenting*" have an ECR relation.

Event Narrative relation When an event's **Object** argument is the outcome of another antecedent event, we say that there exists a narrative relation between them. To reduce annotation costs, we apply multiple tagging on the trigger entity of the antecedent, regarding it as the **Object** argument of the subsequent event.

For instance, the text shown in Figure 3 contains two events, triggered by predicates "*turn*" and "*mix*", respectively. The first event ("*turn*") takes "the soil" as its object, represented by an *ARG* link between the trigger and the argument. The second event ("*mix*") refers to the mixing of the fertilizer and the soil that has been tuned over in the first event. Therefore, the second trigger forms an argument link to the trigger of the first trigger, indicating the end product of the first event being one of the objects of the second event.

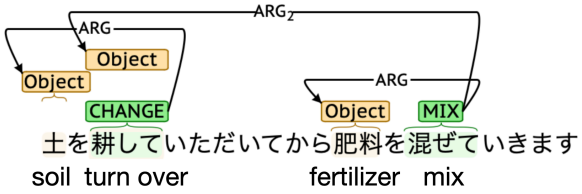


Figure 3: An example of multiple tagging for trigger (*Turn over the soil, then mix the fertilizer with it*).

Counterpart (CP) relation When mentioning a process that might replace the original one, CP is used to relate the original element to the alternative:

- 殺虫剤 で虫を退治します。ない場合は酢 を使ってください。 (*Repel bugs with an insecticide. In case you don't have it, use a vinegar.*)
- 卵液をオーブンで焼きます。オムレツを焼くようにフライパンで焼いて もいいです。 (*Bake the egg mixture in the oven. You can also cook it in a frying pan like an omelet.*)

4. Dataset Construction

4.1. Dialogue Data Collection

We collected expert interview dialogues of different domains to facilitate further experiments and analysis on domain transferability. We focused on the culinary and gardening domains. Both domains contain unique domain knowledge that takes years of practice to acquire and accumulate. Moreover, the widespread popularity of both culinary and gardening interests allowed us to amass a considerable amount of interview data.

For the culinary interview data, we used the dataset collected by Okahisa et al. (2022), which contained 308 Japanese dialogues. Following the same procedures introduced in Okahisa et al. (2022), we collected 100 Japanese dialogues between gardening experts and interviewers.

All interview dialogues were collected using Zoom, a video conferencing tool (Figure 4). In each dialogue, an interviewer and a domain expert discussed a domain-specific topic, such as

	Culinary			Gardening		
	Experts	Interviewers	Total	Experts	Interviewers	Total
# of dialogues	-	-	308	-	-	100
Avg. duration of dialogue video (min.)	-	-	12.6	-	-	14.4
Avg. # of utterances per dialogue	115.2	93.1	208.3	84.9	78.5	163.4
Avg. # of characters per utterance	24.5	28.0	20.4	34.8	44.4	24.4

Table 1: Statistics of the datasets.

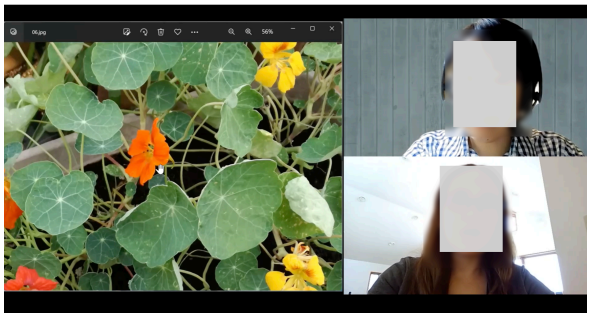


Figure 4: A screenshot of a gardening expert interview via Zoom.

how to make a specific dish or how to grow a specific type of plant. In the process, the expert was instructed to use static images to help the illustration of ideas. Each dialogue was transcribed into textual form. For the details of preprocessing (e.g., excluding backchannel and parenthesizing filler), refer to Okahisa et al. (2022).

