Shedding Light on Software Engineering-specific Metaphors and Idioms

Mia Mohammad Imran
Virginia Commonwealth University
Richmond, Virginia, USA
imranm3@vcu.edu

Preetha Chatterjee
Drexel University
Philadelphia, Pennsylvania, USA
preetha.chatterjee@drexel.edu

Kostadin Damevski
Virginia Commonwealth University
Richmond, Virginia, USA
kdamevski@vcu.edu

Proceedings of the IEEE/ACM 46th International Conference on Software Engineering
Article No.: 207, Pages 1 - 13

発表者：松尾 和輝

背景

- ソフトウェア工学（SE）では，さまざまな独特の比喩表現が使用される
  - garbage code
  - “Hi @[USER], thanks for pinging me on this issue.”
- 比喩表現が，言語理解系のタスクに影響を与えることが示唆されている[8][9][10][11]
  - 特に感情分析
- どの程度影響を与えるかは明らかにされていない

[8] N. Novielli, F. Calefato, F. Lanubile, and A. Serebrenik, “Assessment of offthe-shelf se-specific sentiment analysis tools: An extended replication study,” Empirical Softw. Engg., vol. 26, no. 4, jul 2021.
[9] N. Novielli, F. Calefato, D. Dongiovanni, D. Girardi, and F. Lanubile, “Can we use se-specific sentiment analysis tools in a cross-platform setting?” in 2020 IEEE/ACM 17th International Conference on MSR. USA: IEEE Computer Society, may 2020, pp. 158–168.
[10] M. M. Imran, Y. Jain, P. Chatterjee, and K. Damevski, “Data augmentation for improving emotion recognition in software engineering communication,” in Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering. USA: ACM, 2023.
[11] Z. Chen, Y. Cao, H. Yao, X. Lu, X. Peng, H. Mei, and X. Liu, “Emoji-powered sentiment and emotion detection from software developers’ communication data,” ACM TOSEM, vol. 30, no. 2, pp. 1–48, 2021.

目的・リサーチクエスチョン

- 目的
  - 比喩表現がLLMに与える影響の調査

- リサーチクエスチョン
  - RQ1：既存のLLMはSEで使用される比喩表現をどの程度解釈できるか？
  - RQ2：SEに特化した感情分析の性能は，比喩表現に対する理解を深めることで向上するか？
  - RQ3：比喩表現をよりよく理解することで，感情が役割を果たすSEの自動化を促進できるか？

データセット

skylot/jadx
laravel/laravel
microsoft/PowerToys
rails/rails
redis/redis
facebook/react
tensorflow/tensorflow
huggingface/transformers
microsoft/vscode

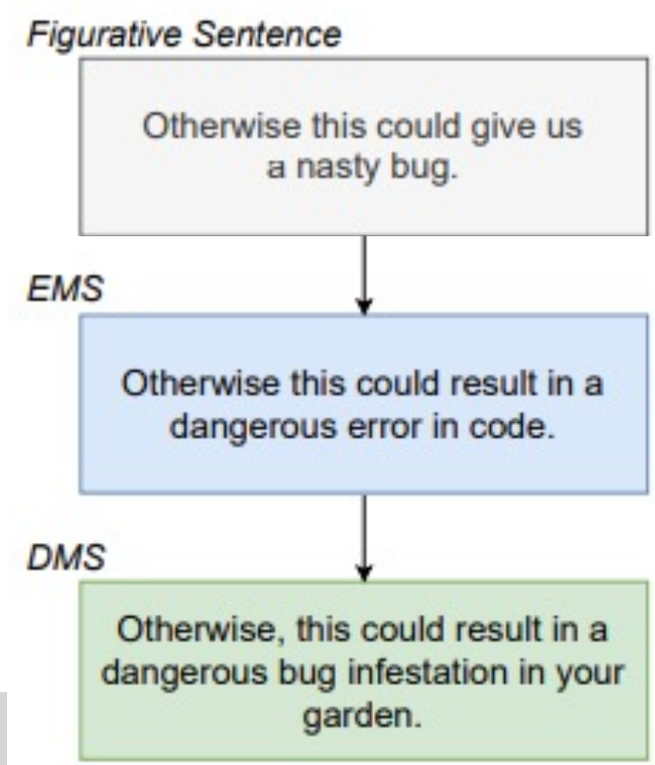
- データ収集
  - 9つの人気のあるGitHubリポジトリからそれぞれ10,000件のコメントを収集
  - 5単語以上で構成された202,000件の文章を取得
  - SEに特化した感情分析ツール[39]を使用して中立的な文章を除外
  - 一般的なメタファー検出[23]とイディオム検出ツール[24]を適用して，それらを含まない文章を除外
  - 残りの文章からそれぞれ1,000文ずつランダムに選択
  - 計2,000文のデータセットを取得

