卒 業 研 究 報 告 書

題　　目

深層学習に基づく 4 コマ漫画の感情推定と
マルチモーダル化への検討

研 究 グ ル ー プ　知能情報第 1 グループ

指 導 教 員　森 直樹 教授

平 成　32　年（　2020　年）度 卒 業

（No.　1171201102　）　高山 裕成

大 阪 府 立 大 学 工 学 部 知 能 情 報 工 学 科

目 次

図 目 次

表 目 次

1　はじめに

近年, 深層学習を始めとする機械学習技術の大きな発展を受けて, 人工知能を用いた創作物理解が注目されている. しかし, 創作は高次の知的活動であるため, いまだに実現が困難なタスクである. 人の創作物の理解に関する分野の中でもコミック工学 [1] など漫画を対象とした研究は, 絵と文章から構成される漫画を対象とするため, 自然言語処理と画像処理の両方の側面を持つマルチモーダルデータを扱う分野である. コミック工学の分野では様々な研究が報告されているが, その多くは画像処理に基づいた研究であり, 自然言語処理による内容理解を目指した研究は少ない. その大きな原因のひとつとしてデータが十分ではないという点が挙げられる. また, 漫画に含まれるテキストには, 口語表現, 擬音語, 表記揺れといった漫画特有の言語表現を含み, これらの扱いについて考慮する必要がある. そして, 漫画が著作物であることに起因する研究用データの不足も課題となっている.

本研究では人工知能を用いた漫画の内容理解のために, 漫画におけるキャラクタのセリフのマルチモーダルな感情推定を目的とする. まず自然言語処理を用いた漫画のセリフの感情を推定して, その上で漫画のコマの画像情報を加えたマルチモーダル化について検討する.

以下に本論文の構成を示す. まず, 2 章ではコミック工学に関連するデータセットについて, また 3 章では本研究で用いる要素技術について概説する. 次に, 4 章では漫画のセリフのマルチモーダルな感情推定を行うための提案手法について述べる. そして, 5 章において, 実験手法とその考察を示す. 最後に, 6 章で本研究の成果をまとめた上で, 今後の課題について述べる.

2　コミック工学に関連するデータセット

本章では, コミック工学に関連するデータセットについて説明する.

2.1　Manga109

Manga109 [2] は, 漫画の研究のために相澤らにより作られたもので, 日本のプロの漫画家によって描かれた 109 冊の漫画で構成されており, これらは 1970 年代から 2010 年代に公開された漫画であり, 対象読者層やジャンルも幅広く網羅している. Manga109 には 109 冊の漫画の画像データや登場人物の名前, 画像内における登場人物の顔, 全身, コマ, テキストの座標などのアノテーションデータが含まれている. しかし, Manga109 にはセリフの発話者に関する情報や, 本研究の趣旨であるセリフの感情に関する情報は付与されていない.

2.2　4 コマ漫画ストーリーデータセット

4 コマ漫画ストーリーデータセット [3] は, 上野によって作られた, コミック工学発展のために研究者が一から開発に関わった世界初の研究用のデータセットであり, 画像データのレイヤー分けや作者によるセリフの感情アノテーションなど, いくつかの特徴がある.

Manga109 といった市販コミックによって構成されたデータセットとは異なり, 4 コマ漫画ストーリーデータセットのデータは本データセットのために幾人かの漫画家によって描き下ろされている. 市販されたコミックをデータとした場合, 著作権などの問題に加え, 計算機上で扱うためのデータが少なく, コミックの意味理解を目的とした研究には適さないという問題がある. 例えばコミックに登場するキャラクターの感情は明示されていないため, 読者によるアノテートによってラベルを付与する必要があるが, アノテートされたラベルが漫画家の意図とは異なる可能性を否定できない. また, マルチモーダルでストーリーの解析をする際にオリジナリティの観点から同一プロットを複数の漫画家が描くことは稀有なため, そういったデータの収集に基づく研究は困難である. 4 コマ漫画ストーリーデータセットはそういった問題点を解決するために作られたデータセットである.

上野は 4 コマ漫画の構造を,

- 一般：標準的な起承転結をもつ
- 繰り返し：1, 2 コマ間の類似が 3, 4 コマ間でも起きる
- 出オチ：1 コマ目におかしな絵が描かれてオチがある
- タイトルオチ：最後にタイトルを見返してオチがわかる
- 再帰：4 コマ目から 1 コマ目に戻り話として成立する
- 参照：1 つ以上前の話の続きの話となる
- 連続した 4 コマを 2 話並べて 8 コマで話となる

と定義し, これに従ってデータセットを作成している. 現在は, 同一のストーリーを 4 コマ目がオチとなる一般と出オチの 2 つの構造から描いたものがデータとして存在している.

