Vol.2020-NL-245 No.4
2020/9/30

文脈化単語埋め込みを用いた慣用句判定

高橋 良輔[1,a)]　笹野 遼平[1,b)]　武田 浩一[1,c)]

概要：慣用句として使用される表現の中には「足を洗う」などのように，慣用句として用いられる場合（idiomatic）と，文字通りの意味で用いられる場合（literal）の両方があるものが存在する．本研究では，このような表現を対象に，文脈を考慮した単語表現である文脈化単語埋め込みを用いて，与えられた文脈における用法が idiomatic であるか，literal であるか判定する手法を提案する．実験は表現ごとに慣用句判定を行う設定と表現横断的に判定を行う設定の２つの設定で行い，いずれの設定においても先行研究の手法と比較して大幅に高い精度を達成できることを確認した．

1. はじめに

慣用句として使用される表現の中には，同じ表現でも慣用句として用いられる場合（idiomatic）と，文字通りの意味で用いられる場合（literal）の両方があるものが存在する．たとえば「足を洗う」という表現の場合，「やくざな稼業から足を洗った」という文では『好ましくない生活をやめる』という慣用句としての意味で使われているのに対し，「散歩の後は犬の足を洗ってから家に入れている」という文では『足をきれいにする』という文字通りの用法で使われている．このような idiomatic な用法でも literal な用法でも使われる表現の用法を正しく捉えることは，文の意味理解において重要となる．本研究では，このような literal な用法も一般的に用いられる慣用表現を対象に，idiomatic な用法であるのか，literal な用法であるのかを判定する慣用句判定に取り組む．

慣用句判定の先行研究では，様々な文脈を反映した情報が特徴量として使用されており，着目している表現の周辺に出現する単語の埋め込みを用いる手法 [1] や，その表現が出現する文の分散表現を用いる手法 [2] などが提案されている．近年では，ELMo [3] や BERT[4] のような文脈を考慮した単語の埋め込みモデルが提案されており，幅広い自然言語処理タスクに有用であることが示されている．これらの文脈化単語埋め込みモデルでは，同じ単語に文脈に応じて異なる埋め込みを与えることから，慣用句判定にも有用であると考えられる．そこで本研究では，日本語および英語を対象に，文脈化単語埋め込みを用いた慣用句判定に取り組む．

本研究では，表現ごとに判定器を構築する設定と，表現横断的に判定器を構築する設定の 2 つの設定を考える．前者の設定では，注目している表現に対する学習データが存在していることを前提に，その表現専用の判定器を構築する．後者の設定では，汎用的な慣用句判定モデルの構築を目的とし，複数の慣用表現で構成される学習データを用い，未知の表現にも適用可能な汎用慣用句判定モデルを構築する．汎用的な慣用句判定モデルの構築を目指す後者の設定では，慣用句を構成する単語の文脈化単語埋め込みが，慣用句の間で共有される何らかの特徴を捉えていることを期待している．

2. 関連研究

慣用句判定のタスクには，多くの先行研究が存在する．橋本ら [5] は，曖昧性を持つ 146 種類の慣用句を含む 102,846 の日本語文に対して，idiomatic であるか literal であるかのラベルを付与したデータセットを作成し，そのデータセット中の 90 種類の慣用句に対して，サポートベクターマシン (SVM) を用いた慣用句判定を行った．判定のための特徴量としては，語義曖昧性解消（WSD）に一般的に用いられる特徴量を用いており，実験を通して WSD の特徴量が慣用句判定にも有効であることを示した．本研究では，橋本らの手法を日本語の慣用句判定のベースラインとして使用する．

Fazly ら [6] は，慣用句は構文的な制約が強いことに着目し，英語の慣用句を対象に，慣用句の局所的な構文のパターンから正準形を定義し，正準形として出現するか否かで idiomatic か literal かの判定を行う手法を提案した．Sporleder ら [7] は，大局的な文脈の語彙情報に基づくモデ

[1] 名古屋大学大学院情報学研究科
[a)] takahashi.ryosuke@e.mbox.nagoya-u.ac.jp
[b)] sasano@i.nagoya-u.ac.jp
[c)] takedasu@i.nagoya-u.ac.jp