データセット

- アノテーション
  - 2人の大学院生と2人の学部生により実施
  - メタファー： Pragglejaz Group[42]によるMIPガイドラインに基づき分類
  - イディオム：辞書や検索エンジンで調べる
  - SE特有の比喩表現は，FDAやGoogleのソフトウェア工学用語集をアノテーターに提供
  - 2,000文のうち比喩表現を含む1,661文を特定
    - 1,741件のユニークな比喩表現
    - 445件はSE特有，1,296件は一般的な表現

データセット

- 言い換え文の構築（EMSとDMS）
  - EMS：比喩表現を意味が同じ用語に置き換えた文章
    - アノテーターに作成してもらう
  - DMS：元の文とは意味が大きく異なるよう変更した文章
    - タイプ1：比喩表現を文字通りの単語に置き換える
    - タイプ2：比喩表現とその文脈を異なる単語で置き換える
  - ChatGPTを用いて各タイプにつき2つ生成し，最も良いものを選択
  - 1,661文全てでEMSとDMSを生成



予備調査

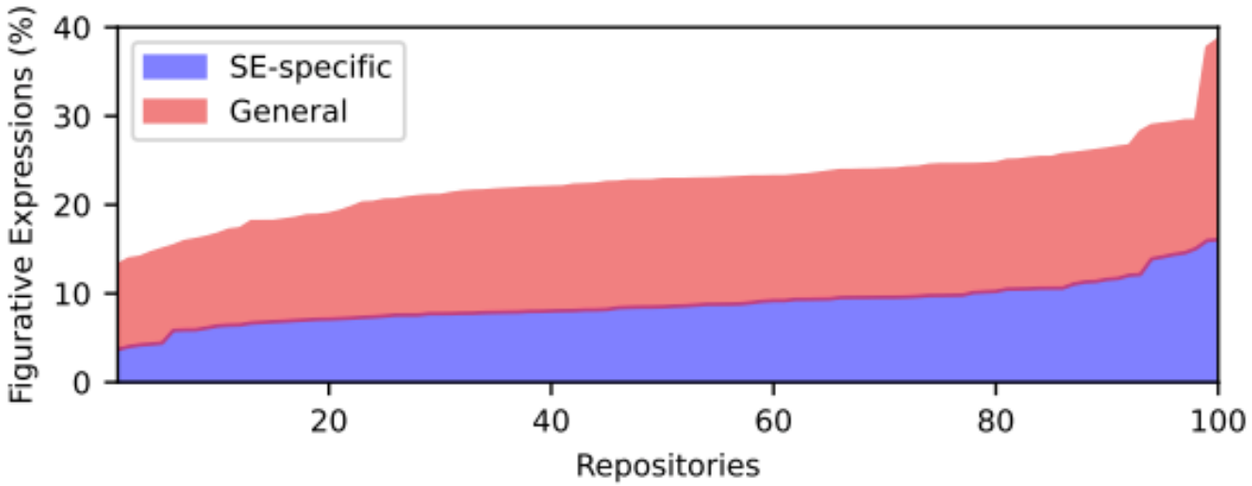
- 目的
  - SE特有の比喩表現がどの程度頻出するのか明らかにする
- アプローチ
  - GitHub上の文章を用いて，比喩表現の出現頻度を調査

