Idiom Corpora Construction via Large Language Models

Gülşen Eryiğit and **Doğukan Arslan**
ITU NLP
Dep. of Artificial Intelligence & Data Eng.,
Istanbul Technical University
gulsen.cebiroglu@itu.edu.tr
arslan.dogukan@itu.edu.tr

This research aims to provide insights into the feasibility of using large language models for efficient and inclusive idiom corpus construction across multiple languages. In addressing this issue, the methodology involves employing large language models to create sentence samples containing idioms. These samples form the basis for training an idiom detection model, with the final phase evaluating trained models against a set of gold-labeled test data. As a future study we plan to extend idiom generation to additional languages. The primary objective is to assess the effectiveness of artificially generated datasets by training state-of-the-art models on idiom detection tasks.

Relevant UniDive working groups: WG1, WG3

1 Introduction

An idiom can be described as a linguistic construct where the collective meaning is distinct and cannot be inferred directly from the meanings of its constituent words. Due to this distinctive composition, idioms negatively affect the performance of models in various tasks, such as machine translation (Isabelle et al., 2017). Traditional idiom annotation approaches (Cook et al., 2008), relying on the annotation of natural text, suffer from unbalanced distributions of idiomatic and nonidiomatic samples, lack of diversity in terms of surface forms and the data scarcity problems. Alternative data collection and annotation approaches have been proposed such as crowd-creating (Eryiğit et al., 2023) and human-collecting (Tayyar Madabushi et al., 2022).

A number of datasets are proposed focusing on sentence samples that include idioms, as can be seen in Table 2. Most of the existing datasets contain only English idioms and lack sufficiently diverse examples. This deficiency is attributed to the high cost and extensive time required for data labeling. In this work, we introduce an ongoing study investigating an economical and efficient alternative to human labeling: large language models. This study primarily concentrates on rapidly generating idiomatic instances that are inclusive and applicable to a variety of languages, utilizing large language models. The outcomes of this approach will be benchmarked against existing datasets. Consequently, this research aims to evaluate whether the corpora produced by these large language models are as effective as those generated through human labeling, in teaching idioms to the language models.

2 Methodology

As of now, sentence samples featuring idioms were generated specifically for the Turkish language, using ChatGPT4. A two-step methodology was employed for this generation process. Initially, the language model was prompted with the idiom itself and asked about the various contexts in which the idiom could be appropriately used. Subsequently, for each identified meaning, ChatGPT4 was tasked with creating distinct sentences both for literal and figurative usage of the idiom, further enriching these sentences by diverse grammatical structures (i.e., declarative-interrogatory, affirmative-negative, short-long sentences) (Table 1). For 36 distinct idioms from Dodiom dataset (Eryiğit et al., 2023), a total of 7200 sentence samples were generated, with each idiom having 200 individual sentence examples in Turkish.

Later, to evaluate performance of the generated dataset in an automatic idiom identification system, a sequence labeling task is defined as assigning labels to each token in a sentence such as idiom (I), literal (L), and other (O). On par with many recent studies within the literature (Eryiğit et al., 2023; Saxena and Paul, 2020; Ehren et al., 2020), a BiLSTM-CRF model is trained using the generated ChatGPT4 dataset. To test this model, 20% of a human-generated and labeled dataset from Eryiğit et al. (2023) is utilized and the remaining part of

Prompt #1	'[DEYİM]' bir Türkçe deyimidir. Bu deyimi hem gerçek hem de mecaz anlamlarında kullanabiliriz. Lütfen bu deyimin mecaz anlamda kullanıldığı durumları listeleyin. *'[DEYYIM]' is a Turkish idiom. We can use this idiom both literally and figuratively. Please list the cases where this idiom is used figuratively.*
Prompt #2	Belirtilen deyimi, farklı bağlamlarda ve nüanslarda kullanarak yukarıdaki her bir kategori için dört farklı cümle oluşturun. Bütün cümleler mecaz anlamı yansıtmalıdır. İlk cümle kısa ve öz, ikinci cümle uzun ve detaylı, üçüncü cümle soru formunda ve dördüncü cümle ise olumsuz bir yapıda olmalıdır. Deyimin kelime köklerini değiştirmeyerek. Deyim: '[DEYİM]' *Create four different sentences for each category above using the given idiom in different contexts and nuances. All sentences should reflect the figurative meaning. The first sentence should be short and concise, the second sentence long and detailed, the third sentence in the form of a question and the fourth sentence in a negative form. Keeping the lemmas of the idiom unchanged. Idiom: '[IDIOM]'*

Table 1: Examples of prompts used to generate sentence samples containing idioms from ChatGPT4.

this dataset is used to train another model with the same architecture for comparison. The final phase includes evaluating the performance of both these trained models against the set of gold-labeled test data. Results can be seen in Table 3.

