Definition Generation for Automatically Induced Semantic Frame

Yi Han **Ryohei Sasano** **Koichi Takeda**
Graduate School of Informatics, Nagoya University
han.yi.u2@s.mail.nagoya-u.ac.jp
{sasano,takedasu}@i.nagoya-u.ac.jp

Abstract

In a semantic frame resource such as FrameNet, the definition sentence of a frame is essential for humans to understand the meaning of the frame intuitively. Recently, several attempts have been made to induce semantic frames from large corpora, but the cost of creating the definition sentences for such frames is significant. In this paper, we address a new task of generating frame definitions from a set of frame-evoking words. Specifically, given a cluster of frame-evoking words and associated exemplars induced as the same semantic frame, we utilize a large language model to generate frame definitions. We demonstrate that incorporating frame element reasoning as chain-of-thought can enhance the inclusion of correct frame elements in the generated definitions.

1 Introduction

Semantic frames are conceptual structures that describe specific types of situations or events, and FrameNet (Baker et al., 1998) is one of the representative semantic frame resources. Each frame in FrameNet comprises a frame name, a definition, a set of core and non-core frame elements (FEs), a set of frame-evoking words (lexical units), and exemplars in which lexical units appear. Table 1 exhibits CUTTING frame in FrameNet as an example.

While FrameNet has expanded to include multiple languages, such as Spanish (Subirats and Sato, 2004) and Japanese (Ohara et al., 2004), it remains absent in many low-resource languages. There may also be cases where frame knowledge of different granularity or specific to a particular domain is required. Accordingly, new frame resources need to be developed in some cases, but the costly annotation process hinders their construction. To address this issue, a number of efforts have been made to automate the process of building frames from large corpora. For instance, the frame induction task aims to automatically group frame-evoking words (typically verbs), along with their associated exemplars, on the basis of the semantic frame they evoke. The mainstream methods are mainly based on contextualized word embeddings (QasemiZadeh et al., 2019; Yamada et al., 2021b) such as BERT (Devlin et al., 2019). These methods leverage the observation that words evoking the same semantic frame tend to appear in similar contexts, resulting in their embeddings being grouped into the same cluster (Yamada et al., 2021a, 2023).

Frame: CUTTING
• **Frame definition**:
An AGENT cuts an ITEM into PIECES using an instrument.
• **Frame elements (core)**:
AGENT, ITEM, PIECES
• **Frame evoking words**:
slice, cut, chop, dice, fillet, mince, · · ·
• **Exemplars**:
◇ I carefully sliced the tomatoes for the salad.
◇ She cut into the melon with a knife.
◇ Chop the onions finely.

Table 1: CUTTING frame in FrameNet.

However, while the frame induction task provides clusters of frames, it lacks interpretability because definitions of these clusters are not provided. To make frame resources intuitive and understandable to humans, it is desirable to analyze each cluster and assign a definition sentence that describes the frame, but manually adding definition sentences is labor-intensive and time-consuming. Therefore, in this study, we address a new task of generating frame definitions from a set of frame-evoking words.

2 Frame Definition Generation

Our formulation of the semantic frame definition generation task is as follows: given a set of frame-evoking words and associated exemplars clustered as one frame, our objective is to produce a cohesive natural language definition that accurately captures

Findings of the Association for Computational Linguistics: ACL 2024, pages 11112–11118

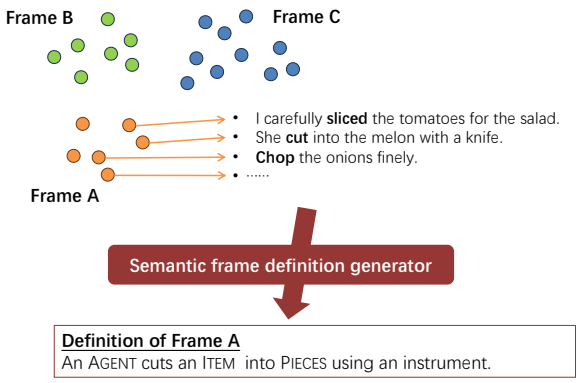

Figure 1: Illustration of semantic frame definition generation task.

the essence of the frame they evoke.[1] Figure 1 illustrates our frame definition generation task.