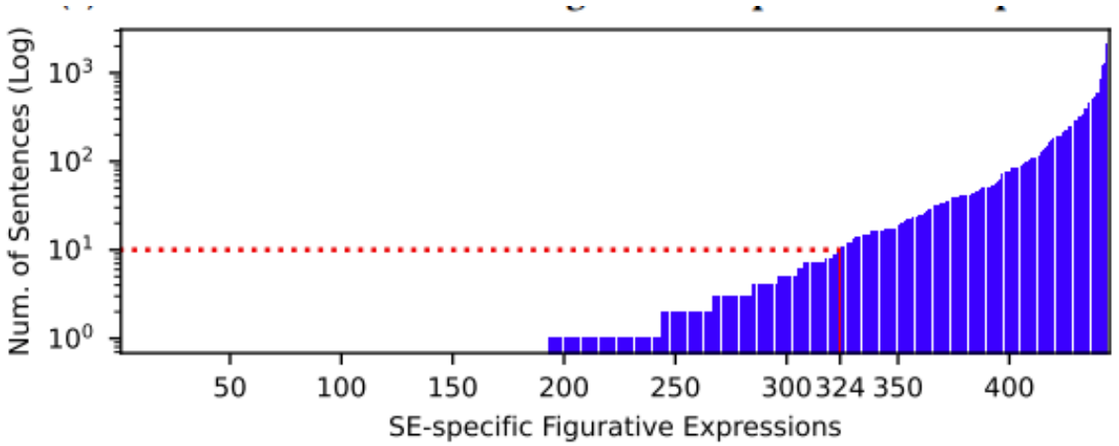
- データセット
  - GitHubのスター数上位100リポジトリそれぞれで，課題コメント1,000件，PRコメント1,000件，計200,000件収集
  - NLTKを用いて484,000件の文章に分割

  - p5で特定したユニークな比喩表現が含むか調査

予備調査

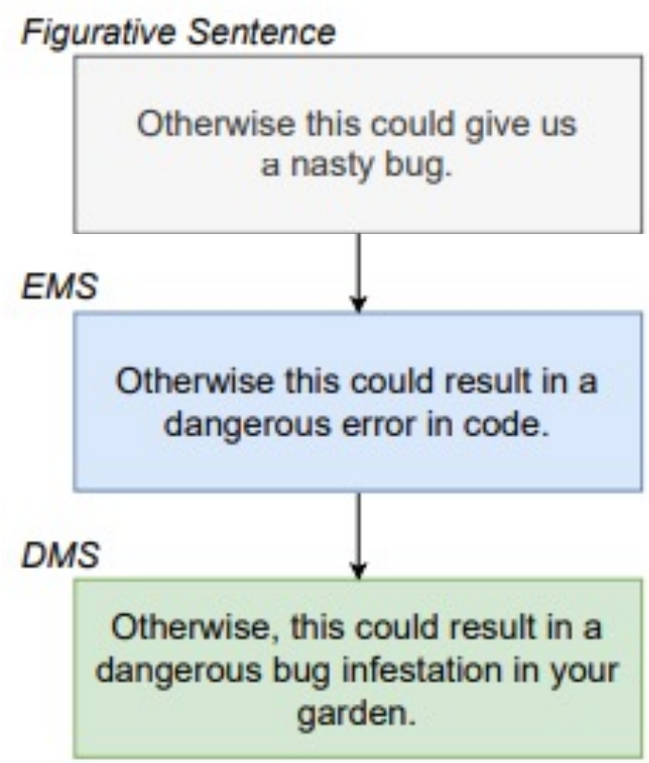
- 結果
  - SE特有の比喩表現は44,000文（9%），一般的な比喩表現は107,000文（22%），両方を持つ文は（2.67%）
  - SE特有の比喩表現は一般的な比喩表現と比較して，あまり出現しない（左図）
  - SE特有の比喩表現のうち324件は（72.8%）は10回以上出現しておらず，193件は一回も出現しない
    - プロジェクト固有の表現（‘ghost highlight’，‘ghost monitor’）
    - 特殊なシナリオを説明（‘dead fork’，‘magic code’）