ルを提案した．Sporleder らは，正準形で出現しない慣用句であっても，idiomatic な用法であると判定できることを示した．Li ら [8] は，慣用句判定のために，大局的および局所的な文脈の語彙情報，談話構造，依存構造などの様々な素性を検証し，慣用句判定には，大局的な文脈の語彙情報と談話構造が最も効果的であると報告している．Peng ら [9] は，慣用句の判定を異常検出の問題として扱い，idiomatic な意味の用法を含む段落と literal な意味の用法を含む段落から，Latent Dirichlet Allocation (LDA) を用いてトピックを抽出し，慣用句判定に利用した．実験の結果，トピック表現は idiomatic な用法であるか literal な用法であるかの判定に有用であると報告している．

ニューラルネットワークを用いた研究としては，Gharbieh らの研究 [1] や Salton らの研究 [2] などがある．Gharbieh ら [1]. は，単語埋め込みを用いた慣用句判定手法を提案している．Gharbieh らは，対象表現の周辺単語の Skip-gram ベクトル [10] を平均したものを素性として用いることで，Fazly らの手法よりも高い性能を実現している． Salton ら [2] は，慣用句判定に対する文の分散表現モデルの有効性を検証した．Salton らは，Skip-Thought モデル [11] により生成された文の分散表現を入力素性として，SVM に基づく分類器を構築した．分類器は，表現ごとの分類器と，表現横断的な分類器をそれぞれ作成し，実験を通して，慣用句判定における Skip-Thought Vectors の有効性を示した．本研究では，Salton らの手法を英語の慣用句判定のベースラインとして使用する．

また近年，ELMo[3] や BERT[4] などの文脈を考慮した単語埋め込みモデルが提案されており，慣用句判定と類似したタスクに応用されている．たとえば，Shwartz ら [12] や Nandakumar ら [13] は，英語の Multiword Expression (MWE) の意味曖昧性解消タスクにおいて，事前学習済みの文脈化単語埋め込みモデルによって生成された単語埋め込みを利用する手法を提案している．しかし，本研究のように動詞とその項で構成される慣用表現に着目し，複数の言語を対象に文脈化単語埋め込みの有効性を検証した研究は報告されていない．

3. 提案手法

本研究では，「足を洗う」などのように，文脈によって慣用句として用いられる場合（idiomatic）と，文字通りの意味で用いられる場合（literal）の両方がある表現を対象として，与えられた文脈における用法が idiomatic であるか，literal であるかの判定をすることを目的とする．その手法として本研究では，文脈を考慮した単語表現である文脈化単語埋め込みを用いた判定手法を提案する．対象とする慣用表現は基本的に「足を洗う」などのように 1 つの動詞 (e.g., 洗う) とその項となる名詞 1 つ (e.g., 足) で構成される慣用表現とする．提案手法の概要は以下の通りである．まず，対象表現を含む文を文脈化単語埋め込みモデルへ入力し，対象表現を構成する動詞とその項である名詞の文脈化単語埋め込みを獲得する．続いて，獲得した埋め込みを入力素性として，その用法が idiomatic か literal かの二値分類を行う単純な分類器を学習する．