We employ a large language model as backbone of the definition generator for its strong performance across various natural language generation (NLG) tasks (Zhang et al., 2023). Specifically, we employ Llama2-70b-chat[2], an open-source large language model that has demonstrated strong performance among large language models (Touvron et al., 2023). The task of the definition generator is to map the clusters of frame-evoking words and exemplars to their respective definitions.

2.1 In-context Learning

Recent studies have shown that large language models have strong few-shot performance on a wide range of downstream tasks, known as in-context learning (ICL) (Brown et al., 2020). In standard ICL, the LLM is prompted with a set of input-output pairs, termed demonstrations, to output new predictions. In our approach, we utilize frame-evoking words and exemplars as input queries, presenting the LLM with query-definition pairs to learn the underlying mapping and generate corresponding definitions. Leveraging ICL enables the integration of frame knowledge into LLMs using a limited number of demonstrations, making our method more practical in low-resource language scenarios. Moreover, ICL does not require parameter adaptation, thereby significantly reducing computational costs compared to other task adaptation methods such as fine-tuning.

[1]As our primary focus lies in frame definition generation rather than frame induction task, we extract data directly from the ground truth semantic frame dataset and input them into definition generator.

[2]https://huggingface.co/meta-llama

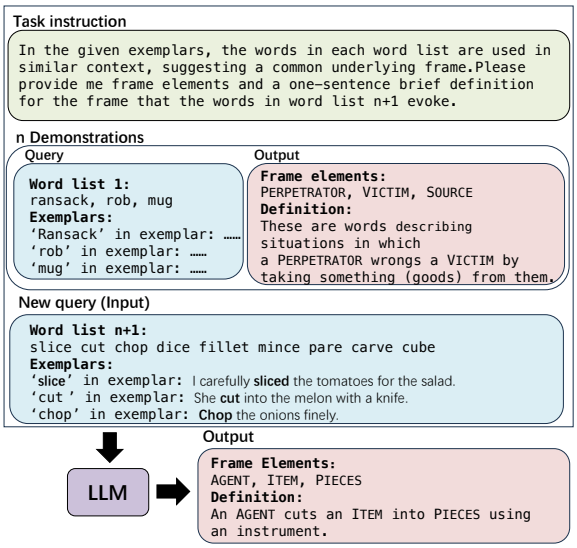

Figure 2: The framework of the prompt of frame definition generation.

Frame elements provide contexts for understanding the relationships between various concepts within the frame. By identifying frame elements, people can better interpret and infer information related to the frame. Related study shows that including Chain-of-Thought (CoT) in the prompt can benefit the ICL performance (Wei et al., 2022). Aiming to increase the inclusion of correct frame elements, we added frame element reasoning between query and definition to construct demonstrations. By generating frame elements and frame definition subsequently, LLM is encouraged to improve the incorporation of accurate frame elements within the generated definitions.

2.2 Prompt for Frame Definition Generation

The framework for the prompt of frame definition generation is illustrated in Figure 2, consisting of three main components: task instruction, demonstrations, and new query. The task instruction provides a concise description of the task objective. Demonstrations are key to the prompts and we evaluate two strategies for selecting demonstrations, as detailed below.

Random demonstrations We use this as our main setting, where demonstrations are randomly extracted from the training set consisting of ground truth semantic frames. Note that once extracted, the demonstrations are fixed when mapping different clusters into definitions. Due to the random extraction of data from FrameNet, performance fluctuations occur due to randomness. To address this, we

conduct multiple experiments using diverse sets of demonstrations to provide a comprehensive assessment of their overall performance.

Similar demonstrations It has been shown that demonstrations that are similar with new query can benefit ICL performance in many tasks (Liu et al., 2022). Given the cluster of one frame, we encode all exemplars within the cluster using a sentence-encoder[3], then we take average of their exemplar embeddings as the frame embedding. We search for similar demonstrations by measuring the cosine similarity between the target frame embedding and other candidates within the training set, and extract the nearest neighbors.

3 Experiment

3.1 Dataset

In this work, we extract frame data from the publicly available Berkeley FrameNet 1.7 dataset[4]. We use frames evoked by verbs, splitting them into an 80 percent training set (513 out of 641 verb-evoking frames) and a 20 percent test set with the remaining 128 frames. Note that for random demonstrations setting, we extract only a small subset of the training data rather than utilizing the entirety. While in similar demonstrations and fine-tuning settings, we utilize the entire training dataset. For frame elements, we mainly focus on core frame elements, without which the frame cannot be instantiated.

