Idiom Identification

Jacob Marano and Dan Weil

Given a dataset of excerpts with idioms, how can a large language model, or other form of natural language processing system, identify and label an idiom?

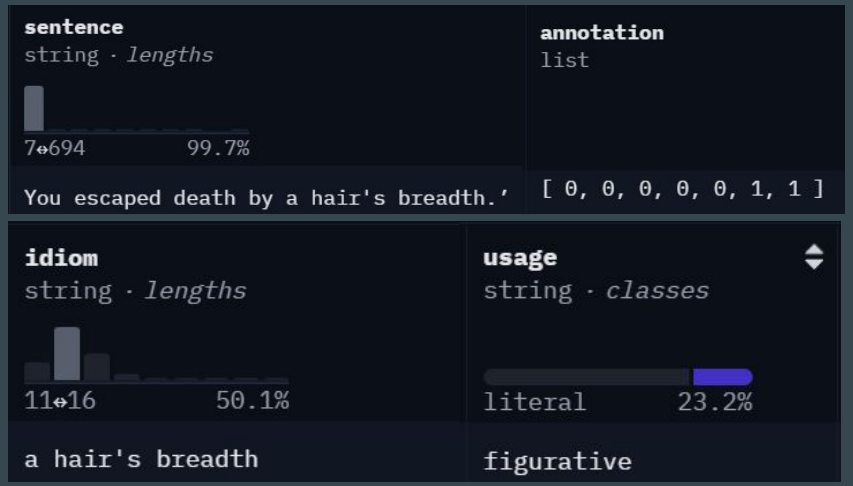
The model

3 Step Process

Preprocessing -> Training -> Evaluation

MAGPIE Preprocessing

- Problem: gsarti/magpie contains only one split and we need the traditional 3 splits
- Solution: Randomly split the dataset into an 80 train/10 dev/10 test split
- Tools:
  - Python 3.11
  - datasets (3.4.0) to load and process the data
  - os to save the data to disk




```python
# Random split into 80/10/10
train_test_split = train_data.train_test_split(test_size=0.2, seed=67)
train_ds = train_test_split['train']

dev_test_split = train_test_split['test'].train_test_split(test_size=0.5, seed=67)
dev_ds = dev_test_split['train']
test_ds = dev_test_split['test']
```

Fine Tuning

- Fine Tuning method:
  - Fine-tuning bert-base-uncased for token classification
  - Binary classification task
  - Supervised learning using labeled annotations from the dataset
  - 3 epochs with learning rate of 2e-5
  - Batch size of 8
  - Cross-entropy loss function
- Tools:
  - Hugging Face Transformers Library
  - PyTorch (CUDA enabled)
  - NVIDIA RTX 4070 Ti SUPER -
- Problems:
  - Processing power - takes 34 minutes to fine tune one model on 3 epochs
  - This issue will propagate with potential Voting Classifier approach
- Solutions:
  - Automated training script to run overnight
  - Contact RIT Research Computing

```
# Training arguments
training_args = TrainingArguments(
    output_dir="./results",
    eval_strategy="epoch",
    learning_rate=2e-5,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=3,
    weight_decay=0.01,
    save_strategy="epoch",
    logging_dir="./logs",
    logging_steps=100,
)

# Trainer
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_ds,
    eval_dataset=dev_ds,
    data_collator=data_collator,
)
```

Evaluation

- Each model was evaluated on its respective test split
- Average Model Performance:
  - 0.9806666667
- Potential Issues:
  - Potential overfitting
- Potential Solutions:
  - Find a larger dataset
  - Voting Classifier approach
  - Embed Encode Predict Model

```
Enter sentence: I am feeling under the weather

Sentence: I am feeling under the weather

Token predictions:
i: 0
am: 0
feeling: 0
under: 1
the: 0
weather: 1

Detected idiom tokens: under weather
```

```
Enter sentence: A black cat in a coal cellar

Sentence: A black cat in a coal cellar

Token predictions:
a: 0
black: 1
cat: 1
in: 1
a: 0
coal: 0
cellar: 1

Detected idiom tokens: black cat in cellar
```

Model A	Token Level Accuracy		0.9816
	Percision	Recall	F1-Score
Non-Idiom	0.99	0.99	0.99
Idiom	0.91	0.91	0.91
Accuracy			0.98
Macro Avg	0.95	0.95	0.95
Weigthed Avg	0.98	0.98	0.98
Model B	Token Level Accuracy		0.9802
	Percision	Recall	F1-Score
Non-Idiom	0.99	0.99	0.99
Idiom	0.90	0.90	0.90
Accuracy			0.98
Macro Avg	0.95	0.95	0.95
Weigthed Avg	0.98	0.98	0.98
Model C	Token Level Accuracy		0.9802
	Percision	Recall	F1-Score
Non-Idiom	0.99	0.99	0.99
Idiom	0.90	0.90	0.90
Accuracy			0.98
Macro Avg	0.95	0.94	0.94
Weigthed Avg	0.98	0.98	0.98

The interface

Idiom Identifier

Idiom Search Idiom Dictionary

speak of the devil

Speak of the devil means the person you were avoiding suddenly appears just as you were about to do something. The idiom dates back to the 17th century, originating from the metaphorical use of "devil" to denote a troublesome person. Check the idiom dictionary for more information.

Using another model to provide definitions/context

kick the bucket

"Kick the bucket" means to die. It is a colloquial expression indicating the end of life. This phrase likely originated in the 18th or 19th century, with the bucket symbolizing death due to historical associations with burial practices. Check the idiom dictionary for more information.

kick the bucket

This idiom means to die. People use it casually to say someone is going to pass away. It comes from 19th-century slang where "bucket" was part of informal expressions for death, though the exact origin is unclear. Check the idiom dictionary for more information.

The phrase first appeared in print in the “Dictionary of the Vulgar Tongue” in 1785. Its origin is unclear, though there are several theories. One common theory is of hanging, when a person standing on a bucket with a noose over the head kicks the bucket and hence, dies. There is no evidence to support this claim, and it appears rather implausible. Another, more plausible, theory refers to the archaic meaning of the word bucket, which used to mean beam in 16th century England. A bucket, or beam, was used to hang animals by the feet for slaughter, and they would kick it while dying. A third theory refers to the Catholic practice of placing the holy water bucket at the feet of a person who has died, so that visitors could sprinkle the holy water on the body.

Using another model to provide definitions/context

Red cent

That phrase **is not an idiom**—it refers to a specific historical British coin (a copper penny minted from 1809–1817 with a red copper color). There is no figurative or idiomatic meaning associated with "red cent" in English usage. Check the idiom dictionary for more information.

```
The idiom "we've been gypped" means that a group of people has been tricked, scammed, or defrauded. It originates from African American Vernacular English (AAVE) and is commonly used in informal US English to describe being a victim of deception.

This idiom is **not** considered culturally offensive in English-speaking contexts.
```

Next steps

For the class

- Evaluate other model performances
- Automate the preprocess, train, and evaluate process and run over the weekend
- Connecting the model to the interface
- Implement different loss function

$$C_{\mathcal{S}} = \frac{1}{N} \sum_{w_i, w_j \in \mathcal{S}, i \neq j} \text{sim}(V(w_i), V(w_j))$$

For the future

- Find a dataset of idiom definitions + acceptability
  - Or find a different LLM
- Create versions of this for other languages
  - Use mBERT for classification

Questions, Comments, Concerns?