PIE: A Parallel Idiomatic Expression Corpus for Idiomatic Sentence Generation and Paraphrasing

Jianing Zhou, Hongyu Gong and Suma Bhat
University of Illinois, Urbana-Champaign
Outline

- Introduction
- Dataset
- Experiments
Introduction
Introduction

Hey Siri, I’m under the weather today. Can you help?

I feel sick today.

- Non-compositional
- Stylistic enhancement
- A ubiquitous part of daily language and social communication
Joe, being one who is here today and gone tomorrow, stayed the night, ate some food and left early next morning.

Joe, being the bird of passage he is, stayed the night, ate some food and left early next morning.

Use of IE$s often conveys stylistic enhancement and makes language more natural
Due to non-compositionality, Idiomatic Expressions (IEs) are difficult to process in NLP.
• Motivation

1. Computer-aided style enhancement

2. Combination of identification and paraphrasing as a preprocessing step
Joe, being one who is here today and gone tomorrow, stayed the night, ate some food and left early next morning.

Idiomatic Sentence Generation (ISG)

Joe, being the bird of passage he is, stayed the night, ate some food and left early next morning.
Vote them out!

Remove them from office!
Introduction

**Contribution**

- Idiomatic Sentence Generation (ISG)
- Idiomatic Sentence Paraphrasing (ISP)

**PIE**

**Benchmarks**

- Most of models are limited
- BART is not enough
- Automatic Evaluation is limited
Task Definition
Idiomatic Sentence Generation

- Transform a literal sentence without idioms into a sentence with idiomatic expressions

Joe, being one who is here today and gone tomorrow, stayed the night, ate some food and left early next morning.

Joe, being the bird of passage he is, stayed the night, ate some food and left early next morning.

- Context surrounding the literal phrase will be retained
- Only literal phrase will be replaced with IE
Idiomatic Sentence Paraphrasing

• Transform idiomatic expressions within a sentence into literal expressions.

Joe, being the bird of passage he is, stayed the night, ate some food and left early next morning.

Joe, being one who is here today and gone tomorrow, stayed the night, ate some food and left early next morning.

• Most context surrounding IEs will be retained
• Only IEs will be replaced with literal counterparts
• Semantic simplification instead of syntactic or lexical simplification
Dataset
| Idiom       | Tick Off                                      |
|-------------|----------------------------------------------|
| Sense       | to complete an item on a list                |
|             | to make someone angry or offended           |
| Idiomatic Sentence | I would like to **tick off** some more items on my list before going home |
|             | My decision is going to **tick off** my entire family. |
| Idiomatic Labels | O O O O B I O O O O O O O O O          |
|             | O O O O B I O O O O                      |
| Literal Sentence | I would like to **cross out** some more items on my list before going home |
|             | My decision is going to **anger** my entire family. |
| Literal Labels | O O O O B I O O O O O O O O O         |
|             | O O O O B O O O.                         |

- IEs
- Definitions
- Sentences with IEs
- Position information about IEs
- Sentences with literal expressions
- Position information about literal expressions
• Idiom-oriented
• Longer sentences
• Most contexts same but target parts different

Pose challenges to the text generation models performing the proposed tasks.

| Statistics      | # of instances | Avg. # of words |
|-----------------|----------------|-----------------|
| Idioms          | 823            | 3.2             |
| Sense           | 862            | 7.9             |
| Idiomatic sent  | 5170           | 19.0            |
| Literal sent    | 5170           | 18.5            |
Experiments
Baselines

- Both new tasks are text generation tasks

Translation Models
- Seq2Seq-LSTM
- Transformer

Copy Models
- Seq2Seq-LSTM + Copy
- Transformer + Copy

BART

Pipeline Model
1. Retrieve: Select Idiom
2. Delete: Remove idiom or literal phrase
3. Generate: Insert literal phrase or idiom

- High similarity
- Much context is unchanged

- State-of-the-art performance of BART

- Our tasks are similar to paraphrasing and style transfer
- Good performance of pipeline model
Automatic Evaluation:
- ROUGE, BLEU, METEOR
- SARI
- Perplexity and GRUEN

Human Evaluation:
1. Context Preservation
2. Target Inclusion
3. Fluency
4. Overall
Results & Analysis
Results

- **ISG & ISP:**
  - Copy-enriched transformer, pretrained BART and Pipeline model
  - BART model achieved best performance in BLEU and GRUEN
  - Pipeline model achieved best performance in SARI

