A Few-shot Learning Approach for Lexical Semantic Change Detection Using GPT-4

Zhengfei Ren, Annalina Caputo, Gareth J. F. Jones,

ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland
zhengfei.ren3@mail.dcu.ie, annalina.caputo@dcu.ie, gareth.jones@dcu.ie

Abstract

Lexical Semantic Change Detection (LSCD) aims to detect language change from a diachronic corpus over time. We can see that over the last two decades there has been a surge in research dealing with the LSC Detection. Recently, a series of methods especially contextualized word embeddings have been widely established to address this task. While several studies have investigated LSCD using large language models (LLMs), an evaluation of prompt engineering techniques, such as few-shot learning with different in-context examples for improving the LSCD performance is required. In this study, we examine the few-shot learning ability of GPT-4 to detect semantic changes in the Chinese language change evaluation dataset ChiWUG. We show that our LLM-based solution improves the GCD evaluation metric on the ChiWUG benchmark compared to the previously top-performing pre-trained system. The result suggests that using GPT-4 with three-shot learning with hand-picked demonstrations achieves the best performance among our different prompts.

1 Introduction

Lexical semantic change detection (LSCD) aims to address the problem of automatically identifying meaning change in target words between the current period and earlier time periods (Kim et al., 2014), (Kulkarni et al., 2015), (Giulianelli et al., 2020) (Schlechtweg et al., 2020). The majority of current work on LSCD uses deep contextualized models, such as BERT (Devlin et al., 2018) or EMLO (Peters et al., 2018), to model the semantics of target words from different time-sliced corpus (Periti and Tahmasebi, 2024) (Kutuzov and Giulianelli, 2020) (Hamilton et al., 2016), (Giulianelli et al., 2020). Semantic change can then be detected by vector similarities between word embeddings using these models.

Recently, Large Language Models (LLMs) have showcased remarkable capabilities in solving natural language processing tasks based on zero-shot predictions (Karjus, 2023), (Karanikolas et al., 2023). Recent work has shown that LLMs can even excel in an wider range of applications with appropriate prompt instructions (Hou et al., 2024), (Marvin et al., 2023), (Chen et al., 2023a). However, current work on the LSCD using LLMs lacks a proper method that uses prompt engineering to build LSCD model, such as using example retrieval algorithm to find the most similar language change context pairs compared to input pairs.

In this paper, we apply prompt engineering on an LSCD task, where few-shot learning using GPT-4 is applied with in-context demonstrations of prompts based on manual selection or machine retrieval algorithms. Our proposed method is systematically tested on a Chinese evaluation dataset ChiWUG (Chen et al., 2023b) following the Diachronic Word Usage Graph (DWUG) annotation and evaluation framework. Our methods serve as an exploratory examination of LLM performance for LSCD with various prompting strategies. This may be applied to other LSCD tasks in different language which also follow the DWUG framework, such as English (EN), German (DE), Swedish (SW) and Latin (LA) (Schlechtweg et al., 2020) and Norwegian(NO) (Kutuzov et al., 2022).

2 Related Work

Lexical semantic change has been evaluated by both static models, such as skip-gram (Kim et al., 2014), (Kulkarni et al., 2015) or contextualized embedding methods, such as BERT (Kutuzov and Giulianelli, 2020), (Giulianelli et al., 2020). To quantitatively evaluate lexical semantic change, Semeval 2020 task 1 defined an evaluation framework for measuring lexical semantic change (Schlechtweg et al., 2020). Two tasks including binary change

August 15, 2024 ©2024 Association for Computational Linguistics

classification and graded change detection (GCD) were developed for evaluating systems seeking to address LSCD. Binary classification simply measures whether the meaning changes or not, while GCD aims to measure the correlation between true scores and change degrees for all the target words. Recent work has shown that LLM models have impressive reasoning and prediction ability on many natural language processing (NLP) including language change detection (Karjus, 2023), (Ziems et al., 2024), (Laskar et al., 2023). Moreover, some evaluations of ChatGPT have been built on a series of NLP tasks including Word Sense Induction (WSI) and Word Sense Disambiguation (WSD) (Laskar et al., 2023).

Meanwhile, one work compared the performances of LLM and pre-trained language models on the shot-term language change dataset TempoWIC and showed that zero-shot GPT-4 achieved superior results (Wang and Choi, 2023). More recently, ChatGPT web interface and the official OpenAI APIs have been evaluated on WSI and LSCD with GCD scores, results show that ChatGPT achieves slightly lower performance than BERT in detecting both long-term and shot-term changes on the HistoWIC dataset and TempoWIC dataset respectively (Periti et al., 2024).

