Submodular-based In-context Example Selection for LLMs-based Machine Translation

Baijun Ji[1], **Xiangyu Duan**[1*], **Zhenyu Qiu, Tong Zhang, Junhui Li**[1], **Hao Yang**[2], **Min Zhang**[1]
[1]School of Computer Science and Technology, Soochow University, Suzhou, China
[2]Huawei Translation Services Center, Beijing, China
begosu@foxmail.com, {xiangyuduan, lijunhui, minzhang}@suda.edu.cn
yanghao30@huawei.com

Abstract

Large Language Models (LLMs) have demonstrated impressive performances across various NLP tasks with just a few prompts via in-context learning. Previous studies have emphasized the pivotal role of well-chosen examples in in-context learning, as opposed to randomly selected instances that exhibits unstable results. A successful example selection scheme depends on multiple factors related to the task, while in the context of LLMs-based machine translation, the common selection algorithms only consider a single factor, which is the similarity between the example source sentence and the input sentence. In this paper, we introduce a novel approach to use multiple translational factors for in-context example selection by using monotone submodular function maximization. The factors include surface/semantic similarity between examples and inputs on either source side or target side, as well as the diversity within examples. Importantly, our framework mathematically guarantees the coordination between these factors, which are different and challenging to reconcile, due to the unique properties of submodular functions. Additionally, our research uncovers a previously unexamined dimension: unlike other NLP tasks, the translation part of an example is also crucial, a facet disregarded in prior studies. Experiments conducted on two LLMs, BLOOMZ-7.1B and LLAMA2-13B, demonstrate that our approach significantly outperforms random selection and robust single-factor baselines across various machine translation tasks.

Keywords: In-context Learning, Submodular Function, LLMs-based Machine Translation

1. Introduction

In-context learning represents a novel paradigm that leverages large language models (LLMs) to perform NLP tasks, even those on which LLMs were not explicitly fine-tuned. These capabilities do not rely on gradient updates but can be readily triggered through task-specific instructions and demonstration examples. While in-context learning serves as a training-free learning methodology, its performance demonstrates inherent instability, chiefly attributable to several features of the example selection such as quantity, sequence order, and choice criteria(Lu et al., 2021; Kim et al., 2022; Rubin et al., 2021).

Prior research of the example selection has predominantly centered around Sentiment Analysis (Liu et al., 2021; Kim et al., 2022), Natural Language Inference (NLI) (Kim et al., 2022), Semantic Parsing(Liu et al., 2021) and other Natural Language Understanding tasks (NLU) (Wang et al., 2022; Lu et al., 2021). There has been a limited focus on developing methodologies tailored explicitly for machine translation, and the prevalent selection criteria for translation examples primarily involves a single factor of similarity between the example source sentence and the input sentence, either in terms of surface characteristics or semantic content(Agrawal et al., 2022; Kumar et al., 2023).

From our perspective, the machine translation task possesses distinct characteristics, which require the example selection algorithm to consider multiple factors. Firstly, the target translation of an example also carries inherent significance and establishes a resilient linkage with the source input when employed in retrieving examples in in-context learning, as opposed to NLU tasks where the target is a classification label not suitable for retrieving examples. Secondly, it is crucial to consider both surface similarity and semantic similarity, as prior research frequently treated them as separate facets. Thirdly, *diversity* plays a pivotal role in machine translation owing to the widespread issue of polysemy, which is a common challenge in translation. Relying solely on a single selection method would result in a dearth of diverse sample inputs, ultimately impeding the model's ability to glean more effective information during the translation process.

In this paper, we propose a submodular-based framework to consider multiple factors in in-context example selection, which can effectively addressing the aforementioned issues. Submodular functions(Fujishige, 1991) are a type of discrete set functions that exhibit a concept called "diminishing returns." In our framework, we develop a set of submodular functions specifically tailored for the machine translation task. These functions take

[*] Corresponding author

LREC-COLING 2024, pages 15398–15409
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

into account the multiple factors such as the relationship between the target sentences in examples and the test input, as well as the similarities in both their surface and meaning, while also ensuring diversity among the examples. Our framework also provides mathematical guarantees that treat the multiple factors in harmony, even when employing a straightforward greedy search approach.

To conclude, the main contributions of our paper are three-fold:

1) We propose a novel submodular-based framework for in-context example selection to enhance the machine translation capabilities of LLMs. This framework is interpretable and amenable for efficient optimization.

2) A class of submodular functions has been developed to assess the quality of a candidate set. These submodular functions take into consideration multiple factors that influence translation performance and can be effectively combined.

3) We conducted comprehensive experiments across various translation tasks and reaffirmed that in-context example selection plays a pivotal role in enhancing the performance of LLMs-based machine translation. Furthermore, our approach outperforms strong baseline models on multiple evaluation datasets, demonstrating superior results.