3.2 Evaluation

Evaluation of definition generation tasks typically relies on standard NLG metrics such as BLEU, Rouge, or BERTScore. However, assessing frame generation tasks demands diversity, where traditional reference-based metrics often fall short. To address this, our evaluation method, represented as Def-Eval, employs GPT-4 to evaluate the generated frame definitions rated on a 1-5 scale. As shown in Figure 3, the prompt consists of task instruction, evaluation criteria, reference definition and generated definition. To validate the efficacy of our evaluation method, we manually annotate 30 frame definitions and assess their correlation with the aforementioned metrics. Results in Table 2 demonstrate that our evaluation approach exhibits the highest correlation with human judgement.

[3] https://huggingface.co/distilbert/distilbert-base-uncased

[4] https://framenet.icsi.berkeley.edu/

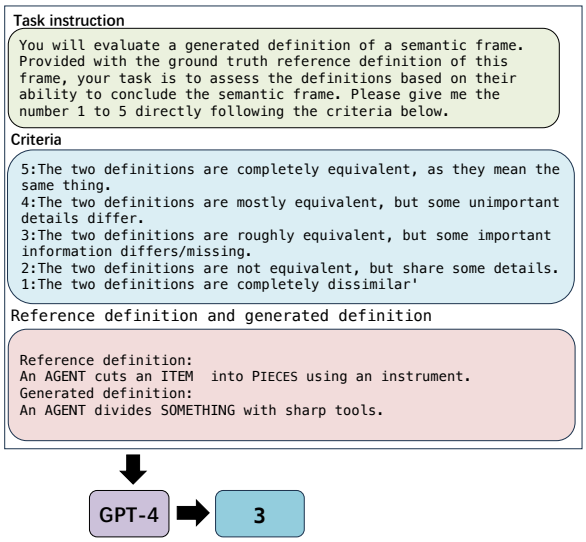

Figure 3: Prompt of evaluating generated definitions.

Metrics	Correlation coefficient
Sacre Bleu	0.09
Rouge-L	-0.07
BERTScore (f1)	0.13
Def-Eval	0.48

Table 2: Correlation coefficient scores between manually annotated scores and evaluation metrics.

Additionally, we leverage an auxiliary evaluation method called FE-Eval to assess the inclusion of frame elements. This involves extracting and concatenating frame elements, followed by calculating BERTScore between the concatenated elements and the reference.

3.3 Experimental Settings

We utilize Llama2-70b-chat as the backbone of definition generator and set the parameters as follows: temperature=0.1, max_new_tokens=1024, repetition_penalty=1.2, top_p=0.9, and top_k=50. We explore the proposed method in three groups of experiments, represented as follows.

- **Impact of Demonstration Components**: In this group of experiments we test the impact of demonstration components. Since frame-evoking word list and frame definition are indispensable for the prompt, we experiment with the effect of exemplars and frame elements by removing them from prompts. We randomly choose 3 demonstrations (3-shot) from FrameNet and report their Def-Eval and FE-Eval scroes.

	Def-Eval	FE-Eval
Word list	3.40±0.11	0.69±0.02
+Exemplars	3.37±0.15	0.67±0.03
+FEs	3.43±0.17	0.79±0.03
+FEs +Exemplars	3.44±0.17	0.80±0.03

Table 3: Results when using different demonstration components.

	Def-Eval	Num
Zero-shot	2.67	0
Random demonstrations	3.59	5
Similar demonstrations	3.62	n
Fine-tuning	3.57	n

Table 4: Def-Eval and Num values of different methods.

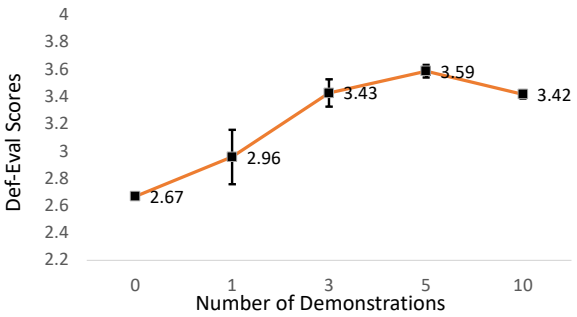

Figure 4: Average and standard deviation of Def-Eval scores across different numbers of demonstrations.

