漢文読解問題の単語分散表現による解法

不破大智（福井大学工学部）
小高知宏・黒岩丈介（福井大学大学院工学研究科）
白井治彦（福井大学工学部）・諏訪いずみ（仁愛女子短期大学）

1. はじめに

国立情報学研究所の「ロボットは東大に入れるか」プロジェクト[1]において，「東ロボくん」はセンター試験における英語や国語の偏差値を上げることに苦戦したという．そこで国語の中でも点数の低かった古典分野のうち漢文に対象を絞り，自然言語処理によって漢文の問題を解くためにはどのような方法があるか，どのような補助を行えば正答率が向上するかなどについて調べることにした．今回は「選択肢の中から本文中にある下線部の正しい解釈を選ぶ」という読解問題について，単語分散表現を用いて解くことが出来ないかを考える．

2. 文章の単語リスト化と単語の分散表現

単語リスト作成の流れを図 1 に示す．UD-Kanbun[2]は形態素解析の結果から各単語を対応する英単語に翻訳してくれるが，これをさらに日本語に訳すと二重翻訳になるうえ，漢文単語から英単語への翻訳は単語ごとに行われるため文章として翻訳することが出来ない．そこで選択肢側を英訳してから単語に分割することで,「下線部の単語リスト(英語)」と「選択肢の単語リスト(英語)」を用意する．なお選択肢の英訳は，作為的な意訳にならないよう機械翻訳(今回は Google 翻訳)を使って行う．

完成した単語リストから fastText の学習済みモデル[3]を用いて各単語の分散表現を取得し,コサイン類似度を使って単語間の類似度を計算する．

3. コサイン類似度を用いた得点付け

図 2 のように下線部の単語リストと各選択肢の単語リストから総当たり的に単語間のコサイン類似度を求め，その結果をもとに得点付けをする．なお，この得点付けは 2 つの方法で行う．1 つ目の方法は，下線部の文を構成するそれぞれの単語について，その単語に対する類似度の最大値を得点に加算する方法である．2 つ目の方法はその単語に対する類似度の平均値を得点に加算する方法である．得点の指標とするべく下線部の文と同一の単語で構成される単語リスト F を追加して各選択肢の得点を計算したところ，次のような結果が得られた．

表 1：各選択肢の方法 1,方法 2 それぞれでの得点

	単語数	方法 1 での得点	方法 2 での得点
A	19	4.83	2.58
B	22	4.54	2.83
C	15	4.38	2.68
D	17	4.78	2.77
E	15	4.67	2.68
F	10	10.0	3.38

表 1 において青は正答である選択肢 E，赤は得点が最も高かった選択肢である．どちらの方法においても選択肢 E の得点は 1 位にならなかった．

4. 考察とまとめ

結果から，この方法で正答を目指すにはなんらかの改良が必要であるという知見が得られた．ほかの問題でも同様の結果になるのかの確認は必要だが，正答の得点が高くならなかったことについて，英語に変換するという手法をとっていることが原因の一つとして考えられる．より原文に近い状態のまま類似度を求められないか，コサイン類似度以外の類似度計算手法で解くことができないか検討することを今後の課題とする．

[1]国立情報学研究所,ロボットは東大に入れるか
https://21robot.org/index.html（参照 2021-04-13）
[2]漢字情報研究センター,古典中国語のコーパスの研究
http://kanji.zinbun.kyoto-u.ac.jp/~yasuoka/kyodokenkyu/（参照 2021-07-06）
[3]fastText, Word vectors for 157 languages
https://fasttext.cc/docs/en/crawl-vectors.html
（参照 2021-07-06）

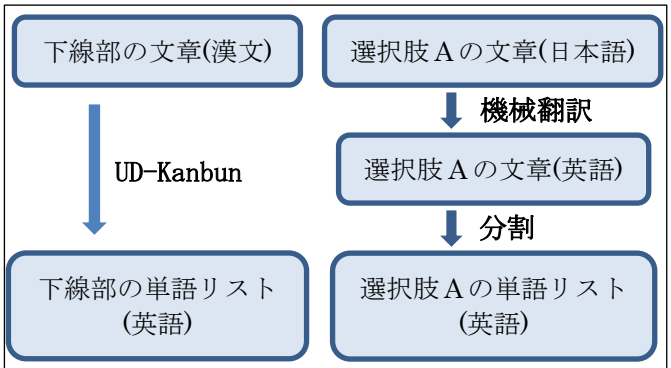

図 1：単語リスト作成の流れ

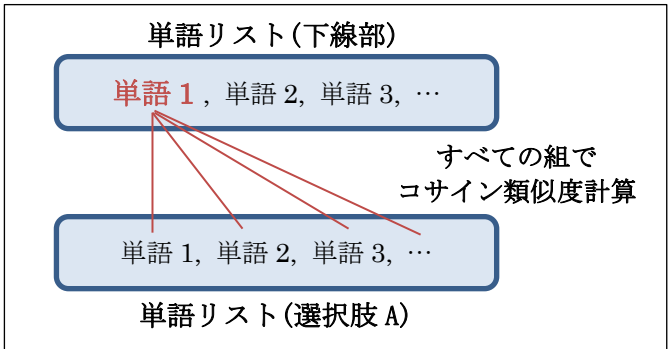

図 2：コサイン類似度計算