- High scores are due to the high similarity between idiomatic sentences and literal sentences

| Model                  | BLEU   | SARI   | GRUEN  |
|------------------------|--------|--------|--------|
|                        | ISP    | ISG    | ISP    | ISG    | ISP    | ISG    |
| Transformer with copy  | 59.56  | 57.91  | 39.93  | 45.10  | 59.27  | 52.25  |
| Pretrained BART        | **79.32** | **78.53** | 62.30  | 61.82  | **77.49** | **78.03** |
| Pipeline               | 65.56  | 70.03  | **67.64** | **62.45** | 67.27  | 74.16  |
Results

• Limitation

  • Mainly measure overlapping tokens, some synonymous idioms or literal phrases are ignored while they are still appropriate

  • IEs are non-compositional, which will have high perplexity scores
Results

- Overall inter-annotator agreement score is 0.76

- Human Evaluation:
  - BART best-performing model in Context Preservation and Fluency
  - Pipeline best-performing model in Target Inclusion and Overall

| Model                | Context ISP | Context ISG | Target ISP | Target ISG | Fluency ISP | Fluency ISG | Overall ISP | Overall ISG |
|----------------------|-------------|-------------|------------|------------|-------------|-------------|-------------|-------------|
| Transformer with copy| 5.4         | 5.3         | 1.2        | 1.6        | 4.6         | 4.6         | 3.9         | 4.2         |
| Pretrained BART      | **5.9**     | **5.9**     | 1.5        | 2.1        | **5.9**     | **5.9**     | 4.4         | 5.0         |
| Pipeline             | 5.6         | 5.8         | **1.7**    | **2.2**    | 5.1         | 5.3         | **4.5**     | **5.1**     |
Limitation of Automatic Evaluation

- Correlation scores between automatic evaluation and human judgements are not high enough.

| Corr | Context | Target | Fluency | Overall |
|------|---------|--------|---------|---------|
|      | ISP   | ISG   | ISP   | ISG   | ISP  | ISG  | ISP  | ISG  |
| BLEU | 0.27  | 0.17  | 0.56  | 0.28  | 0.09 | 0.02 | 0.64 | 0.29 |
| SARI | 0.21  | 0.17  | 0.61  | 0.40  | -0.02 | -0.01 | 0.61 | 0.39 |
| GRUEN| -0.18 | -0.07 | -0.11 | 0.12  | 0.23 | 0.15 | -0.18 | 0.11 |
• Copy mechanism benefit from the ability of explicitly retaining the context

• BART and Pipeline model greatly outperform other baselines
  • BART leveraged large pretrained corpora.
  • Pipeline model utilized the selective idiomatic expression information.

| Model                  | BLEU ISP | BLEU ISG | SARI ISP | SARI ISG | GRUEN ISP | GRUEN ISG |
|------------------------|----------|----------|----------|----------|-----------|-----------|
| Seq2Seq                | 25.16    | 42.96    | 24.13    | 33.89    | 32.25     | 33.45     |
| Seq2Seq with copy      | 38.02    | 47.58    | 43.02    | 49.69    | 27.79     | 32.84     |
| Transformer            | 45.58    | 46.65    | 36.67    | 38.62    | 44.05     | 44.06     |
| Transformer with copy  | 59.56    | 57.91    | 39.93    | 45.10    | 59.27     | 52.25     |
| Pretrained BART        | 79.32    | 78.53    | 62.30    | 61.82    | 77.49     | 78.03     |
| Pipeline               | 65.56    | 70.03    | 67.64    | 62.45    | 67.27     | 74.16     |
Model Comparison

- **BART:**
  - Better at retaining the contexts surrounding idiomatic expressions
  - Better at copying instead of inserting idioms or literal paraphrases.

- **Pipeline:**
  - Because of retrieve and delete stage, pipeline model is better at including target idioms or literal phrases.
  - Suffer from error propagation.

| Attribute               | low non-compositionality |
|-------------------------|--------------------------|
| Literal sentence        | Finding the ruins of Babylon was the archaeologist ’s greatest find . |
| Reference               | Finding the ruins of Babylon was the archaeologist ’s treasure trove . |
| Seq2Seq                 | Missing the aftermath of pouring down the cake ’s share of the city . |
| Transformer             | catching up with silver lining of the challenges ’s volatility . |
| Transformer with copy   | finding the ruins of unk was the ’s ’s trove . |
| Seq2Seq with copy       | finding the ruins of babylon was the archaeologist ’s greatest silver spoons . |
| Transformer with copy   | |
| Pretrained BART         | Finding the ruins of Babylon was the archaeologist’s greatest find . |
| Pipeline                | Finding the ruins of babylon was the archaeologist’ treasure trove . |
Summary

- Idiomatic Sentence Generation (ISG)
- Idiomatic Sentence Paraphrasing (ISP)

**PIE**

**Benchmarks**

- Most of models are limited
- BART is not enough
- Automatic Evaluation is limited
