Idiom Identification

Jacob Marano & Dan Weil
Group 7

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?

Keywords: Idiom, Large Language Model, Information Extraction, Classification, Span Identification

Idioms: An approach to identifying major pitfalls for learners

Research Questions:

1. When teaching a language, should idioms be taught separately or should they just be dealt with as they turn up in texts?
2. Should the idioms that are taught be selected based on frequency?
3. Are certain idioms only appropriate in certain contexts and how does this affect their teaching?
4. How do language learners feel about idioms?
5. What factors influence the confusion caused by idioms?

Relevance:

- While it can be useful to teach idioms separately, it is more practical to address them as they come up
- Language learners do not have the cultural context necessary to effectively identify or guess what certain idioms mean
- Often less skilled language learners will try to guess the meaning of a phrase without recognizing that it is an idiom
    - If there is a plausible literal meaning
    - If all of the words are familiar
    - If the words are unique to the idiom

Alan Cornell, 1999

The Impact of Context on Learning Idioms in EFL Classes

Research Questions:

1. Will the participants who learn idioms with the most context do better on immediate tests than participants who learn idioms with lower context and no context?
2. Will the participants who learn idioms with the most context do better on later tests than participants who learn idioms with lower context and no context?

Method:

- Three groups of upper intermediate English students were taught English idioms with different levels of context
- These groups were tested on their comprehension of the idioms immediately after and two weeks after learning them

Relevance:

- Having greater context when learning idioms in a foreign language improves a person’s ability to understand and remember them

Fatemeh Mohamadi 2013

Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation.

Contributions:

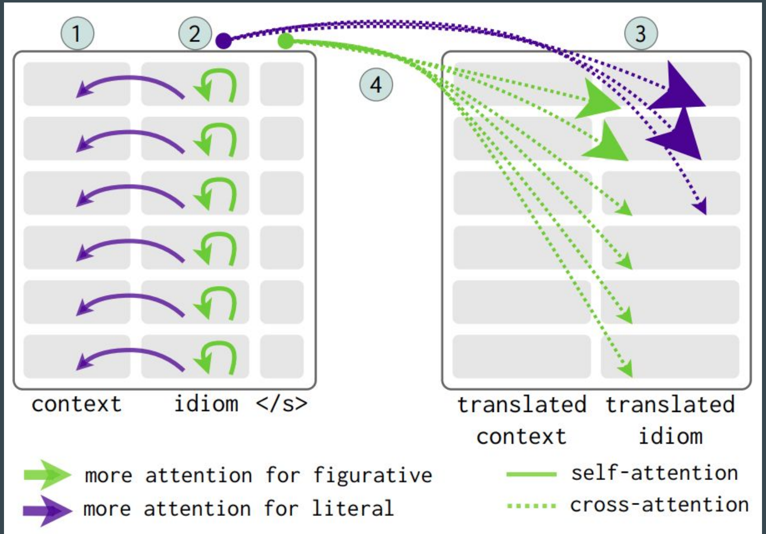
- NMT Transformer translates idioms too compositionally
- >= 76% of figurative cases were translated literally

Method:

- Seven European language pairs with English as the source language
- MAGPIE corpus (labeled as non-literal or literal)
- Translate MAGPIE to target language, see if the translation is literal or paraphrased

Relevance:

- While transformers are the go to solution for MT, we may need to consider another method or usage involving transformers



Attention patterns of figurative PIEs

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022

Leveraging Three Types of Embeddings from Masked Language Models in Idiom Token Classification (1/2)

Contributions:

1. Contextualized token embeddings
2. Uncontextualized token embedding
3. Masked token embedding

Method:

- English dataset: VNC-Tokens
- Japanese dataset: OpenMWE Corpus
- Evaluate performance of embeddings generated from an English and Japanese model against a uncontextualized baseline

Relevance:

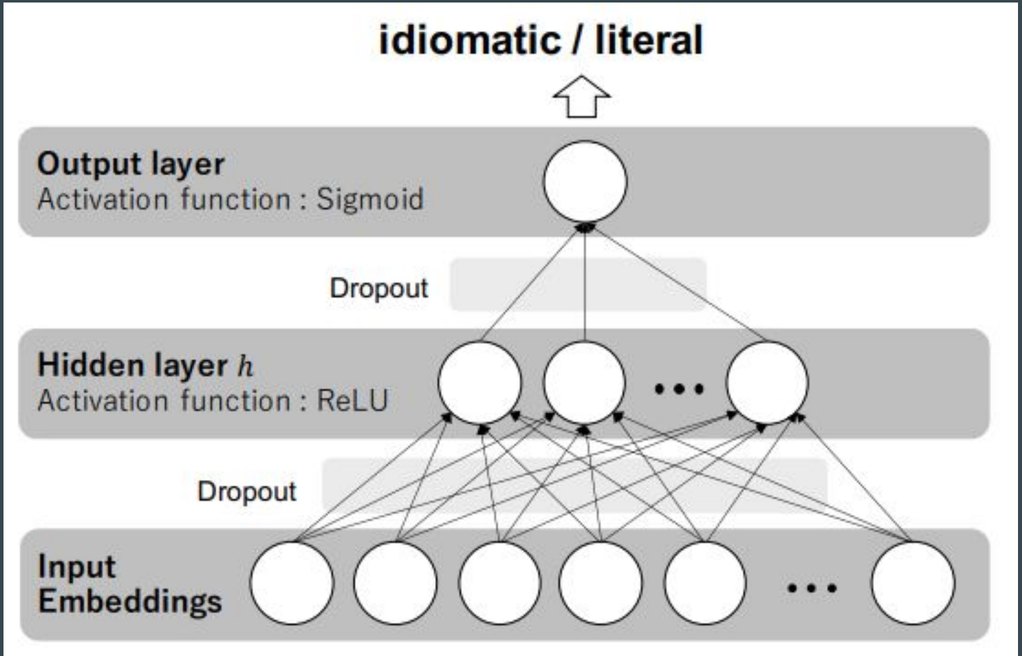
- Uncontextualized token embeddings and masked token embeddings improve idiom token *classification* in a zero-shot setting

Embeddings	English	Japanese
$\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{N}}$	0.840	0.823
$\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{V}\_t}; \boldsymbol{v}_{\mathrm{N}}; \boldsymbol{v}_{\mathrm{N}\_t}$	0.859	0.842
$\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{V}\_m}; \boldsymbol{v}_{\mathrm{N}}; \boldsymbol{v}_{\mathrm{N}\_m}$	0.852	0.829
$\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{V}\_t}; \boldsymbol{v}_{\mathrm{V}\_m}; \boldsymbol{v}_{\mathrm{N}}; \boldsymbol{v}_{\mathrm{N}\_t}; \boldsymbol{v}_{\mathrm{N}\_m}$	**0.865**	**0.847**

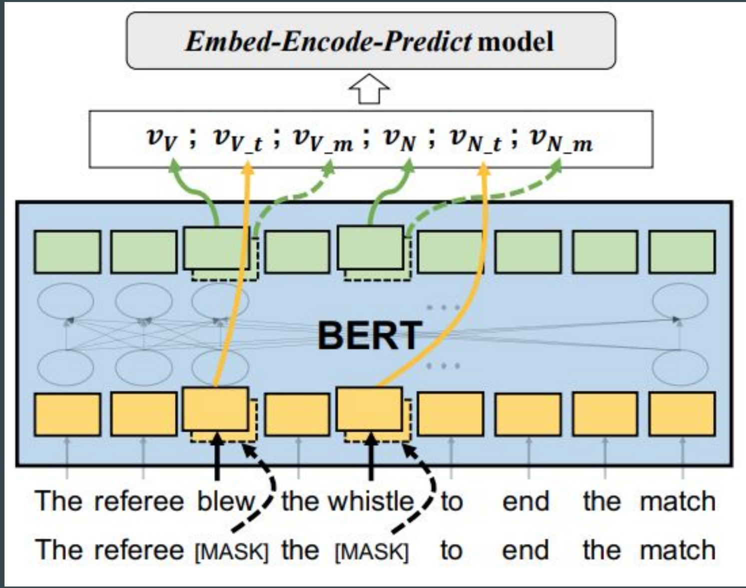
Macro-averaged accuracy for different combinations of input embeddings.