2. Background

2.1. In-context Learning for Machine Translation

In-context learning is a powerful paradigm that enables LLMs to effectively learn tasks by demonstrating limited examples. Formally, considering a large language model denoted as *LLM*, accompanied by n in-context examples denoted as $S = \{(x_i, y_i)\}_{i=1}^{n}$, and given a test input x_{test}, the prediction for x_{test} is generated through the following process:

$$\arg\max \mathcal{P}_{LLM}(y|x_1 \oplus y_1 ... x_n \oplus y_n \oplus x_{test}), \quad (1)$$

where y is the translated sentence via greedy search in the case of machine translation and $\oplus$ represents the concatenation operation according to some predefined templates like Table 1. The example set S represents a collection of translation pairs. Generally, the source part $S_x = \{x_i\}_{i=1}^{n}$ of S is preferable to have a high coverage of n-grams or to be closely aligned with x_{test} in the semantic space (Agrawal et al., 2022).

Previous studies have demonstrated that the effectiveness of in-context learning heavily relies on various characteristics associated with the example demonstrations, including the format of the examples, the order of demonstrations, and other related aspects (Zhao et al., 2021; Lu et al., 2021). In this paper, our primary focus is on the selection of examples, while we postpone the discussion of order and format to future investigations.

[language1]: [x_1] = [language2]: [y_1]
###
...
###
[language1]: [x_k]= [language2]: [y_k]
###
[language1]: [x_{test}] = [language2]:

Table 1: Prompt Template for Machine Translation.

2.2. Submodular Functions

Formally, given a set V and its function $F : 2^V \to \mathbb{R}$, where 2^V denotes the power set of V, F is considered submodular if it satisfies the diminishing returns property for any $A \subseteq B \subset V$ and $v \in V \setminus B$:

$$F(A \cup \{v\}) - F(A) \geq F(B \cup \{v\}) - F(B). \quad (2)$$

This inequality implies a decreasing incremental gain of inserting element v into the sets of A and B. Moreover, such a function is termed monotone if $F(A) \leq F(B)$. Monotone submodular functions exhibit a rich set of properties. Among them, we introduce three properties utilized in our approach:

Theorem 1: If $\forall i$, F_i is a monotone submodular function and $\alpha_i \geq 0$, then $\sum \alpha_i F_i$ is also a monotone submodular function.

Theorem 2: If F' is a monotone submodular function and $g : \mathbb{R} \to \mathbb{R}$ is a non-decreasing concave function, then $F = F' \circ g : 2^V \to \mathbb{R}$ is also a monotone submodular function.

Theorem 3: When maximizing a monotone submodular function F under cardinality constraints, i.e., the result subset S is constrained to have K elements, where $K \ll |V|$, greedy algorithm has a worst-case guarantee: $F(S) \geq (1 - \frac{1}{e})F(\hat{S}) \approx 0.63F(\hat{S})$, where $\hat{S}$ represents the optimal set.

Application for In-context Example Selection: Our objective is to find optimal set of examples for in-context learning. The set can be built by selectively inserting elements of new examples into the current set iteratively. Monotone submodular functions provide a solution for such insertion according to the above properties. Firstly, we can seamlessly integrate multiple translational factors embedded in different monotone submodular functions by employing a linear combination of them

(Theorem 1). Secondly, we can design monotone submodular functions with flexible functions tailored for machine translation (Theorem 2). Thirdly, this objective can be efficiently optimized through the greedy algorithm with a stable worst-case guarantee (Theorem 3). Despite the situation that some factors such as similarity and diversity are hard to reconcile, the scheme of monotone submodular functions can accommodate these factors with the mathematical guarantee.

3. Proposed Methodology

Given an input source sentence denoted as x_{test} and a translation database represented by V, the objective of in-context learning is to search a subset $S \subset V$ ($|S| \ll |V|$) that enhances translation quality. Given the expansive nature of the search space, it is impractical to exhaustively explore all conceivable combinations. However, we can employ monotone submodular functions to evaluate the quality of the candidate set and apply a greedy algorithm to progressively insert elements into the set. In the following section, we will introduce a series of functions based on distinct criteria for machine translation.

3.1. Surface Coverage Submodular Functions

Prior research suggests that selecting contextual examples based on their semantic similarity to the test sample in the embedding space is an effective strategy(Liu et al., 2021). Nonetheless, while this global sentence representation holds its value, it falls short in capturing the subtler nuances of alignment at the lexical or phrase level. For instance, the sentences "Nominations flooded in for the prestigious award, but only a few were received." and "She received multiple nominations for her outstanding performance in the film." do not possess identical meanings. However, the shared terms "received/nominations" can still offer alignment information, especially when the target of the extracted example is likely to encompass partial translations of the source input. We argue such surface coverages at the lexical or phrase level are equally crucial for enhancing the performance of LLMs in translation tasks. Hence, we propose a pair of monotone submodular functions that take surface coverage into account.

Source-specific Coverage

A recall-based n-gram overlap score was introduced by (Agrawal et al., 2022) to quantify the degree of coverage between the example and the test input x_{test}. We extends the application of this scoring function to a monotone submodular function. Given a set S and its source part S_x, we define the coverage score of S on the source side using the following equation:

$$R_{SRC}(S) = \frac{\sum_{e \in x_{test}} min(C_e(S_x), C_e(x_{test}))}{\sum_{e \in x_{test}} C_e(x_{test})}. \quad (3)$$

$C_e(\cdot)$ represents the frequency of occurrence of an n-gram e within a given set, evidently functioning as a monotone submodular function for each e. Given that $min(x, a)$ is a non-decreasing concave function of x and $C_e(S)$ is a monotone submodular function, we can readily deduce that $R_{SRC}(S)$ is a monotone submodular function(Theorem 2).