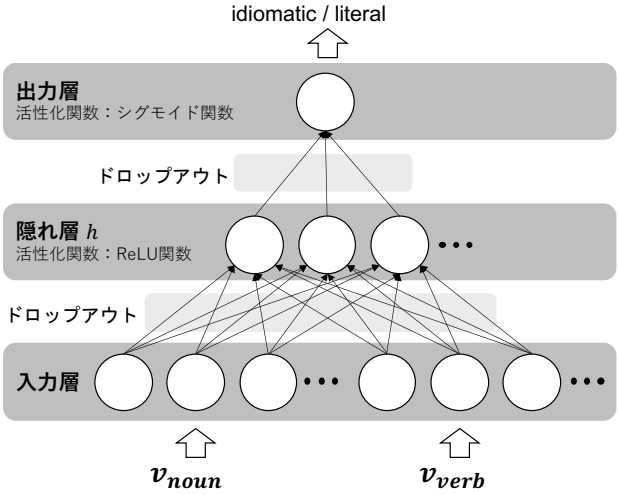

図 1 慣用句判定モデルの概要

対象表現の埋め込みを得るために，まず対象表現を含む文 $\boldsymbol{s} = \boldsymbol{w}_1 \ldots \boldsymbol{w}_n$ を文脈化単語埋め込みモデルへ入力し，文中の各単語の埋め込み $\boldsymbol{v}_i$ を獲得する．この際，日本語文は事前に Juman++[*1] を用いて単語に分割した．また，BERT に入力する文は最長で対象表現の前後 50 単語ずつまでとし，50 単語以上離れた部分は削除した．続いて，獲得された対象表現を構成する動詞と名詞それぞれの埋め込み $\boldsymbol{v}_{verb}$ および $\boldsymbol{v}_{noun}$ を，図 1 に示すような多層パーセプトロンへの入力とし，慣用句判定を行う．図 1 において，2 つのドロップアウト層のドロップアウト率はいずれも 50%，隠れ層 h の次元数は入力次元数の半分とした．

前述の通り本研究では，表現ごとに慣用句判定器を構築する設定と，表現横断的に判定器を構築する設定の 2 つの設定を考える．表現ごとの慣用句判定器の構築では，表現ごとに正解情報付きデータがあるという前提で，モデルの学習と評価を行う．たとえば，いくつかの「足を洗う」の用例で判定器を学習し，「足を洗う」の別の用例で評価を行う．表現横断的な慣用句判定を構築する設定では，複数の慣用表現の正解情報付きデータがあるという前提で，学習データには含まれていない慣用表現にも適用可能な 1 つの汎用的な慣用句判定器を構築する．たとえば，「足を洗う」や「頭が痛い」などの複数の表現の用例で判定器を学習し，学習時に使用していない「足が付く」や「手がない」などの表現で評価を行う．

慣用句として使われる「足を洗う」における「足」や「洗う」の意味は，一般的な用法における「足」や「洗う」の

[*1] https://github.com/ku-nlp/jumanpp

意味とは異なると考えられるが，このような差異には表現横断的な傾向が存在する可能性がある．このような表現横断的な傾向を捉えることを狙い，表現横断的な慣用句判定モデルでは入力素性として，構成単語の文脈化単語埋め込みに加えて，慣用句を構成する動詞と名詞それぞれの一般的な用法における単語埋め込み $\overline{\boldsymbol{v}}_{verb}$，$\overline{\boldsymbol{v}}_{noun}$ も利用する．もし，表現横断的な傾向が存在するならば，各単語の文脈化単語埋め込みと，一般的な用法における単語埋め込みを比較することで，慣用句判定に有用な情報が得られることが期待できる．

本研究では，一般的な用法における単語埋め込みを，対象の単語の複数の用例における文脈化単語埋め込みの平均を取ることで近似する．用例の中には慣用句として使用されているものも含まれる可能性はあるが，その割合は限定的であり，多くの用例の平均を取ることでこのような用例の影響を無視できると考えた．以下，本稿では，$\overline{\boldsymbol{v}}_{verb}$，$\overline{\boldsymbol{v}}_{noun}$ を平均単語埋め込みと呼ぶ．たとえば，「足」の平均単語埋め込みは，「足」が出現する文を収集し，各文における「足」の文脈化埋め込みの平均を取ることで得られる．

4. 実験

慣用句判定タスクにおける文脈化単語埋め込みの効果を確認するため，文脈によって idiomatic な用法か literal な用法か変化する表現を対象に慣用句判定実験を行った．

4.1 データセット

実験は，日本語と英語のデータセットを用いて行った．日本語の慣用句データセットとして OpenMWE Corpus[5] を，英語のデータセットとして VNC-Tokens data [14] をそれぞれ用いた．いずれのデータセットも判定対象の表現を含む文に対してそれぞれ idiomatic か literal かのラベルが付与されたものとなっている．

OpenMWE Corpus は，橋本ら [5] により作成された日本語慣用句データセットであり，名詞と動詞で構成[*2]された慣用句の可能性のある表現 146 種類からなる．前述の通り，合計で 102,846 文の用例が含まれており，それぞれの文に “I”(idiomatic) または “L”(literal) のラベルが付与されている．本研究では，橋本らと同様に，50 文以上の用例を持つ 90 種類の慣用句を実験に使用した．