To the best of our knowledge, only one study has employed a series of contextualized models to implement language change detection on all LSCD datasets including the ChiWUG task (Periti and Tahmasebi, 2024). The XL-LEXEME (Cassotti et al., 2023) with the average pairwise distance (APD) performs best among their models. The performance of GPT-4 was comparable to XL-LEXEME on three tasks relevant to LSCD: Word-in-Context (WIC), WSI and GCD task. GPT-4 and XL-LEXEME achieve close to human-level while other contextualized embeddings perform in a low-moderate level, the performance of GPT-4 was only slightly lower than the BERT model. However, their GPT-4 model was only evaluated on an English dataset, and not for any other language dataset for the LSCD task. In our study, we compare our approach to this system using GCD scores on the ChiWUG evaluation dataset.

3 LSCD using LLM

To implement LSCD using LLM, we use the official GPT-4 API to conduct our experiments, other versions of GPT-4 can be found in the OpenAI documentation [1]. Our basic prompt is to predict whether the meaning of a target word changes or not given two context sentences. The task instruction leverages a similar prompt template proposed in (Karjus, 2023). We show a prompt example of a one-shot learning method with this template in Appendix B. In this paper, we propose to use few-shot learning using GPT-4 with different methods to select the demonstration examples for further improvements in performance of LSCD prediction.

下海 *xiahai*
原本在大学担任生物学教授的他, 决定下海创办了一家生物科技公司. A professor of biology in a university decided to set up a biotechnology company
她曾是一名成功的时尚设计师， 后来选择下海，开设了自己的时装品牌 She was as a successful fashion designer before she chose to go to business and start her own fashion brand

Table 1: Our hand-picked context example in one-shot learning with label *Related*.

3.1 Prompt Engineering

To increase the prediction ability, we use the few-shot learning approach to enrich the LLM's representation ability for semantic change. Meanwhile, we set the temperature of the GPT-4 model to zero and to reduce the randomness of the generated language change results to improve the performance.

To construct the in-context example, we first develop our hand-picked examples and then design a method to select an example from the training corpora for providing similar semantic knowledge directly from the ChiWUG dataset and inject it into a prompt. In following subsection, we provide details of the selections of demonstration learning examples using both methods.

3.2 Manual Selection

Our manually selected examples are developed from searching online linguistic resources from the internet containing two context sentences of a target word. We show one of these examples in Table 1. This manually selected example contains a Chinese target word 下海, which means 'go into the sea' or 'to venture'.

[1] https://platform.openai.com/docs

This sample is labeled by the change type *Related*, in which the meanings between the two text inputs are basically similar, but with different background contexts. We suppose such information could improve representation of GPT-4 model for inferring related semantic change types.

3.3 Example with Retrieval

As well as manual selection, we explore selection of demonstration examples by retrieval from a corpus with similar semantic representation with input queries. The retrieval process relies on the Chinese Bert model from the Huggingface [2].

Specifically, the last four hidden state embeddings of the Chinese BERT base model for the target word in two input sentences from two time periods are extracted for constructing the word embedding. For computing the similarity, two context sentences are concatenated to form a single vector representation, then cosine similarity is calculated between the representations for the input context pairs and the sample context pairs from the dataset. Two contexts in the dataset with the highest similarity are used as the retrieved examples to construct the prompt demonstrations. The retrieval corpus was generated from the first 40 sentences among the whole dataset for each word. An example of retrieved and original context pairs is shown in Appendix B with input and retrieved sentence pairs.

Our idea of example retrieval is that the greater the similarity between the input context and the demonstration example, the higher probability that the model will improve the performance, such in-context information could provide LLMs with better representation ability for detecting similar semantic changes.

4 Experiments

In this section, we introduce the dataset used for our experiments, give details of our experiments with results and analyze our findings.

4.1 Dataset

The dataset used for our investigation is ChiWug (Chen et al., 2023b). This consists of 6,100 human semantic relatedness judgments for 40 target words. The ChiWUG dataset follows the DWUG framework for LSCD tasks (Schlechtweg et al., 2021). Moreover, the context pairs are annotated with the relatedness between them with a four-scale degree with 1 to 4 referring to semantic proximity from unrelated to the identical usages. The examples are represented in a DWUG with related semantic relations between target words, figure 1 shows one such word example for a target word 'xiahai'.