The findings indicate that while data generated using ChatGPT4 does not match the efficacy of human-generated and labeled data in training idiom identification systems, the proposed approach still holds promise. Additionally, we observed that the produced sentences consist of mostly uniform idiom surface forms with adjacent idiom components.

Training Dataset	**Macro-Avg. F1 Score**
ChatGPT4	0.65
Dodiom	0.75

Table 3: Macro-Avg. F1 Scores of models tested on Dodiom test split.

3 Future Work

Future works of this study will focus on extending the idiom generation to additional languages. The efficacy of these artificially generated datasets will be assessed by training state-of-the-art models on idiom identification task to measure their performance. Also, the study will explore enhancements in data generation processes, particularly focusing on the application of prompt engineering techniques, with the aim of improving the quality of the synthetically generated data. For example, the prompts may be enhanced similar to Eryiğit et al. (2023) where the crowd is encouraged to also produce sentences containing idioms composed of non-adjacent components (i.e., other words intervene between the components of the idiom).

Dataset	**# of sentences**	**# of idioms**	**Language**
AStitchInLanguageModels (Tayyar Madabushi et al., 2021)	6430	336	Multilingual
Dodiom (Eryiğit et al., 2023)	12706	73	Multilingual
EPIE (Saxena and Paul, 2020)	25206	717	English
Gigaword (Li and Sporleder, 2009)	3964	17	English
ID10M$_{gold}$ (Tedeschi et al., 2022)	800	470	Multilingual
ID10M$_{silver}$ (Tedeschi et al., 2022)	262781	10118	Multilingual
IDIX (Sporleder et al., 2010)	5836	78	English
MAGPIE (Haagsma et al., 2020)	56622	1756	English
OpenWME Japanese (Hashimoto and Kawahara, 2009)	102856	146	Japanese
PARSEME (Savary et al., 2015)	274376	13755	Multilingual
PIE (Haagsma et al., 2019)	2239	591	English
SemEval-2013 Task 5b (Korkontzelos et al., 2013)	4350	65	English
SemEval-2022 Task 2 (Tayyar Madabushi et al., 2022)	8683	50	Multilingual
TroFi (Birke and Sarkar, 2006)	1298	25	English
VNC-Tokens (Cook et al., 2008)	2984	53	English

Table 2: Some corpora containing idiom samples.

References

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vnc-tokens dataset.

Rafael Ehren, Timm Lichte, Laura Kallmeyer, and Jakub Waszczuk. 2020. Supervised disambiguation of german verbal idioms with a bilstm architecture. In *FIGLANG*.

Gülşen Eryiğit, Ali Şentaş, and Johanna Monti. 2023. Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering*, 29(4):909–941.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. Magpie: A large corpus of potentially idiomatic expressions. In *International Conference on Language Resources and Evaluation*.

Hessel Haagsma, Malvina Nissim, and Johan Bos. 2019. Casting a wide net: Robust extraction of potentially idiomatic expressions. *ArXiv*, abs/1911.08829.

Chikara Hashimoto and Daisuke Kawahara. 2009. Compilation of an idiom example database for supervised idiom identification. *Language Resources and Evaluation*, 43:355–384.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. Semeval-2013 task 5: Evaluating phrasal semantics. In *International Workshop on Semantic Evaluation*.

Linlin Li and Caroline Sporleder. 2009. Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 315–323, Singapore. Association for Computational Linguistics.

Agata Savary, Manfred Sailer, Yannick Parmentier, Mike Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, M. Constant, Petya N. Osenova, and Federico Sangati. 2015. Parseme – parsing and multiword expressions within a european multilingual network.

P. Saxena and Soma Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions. *ArXiv*, abs/2006.09479.

Caroline Sporleder, Linlin Li, Philip John Gorinski, and Xaver Koch. 2010. Idioms in context: The idix corpus. In *International Conference on Language Resources and Evaluation*.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. Id10m: Idiom identification in 10 languages. In *NAACL-HLT*.