中国語と日本語の同形異義語の意味の違いの可視化

理学専攻　情報科学コース　2240678　HAN HE　(**指導教員：伊藤 貴之**)

1　はじめに

中国語と日本語は多くの漢字を共有し，同じ漢字の熟語も多い．例えば,「検討」「質問」「連想」などがある．だが，同じ漢字の熟語でも，中国語と日本語で意味が同じとは限らない．このような単語は同形異義語と呼ばれる．本研究の目的は，中国語と日本語の同形異義語の意味の違いにどのようなパターンがあるかを観察することであり，そのために可視化を適用する．

本研究は3つのステップで構成される．まず，中国語と日本語の同形異義語とその意味をリストアップする．次に，SimCSE モデルを使って，これらの意味の文埋め込みを作る．そして，これらの文埋め込みを次元削減し，可視化する．この可視化を通じて，中国語と日本語の同形異義語の意味の違いにどのようなパターンがあるかを理解しやすくなる．さらに，可視化した結果をもとに，クラスタ分析を行い，その結果を分析する．本報告では，可視化とクラスタ分析から得られた面白い事例を紹介する．

2　処理手順

2.1　データの収集

本研究では,「日中同形異義語 1500」という重要な文献を活用して，中国語でも日本語でも使われる熟語を列挙する．この文献は，日本語と中国語で同形でありながら意味が異なる約 1500 の熟語を紹介している．本研究ではその中から，本研究で最も関連性が高いと思われる 352 の例を選び出し，詳細に分析している．その中から 5 組の代表的な例を表 1 に示す．この表では，中国語の意味と日本語の意味をいずれも日本語で記述している．

表 1: 日中同形異義語の例

単語	中国語の意味	日本語の意味
大丈夫	A man on his own	No problem
得体	Behavior appropriate for the occasion	True nature or appearance
汽車	A vehicle that runs on the road by engine	A railway vehicle powered by steam
前年	The year before last year	The year before a certain year
入口	Putting leaves into the mouth	The initial stage of something

2.2　「文埋め込み」の計算

本研究では，単語ベクトル加重平均法 [1]，BERT[2]，SimCSE 対比学習法 [3] の 3 つの方法で「文埋め込み」の算出を検討した．結果として本研究では，単語ベクトルの加重平均法を採用しなかった．これは文の構造を無視するためである．

BERT 法と SimCSE 法の検討結果は以下のとおりである．BERT 法は Google によって開発された多言語対応のオープンソース言語モデルで，実装と学習が容易である．SimCSE 法はより高次元の文埋め込みを提供していることから，本研究では SimCSE 対比学習法を適用して「文埋め込み」を算出する．しかし，現在の SimCSE 法の公開モデルは英語のみをサポートしているため，全てのテキストを英語に翻訳して処理している．中国語や日本語への対応は今後の課題である．

2.3　次元削減と可視化

本研究では，同形異義語の分析において重要な要素であるデータの次元削減方法について，t-SNE（t 分布型確率的近傍埋め込み）と PCA（主成分分析）の 2 つを比較検討した．t-SNE は，特に高次元データの可視化に適しており，データの局所的な構造を保持する能力が高い．t-SNE の大きな利点の一つは，その非線形性により，高次元データの複雑なパターンや構造を低次元空間にマッピングする際に，その局所的な構造を保持しやすい点にある．この特性は，特に同形異義語のような微細な意味の違いを可視化する場合に有用である．以上の理由から，本研究では t-SNE を次元削減手法として採用することに決定した．

クラスタリング手法としては．k-means を使用する．

本研究では次章にて，以下のような可視化を紹介する．

- i 番目の単語における中国語の意味の m 次元ベクトルを c_i とし，日本語の意味の m 次元ベクトルを d_i としたときに，中国語の意味のベクトルの集合 $C = \{c\_1, c\_2, \ldots, c\_n\}$ と日本語の意味のベクトルの集合 $D = \{d\_1, d\_2, \ldots, d\_n\}$ を 1 つのデータセットとして混合して次元削減を適用する．ここで n は単語数とする．
- 必要に応じて，c_i と d_i をエッジで連結する．
- 必要に応じて，中国語と日本語の同一単語の意味のベクトルを結合した 2m 次元のベクトルに対して，Kmeans 法によるクラスタリングを適用する．

3　実行例

3.1　「文埋め込み」のコサイン類似度

表 1 に示した 5 個の熟語に対して，中国語の意味と日本語の意味の文埋め込み結果のコサイン類似度を SimCSE 対比学習により算出した．結果を表 2 に示す．コサイン類似度が高いほど，その意味が似ていないことを示している．その中で最も意味の似ていない熟語が「入口」，最も意味の近い熟語が「前年」である．

