LREC-COLING 2024

# BERT-based Idiom Identification using Language Translation and Word Cohesion

Arnav Yayavaram, Siddharth Yayavaram,
**Prajna Upadhyay**, Apurba Das

BITS Pilani Hyderabad, India

# Idioms

- Multi-word expression (MWE).
- Figurative meaning – contextual.
- Hard to deduce from literal interpretations.

# Challenges with Idioms

- Alters default semantic roles of syntactic categories.
- This poses significant challenges for Natural Language Processing (NLP) systems.
- Expression covers literal uses as well – Potentially idiomatic expressions (PIE).

# Technical Challenges

- Automatically detect if an idiomatic expression is present in a sentence.

- If yes, identify the idiomatic tokens.

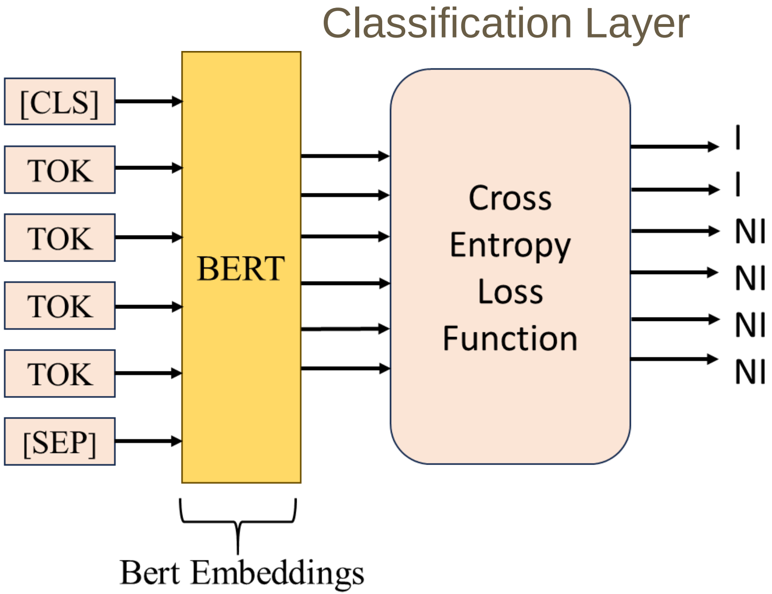
  - *Oh – for about four years, **on and off**, he said vaguely. (Figurative)*
  - *Participate in training, both **on and off** station. (Literal)*

# Problem Statement

- **Input**:
  - A sentence S with n tokens $w_1, w_2, \ldots, w_n$ where each wi represents a tokenized unit. S is a syntactic ordering over $w_i$'s.
  - Labels L = {I, NI} where
    - **I** represents **idiom.**
    - **NI** represents **not idiom.**
- **Output:** Generate a sequence of class labels $Z = z_1, z_2, \ldots, z_n$ where $\mathbf{z_i = f(w_i)}$**.**
- **Objective:** Learn the function **f().**

# Deep Learning for Idiom detection

- Fine-tune BERT model on the idiom detection task.
- Cross Entropy Loss Function ( $\mathcal{L}$ )