Intuitively, this function encourages selecting examples covering specific n-gram e in x_{test}, but the advantage of such selection diminishes as soon as its frequency in S_x matches that in x_{test}. In other words, when $R_{SRC}(S \cup v) - R_{SRC}(S)$ equals zero, it signifies that the value has reached a point where further inclusion of such sentences does not yield additional benefit.

Target-specific Coverage

In addition to ensuring good coverage from the source side, we also let the example set S to encompass as many partial translations of the source input x_{test} as feasible. Specifically, we construct the pseudo translation of x_{test} by looking up the ground truth dictionary or the artificial dictionary generated by MUSE(Lample et al., 2017). We denote the result of translating word-by-word as $T(x_{test})$. The coverage score of S on the target side can be defined as follows:

$$R_{TGT}(S) = \frac{\sum_{e \in T(x_{test})} min(C_e(S_y), C_e(T(x_{test})))}{\sum_{e \in T(x_{test})} C_e(T(x_{test}))}, \quad (4)$$

where $S_y = \{y_i\}_{i=1}^{n}$ denotes the target part of S. In cases where words have multiple translations, we select all possible word translations. As the process of word-by-word translation does not preserve sentence structures and grammatical rules, the resultant translation $T(x_{test})$ is a collection of words devoid of any specific order. Therefore, our practical computation is restricted to the occurrence of 1-gram exclusively when calculating the target-specific coverage. Moreover, it is obvious that R_{TGT} is a monotone submodular function as R_{SRC}. Figure 1 shows the detailed calculation of R_{TGT}.

3.2. Diversity Submodular Functions

Diversity constitutes another pivotal factor in the process of example selection. Ye et al. (2022); Li and Qiu (2023) proposed a

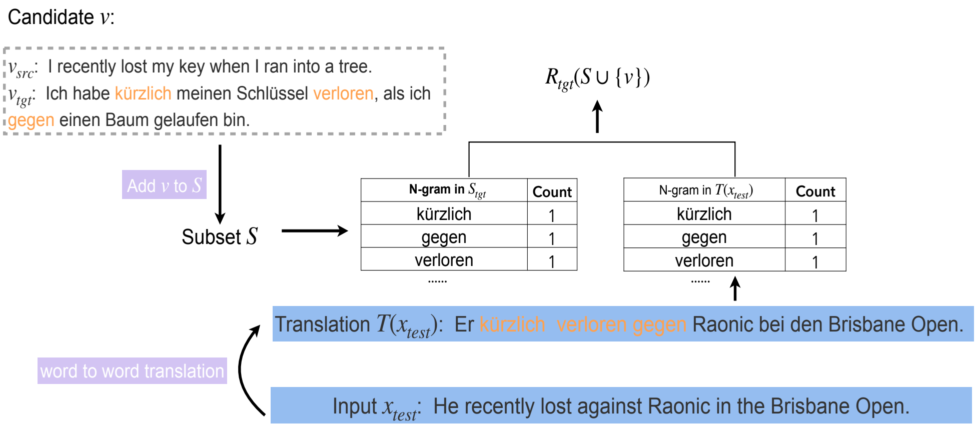
Candidate v:
v_{src}: I recently lost my key when I ran into a tree.
v_{tgt}: Ich habe kürzlich meinen Schlüssel verloren, als ich gegen einen Baum gelaufen bin.
Add v to S
Subset S
$R_{tgt}(S \cup \{v\})$
N-gram in S_{tgt} | Count
kürzlich | 1
gegen | 1
verloren | 1
......
N-gram in $T(x_{test})$ | Count
kürzlich | 1
gegen | 1
verloren | 1
......
Translation $T(x_{test})$: Er kürzlich verloren gegen Raonic bei den Brisbane Open.
word to word translation
Input x_{test}: He recently lost against Raonic in the Brisbane Open.

Figure 1: The Calculation Procedure for R_{tgt}. It begins by translating the input sentence x_{test} word-by-word. Then the gain of inserting v into to the candidate set S is computed. The calculation for R_{src} is identical to that of R_{tgt}, with the exception of the word translation.

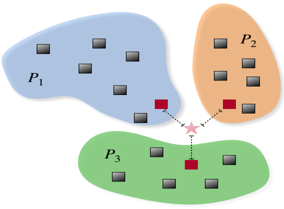

Figure 2: The Illustration of Diversity in Submodular Functions: The candidates should be as closely related to the input (the star) as possible while also belonging to different classes to the greatest extent possible.

Algorithm 1 In-context Example Selection for Machine Translation based on Monotone Submodular Functions

Input: source sentence x_{test}, candidate set $V = \{x_i, y_i\}_{i=1}^{N}$, number of in-context examples K.