Takahashi, R., Sasano, R., & Takeda, K. (2022)

Leveraging Three Types of Embeddings from Masked Language Models in Idiom Token Classification (2/2)



The Embed-Encode-Predict Model



Masked embeddings

Takahashi, R., Sasano, R., & Takeda, K. (2022)

BERT-based Idiom Identification using Language Translation and Word Cohesion(1/2)

Contributions:

1. A new loss function using language translation and word cohesion

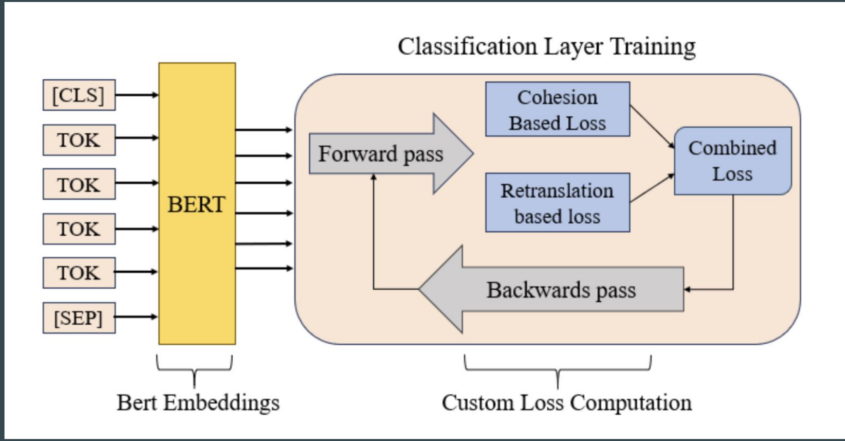
Method:

- Tested with bert-based-uncased
- Language translation metric-METEOR score
- Word cohesion metric-

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

Relevance:

- Incorporation of language translation and word cohesion into the loss function improves accuracy



Architecture of the model

Arnav Yayavaram, Siddharth Yayavaram, Prajna Devi Upadhyay, and Apurba Das. 2024.

BERT-based Idiom Identification using Language Translation and Word Cohesion(2/2)

MAGPIE	Regular Cross Entropy Loss	98.74
	Translation Retranslation Loss	98.76
	Cohesion based Loss	98.76
	Combination	**98.8**
VNC	Regular Cross Entropy Loss	99.43
	Translation Retranslation Loss	99.66
	Cohesion based Loss	99.62
	Combination	**99.7**

theidioms	Regular Cross Entropy Loss	95.73
	Translation Retranslation Loss	97.5
	Cohesion based Loss	**97.62**
	Combination	97.61
formal	Regular Cross Entropy Loss	97.83
	Translation Retranslation Loss	98.75
	Cohesion based Loss	98.67
	Combination	**98.83**

gtrans	Regular Cross Entropy Loss	92.61
	Translation Retranslation Loss	**94.91**
	Cohesion based Loss	94.87
	Combination	94.79
gpt&gtrans	Regular Cross Entropy Loss	96.07
	Translation Retranslation Loss	97.23
	Cohesion based Loss	**97.25**
	Combination	97.22
theidioms 1-1	Regular Cross Entropy Loss	89.44
	Translation Retranslation Loss	90.56
	Cohesion based Loss	90.75
	Combination	**90.85**

Accuracy score for each type of model and each dataset

Arnav Yayavaram, Siddharth Yayavaram, Prajna Devi Upadhyay, and Apurba Das. 2024.

Our Model & Methods

Dataset - MAGPIE

- Sense-annotated corpus of potentially English idiomatic expressions
- 44.5k samples of 1,482 idioms
- Includes: **the sentence** , **annotation** , **idiom** , **usage** , variant, and pos_tags
- Problem:
    - Dataset only has a train split
- Solution:
    - 80/10/10 split

sentence string · *lengths* 7 6.87k	annotation list	idiom string · *lengths* 6 47	usage string · *classes* 2 values
And she had an incoherent sense…	[0, 0, 0, 0,… {...}	across the board	literal
Similar signs of progress and…	[0, 0, 0, 0,… {...}	across the board	figurative
An increase in P is across the…	[0, 0, 0, 0,… {...}	across the board	figurative
There's always a demand for jokes…	[0, 0, 0, 0,… {...}	across the board	figurative
While sovereign credit quality ha…	[0, 0, 0, 0,… {...}	across the board	figurative
‘ Are acts of God designed to show…	[0, 0, 1, 1,… {...}	act of god	figurative

Example Data

The old man **kicked the bucket**.