表 2:「文埋め込み」のコサイン類似度

単語	中国語の意味と日本語の意味の「文埋め込み」のコサイン類似度
大丈夫	0.5424
得体	0.3491
汽車	0.5115
前年	0.8113
入口	0.3419

3.2　5 個の例に対する可視化結果

図 1 は，5 個の単語を対象にして次元削減を適用して可視化した結果である．同じ色のドットは，同じ単語の中国語と日本語の熟語を示している．意味の違いが大きい単語は，2 つのドットからも離れていることが確認できる．

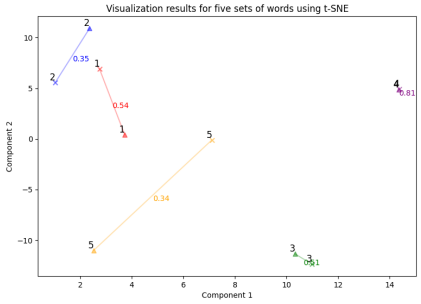

図 1: 5 つの例に対するの可視化結果

3.3　352 個の単語に対する可視化結果

続いて，文献 [4] から 352 組の同型異義語を選出し，これらの文埋め込みに対して次元削減を適用した．日本語の意味と中国語の意味に対応する 2 点を連結し，クラスタ分析を実施した．その結果を図 2, 3 に示す．

各クラスタの結果を詳細に分析したところ，いくつかの興味深い結果が得られた．クラスタ 24 には，「書法」,「筆頭」,「翻訳」,「流暢」の 4 つの語が含まれている．これらの語が同一クラスタに分類されたことは，両言語における言語能力や表現に関する概念の共通点を反映しているのに対して，中国語では具体的な能力や行動に焦点を当てる意味が強く，日本語ではより広範な概念を含むことを示している．

これらの結果から，同型異義語の文脈的な使用法や意味の違いに関する新たな洞察が得られ，言語理解の深化に寄与することが期待される．

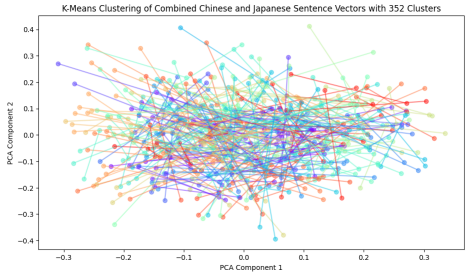

図 2: 352 つの単語に対するの可視化結果

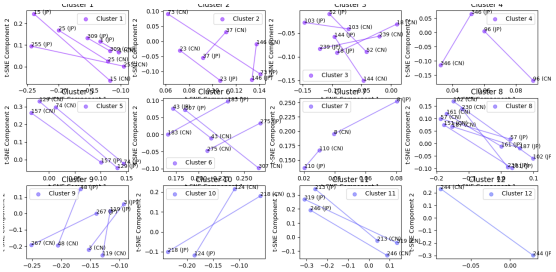

図 3: 一部のクラスタを抽出した可視化結果

4　まとめ

本研究では，中国語と日本語の同形異義語の意味の違いのパターンを可視化する試みを提案した．本研究では，中国語と日本語の同形異義語の意味を列挙し，SimCSE モデルを用いて，これらの意味を表す「文埋め込み」を形成する．そしてこれらの文埋め込みを次元削減して可視化する．今後の課題として，同形異義語の中には，中国語と日本語で異なる感情を表すものも存在する． 例えば,「緊張」という言葉は，日本語ではネガティブな意味，中国語ではポジティブな意味を持っている． 可視化結果にこのような次元を加えることができれば，さらに新しいパターンを発見できることが期待される．

参考文献

[1] Sanjeev Arora, Yingyu Liang, and Tengyu Ma, ”Simple but Tough-to-Beat Baseline for Sentence Embeddings”, International Conference on Learning Representations, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al., ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, arXiv preprint, arXiv:1810.04805, 2018.

[3] Tianyu Gao, Xingcheng Yao, and Danqi Chen, ”SimCSE: Simple Contrastive Learning of Sentence Embeddings”, arXiv preprint, arXiv:2104.08821, 2021

[4] Minghui Guo, Mieko Taniuchi, Yuko Isobe, and Shunsuke Kotani (eds.), ”1500 Chinese-Japanese Homographic Words”, Kokusai Gogaku-sha, 2011.