Output: selected set $S = \{x_i, y_i\}_1^K$

```
Let $S \leftarrow$ empty list.
while $|S| < K$ do
    for $v \in V \setminus S$ do
        $Score[v] \leftarrow F(S \bigcup \{v\})$
    end for
    $S.append(argmax_v(Score))$
end while
return $S$
```

maximal-marginal-relevance(MMR)(Carbonell and Goldstein-Stewart, 1998) method for example selection. It chooses instances that are relevant to the query and penalizes the redundant information in the mean time to foster collaborative outcomes. The greedy algorithm is employed to select the subsequent example $v \notin S$ based on the following score with the highest value:

$$f_{MRR}(v) = \lambda sim(x, v) - (1-\lambda) \max_{q \in S} sim(v, q), \quad (5)$$

where x denotes the query, S is the current set and sim is the sentence-level similarity function. Moreover, we can create a function $F(S)$ that fulfills the condition $F(S \cup \{v\}) - F(S) = f_{MRR}(v)$, and this function also exhibits submodularity. However, it is important to note that $F(S)$ is not monotone since the value of $f_{MRR}(v)$ is not always postive and employing a greedy search would not ensures worst-case guarantees.

Inspired by (Lin and Bilmes, 2011), we use the diversity reward instead of a redundancy penalty. We cluster the source sentences within the set V and divide them into M distinct clusters, denoted as $\{P_i\}_{i=1}^{M}$. Subsequently, we calculate the diversity score using the following equation:

$$D_{SRC}(S) = \sum_{i=1}^{M} \log(1 + \sum_{v \in S_x \cap P_i} sim(v, x_{test})). \quad (6)$$

$D_{SRC}(S)$ promotes diversity by offering greater values when opting for a sentence from a cluster that has not had any of its sentence selected. For example, the reward score is larger if two selected sentences are from two different clusters, since $\log(1+a) + \log(1+b) > \log(1+a+b)$, where a, b represent the non-negative similarity score respectively. Within every non-decreasing concave logarithmic function, there exists a modular function defined by non-negative weights to ensure monotonicity. Applying the logarithmic function to such a monotone submodular function results in another submodular function and summing up these transformed functions maintains their submodularity for D_{SRC}. Figure 2 illustrates our main idea.

In addition to the source diversity, we also consider the diversity of the target side:

$$D_{TGT}(S) = \sum_{i=1}^{M} \log(1 + \sum_{v \in S_y \cap P_i} sim(v, T(x_{test})). \quad (7)$$

3.3. Methodology Overview

To summarize, we can formulate the quality of the chosen set S as follows:

$$\begin{aligned} F(S) = \lambda_1(R_{SRC}(S) + R_{TGT}(S)) + \\ \lambda_2(D_{SRC}(S) + D_{TGT}(S)), \end{aligned} \quad (8)$$

where $\lambda_1, \lambda_2 > 0$ are trade-off coefficients. Then we use the standard greedy search as shown in Algorithm 1, which offers performance guarantees since $F(S)$ is still monotone submodular(Theorem 1 and 3). Note that the training corpora for machine translation are usually quite large. To streamline the process, we initially retrieve the top 50 sentences using BM25, followed by selecting sentences from the top 50 sentences to insert into S using our proposed approach. The appendix contains an analysis of the impact of different values of N on the results.

	Direction	Datastore	#Pairs
Low-resource	bn↔en	Samanantar	8.6
	gu↔en	Samanantar	3.1
	hi↔en	Samanantar	10.1
High-resource	fr↔en	Europarl	1.9
	es↔en	Europarl	1.9
Domain-adaptation	de→en	IT	0.2
	de→en	Medical	0.2
	de→en	Law	0.4

Table 2: The datasets employed for retrieving in-context examples, along with the respective number of sentence pairs per language (in millions).

4. EXPERIMENTS

4.1. Settings

Dataset. We evaluate the performance of our proposed framework in three scenarios: 1) Low-resource translation, where the model faces the challenge of translating text to/from a language with a restricted amount of training data. 2) High-resource translation, where the model has exhibited strong translation capabilities due to extensive training on these languages. 3) Domain adaptation, where the model translates domain-specific sentences by utilizing in-domain examples. For the low-resource and high-resource translation tasks, we present the results on the devtest set of FLORES-101 (Goyal et al., 2022). For the domain adaptation task, we utilize the multi-domain dataset(Koehn and Knowles, 2017) and evaluate on the test sets of IT, Medical, and Law domains in our experiments. Table 2 presents detailed information regarding the specific datastore for retrieving in-context examples.

Models and Metrics. We mainly evaluate the renowned LLMs *BLOOMZ-7.1B*(Muennighoff et al., 2023) and *LLAMA2-13B*(Touvron et al., 2023). Following previous work(Kumar et al., 2023; Agrawal et al., 2022), the primary evaluation metric employed in our experiments is COMET(Rei et al., 2020), which is calculated using the wmt20-comet-da model, due to its better consistency with human evaluations. For the low-resource translation task, we conduct experiment with BLOOMZ-7.1B due to its better generalization ability across low-resource languages and its broader coverage of languages. For the high-resource translation task, we experiment with both BLOOMZ-7.1B and LLAMA2-13B. For the domain adaptation task, we observe that BLOOMZ-7.1B fails to translate German, so we experiment with LLAMA2-13B. We have published our code[1].

Gerneration Configs. In in-context example selection, we select four sentence pairs from the datastore as examples, and arrange them in descending order. The generated outputs from the LLMs are truncated to twice the length of the source text, and the maximum output length is limited to 250 tokens. To calculate semantic similarity and perform clustering, we utilize the stsb-xlm-r-multilingual[2] model to obtain sentence embeddings. We employ the K-means algorithm for clustering and set the number of clusters to be 10. During the decoding process, we employ a greedy search with a batch size of 5. In all of our experiments, we fix both λ_1 and λ_2 in Equation 8 to be one.