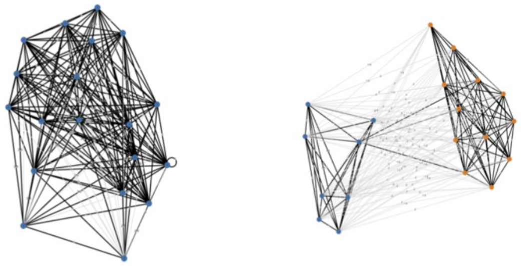

Figure 1: Word Usage Graph for word '下海' (xiahai). Nodes represents the word usages, the edges represent the usage relatedness between word usages (Chen et al., 2023b).

In ChiWUG, the corpora are divided into two sub-corpora, the EARLIER is from 1953 to 1978 and the LATTER is from 1979 to 2003. Three metrics are set within the dataset (Chen et al., 2023b): binary change, Jensen-Shannon Distance (JSD) and COMPARE. Binary change is the same as that used in the Semeval 2020 subtask1 and JSD can be regarded as the graded change scores.

4.2 Evaluations

In our system, we use two metrics to evaluate our method: binary change and the GCD score. To detect the binary change, we label target changed that contain more than 4 labels *unrelated* following similar criteria in (Karjus, 2023).

Moreover, we compute the GCD scores by calculating the Spearman correlation between the sum of all the change scores from 1 to 4 for a target word to ground truth scores. We evaluate these two metrics based on a sample of ChiWUG with solely 40 sentences pairs among 1,560 for each target word, which was shown to be a sufficient number of samples to predict correct change scores.

4.3 Zero-shot vs One-shot vs Few-shot

Zero-shot can be built directly with an initial prompt, where the instruction leveraged prompts used in (Karjus, 2023), a one-shot learning example with the same task introduction of the language change task is shown in the Appendix A.

As shown in Table 2, the three-shot model with hand picked examples shows the best re-

[2] https://huggingface.co/google-bert/bert-base-uncased

Approaches	Binary Change	**GCD**
XL-LEXEME	/	0.73
Zero-shot	0.65	0.65
Two-shot	**0.83**	0.73
Three-shot	0.70	**0.79**
One-shot Retrieval	0.70	0.72

Table 2: GCD predictions with different zero-shot or few-shot settings of GPT-4 models, XL-LEXEME is the previous best-performing model evaluated on ChiWUG (Periti and Tahmasebi, 2024), (Cassotti et al., 2023).

sults for both GCD scores and binary classification among these methods, which also outperforms the current best GCD prediction scores by XL-LEXEME model on ChiWUG benchmark dataset with smaller corpus for change prediction. Results for one-shot and three-shot models demonstrate improvement compared to zero-shot learning and two-shot-learning models. However, we do not see any improvements from two-shot learning compared to the one-shot learning method, the performance of one-shot learning model, with or without the machine selected demonstration example, is shown in the Table 2. Results show that they achieve the same scores, we leave the detailed discussion of this to the next section.

4.4 Discussion

Overall, we can see an upward trend of performance as the number of in-context demonstration examples increases. The three-shot method is better than all other established methods including zero-shot, one-shot and two-shot models. We can also see that few-shot method can benefit from our meticulously selected examples. Moreover, as shown in Table 2, our three-shot learning model outperforms the previously best contextualized word embeddings and achieves a new state-of-art performance on ChiWIG evaluation dataset, two-shot learning model with manually selected examples also shows superior change detection predictions over a pre-trained language model. We infer that few-shot learning with typical semantic change examples can improve LLMs in-context ability for language change detection.

Nevertheless, we get relatively similar results with one-shot learning using a manually selected demonstration example and automatically selected example, although the retrieved example is sharing similar semantic change context with input pairs, this method does not provide any improvements as we expected. We show one such example retrieved from the sample dataset in the Appendixm B to illusrate retrieved contexts and original inputs. Though they are most similar context pairs among our sample dataset according to BERT retrieval, the in-context learning may not improve from this directly. Our manually selected example in one-shot learning may be representative enough to provide semantic changes knowledge for GPT-4. Moreover, results show that two-shot learning with hand-picked examples may not provide further improvements in predicting language change results. This may also be because the quality of the added demonstration examples in two-shot learning may be poorer than other examples.

In the next stage of our work, we will examine different combinations of examples manually selected and retrieved for any improvements in performance. We leave the detailed reasons for the relation between the detection performance and the example similarity with the original query to the future work.