また, 上野は異なる作者によって描かれた 4 コマ漫画を, そのタッチを基に

- ギャグタッチ
- 少女漫画タッチ
- 少年漫画タッチ
- 青年漫画タッチ
- 萌えタッチ

と分類した.

3　要素技術

本章では，本研究に関連する要素技術について説明する．

3.1　自然言語処理に関する要素技術

自然言語の単語や文を計算機上で表現するための分散表現獲得手法について説明する．

3.1.1　形態素解析

形態素とは日本語などの自然言語において意味を持つ最小の単位のことであり, 文を形態素に分割し, 各形態素の品詞などを判定する技術を形態素解析という．英語の文では, 単語と単語の区切りがほとんどの箇所で明示的に示される．このため、形態素への分割処理は簡単なルールに基づいて行われる場合が多い．一方で, 日本語の文は単語間の区切りが英語ほど明確でないため, 形態素への分割は困難かつ重要である．

形態素解析器としては, MeCab [4] や Juman++ [5], SentencePiece [6] などが存在する．

3.1.2　局所表現, 分散表現

自然言語の単語を計算機上で表現する手法として, 最もシンプルなものが局所表現である．単語の代表的な局所表現の 1 つに One-hot 表現がある．One-hot 表現は単語をベクトルの各次元に 1 対 1 対応させる表現方法である．非常に単純な手法であり, 実装が容易であるという利点がある．一方で, One-hot 表現では語彙数とベクトルの次元数が等しくなるため, 語彙数の増大とともにベクトルの次元数も増大し, ベクトル空間がスパースになってしまう問題がある．また, 各単語がベクトル空間上で等距離に配置されてしまうため, 単語間の意味的な関係性については定義できないことも大きな問題である．

局所表現の問題点を解決するために考案された手法が分散表現である．分散表現は各概念をベクトルの単一次元ではなく複数次元の実数で表す．単語の分

散表現は，類似した文脈で使用される単語は類似した意味をもつ，という分布仮説を基盤としている．単語を実数値密ベクトルで表現することにより, 単語間の意味的な関係性をベクトル空間上での類似度として定義できるという大きな利点がある．

3.1.3 Word2Vec, Doc2Vec

Word2Vec [7] は単語の分散表現を獲得する手法の 1 つである．この手法は，同じ文脈で出現する単語は類似した意味を持つと予想されることに基づいており, 写像されたベクトルは, One-hot 表現のような局所表現と異なり, 単語間の意味を考慮した類似度測定や,「王様」−「男」+「女」=「女王」のような単語間の意味における演算などができるようになる．

Word2Vec では, 自己から周りの単語あるいは周りの単語から自己を予測することにより分散表現を獲得する．前者の手法を Skip-gram といい, 後者の手法を Countinuous Bag-of-Words (CBOW) という．

Doc2Vec [8] は Word2Vec をベースとした, 文書をベクトル空間上に写像して分散表現を得る自然言語処理の手法である．Paragraph ID は各文書と紐づいており, 単語の学習時に一緒にこの Paragraph ID を学習することで文書の分散表現を獲得する．このベクトルを用いると文書間の類似度の算出や文書間での加減算が可能になる．

CBOW を拡張したモデルを Distributed Memory モデルといい, Skip-gram を拡張したモデルを Distributed Bag-of-Words という．

3.1.4　Attention

機械翻訳のタスクに対して考案された, LSTM を用いる Encoder-Decoder モデルは可変長の文を固定長のベクトルにエンコードするため, 長い入力文になるほど隠れ層のノード数が不足し, 学習が難しくなる問題がある. そこで Bahdanau らにより提案されたのが Encoder 側で入力文の各単語の荷重を決定してエンコードすべき場所を制御する Attention 機構 [9] である. Luong らはこの Bahdanau らによるモデルを単純化したモデルとして, 入力されたすべての単語を使用する Global Attentional Model [10] を提案した. 図 3.1 にその概略図を示す.