4.2. Semantic Frame Annotation

We used the web-based tool Brat[1] as the platform for collecting semantic frame annotations. The Brat interface enables the annotators to select any span of text as a tag (event or entity) and assign corresponding labels and attributes to it. Additionally, annotation of relations between tags is also possible. In Brat, triggers were considered events, whereas arguments and specifiers were regarded as entities. Brat's relation links were used to annotate the ECR relations, CP relations, and the connections between an entity and the tag it modifies.

The annotators reviewed the transcription of the interview dialogue and labeled the semantic frame entities present in the text. Moreover, they had access to the video and audio data from the interview in case the textual data alone proved insufficient for resolving any ambiguities. All annotators were native speakers of Japanese. They were lectured on the concept of semantic frames and label specifications prior to the annotation process. First, they were instructed to look for triggers that indicate semantic frames. Next, they were asked to find arguments corresponding to the trigger. Finally, they were asked to find elements that specify the details of the triggers and arguments. After these processes, they identified relations between tagged elements based on the annotation scheme in Section 3.2.

[1] https://brat.nlplab.org/

4.3. Statistics

Interview Dialogues Statistics The statistics of the collected dialogues are summarized in Table 1. The dataset consisted of 308 interview dialogues collected by Okahisa et al. (2022) and 100 gardening dialogues we newly collected. The average duration of an interview was 12.6 minutes for the culinary domain and 14.4 minutes for the gardening domain. Compared to culinary interviews, gardening interviews had longer utterances in terms of average character count. We speculate that this difference primarily arises from variances in the domain and the speakers' individual characteristics. Speaker-wise, experts typically delivered more utterances than interviewers in both domains. Furthermore, expert utterances tended to be longer than interviewer utterances.

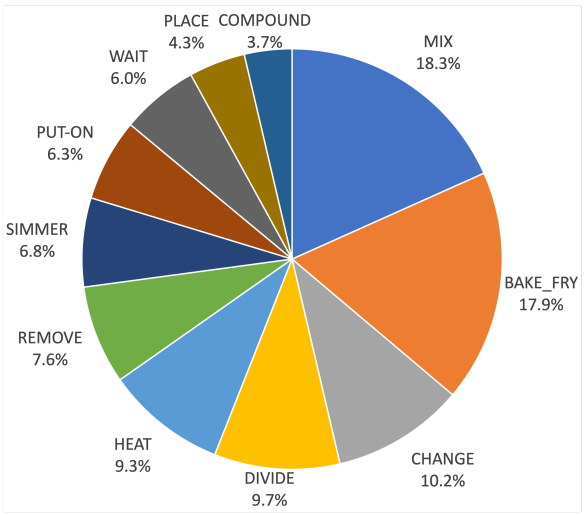
Label Distributions The label distributions of triggers (frame types) are summarized in Figure 5. We obtained $21,467$ frame type labels for the culinary domain and $4,227$ labels for the gardening domain.

Figure 6 shows the distributions of argument and specifier labels. It is evident that the most prevalent category was the **Object** arguments, followed by the **Manner** arguments. There were disparities in both argument label and specifier label distributions, reflecting the domain differences.

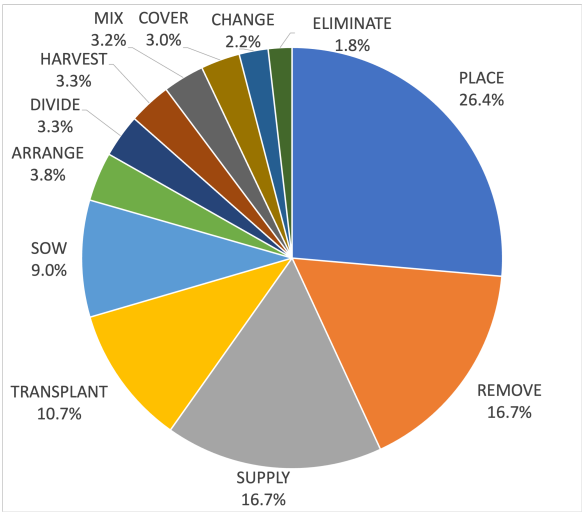
4.4. Inter-Annotator Agreement

We measure the inter-annotator agreement of the semantic frame annotation task. Specifically, we randomly selected 5 dialogues from the culinary domain and the gardening domain, respectively. Each dialogue was annotated by two different annotators.