VNC-Tokens dataset は名詞と動詞で構成された慣用句の可能性のある表現 53 種類を収集した英語のデータセットである．British National Corpus (BNC) から抽出された合計 2,984 文の用例が含まれており，それぞれの文に “I” (idiomatic) か “L” (literal) のラベルが付与されている．本研究では，先行研究に倣い，“I” と “L” のラベルの偏りの小さな 28 種類の慣用句を実験に使用した．

4.2 比較モデル

提案手法と比較するため，日本語については橋本らの手法 [5]，英語に対しては Salton らの手法 [2] で使われている素性を用いたモデルを作成し，提案手法と比較した．

橋本らの手法では素性として，N グラムに加え，構文解析器 KNP[*3] を使うことで得られる品詞，原形，カテゴリ・ドメイン属性，態などの情報が用いられている．本研究ではこれらの情報を素性として使用した SVM に基づくモデルを構築し，日本語のベースラインモデルとして使用した．Salton らの手法では素性として，Skip-Thought Vectors[11] が用いられている．本研究では先行研究と同様に，事前学習済みの Skip-Thought Vectors[*4] を素性として使用した SVM に基づくモデルを構築し，英語のベースラインモデルとして使用した．

また，表現ごとに判定器を構築する設定では，慣用表現を構成する動詞，および，名詞，それぞれの文脈化単語埋め込みの有用性を検証するため，いずれか片方だけ用いたモデルとの比較も行った．

4.3 実験設定

表現ごとに慣用句判定器を構築する設定では，各表現ごとにデータセット中の文を無作為に 10 分割し，10 分割交差検証を行った．この際，10 分割したデータセットのうち，8 つを訓練データ，1 つを開発データ，1 つをテストデータとして使用した．表現横断的な慣用句判定器を構築する設定では，データセットに存在する表現を無作為に 10 分割し，同様に 10 分割交差検証を行なった．すなわち，テストに使用する表現は，学習データに含まれていない表現となっている．

文脈化単語埋め込みモデルには，日本語と英語いずれについても事前学習済みの BERT[4] を用いた．日本語の BERT の実装[*5] には，日本語の Wikipedia で学習した BERT base モデルを，英語の BERT の実装[*6] には，英語の Wikipedia と BookCorpus で学習した BERT base モデルを使用した．BERT base モデルの次元数は 768 であり，予備実験の結果に基づき，全 12 層のうち 11 層目のベクトルを，対象の単語の文脈化埋め込みとして使用した．

また，平均単語埋め込み $\overline{\boldsymbol{v}}_{verb}$ と $\overline{\boldsymbol{v}}_{noun}$ を獲得するため，日本語データを用いた実験では，CommonCrawl に含まれる日本語文を Juman++を用いて形態素解析し，各対象表現を構成する名詞および動詞を含む文を 100 個ずつ抽出し，各文における対象の単語の文脈化単語埋め込みの平均を算

[*2] 「一から十まで」などのように，一部，動詞を含まない表現も含まれているが，本研究ではこのような表現があった場合，2 つ目の名詞を動詞として扱う．

[*3] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
[*4] https://github.com/ryankiros/skip-thoughts
[*5] http://nlp.ist.i.kyoto-u.ac.jp/index.php?BERT
[*6] https://github.com/google-research/bert

表 1　表現ごとの慣用句判定実験の結果 (日本語)

モデル	正解率のマクロ平均
Majority Baseline	0.740
橋本ら [5]	0.870
BERT($\boldsymbol{v}_{verb}$)	0.935
BERT($\boldsymbol{v}_{noun}$)	0.940
BERT($\boldsymbol{v}_{verb};\boldsymbol{v}_{noun}$)	**0.941**

表 2　表現ごとの慣用句判定実験の結果 (英語)

モデル	正解率のマクロ平均
Majority Baseline	0.710
Salton ら [2]	0.792
BERT($\boldsymbol{v}_{verb}$)	0.898
BERT($\boldsymbol{v}_{noun}$)	0.919
BERT($\boldsymbol{v}_{verb};\boldsymbol{v}_{noun}$)	**0.936**

