日英字幕付き音映像コーパスによる英語ライティング学習

田淵 龍二

アブストラクト: ライティングの自律的学習法としてコーパスによるデータ駆動型学習（corpus-based data-driven learning）について発表する。Google 翻訳は手軽だが，場面や文脈に即した表現が欲しい時，分野別コーパスが有力である。今回は，映画とプレゼンによる会話と講演の２種類の日英字幕付き音映像コーパス活用法を提案する。和英辞典用例が音映像付きで大規模になったことにより，多様な場面を見聞きしながら最適な表現を学習可能とする。本研究は AI の深層学習を言語学習に応用するパイロット研究である。

キーワード: 対訳コーパス，データ駆動型学習，ライティング，AI 深層学習，機械翻訳

１　はじめに

人工知能（AI）の活用が広まってきた。いまやプロの囲碁棋士が AI の打つ手を真似するほどである。言語分野でも読み上げや翻訳で実用化領域を拡大している。廉価な費用で言葉の壁が取り払われるのはまもなくだろう。

２　AI の機械翻訳手法を言語学習に使う

AI は如何にして読み上げや翻訳を行っているか。それは深層学習と言われる。しかし深層学習はブラックボックスである。そこで AI に至る機械翻訳の流れを整理した。ヒトの言語知識（文法と辞書）を人手で集大成してプログラミングする手法（ルールベース，1980 年代以降），対訳文の対応付けからルールを自動生成する手法（統計翻訳，コーパスベース，1990 年代以降）を経て，今日のニューラル翻訳に至ったのは 2010 年代になってからである。成功の要因はニューラル翻訳に必要な膨大な対訳資料をウェブから手軽に入手できるようになったことである。

本研究の目的は，ニューラル翻訳の手法を言語学習，特にライティングに応用する学習法・教授法の開発である。対象は基礎的な言語知識を習得した学習者（中学生以上）とする。

ミント音声教育研究所（tabuchiryuji@nifty.ne.jp）

３　方法

人とコンピュータの違いは速度と容量にある。人の場合一度（ひと目）で多くても 10 以下にする必要がある。つまり，大きな対訳コーパスと適切なフィルタによる所望情報の絞り込みだ。

表１　CORPORA 基本情報

		Seleaf	TED Talks
1	作品数	25 films	2,439 talks
2	収録時間	46 時間	610 時間
3	字幕数	43,659	661,787
4	総単語数	0.3M	5M words
5	異なり語数	13K	67K tokens
6	見出し語数	6K	16K lemmas
7	対訳形式	フレーズ訳	文訳（ズレあり）
8	適応学年	中２～高２	高１～高３

http://www.mintap.com/talkies/pac/corpora.html

3.1　対訳コーパス

対訳コーパスとしてコーポラを使った。コーポラ（CORPORA）は，映画と講演のビデオシーン（20 秒程度）を字幕検索するオープンサイトである。映画には Seleaf コーパス，講演には TED コーパスを収録し，字幕は 2 言語（英語と日本語）である。基本情報を表１に示す。

3.2　検索機能

英語あるいは日本語で検索する。英語での検索には正規表現と見出し語検索に対応している。

3.3　フィルタ機能

検索結果を絞り込むフィルタには共起フィルタと対訳フィルタがある。

3.3.1　共起フィルタ

検索結果に含まれる単語を度数順に示し，検索結果を再配列する。検索語とかかわり合いの深い単語(共起語)や表現を知ることができる。

3.3.2　対訳フィルタ

検索結果の対訳に含まれる単語を度数順に示し，検索結果を再配列する。検索語とかかわり合いの深い訳語や表現を知ることができる。

3.4　語義探索と語用探索

英単語検索結果から共起フィルタを使うと，一般的な用例・用法を知ることができる。対訳フィルタを使うと，その語の語義・語感・訳語を知ることができる。表 2 と図 1 に単語 ride での検索結果とフィルタ例を示す。

4　結果

対訳フィルタ（表 1）の最上位は「乗」である。そしてその下位に「車」や「旅」がある。動作・動作対象・目的などが読み取れた。また検索結果テキスト上位に” We ride our bikes”のような典型表現が音映像付き文脈とともに示された。

表 2　ride による探索事例

	検索語			ride
	コーパス			TED Talks
	ヒット数			319 views
	共起フィルタ		対訳フィルタ	
1	BIKE	26	乗	118
2	SHARE	12	車	60
3	CAR	11	自転車	31
4	BICYCLE	10	旅	14
5	SUBWAY	10	馬	14

フィルタのリストは，フィルタ語彙から機能語などをマスクしたあと，特徴語を人手で抽出したもの。

5　考察

データ駆動型学習の要点は，適切な情報を適切な分量だけ提示することであり，その点で十分な性能を達成していることがわかった。

参考資料

CORPORA (2018). Parallel Corpus Search Engine，ミント音声教育研究所，田淵龍二. http://www.mintap.com/talkies/pac/corpora.html

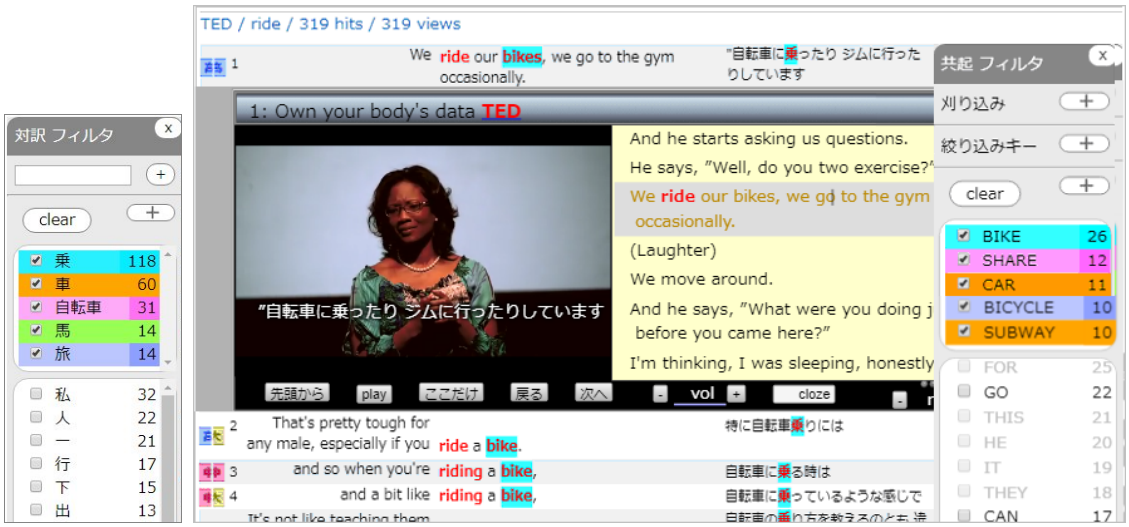

図 1　ride による検索結果と共起フィルタと対訳フィルタ（左）／結果を文字と音声と映像で確認