5 Conclusions

Overall, we have demonstrated higher performance of the proposed GPT-4's few-shot learning model on the LSCD task following the Semeval 2020 task 1 evaluation, compared to the previous contextualized embedding model. We tested the effectiveness of few-shot learning with hand-picked examples and the most similar samples from corpora with our retrieval method utlizing BERT. Our model, utilizing three-shot leaning featuring manually selected demonstration examples for semantic change detection, achieves the current highest GCD scores on the ChiWUG evaluation dataset. We show that few-shot learning with representative examples in prompts has the potential to increase the semantic representation ability of the LLM for this task. However, there is no evidence that one-shot learning with example retrieval increases GPT-4's prediction performance on the LSCD task. We leave developing explanations for the effect of retrieval on LSCD performance to future work.

Acknowledgments

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223, and partially as part of the ADAPT Centre at DCU (Grant No.

13/RC/2106_P2) (www.adaptcentre.ie). For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

References

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023a. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv preprint arXiv:2310.14735*.

Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. 2023b. ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 93–99, Singapore. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pages 364–381. Springer.

Nikitas Karanikolas, Eirini Manga, Nikoletta Samaridi, Eleni Tousidou, and Michael Vassilakopoulos. 2023. Large language models versus natural language understanding and generation. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, pages 278–290.

Andres Karjus. 2023. Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence. *arXiv preprint arXiv:2309.14379*.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on world wide web*, pages 625–635.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Ranveig Enstad, and Alexandra Wittemann. 2022. Nordiachange: Diachronic semantic change dataset for norwegian. *arXiv preprint arXiv:2201.05123*.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2023. Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics*, pages 387–402. Springer.

Francesco Periti, Haim Dubossarsky, and Nina Tahmasebi. 2024. (chat)GPT v BERT dawn of justice for semantic change detection. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 420–436, St. Julian's, Malta. Association for Computational Linguistics.

Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of* *the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A large resource of diachronic word usage graphs in four languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ruiyu Wang and Matthew Choi. 2023. Large language models on lexical semantic change detection: An evaluation. *arXiv preprint arXiv:2312.06002*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, pages 1–55.

A Prompts for One-shot Learning

Initial prompt for the the system introduction:

```
You are a expert in multilingual language change detection, determine whether the target word has changed its semantic meaning between given sentences, answer with Same, Related, Linked or Distinct.
```

Prompt for one-shot learning:

```
This is very important to my career. Consider the use of target word in two contexts of sentences, determine whether the target word has changed its semantic meaning between those sentences. Do the refer to the Same, different but Related, distant Linked or unrelated objects.

Determine the meaning change of target word target in following sentences:
1. [sentence1]
2. [sentence2]
Answer: choose from (Same, Related, Linked, Distinct).
Provide your answer without any illustration
```

B Retrieved Example

Target word, 下海 xiahai, `go to the sea or join a business`

Query input:

1. 辽西省在春汛下海时即已组成九十个渔业生产合作社，一百四十五个互助组。
When the spring floods hit the sea, Western Liaoning Province had already formed 90 fishery production cooperatives and 145 mutual aid groups.

2. 福建省沿海各地民兵积极配合人民解放军加强海防巡逻和解放台湾的斗争，并组织了武装护渔队、巡逻队，保卫渔民下海捕鱼。
He militias in various coastal areas of Fujian Province actively cooperated with the People’s Liberation Army in strengthening coastal defense patrols and the struggle to liberate Taiwan, and organized armed fishing teams and patrols to protect fishermen fishing in the sea.

Retrieved sentence:

1 ”他瞅了瞅他现在穿的新皮大氅，又说：“过去我下海、在家，总是穿一件又腥又破的棉短袄；吃呢，一天挣来的钱连啃窝窝头吃都不够......有一次我们有四个人在葫芦岛下潮（出海），半路遇着大风，一个三丈多高的浪头，打翻了我们的小船，其中一个同伴被打下水以后没有踪影了，剩我们三个人在孤岛上冻饿了好几天，好容易才返回来。
He took a look at the new leather cloak he was wearing now, and said: "In the past, when I went to the sea and at home, I always wore a fishy and torn cotton jacket. When it came to eating, the money I earned in a day even cost me a lot of money. Not enough... One time, four of us went out to sea in Huludao. We encountered strong winds on the way, and a wave more than three feet high capsized our boat. One of our companions was knocked into the water and disappeared without a trace. The three of us froze and starved on the isolated island for several days, and finally returned with great difficulty

2. 他们说：“只要能治好唐山亲人的伤病，别说上山捉毒蛇，就是下海擒蛟龙，俺们也在所不辞。
’They said: "As long as we can cure the injuries and illnesses of our relatives in Tangshan, we will do whatever we can to catch venomous snakes in the mountains or go to the sea to catch dragons."