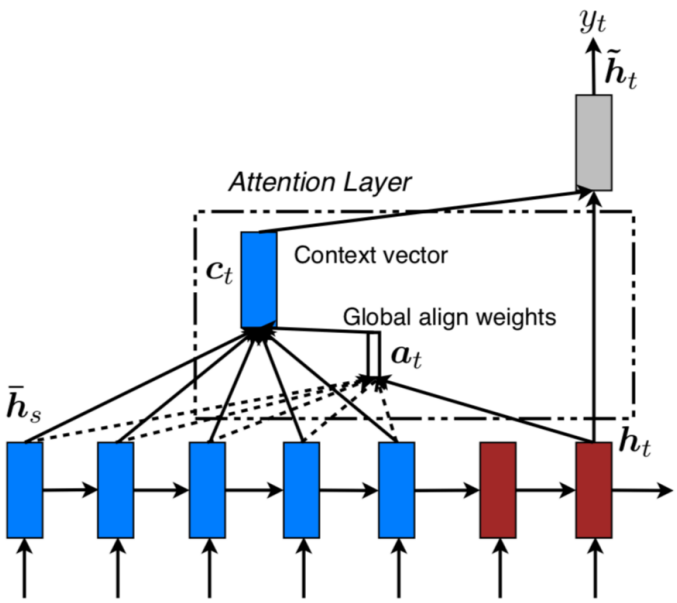

図 3.1: Attention 機構の概略図：文献の図を参照

3.1.5 Transformers

Transformer [11] は 他モデルで頻繁に用いられてきた RNN を用いずに Attention 機構のみを基本構造とする Encoder-Decoder モデルである. 図 3.2 にその概略図を示す. Transformer のエンコーダおよびデコーダはそれぞれ Self-Attention を基本構造とする. Self-Attention とは, Attention 機構の特別な場合である. Attention 機構は Query と Key-Value へのマッピングとして表現することが可能である. 通常 Query はデコーダからのターゲットを, Key-Value はエンコーダからのソースを表す. しかし Self-Attention は下層のすべての位置を参照することができシーケンスの依存関係を獲得できる.

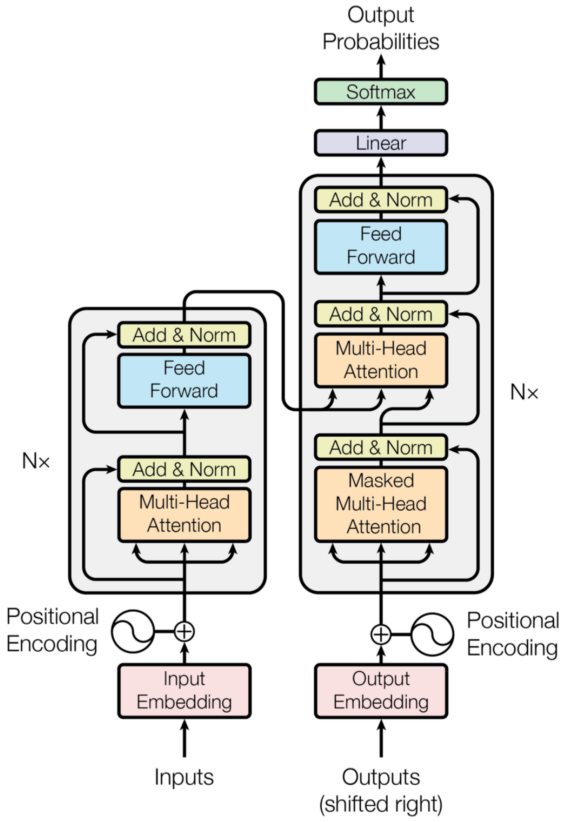

図 3.2: Transformer の概略図 (出典: Attention is All you Need)

3.1.6　BERT

Bidirectional Encoder Representations from Transformers (BERT) [12] は, 2018 年に Google が発表した言語モデルであり, 複数の双方向 Transformer に基づく汎用言語モデルである. これまでの言語モデルは特定の学習タスクに対して 1 つのモデルを用いてきたが, BERT は大規模コーパスに対して事前学習を施して, 各タスクに対して fine-tuning をすることで, さまざまなタスクに柔軟に対応することができる. さらに, 以前はモデルごとに語彙を 1 から学習させるため, 非常に多くの時間とコストがかかっていたが, BERT ではオープンソースで公開されている文脈を既に学習させた Pre-Training BERT モデルを使用することで短時間で学習ができる.