Bilingual Dictionary. We primarily employ the bilingual dictionary generated by MUSE to translate the source sentence word by word. For language pairs not found in MUSE, we derive the top-n dictionary from the parallel corpora using the fastalign[3] tool.

[1] https://github.com/cocaer/submodular-llm-mt

[2] https://huggingface.co/sentence-transformers/stsb-xlm-r-multilingual

[3] https://github.com/clab/fast_align

	en→bn	bn→en	en→gu	gu→en	en→hi	hi→en	average
w/o ICL	44.99	21.02	61.84	29.39	46.18	28.46	38.64
Random	62.22	36.41	43.53	46.20	59.20	47.23	49.11
BM25	64.40	39.16	43.01	47.18	61.23	51.06	51.01
TopK	68.02	42.13	50.56	50.23	62.24	51.76	54.15
MRR	64.53	44.67	51.48	48.22	64.70	54.69	54.71
CTQ	-	42.99	-	41.77	-	50.03	-
				OurMethods			
$R_{src}(S)$	66.81	41.77	55.48	45.74	60.46	47.13	52.89
$R_{tgt}(S)$	72.26	43.80	57.14	49.50	67.30	48.87	56.47
$D_{src}(S)$	69.57	46.95	56.88	51.16	65.80	52.46	57.13
$D_{tgt}(S)$	66.21	48.26	53.22	51.88	64.82	53.26	56.27
$F(S)$	73.31	48.64	58.43	53.07	67.30	53.77	59.08

Table 3: COMET scores on the low-resource translation task based on using BLOOMZ-7.1B.

	BLOOMZ-7.1B					LLAMA2-13B				
	en→fr	fr→en	en→es	es→en	avg.	en→fr	fr→en	en→es	es→en	avg.
w/o ICL	74.44	57.87	67.93	47.05	64.32	56.23	78.31	50.07	69.54	63.53
Random	81.87	76.19	69.21	66.57	73.46	72.16	78.83	66.82	71.10	72.22
BM25	82.69	75.80	67.84	66.49	73.20	73.91	79.47	67.77	71.25	73.10
TopK	82.95	76.14	66.64	70.50	74.05	74.22	78.86	67.61	71.55	73.06
MRR	80.38	75.69	68.01	67.46	72.88	74.05	78.87	67.13	71.61	72.91
$F(S)$	83.00	78.63	68.97	69.38	75.00	75.25	79.42	68.01	71.66	73.60

Table 4: COMET scores on the high-resource translation task based on using BLOOMZ-7.1B and LLAMA2-13B.

	IT	Law	Medical	avg.
w/o ICL	31.18	44.38	43.29	39.62
Random	21.63	46.89	40.25	41.01
BM25	40.46	56.97	47.56	48.33
TopK	38.67	58.34	50.04	49.02
MRR	40.64	58.26	50.49	49.80
$F(S)$	41.54	58.38	51.65	50.52

Table 5: COMET scores on the domain adaptation task based on using LLAMA2-13B.

4.2. Baseline Methods

We compare our methods with baselines as follows:

- **Random Selection.** We randomly select examples from the datastore and report the average scores(number of trials = 3).
- **BM25**(Agrawal et al., 2022). This method retrieves K examples that have source sentences most similar to the input source sentence. They employ the BM25 retriever, which emphasizes the overlap of n-grams.
- **TopK**(Liu et al., 2021). This method aims to retrieve examples that are semantically similar to the input source sentence. We use the stsb-xlm-r-multilingual model to obtain the sentence embedding and measure the similarity using cosine distance.
- **MMR**.Ye et al. (2022) proposed a maximal-marginal relevance(MMR) based selection strategy to choose diverse examples. Although this method was not originally intended for machine translation, we have found that it still can generate competitive results.
- **CTQ**(Kumar et al., 2023). It learns a regression function that selects examples based on multiple features to maximize the translation quality.

4.3. Results on Low-resource Translation

The evaluation results of different methods on the low-resource translation task are presented in Table 3. Based on the results, we can conclude that: **Firstly**, the quality of in-context examples plays a pivotal role in the context of translation performance. Randomly selecting examples significantly lags behind those sophisticated mechanisms. Notably, we observe that employing semantic-based selection methods(TopK), yields competitive results. The results are further improved by using MRR which takes into account the diversity among examples. **Secondly**, by employing various monotone submodular functions, we can achieve further enhancements in translation quality. While both $R_{src}(S)$ and BM25 highlight the importance of the overlapping n-grams, R_{src} yields superior results due to its consideration of n-gram saturation within the candidate set S. The performance of R_{tgt} surpasses that of R_{src}, suggesting that the target coverage may hold greater significance than the source coverage in machine translation. Both D_{src} and D_{tgt} consistently yield highly competitive results, underscoring once more the paramount significance of diversity. D_{src} performs superior to MRR, indicating that our diversity submodular functions are more efficient than the diversity score used in MRR. **Finally**, combining all monotone submodular functions can further enhances the translation quality, outperforming the strong baseline TopK and MRR by 4.93 and 4.37 COMET scores.