出した．同様に，英語データを用いた実験では，品詞付与済みの British National Corpus (BNC) から各対象表現を構成する名詞および動詞を含む文を 100 個ずつ抽出し，各対象表現を構成する名詞および動詞の文脈化単語埋め込みの平均を算出した．学習のエポック数は，200 エポックまで実行し，検証データにおいてもっとも高い精度となったモデルをテストデータに適用した．

4.4　実験結果

日本語および英語のデータセットに対する表現ごとの慣用句判定実験の結果を表 1 および表 2 にそれぞれ示す．Majority Baseline は，表現ごとに学習データにおいて多数派のラベルを出力した場合の精度を表す．BERT を用いて生成された動詞と名詞の埋め込みを連結したものを入力素性として用いた提案手法が，日本語，英語，いずれのデータセットにおいて最も高い性能となっていることが確認できる．

動詞の埋め込みのみを用いたモデル，名詞の埋め込みのみを用いたモデルと比較すると，日本語データセットに対しては，動詞と名詞の埋め込みを両方用いたモデルと比較して，ほとんど精度の低下が確認できないのに対し，英語データセットに対しては，特に名詞の埋め込みを使わなかった場合に精度の低下が確認された．このことから，英語の場合，慣用句を構成する名詞の埋め込みの方により強く慣用表現特有の情報が埋め込まれていると考えられる．また，日本語データセットに対し精度低下が発生しないのは，日本語の場合は，慣用句を構成する動詞と名詞が比較的近くに出現することから，慣用句を構成する動詞と名詞の埋め込みに重複する情報を埋め込まれている可能性が考えられる．

日本語および英語のデータセットに対する表現横断的な慣用句判定実験の結果を表 3 および表 4 にそれぞれ示す．Majority Baseline は，全学習データにおいて多数派となるラベルを出力した場合の精度を表す．表現横断的な慣用句判定においても，BERT に基づくモデルは先行研究より高い性能となっていることが確認できる．また，日本語と英語，いずれのデータセットに対しても，一般的な用法における単語埋め込みに相当する平均単語埋め込みを利用することで，慣用句判定精度が向上することが確認できる．一般的な用法では，literal な用例が支配的であるため，平均単語埋め込みは literal な情報が強く埋め込まれていると考えられる．このことは，慣用句を構成する動詞および名詞の慣用的な用法における単語埋め込みと，一般的な用法における単語埋め込みの間には，表現横断的な特徴がある可能性を示唆している．

表 3　表現横断的な慣用句判定実験の結果 (日本語)

モデル	正解率
Majority Baseline	0.629
橋本ら [5]	0.740
BERT($\boldsymbol{v}_{verb};\boldsymbol{v}_{noun}$)	0.824
BERT($\boldsymbol{v}_{verb};\boldsymbol{v}_{noun};\overline{\boldsymbol{v}}_{verb};\overline{\boldsymbol{v}}_{noun}$)	**0.836**

表 4　表現横断的な慣用句判定実験の結果 (英語)

モデル（英語）	正解率
Majority Baseline	0.672
Salton ら [2]	0.780
BERT($\boldsymbol{v}_{verb};\boldsymbol{v}_{noun}$)	0.852
BERT($\boldsymbol{v}_{verb};\boldsymbol{v}_{noun};\overline{\boldsymbol{v}}_{verb};\overline{\boldsymbol{v}}_{noun}$)	**0.877**

4.5　考慮すべき文脈長について

本研究では，使用する文脈は最長で対象表現の前後 50 単語ずつまでとし，50 単語以上離れた部分は削除して，BERT への入力とした．対象表現に対して考慮する文脈の範囲による慣用句判定への影響を調査するために，BERT へ入力する文脈の窓幅を変化させた慣用句判定実験を行った．実験は，表現ごとに慣用句判定器を構築する設定，表現横断的な慣用句判定器を構築する設定，いずれの設定についても，判定器の入力として，慣用句を構成する動詞と名詞の単語埋め込みを用いるモデルに対し実験を行った．

窓幅を $\{1, 2, 3, 5, 10, 20, 50\}$ の範囲で変化させたときの各実験設定での正解率を表 5 に示す．いずれの実験設定においても，窓幅が大きいほど正解率が高くなることが確認できる．ただし，窓幅が 20 と場合と 50 の場合の正解率の違いは限定的であり，前後 20 単語程度を考慮することで十分であると考えられる．

