英語教育支援のための複単語表現平易化手法の検討

芦原和樹†，高田祥平‡，荒瀬由紀‡，内田諭*
† 大阪大学工学部電子情報工学科，‡ 大阪大学大学院情報科学研究科マルチメディア工学専攻，
* 九州大学大学院言語文化研究院

{ashihara.kazuki, takada.syouhei, arase}@ist.osaka-u.ac.jp
uchida@flc.kyushu-u.ac.jp

1 はじめに

グローバル化に伴い，ノンネイティブ話者の英語習得に対するニーズが高まっている．Laufer [1] は英文を適正に理解するためには95%の単語が既知であることが望ましいと指摘している．英語を学習する際も未知の単語や言い回しが多く含まれるテキストを使用すると，学習効率が低下するため，英語教育の現場においては，学習者にとって英文中の難解な単語やフレーズを教員が手作業で平易な表現に言い換えて教材として使用している．特に，単語の境界を越えて意味を構成する複単語表現 [2] は，複数の単語で構成されているため特定自体が困難であり，英語教育現場においてこれらを平易な表現に言い換える作業は大きな負担となっている．

本研究は英語教育における教材準備支援を目的とし，連続で現れる難解な複単語表現を平易な単語に言い換える手法を検討する．平易化は言い換え対象となる表現 (Target) の検索，言い換え先候補 (Candidate) の選定，Candidate のランク付けの 3 つのプロセスからなる．Paetzold ら [3] は単語の平易化における Candidate の選定に対し，各単語に品詞タグを連結して学習した単語分散表現を用いている．これまで，複単語表現を対象とした平易化の研究はほとんどされていなかった．本研究では Target を複単語表現とし，Candidate のランク付けのため構成単語を連結して学習した単語分散表現を用いる手法を提案する．

複単語表現には，その意味が個々の構成単語から想起できる compositionality をもつものと，そうでない non-compositional なものが存在する [4]．平易化においては，non-compositional な複単語表現の言い換えはより困難と考えられる．そこで本研究では，複単語表現の平易化における compositionality の影響についても分析する．

提案手法の性能評価のため，正解データを作成した．

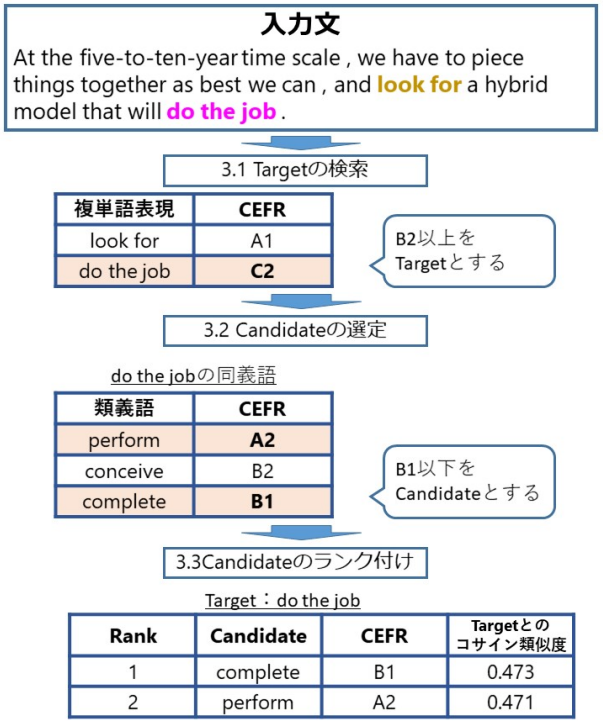

図 1: 提案手法の概要

各 Target に対して Candidate と言い換え可能であるか，及び compositionality を有するかをネイティブ話者に判定してもらった．実験の結果，複単語表現の構成単語を連結し，ベクトルを生成するモデルによって，高い精度で言い換え可能な単語を判定できることが示された．

2 提案手法

提案する複単語表現平易化の概要を図 1 に示す．

2.1 Target の検索

単語及び複単語表現の難易度については Common European Framework of Reference for Languages (CEFR) [5] に準拠する．English Vocabulary Profile (EVP)[*1] では，記載されている単語及び複単語表現に対して平易な順に A1，A2，B1，B2，C1，C2 と難易度が付けられている．本研究では，EVP 及び Thesaurus.com[*2]両方に採録されている難易度 B2 以上である 307 個の複単語表現を Target とする．Stanford CoreNLP[*3]を用いて入力文を原形にし，同順かつ連続で現れる複単語表現を検索し，Target とする．

2.2 Candidate の選定

Thesaurus.com に収録されている複単語表現の類義語のうち，難易度が B1 以下の単語を Candidate とする．単語の難易度は CEFR-J Wordlist Version 1.3[*4] 及び EVP で定義されたものを用いる．

2.3 Candidate のランク付け

Faruqui ら [6] は，類義語辞書を用いることにより類義語のベクトルがベクトル空間で近傍となるように単語分散表現モデルの最適化を行う Retrofitting を提案している．提案手法では分散表現モデルに加え，分散表現モデルに Retrofitting を適用するモデルの 4 つの手法それぞれで Target および Candidate のベクトルを生成する．そして，Target と各 Candidate のベクトルとのコサイン類似度を用いてランク付けを行う．類義語辞書として，Ganitkevitch ら [7] が単語及びフレーズの言い換え表現を収集した Paraphrase Database (PPDB) [*5]を使用する．