Following Kulick et al. (2014), we first perform entity mapping across the two sets of annotations.

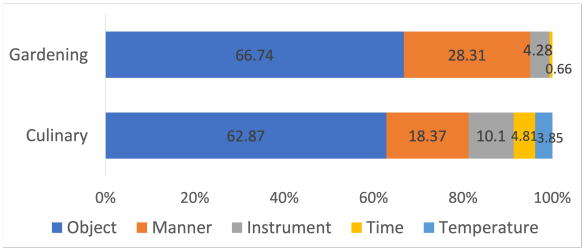


(a) Culinary frame types.

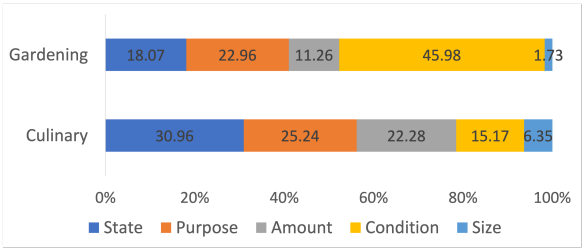


(b) Gardening frame types.

Figure 5: The distributions of frame types in different domains.



(a) Percentage of each argument type.



(b) Percentage of each specifier type.

Figure 6: The distributions of frame elements.

We use the Hungarian algorithm to find the optimal one-to-one mapping between entities in different annotations, with a focus on prioritizing mapping between entities pairs with larger span overlap. We calculate the mapping rate as the proportion of successfully mapped entities among all annotated entities. In addition, we measure the the agreement between annotators by computing Cohen's Kappa coefficient for the mapped entities.

Table 2 shows the inter-annotator agreement measures, including the mapping rate and Cohen's Kappa coefficient for trigger, argument, and specifier annotations. For both domains, the mapping rate decreases in the order of trigger, argument, and specifier annotations. This reflects the increasing complexity of the three tasks. With the exception of the trigger annotation task in the gardening domain, the mapping rates for the other annotation tasks are below 70%, showing the challenging nature of the annotation tasks. On the other hand, Cohen's Kappa coefficients all surpass 80%, indicating a high level of agreement among annotators.

4.5. Qualitative Analysis of the Dataset

In this section, we briefly present a qualitative analysis of corpus data from the following perspectives: distribution of entities and linguistic structure.

4.5.1. Distribution of Entities

As shown in Figure 6, the two domains exhibit contrasting distributions for certain arguments and specifiers. For instance, the gardening domain showed a relatively high frequency of **Manner**, while the frequency of **Time** was drastically lower when compared to the culinary domain. A possible explanation for the greater frequency of **Manner** in the gardening domain relates to the prevalence of references to methods, especially those concerning planting and caring practices:

- 種を3cm 間隔 で撒いてください
  (*Sow seeds 3cm apart*)
- 朝と晩に 1 回ずつ 水をあげてください
  (*Water once each morning and night*)

Conversely, **Time** was rarely mentioned in the gardening domain. In contrast to the culinary domain, the duration of a process in gardening may

	Mapping Rate (%)			Cohen's Kappa (%)		
	Trigger	Argument	Specifier	Trigger	Argument	Specifier
Cooking	67.8	56.8	54.2	84.0	94.1	88.6
Gardening	84.1	65.3	50.4	89.9	90.7	95.0

Table 2: Mapping rate and Cohen's Kappa coefficient between two annotators.