[0, 0, 0, **1**, **1**, **1**]

Figurative

I **kicked the bucket** over.

[0, 1, 1, **1**, **0**]

Literal

Road Map

Training

- bert-based-uncased
- Embed-Encode-Predict
  - uncontextualized and masked embeddings
- MAGPIE 80/10/10 split
- Train three separate models on three independently random splits

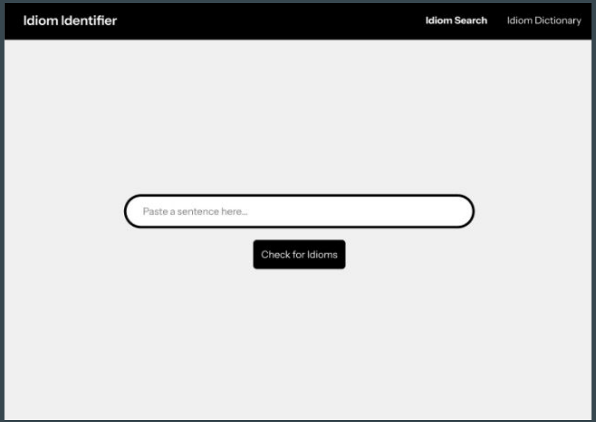
Evaluation

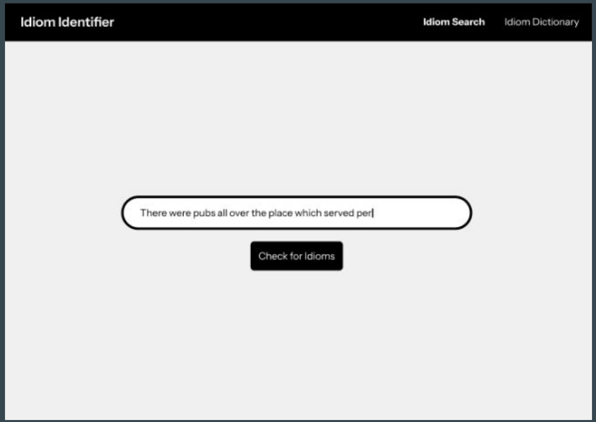
- Run our model and two other models on each test split
- Compare results with golden data to get a "score"

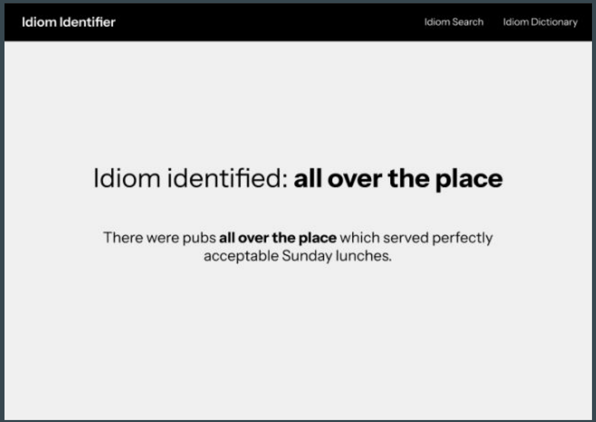
Metrics

- Score of our model(s)
- Score of baseline models
- Goal: > 90% average score across all three of out models

Interface


Idiom Identifier
Idiom Search
Idiom Dictionary
Paste a sentence here...
Check for Idioms


Idiom Identifier
Idiom Search
Idiom Dictionary
There were pubs all over the place which served per
Check for Idioms


Idiom Identifier
Idiom Search
Idiom Dictionary
Idiom identified: **all over the place**
There were pubs **all over the place** which served perfectly acceptable Sunday lunches.

Expected Contribution

- We are combining strategies found in BERT-based Idiom Identification using Language Translation and Word Cohesion (BERT) and Leveraging Three Types of Embeddings from Masked Language Models in Idiom Token Classification (MLM).
- We are also creating a website that will support English language learners in idiom acquisition.

Risks

1. If used for educational purposes, it is important for our models to be highly accurate
   a. Mitigation plan: Hyperparameters, embedding strategy, and training epochs can be tweaked to increase model performance.
2. Some idioms may be antiquated and derived from hate speech
   a. Mitigation plan: TBD, potential idea is to flag these idioms in our interface. We may need another dataset.

Any Questions?

We are all ears