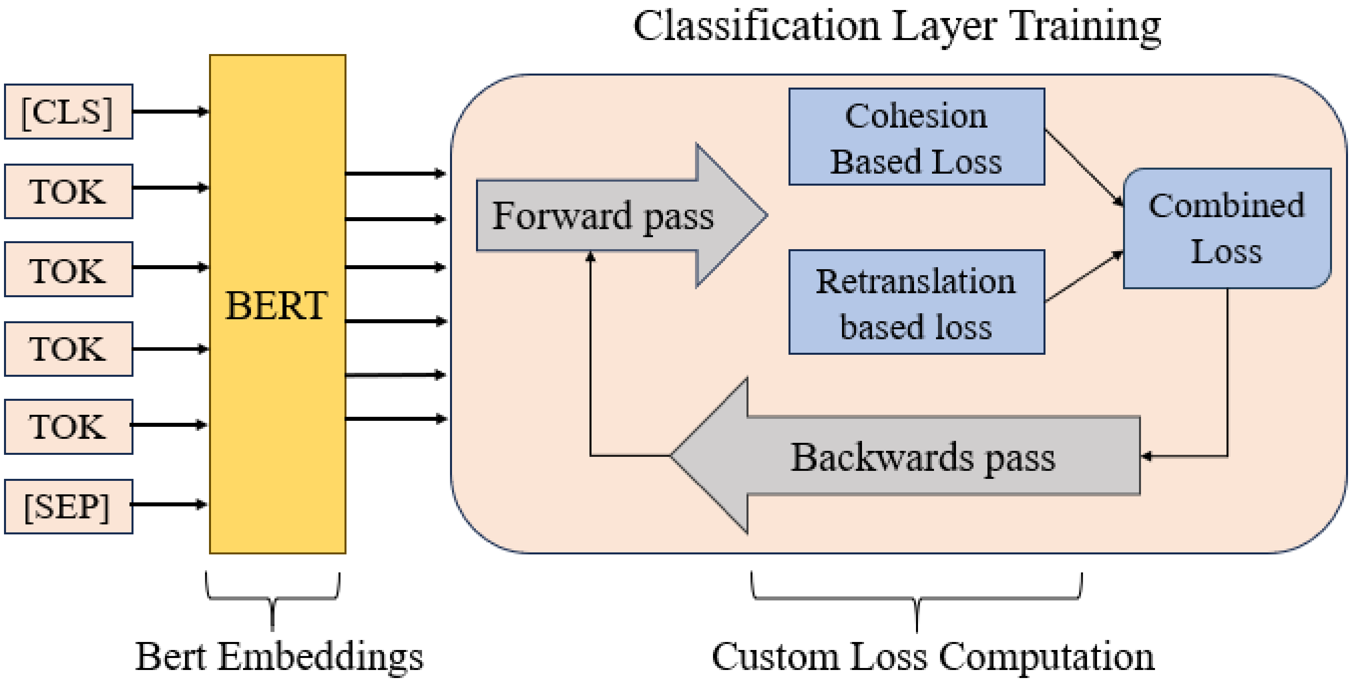
# Deep Learning for Idiom detection: Challenges

- Idiom Token Classification suffers from class imbalance
  - Number of **I** tokens far lesser than number of **NI** tokens

| **Dataset** | **I** | **NI** |
|---|---|---|
| EPIE Formal | 10767 | 70522 |
| EPIE Static | 63012 | 583923 |

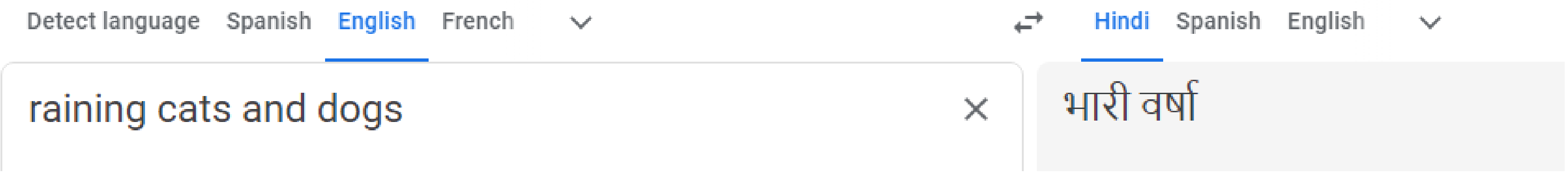
- Cross Entropy not suitable to handle class imbalance
  - Poor accuracy on **I** labels

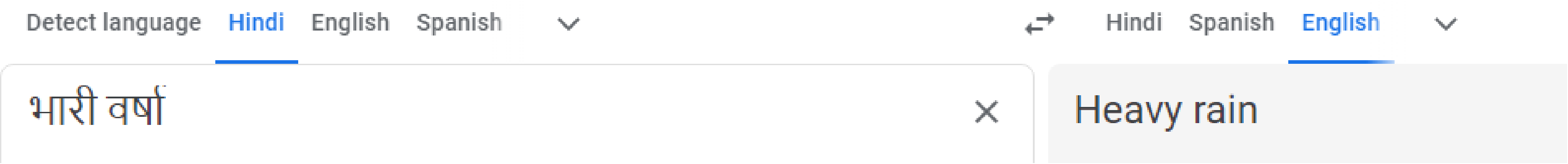
# Our solution: Custom Loss Function



LREC-COLING 2024

# Translation based Loss Function



LREC-COLING 2024

MWE

# Translation based Loss Function

- $S_{L1}$ is the original sentence in L1 **(English).**
- $S_{L1 \rightarrow L2}$ is a translation of $S_{L1}$ in L2 **(Hindi).**
- $S_{L1 \Leftrightarrow L2}$ is a translation of $S_{L1 \rightarrow L2}$ back to L1.
- The retranslation will be substantially different from each other with the presence of an idiom.
- We capture this difference through **METEOR*** score *MS*.
- Idea: ***MS*(S$_{L1}$, S$_{L1 \Leftrightarrow L2}$)** is low when $S_{L1}$ is highly likely to contain idiom