BERT の 事前学習では, 周囲の単語からある単語を予測する Masked Language Model (MLM) と 2 つ目の文章が 1 つ目の文章の次の文章であるかを予測する Next Sentence Prediction (NSP) によりモデルを学習する.

図 3.3 に BERT の事前学習と fine-tuning の概略を示す.

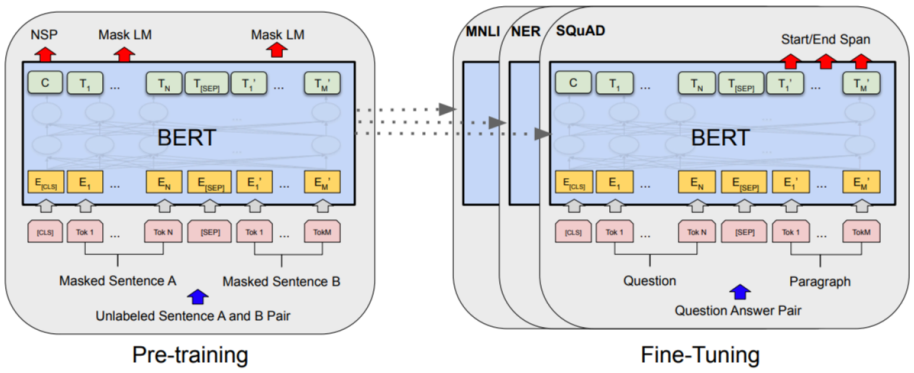

図 3.3: BERT の事前学習と fine-tuning の概略図 : 文献 の図を参照

3.2 画像処理に関する要素技術

画像処理に関する要素技術について説明する.

3.2.1 VGG

VGG16 [13] とは, ImageNet と呼ばれる大規模画像データセットで学習済みの畳み込みニューラルネットワーク (VGGNet) であり, 13 層の畳み込み層と 3 層の全結合層の合計 16 層からなる. 構成する層の数に応じて, VGG11 や VGG16 などと呼ばれることが多い. VGGNet は Oxford 大学に提案された手法であり, 2014 年の画像認識大会で非常に好成績を収めたことからその後のモデルアーキテクチャに広く取り入れられている. あらかじめ学習が済んでいるモデルを転移学習することで事前学習なしでも深いネットワークを学習できる.

3.2.2 illustration2vec

illustration2vec [14] は Saito, Matsui らが提案した VGG をベースとした画像のベクトル化手法であり, 画像リンク集サイトである Danbooru と Safebooru から 100 万枚のイラストを用いて学習した事前学習済みモデルが公開されている. illustration2vec が扱った問題として, イラストに対する画像認識の難しさがあり, VGG などの既存の画像認識モデルのほとんどが ImageNet などの実画像を評価対象にしており, アニメや漫画といったイラストに対して評価をしていなかった. illustration2vec はそれらと比較してイラストのより合理的なベクトル化が期待できる手法である. また, Danbooru と Safebooru でよく使われているタグを正解ラベルとして学習しているため, 簡単にイラストの特徴を検出でき, 大量の画像に対して類似画像を検索出来たり, 画像の意味における画像変換や応用例としてタグの特徴を満たす画像の生成などが可能となっている.

illustration2vec では学習データを下記の要領で作成している.

1. Danbooru と Safebooru から画像とメタデータを収集
2. メタデータを 4 つのカテゴリに分類する
   - general : 一般的な属性 (例 : ”smile”, ”short hair”)
   - copyright : 著作権名
   - character : キャラクタ名
   - rating : X レーティング (”safe”, ”questionable”, ”explicit”)
3. general, copyright, character から最も使われている 512 個のタグをそれぞれ抽出
4. 3 で抽出したタグと rating を連結させた 1539 個のタグをラベルとする

4　提案手法

本章では，本研究の提案手法について説明する．

4.1　漫画のセリフのマルチモーダルな感情推定手法

本研究では，4 コマ漫画ストーリーデータセットを用いて，各セリフにアノテートされた感情ラベルを推定するタスクを解き，その精度を確認する．

図 4.1 にマルチモーダルな推定手法として，提案手法の概要を示す．Text Embedding 層への入力として，あるセリフを選んだ時に，Image Embedding 層への入力をこのセリフが含まれているコマの画像全体とする．そして，それぞれの層から得たものをセリフベクトルとコマベクトルとし，これらを結合したものを識別器への入力とすることでセリフの感情ラベルを推定する．