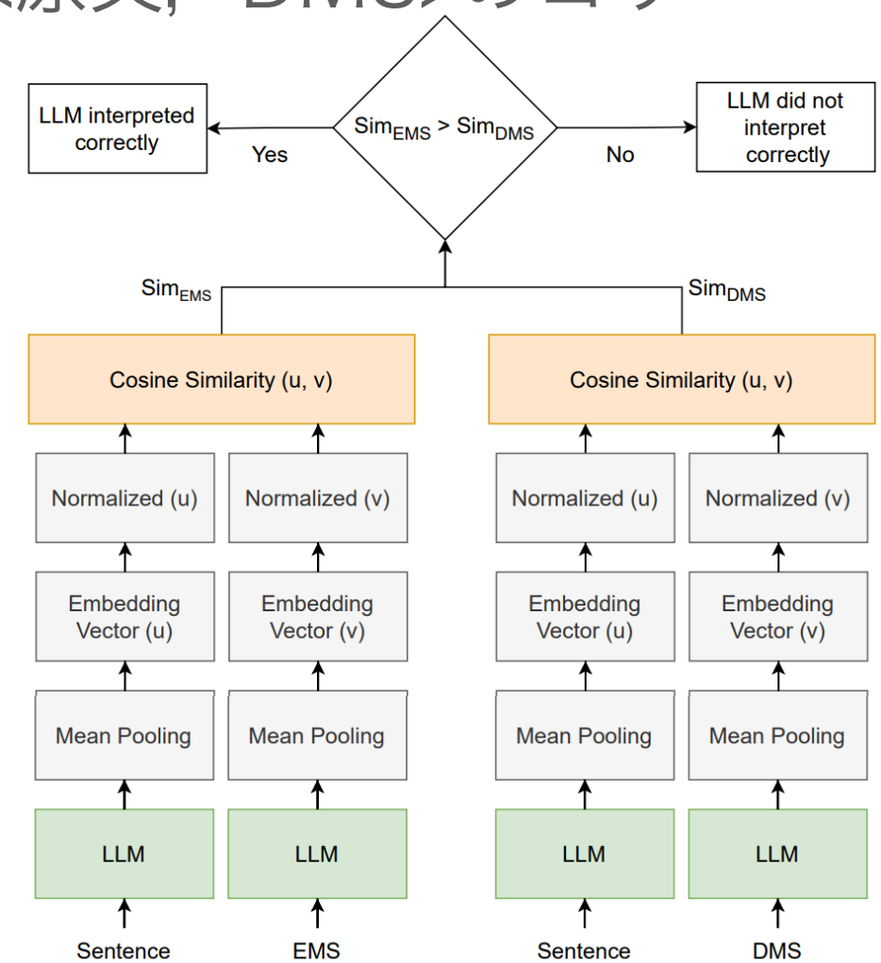




RQ1｜LLMの比喩表現の解釈性

- アプローチ
  - 原文，EMS，DMSの意味的関係を理解できるか調査
  - <原文，EMS>のコサイン類似度が，<原文，DMS>のコサイン類似度よりも高いかを計測
  - 使用LLMは
    - BERT[15]
    - RoBERTa[16]
    - ALBERT[51]
    - CodeBERT[54]





[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proceedings of the 2019 Conference of the NAACL: Human Language Technologies. Minneapolis, Minnesota: ACL, Jun. 2019, pp. 4171–4186.
[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” ArXiv, vol. abs/1907.11692, 2019.
[51] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” ArXiv, vol. abs/1909.11942, 2019.
[54] Z. Feng, D. Guo, D. Tang, N. Duan, X. Feng, M. Gong, L. Shou, B. Qin, T. Liu, D. Jiang, and M. Zhou, “CodeBERT: A pre-trained model for programming and natural languages,” in Findings of the ACL: EMNLP 2020. Online: ACL, Nov. 2020, pp. 1536–1547.

RQ1｜LLMの比喩表現の解釈性

- 結果

Model	*SE-specific*			*General*			*Overall*		
	$Sim_{EMS} > Sim_{DMS}$	*p-value*	|Cliff's δ|	$Sim_{EMS} > Sim_{DMS}$	*p-value*	|Cliff's δ|	$Sim_{EMS} > Sim_{DMS}$	*p-value*	|Cliff's δ|
BERT	84.51%	$p < 0.01$	0.629	87.40%	$p < 0.01$	0.638	86.57%	$p < 0.01$	0.637
RoBERTa	83.70%	$p < 0.01$	0.648	85.21%	$p < 0.01$	0.620	84.95%	$p < 0.01$	0.632
ALBERT	81.79%	$p < 0.01$	0.610	85.80%	$p < 0.01$	0.598	85.00%	$p < 0.01$	0.605
CodeBERT	77.99%	$p < 0.01$	0.498	79.63%	$p < 0.01$	0.493	79.11%	$p < 0.01$	0.495

- 全体的に，EMS＞DMS（＞80%）であることから，LLMは比喩表現を認識する能力が高い
  - 統計的に有意であり（p値），効果量大（δ>0.474）である