5.　まとめと今後の展望

本研究では，慣用句判定に，文脈化単語埋め込みを用いる手法を提案した．提案手法の評価を行うために，日本語と英語のデータセットを用いて，表現ごとの慣用句判定実験と表現横断的な慣用句判定実験を行い，BERT を用いた提案手法を用いることで，各データセットにおいて既存手法よりも大幅に高い精度で慣用句判定を行えることを示し

表 5　BERT への入力文脈の窓幅を変化させた実験結果

実験設定 ＼ 窓幅 n	1	2	3	5	10	20	50
表現ごとの慣用句判定（日本語）	0.791	0.858	0.881	0.912	0.929	0.938	**0.941**
表現ごとの慣用句判定（英語）	0.782	0.844	0.886	0.919	0.928	0.933	**0.936**
表現横断的な慣用句判定（日本語）	0.648	0.671	0.740	0.791	0.817	0.820	**0.824**
表現横断的な慣用句判定（英語）	0.702	0.725	0.798	0.834	0.846	0.850	**0.852**

た．また，表現横断的な慣用句判定実験では，入力素性として対象表現の一般的な用法の埋め込みを平均したものを加えることでモデルの性能が向上することを示した．しかし，特に表現横断的な慣用句判定モデルについては，実際にどのような手掛かりに基づき慣用句判定を行っているかは不明な部分が多く，今後の課題として，文脈化単語埋め込みが慣用句のどのような特徴を捉えているかの分析・解明や，その分析結果に基づく慣用句判定器の改良が挙げられる．

参考文献

[1] Gharbieh, W., Bhavsar, V. and Cook, P.: A Word Embedding Approach to Identifying Verb-Noun Idiomatic Combinations, *In Proceedings of the 12th Workshop on Multiword Expressions (MWE'16)*, pp. 112–118 (2016).

[2] Salton, G., Ross, R. and Kelleher, J.: Idiom Token Classification using Sentential Distributed Semantics, *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, pp. 194–204 (2016).

[3] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L.: Deep Contextualized Word Representations, *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL'18)*, pp. 2227–2237 (2018).

[4] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL'19)*, pp. 4171–4186 (2019).

[5] Hashimoto, C. and Kawahara, D.: Construction of an Idiom Corpus and its Application to Idiom Identification based on WSD Incorporating Idiom-Specific Features, *In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP'18)*, pp. 992–1001 (2008).

[6] Fazly, A., Cook, P. and Stevenson, S.: Unsupervised Type and Token Identification of Idiomatic Expressions, *Computational Linguistics*, Vol. 35, No. 1, pp. 61–103 (2009).

[7] Sporleder, C. and Li, L.: Unsupervised Recognition of Literal and Non-Literal Use of Idiomatic Expressions, *In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09)*, pp. 754–762 (2009).

[8] Li, L. and Sporleder, C.: Linguistic Cues for Distinguishing Literal and Non-Literal Usages, *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pp. 683–691 (2010).

[9] Peng, J., Feldman, A. and Vylomova, E.: Classifying Idiomatic and Literal Expressions Using Topic Models and Intensity of Emotions, *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pp. 2019–2027 (2014).

[10] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, *In Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'13)*, pp. 3111–3119 (2013).

[11] Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A. and Fidler, S.: Skip-Thought Vectors, *In Proceedings of Advances in Neural Information Processing Systems 28 (NIPS'15)*, pp. 3294–3302 (2015).

[12] Shwartz, V. and Dagan, I.: Still a Pain in the Neck: Evaluating Text Representations on Lexical Composition, *Transactions of the Association for Computational Linguistics*, Vol. 7, pp. 403–419 (2019).

[13] Nandakumar, N., Baldwin, T. and Salehi, B.: How Well Do Embedding Models Capture Non-compositionality? A View from Multiword Expressions, *In Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP (RepEval'19)*, pp. 27–34 (2019).

[14] Cook, P., Fazly, A. and Stevenson, S.: The VNC-Tokens Dataset, *In Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE'08)*, pp. 19–22 (2008).