$$\mathcal{L}_{retranslation} = \mathcal{L}(1 + \lambda_1 \mathbb{1}(\mathcal{MS} < \lambda_2))$$


*Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*

# Cohesion based Loss Function

- **Idea**: *Semantic compositionality** among the words in an idiom is low .
- We capture this idea with cohesion score $C_S$ of a sentence S where

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

- $C_{S1}$: cohesion score after removing the idiom from S.
- $C_{S2}$: cohesion score without removing the idiom from S.
- **Our loss function**:

$$\mathcal{L}_{cohesion} = \mathcal{L}(1 + \lambda_3 \mathbb{1}(|\mathcal{C}_{S_1} - \mathcal{C}_{S_2}| > \lambda_4))$$

*Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. Handbook of natural language processing

# Final Loss

Linear combination of **Translation Loss** and **Cohesion Loss**

$$\mathcal{L}_{final} = \tau_1 \mathcal{L}_{retranslation} + \tau_2 \mathcal{L}_{cohesion}$$

$\tau_1$ $\tau_2$ control the effect of both losses

# Experiments

- **bert-base-uncased**
- Training: **80%**, Validation: **10%**, Test: **10%**
- Learning rate: **2e-5**
- Epochs: **3**
- λ1, λ2, λ3, λ4: **999**
- $T_1$ and $T_2$ : **0.01**

# Datasets

| Dataset | total number of sentences | #idioms | #sentences containing idioms | average sentences per idiom |
|---|---|---|---|---|
| MAGPIE | 36192 | 1727 | 27727 | 16.05 |
| VNC-Tokens | 2571 | 48 | 2111 | 43.97 |
| theidioms | 7380 | 1606 | 7830 | 4.87 |
| formal | 3136 | 358 | 3136 | 8.76 |
| gtrans | 440 | 22 | 440 | 20 |
| gpt+gtrans | 880 | 22 | 440 | 20 |
| theidioms 1-1 | 1606 | 1606 | 1606 | 1 |

# Metrics

Precision, Recall, F1

Macro Average

Weighted Average

$$P = \frac{\sum_{i=1}^{N}(TP_i + FP_i) \times P_i}{\sum_{i=1}^{N}(TP_i + FP_i)}$$

$$R = \frac{\sum_{i=1}^{N}(TP_i + FN_i) \times R_i}{\sum_{i=1}^{N}(TP_i + FN_i)}$$

$$F1 - \text{score} = \frac{\sum_{i=1}^{N}(2 \times P_i \times R_i) \times (TP_i + FN_i)}{\sum_{i=1}^{N}(P_i + R_i) \times (TP_i + FN_i)}$$

Where **P**: Precision; $\mathbf{P_i}$: Precision of the $i^{th}$ example; **R**: Recall; $\mathbf{R_i}$: Recall of the $i^{th}$ example; **N**: Number of classes ($2$ in our case); $\mathbf{TP_i}$: True Positives for class $i$; $\mathbf{FP_i}$: False Positives for class $i$; $\mathbf{FN_i}$: False Negatives for class $i$; $\mathbf{TN_i}$: True Negatives for class $i$.