RQ2｜比喩表現を考慮した感情分析

- アプローチ
  - 対象学習により，LLMをファインチューニング

  - 原文とEMSをanchorやpositive
  - DMSをnegativeとし，合計3,322件のトリプレットを作成

  - 損失関数： $\text{InfoNCE Loss}(a, p, n) = -\log\left(\frac{e^{\text{sim}(a,p)}}{e^{\text{sim}(a,p)} + e^{\text{sim}(a,n)}}\right)$
  - optimizer：Adam

  - ファインチューニングは２回実施
    - 1回目：対象学習
    - 2回目：感情分析データセット

RQ2｜比喩表現を考慮した感情分析

- 結果
  - ○○-FL：対象学習を行なったもの

Model	Anger	Love	Fear	Joy	Sad.	Surp.	Mic.Avg.
BERT	0.506	**0.712**	0.536	0.579	0.636	0.594	0.588
BERT-FL	**0.547**	0.709	**0.562**	**0.608**	**0.661**	**0.632**	**0.627**
+/-	+8.10%	-0.42%	+4.85%	+5.01%	+3.93%	+6.40%	+6.60%
RoBERTa	0.525	0.683	0.500	0.613	**0.673**	0.592	0.593
RoBERTa-FL	**0.551**	**0.733**	**0.545**	**0.667**	0.667	**0.617**	**0.632**
+/-	+4.95%	+6.82%	+8.26%	+8.10%	-0.90%	+4.05%	+6.66%
ALBERT	**0.462**	0.658	0.430	0.487	**0.628**	0.564	0.531
ALBERT-FL	0.443	**0.682**	**0.435**	**0.540**	0.624	**0.592**	**0.550**
+/-	-4.11%	+3.52%	+1.15%	+9.81%	-0.64%	+4.73%	+3.63%
CodeBERT	0.484	0.711	**0.507**	0.558	0.575	0.576	0.561
CodeBERT-FL	**0.497**	**0.723**	0.444	**0.605**	**0.645**	**0.617**	**0.583**
+/-	+2.79%	+1.70%	-14.08%	+7.75%	+10.92%	+6.61%	+3.90%
Avg. +/-	+2.93%	+2.90%	+0.04%	+7.67%	+3.33%	+5.45%	+5.20%

  - 比喩表現を考慮することで，ほとんどのケースで精度向上
    - 微妙なニュアンスを理解し，解釈する能力が向上するため

RQ3｜SEタスクの自動化の強化

- アプローチ
  - ソフトウェア工学の主要な研究分野[34, 35, 37, 75]であるバグ報告の優先度検出にフォーカス
  - バグ報告の優先度検出で，RQ2と同様のことを実施
    - Bugzillaのバグ報告では，P1（最高）からP5（最低）までの5つのクラスで優先度を管理
    - [34]で提供された，49,600件のバグ報告を利用

RQ3｜SEタスクの自動化の強化

- 結果

Model	P1	P2	P3	P4	P5	Micro Average
BERT	0.606	0.329	0.833	0.0	0.663	0.716
BERT-FL	**0.632**	**0.359**	**0.842**	0.0	**0.667**	**0.730**
+/-	+4.31%	+9.14%	+1.10%	-	+0.52%	+1.96%
RoBERTa	0.61	0.293	0.827	0.0	**0.677**	0.707
RoBERTa-FL	**0.624**	**0.343**	**0.839**	0.0	0.674	**0.724**
+/-	+1.91%	17.24%	+1.39%	-	-0.51%	+2.40%
ALBERT	0.564	0.288	0.810	0.0	0.670	0.683
ALBERT-FL	**0.602**	**0.299**	**0.827**	0.0	**0.674**	**0.709**
+/-	+6.71%	+3.88%	+2.14%	-	+0.53%	+3.71%
CodeBERT	0.608	0.363	0.830	0.0	0.667	0.714
CodeBERT-FL	**0.636**	**0.373**	**0.839**	0.0	**0.670**	**0.726**
+/-	+4.55%	2.64%	+1.08%	-	0.52%	+1.61%
Avg. +/-	+4.37%	+8.23%	+1.43%	-	+0.27%	+2.42%

  - ほとんどのケースで精度が向上
  - P5の精度向上量が僅かなのは，データ数の少なさが原因

まとめ

- 目的
  - 比喩表現がLLMに与える影響の調査

- 結果
  - 比喩表現を取り入れることで，さまざまなSEタスクにおいてLLMの精度を向上させることが可能

- 所感
  - 比喩表現に着目するのは面白いと思った
  - どんな比喩表現があるのかを見てみたかった（提供されているデータにない？）