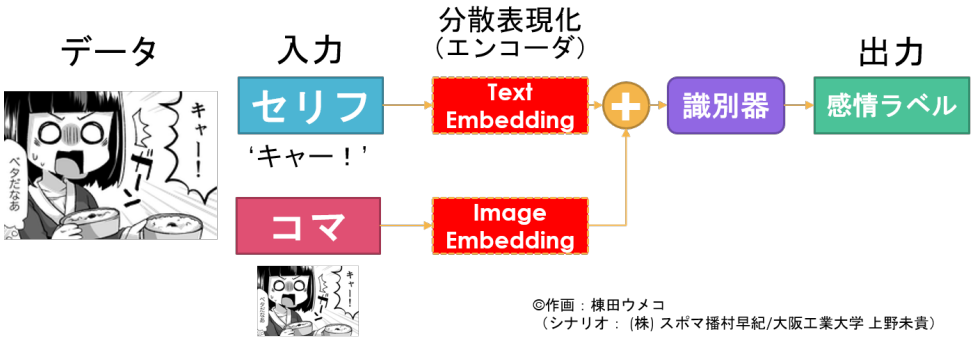

図 4.1: 提案手法の概要

4.2　Data Augmentation

4 コマ漫画ストーリーデータセットの欠点として，データ数が少ないことがあげられる．そこで，本研究では日本語 WordNet [15] のシソーラスを用いてテキストデータを拡張する．

図 4.2 に Data Augmentation の概要を示す．分かち書きされたオリジナルのセリフに対して，日本語 WordNet で類似語を持つ単語について類似語に置き換え，文を生成することでテキストデータを拡張した．ただし，文の中に類

義語を持つ単語が複数あった場合, 類似語に置き換える単語は同時に 1 つまでとし, 英数字・記号のみで表されている類似語は除外した. 例えば, 5 つの単語からなる文章があり, 各単語が 5 つの類似語を持っている場合, その文からは新しく 25 文が生成されることとなる.

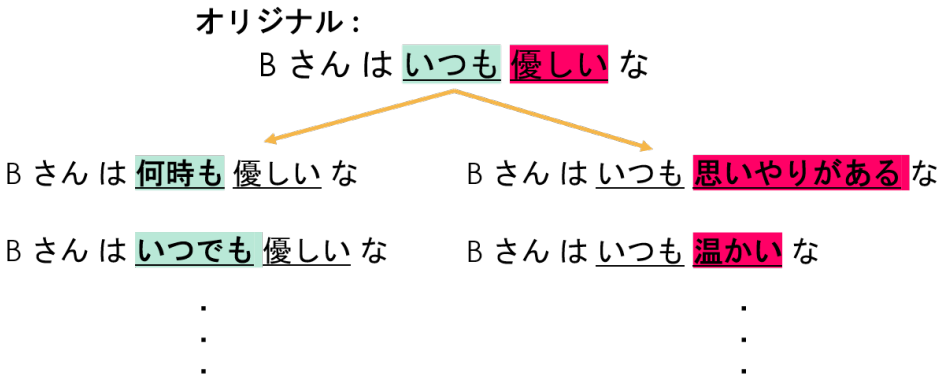

図 4.2: Data Augmentation の概要

5　数値実験

本章では，実験について説明する．

5.1　実験概要

BERT および MLP を用いて新聞記事データの段落間の接続詞の有無を推定する実験をした．

5.2　データセット

5.2.1　使用データ

本稿では叙述的な文章として毎日新聞データセット[1] の新聞記事を用いた．このデータセットにはジャンルごとに 2008 年から 2012 年までの記事がある．そのなかのジャンルが国際のもので本文が 10 行以上ある 5000 記事を用いた．それぞれの文章に対し，“■”や “◇” , また感嘆符といった記号を除去し，“＜＞” や “《》” の間に書かれる注釈や作者名等を除いた．その上で，データセットの性質上，数字の羅列などを含む記事や，箇条書された記事が含まれているので，そのような記事を取り除き，文章として乱れていない記事を使用データとして扱った．

5.3　実験準備

実験では新聞記事データの 1 から 3 面の本文を用いた．まず，新聞記事データの各本文を段落ごとに分けて，すべて分かち書きする．次に，はじめの段落を除いて，各段落のはじめに接続詞があるものについてはその接続詞を “[MASK]” に変換する．そうでないものは段落の始めに “[MASK]” という単語を追加する．続いて，接続詞の直後に “、” があるものは，それを取り除く (接続詞があることが容易に推定できてしまうため)．前後 2 段落を “[SEP]” でつなげたものを 1 データとして扱う．これらのトークンを含めて単語数が 248 個より多いデータを除く．

[1] http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html

これにより各データには 2 段落あり, それらは “[SEP]” でつながっている. さらに, その直後には必ず “[MASK]” がある. その “[MASK]” は, 通常の “[MASK]” のように, 接続詞を隠したものもあれば, ダミー (その部分には単語は入らない) もある.