4.4. Results on High-resource Translation

Table 4 displays the results of the two LLMs on the high-resource translation task. The results reveal that despite LLAMA2-13B having a larger number of model parameters, its translation performance is not superior to that of BLOOMZ-7.1B when translating from English into other languages. We contend that the translation ability of LLMs are influenced not only by their model size but also by the distribution of languages in their training data. It is noteworthy that a substantial majority of the training data for LLAMA2-13B consists of English content, accounting for nearly 89.70% of the dataset. This bias towards English data explains its proficiency in generating English sentences. We also notice that randomly selected examples produce translation performance comparable to that achieved by BM25 and TopK methods. Our argument is that LLMs have excelled in translation for these high-resource languages. Providing informative examples for the translation task becomes more challenging. Even in such challenging situation, our approach exhibits significant improvements over the robust baselines, achieving 1.52 COMET score improvement on BLOOMZ-7.1B and 0.74 COMET score improvement on LLAMA2-13B.

4.5. Results on Domain Adaptation Task

Table 5 shows the results of the domain adaptation task based on using LLAMA2-13B. We notice that using randomly selected examples leads to noticeably lower results in comparison to BM25 and TopK. This indicates that LLMs require domain-specific knowledge to excel in this task. When prompts are consistent with the domain, LLMs can adapt more effectively on-the-fly. An effective in-domain demonstration can offer valuable information and instruct the model to translate domain-specific terms or expressions well. Our method consistently outperforms the strong baseline MRR by an average of 0.7 in COMET scores. We believe that this fact holds practical value since only the specific examples from the in-domain are needed for deploying the cross-domain translation service.

5. Analysis

5.1. Distribution of Selected Examples

When doing in-context learning for machine translation, the selection process consists of two stages. First, we retrieve a small set of examples(about 50-100) from the large datastore using the BM25 algorithm. Then we re-rank this set using methods such as TopK, MRR, or our approach. So it is valuable to understand the distribution of the selected examples across these different methods.

In Figure 3(a), TopK algorithm tends to favor the selection of examples with higher BM25 scores. This phenomenon is expected, as examples with the highest number of overlapping n-grams tend to have similar meanings. Due to the inclusion of diversity in MRR and our approach, we observe a smoother selection of samples. Even those ranked beyond 15 still have the potential to be selected.

Although the diversity of samples needs to be taken into consideration, the similarities between examples and the source input should not be excessively compromised. Figure 3(b) illustrates that the examples chosen by MRR exhibit the lowest similarities between examples and the source input. This suggests that MRR encounters difficulty in maintaining a harmonious equilibrium between diversity and relevance, despite the introduction of a balancing factor. Our approach achieves average similarity scores that fall between those of TopK and MMR, with a notably high lower

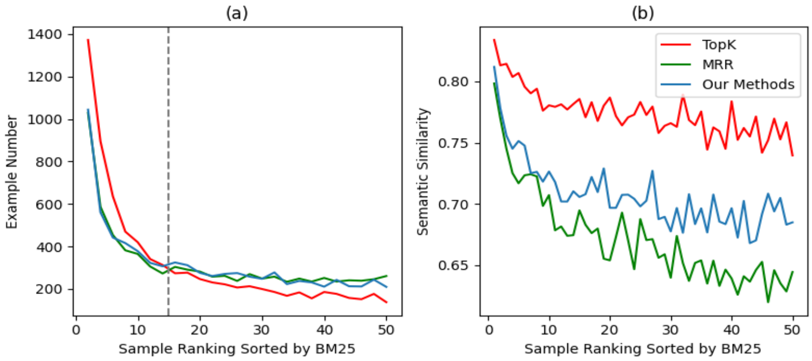

Figure 3: The x-axis in the figure represents the sample ranking sorted by BM25 score. The y-axis in figure (a) indicates the frequency with which the sample in that BM25 ranking was selected. In figure (b), the y-axis represents the average similarity score between the source input and the selected samples at their respective BM25 rankings.

bound (>0.65). This indicates the superiority of our method, supported by the worst-case performance guarantee.

	Informativeness
Random	6.713
BM25	7.186
TopK	7.173
MRR	7.191
Ours	7.222

Table 6: The example informativeness on multi-domain datatsets.

5.2. Informativeness from Selected Examples

A good example should be informative for LLMs to predict. Li and Qiu (2023) proposed **InfoScore** to measure the individual informativeness of one example. We adapt this metric for machine translation to assess the effectiveness of utilizing a prompt consisting of multiple examples in generating accurate translations. The example informativeness metric is calculated as follows:

$$I(Prompt, x_{test}) = \log(\mathcal{P}(\hat{y}|Prompt \oplus x_{test})) \\ - \log(\mathcal{P}(\hat{y}|x_{test})),$$

where $\hat{y}$ denotes the correct translation. Table 6 presents the results on the domain adaptation task. It shows that the example informativeness score exhibits a strong correlation with translation performance, as presented in Table 5. Specifically, when the prompt provides a greater amount of information, there is a notable increase in the COMET score. Furthermore, our approach outperforms other methods in terms of informativeness, implying that the selected examples hold higher value.