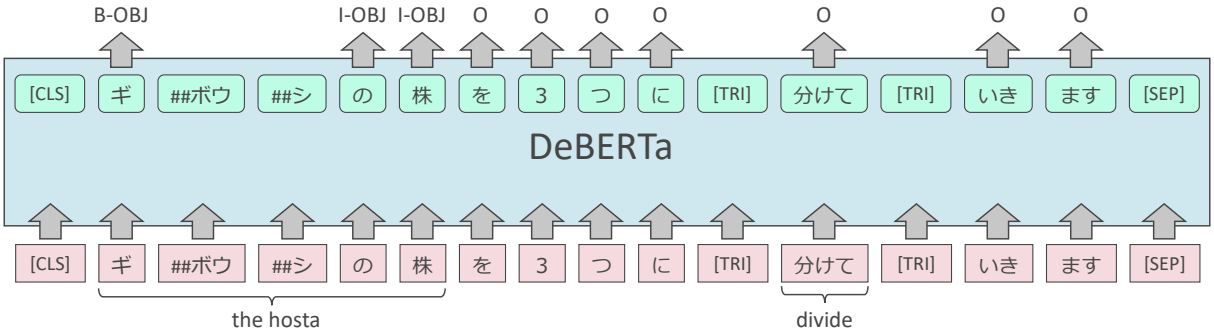


Figure 7: DeBERTa-based argument detector. The first sentence in Figure 2 (*Divide the hosta into three pieces.*) is the subject of analysis. With the gold trigger provided (i.e., *divide*), the detector assigns BIO labels to the tokens. The specifier detector is implemented in a similar manner.

not be a crucial factor for gardening-related skills.

4.5.2. Linguistic Structure

When a series of actions consists of multiple frame types, narrative relations introduced in Section 3.2 were frequently employed. While this relation label was frequently established between clauses within a single utterance, as illustrated in Figure 3, several instances connected multiple utterances. In the example in Table 3, the predicate "スライス" (*slice*) in the first utterance functions as the object of "除去して" (*remove*) in the third utterance, since "除去して" requires two arguments: **what** to remove ("種" (*seeds*) in the second utterance) and from **where**.

Specifiers, which modify other elements (including other specifiers), were frequently conveyed through not only nouns but also noun phrases or clauses. The assignment of a single tag to a long span may impact domain transferability:

- **Purpose**: 凍結防止用 にカバーを覆ってあげます
  (*cover them up for freeze protection*)
- **State**: まずは玉ねぎを炒めます。玉ねぎが透明になってきたら...
  (*First, fry the onion. When it becomes translucent...*)

5. Domain Transfer Experiments

In this section, we conducted domain transfer experiments to verify the efficacy of the proposed annotation scheme.

5.1. Frame Element Detectors

We fine-tuned a DeBERTa$_{LARGE}$ (He et al., 2021) model[2] as frame element detectors. Separate detectors were developed for arguments and specifiers. These detectors approached the semantic frame analysis task as a sequence labeling task, assigning a B (Begin), I (Inside), or O (Outside) label to each token. Figure 7 shows an argument detector. We inserted the special token `[TRI]` before and after the target trigger. In this study, we employed gold triggers for analyzing arguments and specifiers, leaving the use of results from a trigger detector for future research.

5.2. Domain Transfer Settings

We trained and tested the models with the following five conditions.

CU_{CU} Trained and tested on the culinary domain.

CU_{GA} Trained on the culinary domain and tested on the gardening domain.

GA_{GA} Trained and tested on the gardening domain.

$CU + GA_{GA}$ Trained on both the culinary and gardening domains and tested on the gardening domain.

$CU \rightarrow GA_{GA}$ Trained initially on the culinary domain and subsequently on the gardening domain, and tested on the gardening domain.

[2] https://huggingface.co/ku-nlp/deberta-v2-large-japanese

Expert	りんごを 2、3 ミリぐらいの薄さにスライス してもらいます。 *Slice the apples into thin slices of about 2 or 3 mm.*
Interviewer	種 とかはあらかじめで取っておきますか？ *Should I pull out the seeds in advance?*
Expert	いえ，その時に除去して いただければ問題ないです。 *You can remove these after slicing.*

Table 3: Narrative relation across utterances. The predicate "除去して" (*remove*) in the third utterance by the expert requires two arguments: what to remove and from where (both are tagged with **Object**).