5.4　実験

実験準備で得られた各データに対して, そのデータの 2 段落間の接続詞の有無を推定した. 今回は, データの不均等性に対してはデータ数を 1：1 にすることで対応した. また, 今回の実験では, 前回までで精度の良かった “[MASK]” 部分の分散表現に着目する方法のみをした. まず, 先程のデータを BERT の入力として, 得られた各単語ベクトルから “[MASK]” 部分の分散表現を得る. それを 3 層 MLP によって 2 次元にして, 接続詞の有無を推定した. 表 5.1 に実験時のパラメータを示す. 学習は BERT の最終層および MLP に対してした. 学習率および最適化アルゴリズムは optuna によって調整した.

表 5.1: 実験時のパラメータ

パラメータ	値
入力層の次元数	768
隠れ層のノード数	768
出力層の次元数	2
バッチサイズ	2
BERT の学習率	0.000105748
MLP の学習率	0.00004541588
最適化アルゴリズム	Adam
活性化関数 (隠れ層)	ReLU
活性化関数 (出力層)	Softmax function
目的関数	categorical cross entropy
学習終了条件	2 epoch

5.5　実験結果

データ数があまり多いわけではないので, 5 分割検証を行い, その際の精度の平均値, 標準偏差を比較した. また, ベースラインは, すべてをランダムに選択した際の期待値とした. 表 5.2 に 5 分割交差検証をしたときの平均及び標準偏差を示す.

表 5.2: 交差検証の結果

パラメータ	値
正解率	0.7502 (0.0176)
F 値	0.6656 (0.06164)

6　結論

本研究では,

謝辞

本研究を進めるにあたり御指導, 御鞭撻を賜りました森直樹教授に深く感謝申し上げます. 直接御指導頂きました岡田真助教には, 研究のアイデアや方針だけでなく, 論文の書き方や発表の作法に関することなど, 日頃から多岐に渡る御助言を頂きました. 心より御礼申し上げます. 最後に, 研究に関して建設的な意見をしてくださった諸先輩方, ともに支え合いながら研究に取り組んできた同期の皆さんに感謝致します.

2021 年 2 月 26 日

参考文献

[1] 松下光範. コミック工学：マンガを計算可能にする試み. 日本知能情報ファジィ学会 ファジィ システム シンポジウム 講演論文集, Vol. 29, pp. 199–199, 2013.

[2] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, Vol. 76, No. 20, pp. 21811–21838, 2017.

[3] 上野未貴. 創作者と人工知能: 共作実現に向けた創作過程とメタデータ付与 4 コマ漫画ストーリーデータセット構築. 人工知能学会全国大会論文集, 2018.

[4] T. KUDO. Mecab : Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.net/*, 2005.

[5] 京都大学大学院情報学研究科黒橋・河原研究室. 日本語形態素解析システム juman++ version 1.0. 2016.

[6] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, Vol. abs/1808.06226, , 2018.

[7] Greg Corrado Tomas Mikolov, Kai Chen and Jeffrey Dean. Efficient estimation of word rep-resentations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[8] Quoc V. Le and Tomás Mikolov. Distributed representations of sentences and documents. *CoRR*, Vol. abs/1405.4053, , 2014.

[9] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[10] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 30, pp. 5998–6008. Curran Associates, Inc., 2017.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[14] Masaki Saito and Yusuke Matsui. Illustration2vec: A semantic vector representation of illustrations. In *SIGGRAPH Asia 2015 Technical Briefs*, SA '15, New York, NY, USA, 2015. Association for Computing Machinery.

[15] Richard Fothergill Francis Bond, Timothy Baldwin and Kiyotaka Uchimoto. Japanese semcor: A sense-tagged corpus of japanese in the 6th international conference of the global wordnet association (gwc-2012), matsue. 2012.