2nd General Meeting
University of Naples L'Orientale Naples, Italy, 8-9 February 2024
https://unidive.lisn.upsaclay.fr/

IDIOM CORPORA CONSTRUCTION VIA LARGE LANGUAGE MODELS

DOĞUKAN ARSLAN, GÜLŞEN ERYİĞİT
{arslan.dogukan, gulsen.cebiroglu}@itu.edu.tr
Istanbul Technical University, Department of AI & Data Engineering

WG1, WG3

BACKGROUND

High cost and extensive **time requirement** in data labelling for idiom corpora construction.

↓

Datasets contain mostly English idioms and **lack** sufficiently **diverse examples**.

↓

Large Language Models?

OBJECTIVES

Rapidly **generating idiomatic instances** that are inclusive and applicable to a variety of languages, **utilizing large language models**.

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.

METHODOLOGY

Two step approach:

i) **GPT4** was prompted with the idiom itself and **asked about the various contexts** in which the idiom could be appropriately used.

ii) For each identified meaning, **GPT4** 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).

Sample data generation flow:

Prompt: "[IDIOM] is a Turkish idiom. We can use this idiom both literally and figuratively. Please list the cases where this idiom is used figuratively."

Answer: "Here are some figurative uses of the idiom.."

Prompt: 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 uchanged. Idiom: [IDIOM].

↓

	Total # of sentences	Total # of idioms
🇹🇷 Turkish	7,200	36
🇮🇹 Italian	7,400	37

EVALUATION

To evaluate, **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 [1].

A **BiLSTM-CRF** architecture and Dodiom [1] datasets are used for evaluation purpose.

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

RESULTS

Experiments;

i. For Italian and Turkish, **20% of the Dodiom** datasets were reserved for testing purposes and a model was trained with the rest (Dodiom$_{TR}$ and Dodiom$_{IT}$). Similarly, a model was trained with the data generated with GPT4 (GPT4$_{TR}$ and GPT4$_{IT}$). These two models were tested with the reserved dataset.

Dataset	Macro-Avg. F1
Dodiom$_{TR}$	0.79
GPT4$_{TR}$	0.69
Dodiom$_{IT}$	0.76
GPT4$_{IT}$	0.73

ii. Next, GPT3.5 and GPT4 models are tested with the **entire Dodiom** datasets for **a classification task** which is an idiomacity identification task, and compared with a BiLSTM-CRF model trained with GPT4 datasets.

Model / Dataset	Dodiom$_{TR}$	Dodiom$_{IT}$
GPT4	0.66	0.69
GPT3.5	0.35	0.46
BiLSTM+CRF	0.64	0.62

FUTURE WORK

i. Extending the idiom generation to **additional languages**.

ii. Applying additional **prompt engineering techniques**, with the aim of improving the quality of the synthetically generated data.

ITU NLP NATURAL LANGUAGE PROCESSING RESEARCH GROUP
İTÜ
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY
UniDive
Funded by the European Union