6. Related Work

6.1. Submodular Functions for NLP

Submodular functions have extensive applications in various natural language processing (NLP) tasks. Lin and Bilmes (2011) leveraged the coverage function and diversity reward function in the context of statistical extractive document summarization. In a different vein, Kirchhoff and Bilmes (2014) framed the challenge of data selection for statistical machine translation as a submodular programming problem, introducing a class of feature-based functions. To address the balanced clustering problem, Kawahara et al. (2011) sought to regularize cluster sizes using submodular functions. Notably, the objective function for balanced clustering is characterized as a fractional submodular function. In contrast, our method is the first work that utilizes the submodular functions for LLMs-based applications.

6.2. In-context Learning for Machine Translation

The utilization of LLMs for machine translation has garnered increasing interest in recent times, as it effectively addresses the issue of limited parallel data availability. Lin et al. (2021) conducted an evaluation of GPT-3 and XGLM-7.5B across 182 translation directions, while Bang et al. (2023) assessed ChatGPT in 12 distinct translation directions. One of the most recent studies, conducted by Zhu et al. (2023), presents comprehensive experiments involving several widely-used LLMs, including XGLM, BLOOMZ, OPT, and ChatGPT, across 202 different directions and 102 languages. All of these studies have demonstrated the significant potential of LLMs in the field of machine translation. Zhu et al. (2023) claimed that in comparison to semantically-selected examples, randomly-chosen examples yield similar translation performance. However, they arrived at this conclusion under the assumption that they utilized a high-quality development set as their candidate pool. Both Agrawal et al. (2022) and Kumar et al. (2023) discovered that randomly selecting examples yields unsatisfactory results when working with larger datasets. This discovery aligns with our own findings, highlighting the importance of designing an improved in-context learning recipe for machine translation.

7. Conclusion

In this paper, we propose a novel example selection approach for LLMs-based machine translation. The approach leverages monotone submodular function maximization to simultaneously consider multiple translational factors for selecting examples. The factors include similarity between examples and inputs on either source side or target side, as well as the diversity within examples. Experiments on various translation tasks show that our submodular-based approach treats the multiple factors in harmony, achieving significant performance improvements over the random selection and robust single-factor baselines. Future work may explore our approach to other NLP tasks and further refine the selection criteria to optimize the performance of LLMs across diverse applications.

8. Acknowledgments

We would like to thank the anonymous reviewers for the helpful comments. This work was supported by National Natural Science Foundation of China (Grant No. 62276179, 62261160648).

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *ArXiv*, abs/2212.02437.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv*, abs/2302.04023.

Jaime G. Carbonell and Jade Goldstein-Stewart. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Satoru Fujishige. 1991. Submodular functions and optimization.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Y. Kawahara, Kiyohito Nagano, and Yoshio Okamoto. 2011. Submodular fractional programming for balanced clustering. *Pattern Recognit. Lett.*, 32:235–243.

Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang goo Lee. 2022. Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator. *ArXiv*, abs/2206.08082.

Katrin Kirchhoff and Jeff A. Bilmes. 2014. Submodularity for data selection in statistical machine translation.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *NMT@ACL*.

Aswanth Kumar, Anoop Kunchukuttan, Ratish Puduppully, and Raj Dabre. 2023. In-context example selection for machine translation using multiple features. *ArXiv*, abs/2305.14105.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Xiaonan Li and Xipeng Qiu. 2023. Finding supporting examples for in-context learning. *ArXiv*, abs/2302.13539.

Hui-Ching Lin and Jeff A. Bilmes. 2011. A class of submodular functions for document summarization. In *Annual Meeting of the Association for Computational Linguistics*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Ves Stoyanov, and Xian Li. 2021. Few-shot learning with multilingual generative language models. In *Conference on Empirical Methods in Natural Language Processing*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? In *Workshop on Knowledge Extraction and Integration for Deep Learning Architectures; Deep Learning Inside Out*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *ArXiv*, abs/2104.08786.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Rose Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir R. Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Annual Meeting of the Association for Computational Linguistics*.

Ricardo Rei, Craig Alan Stewart, Ana C. Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *ArXiv*, abs/2009.09025.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *ArXiv*, abs/2112.08633.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. In *Annual Meeting of the Association for Computational Linguistics*.

Xi Ye, Srini Iyer, Asli Celikyilmaz, Ves Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2022. Complementary explanations for effective in-context learning. *ArXiv*, abs/2211.13892.

Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *ArXiv*, abs/2304.04675.

Appendix

A. Regarding the size of examples selected from BM25

We experimented with various values of top N, specifically $N = \{10, 20, 30, 40, 50, 100, 200\}$, using the BM25 model in Table A1. Our observations indicate that as N increases, the results exhibit a gradual improvement, saturating when N is bigger than 50. So we choose top 50 to balance the efficiency and performance in the experiments.

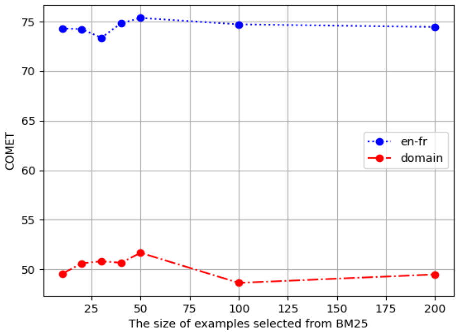

Figure A1: Results with different size of examples selected from BM25

B. Computational Costs

The selection speeds for TopK, MRR, and our method are 10.45, 8.84, and 7.52 sentences per second, respectively on a single A100 GPU. By exploiting parallelism in implementation, we can enhance the overall acceleration, as these factors operate independently. When our method exclusively focuses on surface coverage, the speed increases to 17.26 sentences per second. Considering that TopK and MRR have been seamlessly incorporated into Langchain as standard selection methods, we assert that our method is already practical for real-world applications.