- **Impact of Demonstration Quantity**: We experiment with varying numbers of demonstrations: 0 (zero-shot), 1, 3, 5, and 10. For each number of demonstrations, we experiment with various demonstration sets and report their average and standard deviation values.

- **Method Selection Strategy**: In this experiment, we examine the strategy for selecting methods for generating definitions, considering the variability in FrameNet data availability across different languages.

3.4 Results

Impact of Demonstration Components As illustrated in Table 3, employing frame element reasoning as Chain-of-Thought yields similar Def-Eval scores, while significantly enhancing FE-Eval scores. This suggests that this process assists the language model in identifying and incorporating the correct frame elements into the frame definitions. Moreover, we observe that excluding exemplars has minimal effect on performance. We attribute this to the pretraining of LLMs on extensive raw corpora, enabling them to interpret words without exemplar contexts. In subsequent experiments, we adopt the Word list + FEs demonstration setting in the following experiments.

Impact of Demonstration Quantity As Figure 4 shows, it is evident that compared to zero-shot scenarios, adding demonstrations can significantly enhance performance, highlighting the effectiveness of ICL. As the number of demonstrations increases, the average Def-Eval score also improves. However, The score peaks and experiences a slight decline, indicating that too many demonstrations may have a negative effect on the performance. Furthermore, we note a consistent decline in the variance value of Def-Eval scores, suggesting that as the number of demonstrations increases, the definition generator tends to produce more stable definitions.

Mehod Selection Strategy We utilize the five demonstration setting which yielded the best and stable results in the previous experiment. Since the availability of FrameNet data varies across different languages, we also report the ground truth labels required for each strategy (represented as Num). Because fine-tuning and ICL are two popular task-adaptation strategies for language models, we also evaluate fine-tuning's effectiveness in the context of frame definition generation task. We adopt QRoLA (Dettmers et al., 2024) to finetune LLM.

Table 4 presents the Def-Eval and Num values for various methods. To clarify the Def-Eval scores, we have included generated definition examples of different frames in the Appendix A. It is evident that ICL, with only five ground truth demonstrations, achieves a significant improvement compared to the zero-shot setting. This implies that in constructing FrameNet data for low-resource languages, researchers need only annotate a minimal number of ground truth demonstrations to define all induced frames. Leveraging semantically similar frames as demonstrations can enhance performance, yet necessitates the entire training dataset, rendering it more viable in contexts with abundant FrameNet data. Fine-tuning yields similar results with ICL, but also demands the entire training dataset. Moreover, it is less efficient due

to its higher computational resource requirements for adapting LLM to our frame generation task.

4 Related Work

Definition generation, also named definition modelling, was initially formulated to generate a readable word definition based on word embeddings (Noraset et al., 2017). Initially, its primary objective was to assess word embeddings' effectiveness, later evolving to generating word definitions in context. Mickus et al.; Ni and Wang treat the definition generation as sequence-to-sequence task. Given a target a word highlighted in the exemplar, they utilize transformer-based model to generate word definition based on contextual information (Giulianelli et al., 2023). Unlike definition generation task aimed at generating individual definitions for each word, the frame definition generation task involves defining semantic frame evoked by multiple words.

5 Conclusion and Future Work

We introduce semantic frame definition generation tasks and employ ICL to generate definitions for induced semantic frames. We explore diverse demonstration formats through experiments on FrameNet. Results indicate that even with a limited number of ground truth demonstrations, our approach yields promising performance. Our future endeavors will involve experimenting with various LLMs to assess their efficacy. Additionally, although we currently leverage ground truth demonstrations from FrameNet, the clustering of induced frames may introduce noise. Investigating the impact of the noise on performance will be one of our focus in the future work.

Limitations

Our investigation focused on the Llama2-70b-chat model, without delving into other language models, potentially limiting the generalizability of our research findings. In future work, we plan to explore our method with various language models. Additionally, the study only examined the English language without exploring the efficacy across other languages.