	Object	Instrument	Temperature	Time	Manner	micro	weighted
CU_{CU}	58.9 / 61.4	54.3 / 58.8	45.6 / 53.5	56.4 / 62.6	43.8 / 52.1	54.7 / 58.9	54.6 / 58.9
CU_{GA}	35.8 / 38.7	18.0 / 17.5	- / -	36.1 / 31.7	26.5 / 34.7	32.1 / 36.0	32.6 / 36.9
GA_{GA}	49.6 / 54.1	**49.7** / **47.6**	- / -	**75.5** / **75.7**	37.9 / 48.6	46.4 / 52.4	46.3 / 52.4
$CU + GA_{GA}$	51.6 / 54.7	39.5 / 40.1	- / -	74.3 / 74.3	**40.6** / 49.4	48.1 / 52.7	**48.2** / 52.8
$CU \rightarrow GA_{GA}$	**52.1** / **55.1**	41.4 / 42.2	- / -	70.5 / 73.3	39.4 / **49.5**	**48.3** / **53.1**	**48.2** / **53.2**

Table 4: Results of argument detection. The strict/loose F1 scores are displayed on each cell. Scores are the mean of five runs of the experiment with different random seeds. The bold scores indicate the highest ones over models for the gardening domain.

5.3. Evaluation Metrics

We used two evaluation metrics: the strict F1 score[3] and the loose F1 score. The former is judged as correct only when both the span and label of the frame elements completely match the gold span and label, while the latter awards a partial score when detected tokens partially match the gold-standard tokens.

5.4. Implementation Details

The detectors were provided with the sentence containing the target trigger, along with the five preceding and succeeding sentences. Sentences were separated by the `[SEP]` token. The models were trained for 50 epochs with a batch size of 64. We chose the snapshot with the highest loose weighted F1 score for the validation split. We used AdamW optimizer (Loshchilov and Hutter, 2019) with β_1 = 0.9, β_2 = 0.999, and ϵ = $1e-6$. The initial learning rate was set to $1e-4$ when training on the culinary domain and to $5e-5$ when training on the gardening domain. We also implemented a cosine learning rate scheduler and allocated the first 500 steps in the cooking domain and the first 100 steps in the gardening domain for warmup.

In preliminary experiments for specifiers, we observed a predominance of O labels, with only a minority of instances containing B (or I) labels. In the culinary domain, for example, instances consisting solely of O labels for specifiers accounted for 92.2%, while the corresponding figure for arguments was 14.9%. To alleviate this imbalanced distribution problem, we used the downsampling method, which reduced the O-only samples. Specifically, we adjusted the ratio of samples containing only O to those that did not to 80%:20% for the training data. We did not use this method for the validation and test splits.

[3] https://github.com/chakki-works/seqeval

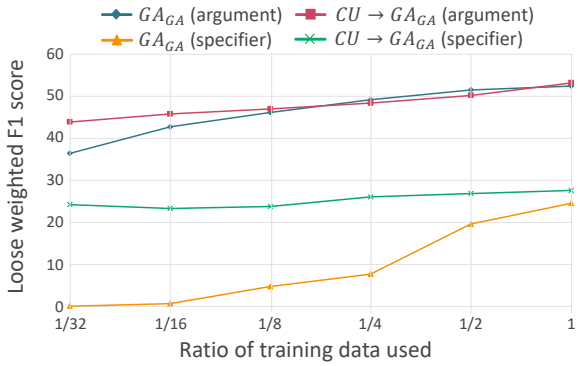


Figure 8: Learning curves for argument and specifier detection.

5.5. Results

Table 4 shows the results of the argument detection. CU_{CU} and GA_{GA} were in-domain settings, with CU_{CU} having a relatively large training dataset and GA_{GA} having a smaller one. CU_{GA} yielded a lower, but still reasonably high, loose weighted F1 score of 36.9, indicating some degree of success in the domain transferable labeling scheme. Also, the combination of the two domains, $CU + GA_{GA}$ and $CU \rightarrow GA_{GA}$, showed slight improvements over GA_{GA} in micro and weighted F1 scores.