Sequence Accuracy

LREC-COLING 2024

# Results

| | | Precision | | | Recall | | | F1 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | Method | Precision | Precision Macro Avg | Precision Weighted Avg | Recall | Recall Macro Avg | Recall Weighted Avg | F1 | F1 Macro Average | F1 Weighted Average | Accuracy |
| gtrans | Regular Cross Entropy Loss | [85.93,93.87] | 89.9 | 92.38 | [72.39,97.27] | 84.83 | 92.61 | [78.54,95.53] | 87.04 | 92.36 | 92.61 |
| | Translation Retranslation Loss | **[86.94,96.71]** | **91.83** | **94.89** | [85.68,**97.03**] | **91.36** | **94.91** | **[86.30,96.87]** | **91.59** | **94.9** | **94.91** |
| | Cohesion based Loss | [86.76,**96.71**] | 91.74 | 94.85 | [**85.69**,96.99] | 91.33 | 94.87 | [86.21,96.85] | 91.53 | 94.86 | 94.87 |
| | Combination | [86.86,96.58] | 91.72 | 94.76 | [85.07,**97.03**] | 91.05 | 94.79 | [85.94,96.80] | 91.38 | 94.77 | 94.79 |
| gpt&gtrans | Regular Cross Entropy Loss | [80.4,97.84] | 89.12 | 96.09 | [80.79,97.78] | 89.29 | 96.06 | [80.53,97.81] | 89.17 | 96.07 | 96.07 |
| | Translation Retranslation Loss | [83.91,98.85] | **91.38** | 97.34 | [89.83,98.06] | 93.94 | 97.23 | [86.74,98.45] | 92.59 | 97.27 | 97.23 |
| | Cohesion based Loss | [83.05,**99.02**] | 91.03 | **97.41** | [**91.37**,97.91] | **94.62** | **97.25** | **[86.99,98.46]** | **92.73** | **97.3** | **97.25** |
| | Combination | [**83.97**,98.83] | 91.4 | 97.33 | [89.64,**98.08**] | 93.86 | 97.23 | [86.70,98.45] | 92.58 | 97.26 | 97.22 |

# Results

| | | Precision | | | Recall | | | F1 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | Method | Precision | Precision Macro Avg | Precision Weighted Avg | Recall | Recall Macro Avg | Recall Weighted Avg | F1 | F1 Macro Average | F1 Weighted Average | Accuracy |
| theidioms | Regular Cross Entropy Loss | [86.61,95.33] | 92.07 | 95.75 | [87.37,97.37] | 92.36 | 95.73 | [86.98,97.45] | 92.21 | 95.74 | 95.73 |
| | Translation Retranslation Loss | [91.60,98.68] | 95.13 | 97.52 | [93.24,98.33] | 95.78 | 97.5 | [92.40,98.50] | 95.45 | 97.51 | 97.5 |
| | Cohesion based Loss | [91.62,**98.83**] | 95.22 | **97.65** | [**94.03**,98.32] | **96.17** | **97.62** | **[92.8,98.57]** | **95.69** | **97.63** | **97.62** |
| | Combination | [**91.76**,98.77] | **95.26** | 97.63 | [93.73,**98.36**] | 96.05 | 97.61 | [92.73,98.56] | 95.65 | 97.61 | 97.61 |
| formal | Regular Cross Entropy Loss | [90.04,99.18] | 94.6 | 97.89 | [95.02,98.29] | 96.65 | 97.82 | [92.46,98.73] | 95.59 | 97.84 | 97.83 |
| | Translation Retranslation Loss | [93.34,99.69] | 96.52 | 98.8 | [98.13,98.86] | 98.49 | 98.76 | [95.68,99.27] | 97.48 | 98.77 | 98.75 |
| | Cohesion based Loss | [92.47,**99.75**] | 96.11 | 98.73 | [**98.51**,98.69] | **98.6** | 98.67 | [95.39,99.22] | 97.31 | 98.68 | 98.67 |
| | Combination | [**93.71**,99.70] | **96.71** | **98.87** | [98.22,**98.92**] | 98.57 | **98.82** | **[95.92,99.31]** | **97.61** | **98.84** | **98.83** |