Acknowledgements

This work was supported by JST FOREST Program, Grant Number JPMJFR216N.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING)*, pages 86–90.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *In Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. QLoRA: Efficient finetuning of quantized llms. *In Advances in Neural Information Processing Systems (NeurIPS)*, 36:10088–10115.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3130–3148.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of the 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures (DeeLIO)*, pages 100–114.

Timothee Mickus, Denis Paperno, and Matthieu Constant. 2019. Mark my word: A sequence-to-sequence approach to definition modeling. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing (DL4NLP)*, pages 1–11.

Ke Ni and William Yang Wang. 2017. Learning to explain non-standard English words and phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP)*, pages 413–417.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 3259–3266.

Kyoko Hirose Ohara, Seiko Fujii, Toshio Ohori, Ryoko Suzuki, Hiroaki Saito, and Shun Ishizaki. 2004. The

Japanese FrameNet project: An introduction. In *Proceedings of LREC-2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora*, pages 9–11.

Behrang QasemiZadeh, Miriam RL Petruck, Regina Stodden, Laura Kallmeyer, and Marie Candito. 2019. SemEval-2019 task 2: Unsupervised lexical frame induction. In *Proceedings of the 13th International Workshop on Semantic Evaluationn (SemEval)*, pages 16–30.

Carlos Subirats and Hiroaki Sato. 2004. Spanish FrameNet and FrameSQL. In *Proceedings of the LREC-2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora*, pages 13–16.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *In Advances in Neural Information Processing Systems (NeurIPS)*, 35:24824–24837.

Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2021a. Semantic frame induction using masked word embeddings and two-step clustering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 811–816.

Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2021b. Verb sense clustering using contextualized word representations for semantic frame induction. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP Findings)*, pages 4353–4362.

Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2023. Semantic frame induction with deep metric learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1833–1845.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. Extractive summarization via chatgpt for faithful summary generation. *arXiv preprint arXiv:2304.04193*.

A Generated definitions

In our evaluation method, we have ensured that the criteria for each point were included in the prompt, providing a reference for assessment, as illustrated in Figure 3. For instance, the majority of our results fell within the range of 3 to 4. Therefore, one could interpret the generated definitions as falling between 4: The two definitions are mostly equivalent, but some unimportant details differ, and 3: The two definitions are roughly equivalent, but some important information differs/missing. To offer a more comprehensive understanding of the generated definitions and their corresponding Def-Eval scores, we present examples of three frames in Tables 5, 6, and 7.

Method	Definition	Def-Eval
Reference	A Communicator communicates a judgment of an Evaluee to an Addressee.	-
Zero-shot	The evaluation of something, whether it be a person, an action, an idea, or a production.	2
Random demonstrations	An Agent expresses a strong evaluation or opinion about a Person, Action, or Thing, often with the intention of influencing others' opinions or behaviors.	3
Similar demonstrations	A Speaker expresses a Judgment about an Evaluee based on a Reason.	4
Fine-tuning	A Speaker expresses an opinion about some Topic by using a word that is either positive or negative to describe it (or sometimes both) and which may also specify the reason for this evaluation.	3

Table 5: Generated definitions and their Def-Eval scores for the Judgment_communication frame.

Method	Definition	Def-Eval
Reference	In this frame, an Agriculturist cultivates Food.	-
Zero-shot	Nurturing or caring for something to promote its growth or development.	2
Random demonstrations	The Agent cultivates Crops on the Land using Tools, with the goal of producing food or other valuable resources.	4
Similar demonstrations	A Farmer intentionally cultivates the Land to produce a Crop, managing the environment to optimize growth and yield.	4
Fine-tuning	A Farmer grows crops on land that they own or lease (the Farm) for sale to others at market prices or for their own use.	3

Table 6: Generated definitions and their Def-Eval scores for the Agriculture frame.

Method	Definition	Def-Eval
Reference	In this frame, an Agent continues to possess a Theme for a Duration of time.	-
Zero-shot	This semantic frame encompasses the idea of holding onto something, whether it be physical objects, information, skills, or relationships.	2
Random demonstrations	An Agent keeps or retains possession of an Object, maintaining its current State without any change or loss.	3
Similar demonstrations	An Agent keeps or retains an Object in a certain State.	3
Fine-tuning	An Agent is responsible for preserving or maintaining an Entity (or Entities) within its bounds and under certain conditions over time.	3

Table 7: Generated definitions and their Def-Eval scores for the Retaining frame.