C. Case Study

Table C1 displays selected examples for the translation direction $en \rightarrow fr$. It is evident that while there are a few identical sentences, the majority exhibit distinctions. This variance indicates that different methods consider a range of factors, yielding diverse results.

Source	Dr. Ehud Ur, professor of medicine at Dalhousie University in Halifax, Nova Scotia and chair of the clinical and scientific division of the Canadian Diabetes Association cautioned that the research is still in its early days.
BM25	1: We also join Mr Andrews in sending our condolences to those who are bereaved and suffering as a consequence of the appalling and tragic Swissair disaster just 10 days ago off the coast of Nova Scotia. 2: I salute the President of the Champalimaud Foundation, Dr Leonor Beleza, who has established stringent criteria for marrying excellence in scientific research with clinical practice. 3: Let me tell you that a few days ago, Professor Huber of Vienna University presented his most recent research achievements and findings. 4: Strengthening the role of clinical and scientific research is vital in the fight against tuberculosis.
TopK	1: Because of both its importance and its interest, research into diabetes is set to continue as the subject of sustained attention in the fifth framework programme of research and technological development which we are currently in the process of getting under way. 2: Through its framework research programmes the Commission has supported diabetes research in the past. 3: I should add that I must declare an interest, as my husband is the Chairman of the UK arm of the Juvenile Diabetes Research Foundation, which supports research on type 1 diabetes. 4: I am a lecturer at Granada University and I know that for more than ten years the Granada Faculty of Medicine has had an excellent research team working on these topics and on certain products that are not mentioned in the resolution, products that are to be used in orthodontic treatments.
MRR	1: Because of both its importance and its interest, research into diabetes is set to continue as the subject of sustained attention in the fifth framework programme of research and technological development which we are currently in the process of getting under way. 2: The moratorium introduced in the North Atlantic twelve years ago has not, I regret to say, had the expected effect. The cessation of fishing has had no effect upon the state of cod stocks off Newfoundland and Nova Scotia and in the Gulf of Saint Lawrence. 3: One was conducted under the leadership of Professor van Ark of Groningen University. 4: Despite evidence, based on research by Aberdeen University, that nandrolene could be produced by a combination of dietary supplements and vigorous training, Mark Richardson is still waiting, eight days before the Olympics, in the Olympic village not knowing whether he is going to compete or not.
Our Method	1: I am a lecturer at Granada University and I know that for more than ten years the Granada Faculty of Medicine has had an excellent research team working on these topics and on certain products that are not mentioned in the resolution, products that are to be used in orthodontic treatments. 2: Because of both its importance and its interest, research into diabetes is set to continue as the subject of sustained attention in the fifth framework programme of research and technological development which we are currently in the process of getting under way. 3: I have invited Professor Weissmann to chair an advisory group of scientific experts, whose members are specialists in BSE and Creutzfeldt-Jacob disease. 4: Despite evidence, based on research by Aberdeen University, that nandrolene could be produced by a combination of dietary supplements and vigorous training, Mark Richardson is still waiting, eight days before the Olympics, in the Olympic village not knowing whether he is going to compete or not.
Source	Fourteen schools in Hawaii located on or near coastlines were closed all of Wednesday despite the warnings being lifted.
BM25	1: I am hopeful that with a little goodwill on all sides the ban can be lifted in the very near future. 2: Schools are being closed, teachers are losing their jobs, researchers are finding themselves out on the street and public investments are being cut or left to stagnate. 3: Those articles were drawn up and adopted with culpable negligence despite the warnings I gave in my counter-report on the Treaty of Nice. 4: Only in the particular circumstances of the early 19th century were all restrictions on fishing lifted.
TopK	1: Fourteen human lives were extinguished and other persons were injured on their way to their holidays. 2: The tunnel was closed last Wednesday. 3: In the space of one and a half years, four schools, with a total of 45 classes, were closed, and a further 107 classes are threatened with closure. 4: Mari language schools are being closed down, and education in the Mari language is only allowed in the primary levels of elementary education.
MRR	1: Fourteen human lives were extinguished and other persons were injured on their way to their holidays. 2: Only in the particular circumstances of the early 19th century were all restrictions on fishing lifted. 3: The tunnel was closed last Wednesday. 4: Despite repeated warnings from the international community, Iran continues its efforts in the area of uranium enrichment. COMET:92.43
Our Method	1: Mari language schools are being closed down, and education in the Mari language is only allowed in the primary levels of elementary education. 2: Fourteen human lives were extinguished and other persons were injured on their way to their holidays. 3: Under the pretext of 'optimisation of the school network', national minority schools, including Polish schools, are to be closed in small towns, and only Lithuanian schools are to remain there. 4: Mrs Schreyer, this transfer took place despite explicit warnings from this House.

Table C1: Case study