	Size	State	Amount	Purpose	Condition	micro	weighted
CU_{CU}	36.9 / 43.1	27.5 / 36.5	48.2 / 56.7	21.0 / 39.2	20.1 / 30.1	31.9 / 42.1	31.5 / 41.6
CU_{GA}	0.0 / 0.0	0.0 / 1.1	35.1 / 40.8	6.4 / 13.7	**17.3** / **24.4**	17.4 / 22.0	18.0 / 23.2
GA_{GA}	0.0 / 0.0	**7.3** / **18.3**	33.0 / 40.9	5.0 / 8.6	4.2 / 13.7	14.7 / 22.2	16.5 / 24.5
$CU + GA_{GA}$	0.0 / 0.0	4.0 / 13.2	**48.5** / **52.3**	**7.8** / **15.1**	11.4 / 21.3	**22.2** / **28.2**	**23.6** / **30.4**
$CU \rightarrow GA_{GA}$	0.0 / 0.0	6.0 / 12.2	45.4 / 47.9	6.5 / 12.4	12.6 / 19.7	20.9 / 25.1	22.8 / 27.6

Table 5: Results of specifier detection. For the notation, please refer to Table 4.

Table 5 shows the results of the specifier detection. Notably, CU_{GA} slightly outperformed GA_{GA} in terms of strict micro/weighted F1 scores, indicating that larger training data outweighed domain differences. Furthermore, $CU + GA_{GA}$ and $CU \rightarrow GA_{GA}$ showed significant improvements, demonstrating the effect of combining different domains.

Lastly, we reduced the amount of training data from the target domain to assess domain transferability in low-resource scenarios, particularly relevant to industrial applications. The results are shown in Figure 8. $CU \rightarrow GA_{GA}$ highlighted the effectiveness of domain transferability for both argument and specifier detection. Even with three dialogues, about 1/32 of the 100 dialogues we have collected, the detectors achieved sufficient performance for argument and specifier detection.

6. Conclusion

In this work, we introduced a domain-transferable annotation scheme designed to address the inherent issue of domain specificity problems in semantic frame analysis. We collected expert interview dialogues from the culinary and gardening domains and annotated them with semantic frame structures. Further, we performed domain transfer experiments based on the acquired annotations. The results showed the feasibility of the proposed annotation scheme, suggesting the possibility of applying it to domains of sparse data, facilitating the elicitation of implicit and tacit knowledge.

7. Bibliographical References

Charles J. Fillmore. 1986. Pragmatically controlled zero anaphora. In *Proceedings of the Twelfth Annual Meeting of the Berkeley Linguistics Society*, pages 95–107.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-Enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

Seth Kulick, Ann Bies, and Justin Mott. 2014. Inter-annotator agreement for ERE annotation. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 21–25, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yuichiroh Matsubayashi, Naoaki Okazaki, and Jun'ichi Tsujii. 2009. A comparative study on generalization of semantic roles in FrameNet. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 19–27, Suntec, Singapore. Association for Computational Linguistics.

Sarah McLeod, Ivana Kruijff-Korbayova, and Bernd Kiefer. 2019. Multi-task learning of system dialogue act selection for supervised pre-training of goal-oriented dialogue policies. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 411–417, Stockholm, Sweden. Association for Computational Linguistics.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker, and Jan Scheffczyk. 2016. FrameNet ii: Extended theory and practice. URL: https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf.

Josef Ruppenhofer, Jonas Sunde, and Manfred Pinkal. 2010. Generating FrameNets of various granularities: The FrameNet transformer. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Tomohide Shibata, Shotaro Kohama, and Sadao Kurohashi. 2014. A large scale database of strongly-related events in Japanese. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*

(LREC'14), pages 3283–3288, Reykjavik, Iceland. European Language Resources Association (ELRA).

Natalia Skachkova and Ivana Kruijff-Korbayova. 2021. Automatic assignment of semantic frames in disaster response team communication dialogues. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 93–109, Groningen, The Netherlands (online). Association for Computational Linguistics.

8. Language Resource References

Taro Okahisa, Ribeka Tanaka, Takashi Kodama, Yin Jou Huang, and Sadao Kurohashi. 2022. Constructing a culinary interview dialogue corpus with video conferencing tool. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3131–3139, Marseille, France. European Language Resources Association.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.