LREC-COLING 2024

# Results

| | | Precision | | | Recall | | | F1 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | **Method** | **Precision** | **Precision Macro Avg** | **Precision Weighted Avg** | **Recall** | **Recall Macro Avg** | **Recall Weighted Avg** | **F1** | **F1 Macro Average** | **F1 Weighted Average** | **Accuracy** |
| MAGPIE | Regular Cross Entropy Loss | [94.1,99.27] | 96.68 | 98.74 | [93.64,99.32] | 96.48 | 98.74 | [93.87,99.3] | 96.58 | 98.74 | 98.74 |
| | Translation Retranslation Loss | [93.96,**99.31**] | 96.64 | 98.76 | [**93.99**,99.31] | **96.65** | 98.76 | [93.98,99.31] | 96.64 | 98.76 | 98.76 |
| | Cohesion based Loss | [94.22,99.28] | 96.75 | 98.76 | [93.77,99.34] | 96.55 | 98.76 | [93.99,99.31] | 96.65 | 98.76 | 98.76 |
| | Combination | [**94.5**,99.29] | **96.89** | **98.79** | [93.78,**99.37**] | 96.58 | **98.8** | [**94.14,99.33**] | **96.73** | **98.79** | **98.8** |
| VNC | Regular Cross Entropy Loss | [97.19,99.64] | 98.41 | 99.43 | [96.14,99.74] | 97.94 | 99.43 | [96.66,99.69] | 98.17 | 99.43 | 99.43 |
| | Translation Retranslation Loss | [97.99,**99.81**] | 98.9 | 99.66 | [**97.99**,99.81] | 98.9 | 99.66 | [97.99,99.81] | 98.9 | 99.66 | 99.66 |
| | Cohesion based Loss | [98.13,99.76] | 98.94 | 99.62 | [97.37,99.83] | 98.6 | 99.62 | [97.75,99.79] | 98.77 | 99.62 | 99.62 |
| | Combination | **[98.45,99.81]** | **99.13** | **99.7** | **[97.99,99.86]** | **98.92** | **99.7** | **[98.22,99.83]** | **99.03** | **99.7** | **99.7** |

# Cross - Domain Results

| | | Precision | | | Recall | | | F1 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Train, Test** | **Method** | **Precision** | **Precision Macro Avg** | **Precision Weighted Avg** | **Recall** | **Recall Macro Avg** | **Recall Weighted Avg** | **F1** | **F1 Macro Average** | **F1 Weighted Average** | **Accuracy** |
| `theidioms, gtrans` | Regular Cross Entropy Loss | [84.73,96.41] | 90.39 | 94.11 | [84.78,96.29] | 90.54 | 94.1 | [84.57,96.35] | 90.46 | 94.11 | 94.1 |
| | Translation Retranslation Loss | [**89.3**,98.21] | 93.76 | 96.51 | [92.46,**97.39**] | 94.93 | 96.45 | [90.85,97.8] | 94.33 | 96.48 | 96.45 |
| | Cohesion based Loss | [**89.3**,**98.39**] | **93.84** | **96.65** | [**93.21**,97.37] | **95.29** | **96.57** | [**91.21**,**97.87**] | **94.54** | **96.6** | **96.57** |
| | Combination | [89.20,97.97] | 93.59 | 96.3 | [91.42,**97.39**] | 94.4 | 96.25 | [90.28,97.68] | 93.98 | 96.27 | 96.25 |

# Comparison with DISC

| Dataset | Method | Sequence Accuracy |
|---|---|---|
| MAGPIE | Regular Cross Entropy Loss | 90.19 |
| | Translation Retranslation Loss | 91.31 |
| | Cohesion based Loss | 91.46 |
| | Combination | **91.51** |
| | DISC[3] | 87.47 |
| VNC | Regular Cross Entropy Loss | 93.75 |
| | Translation Retranslation Loss | **96.88** |
| | Cohesion based Loss | **96.88** |
| | Combination | **96.88** |
| | DISC | 93.31 |

LREC-COLING 2024

# Discussion

| Error Type | Sentence with PIE | Prediction |
|---|---|---|
| Multiple Expressions Predicted | I then walked across to the photographers and lost my temper and then *lost my head*. | lost my temper , lost my head |
| Alternate expression detected | Cantona will have to *kick his heels* on the sidelines if the manager had his way. | had his way |
| Extra tokens surrounding expression | Julia had her *attention caught* by the commotion. | attention caught by |
| Partial | His blistering turn of speed and attitude *made him an instant hit* with the fans. | hit |
| Predicting Nothing | Everyone talks about *hitting a wall* at the 24 mile mark. | Empty String |

Table 5: Different error types along with examples and the incorrect prediction. The ground truth values have been colored blue in sentences.

# Conclusion and Future Work

- We can adjust our own loss functions to refine different architectures effectively
- Systematically address each identified error category.
- Intuitive and efficient tool utilizing these fine-tuned models to detect an idiom in a given sentence

# Thank You