(1) w2v wikipedia の dump データ[*6] を学習データとし，CBOW を用いる word2vec [8] で単語分散表現を学習する．Target となる複単語表現の全構成単語のベクトルを足し合わせたものを Target のベクトルとする．

(2) w2v (PPDB) w2v について，約 12.3GB の PPDB に収録されている単語同士のペアを類義語辞書として用いて Retrofitting を行う．得られた単語ベクトルをもとに，w2v と同様の方法で Target のベクトルを生成する．

(3) w2v_m 全単語を原形にした wikipedia の文章について，複単語表現を構成する単語をアンダーバー（_）で連結して分散表現モデルの学習を行う．これにより，各複単語表現について一つのベクトルを生成する．

(4) w2v_m (PPDB) w2v_m について，PPDB 中の全データを用いて Retrofitting を行う．PPDB 中の複単語表現およびフレーズについては，それらの構成単語をアンダーバー（_）を用いて連結して学習する．

表 1: アノテーションデータ

Target 数	81
compositional	44
non-compositional	24
判定不能	13
Candidate 数	778
正解 Candidate 数	168

3 評価実験

3.1 実験データ

実験の対象テキストとして Rice 大学が公開している教科書データ[*7]を用いる．文中の各 Target について，それぞれの Candidate と言い換え可能かをネイティブ話者が判定した．このとき，判定には文脈を考慮している．アノテータは 30 代後半のイギリス出身のネイティブ話者で，日本での英語教育歴が 15 年以上ある人物である．豊富な英語教育経験を有しているため，アノテータとしての信頼性は高い．

また，各 Target の compositionality についてもアノテーションをを行った．例えば,「take advantage」(利点を生かす) は各単語の意味から複単語表現の意味を推測可能である．しかし,「pop the question」(結婚を申し込む) はそれが容易ではない．各 Target について,「単語から複単語表現の意味を推測できない」かアノテータに判定してもらった．Rice 大学の教科書データから，economics，psychology，sociology の教科書のイントロダクション及び本文中に現れる Target126

[*1] http://www.englishprole.org/wordlists
[*2] http://www.thesaurus.com
[*3] https://stanfordnlp.github.io/CoreNLP/index.html
[*4] http://www.cefr-j.org/download.html
[*5] http://paraphrase.org
[*6] https://dumps.wikimedia.org/enwiki/
[*7] https://cnx.org/

表 2: Target ベースの適合率

n	Rand.	w2v	w2v (PPDB)	w2v_m	w2v_m (PPDB)
1	0.358	0.543	0.519	**0.654**	0.617
2	0.539	0.765	**0.790**	0.778	**0.790**
3	0.658	0.840	**0.864**	0.840	**0.864**

表 3: Candidate ベースの適合率

n	Rand.	w2v	w2v (PPDB)	w2v_m	w2v_m (PPDB)
1	0.358	0.543	0.519	**0.654**	0.617
2	0.334	0.500	0.500	0.526	**0.551**
3	0.316	0.413	0.422	0.444	**0.493**

表 4: compositionality による Target ベースの適合率への影響 ($n = 1$)

	compositional	non-compositional
w2v	0.50	0.50
w2v_m	0.61	0.71

入力文
These leaders work hard to build consensus before choosing a **course of action** and moving forward .
Target : course of action
Candidate : approach, course, position, method ...
出力
W2v : course
W2v_m : approach

図 2: non-compositional な Target の例

入力文
In the 1980s , inflation rates **came down** in the United States and in Europe and have largely stayed down .
Target : come down
Candidate : decrease, fail, fall, reduce, improve ...
出力 : fail

図 3: 出力失敗例

個（Candidate は 1,133 個）をアノテーションの対象とした．

表 1 に示すアノテーションの結果，言い換え可能な Candidate をもつ Target 81 個を評価対象とする．この Target 81 個のうち compositional と判定された複単語表現が 44 個，non-compositional と判定されたものが 24 個，判断が難しく判定不能だったものが 13 個であった．Candidate の総数は 778 個であり，そのうち言い換え可能な Candidate は 168 個である．

3.2 評価指標

教育者支援のための平易化では，教育者にとって提示された Candidate の正誤判定は容易と期待できるため，適切な Candidate を少なくとも一つは含むリストを出力することが重要となる．そのため，Candidate のランク付け手法の評価指標として 2 種類の適合率を測定する．1 つ目は，n 個の Candidate をユーザに提示することを考え，ランク付けした上位 n 件中に正解が 1 つ以上存在する Target の割合を表す Target ベースの適合率を用いる．2 つ目はランク付けした上位 n 件中の正解 Candidate の割合を示す Candidate ベースの適合率を用いる．

3.3 結果

各手法で得られた Candidate のランク付けに対して，Target ベース及び Candidate ベースの適合率を表 2，表 3 に示す．Random は，Candidate をランダムにランク付けした場合の適合率の期待値である．表 2 より，$n = 3$ の場合，w2v (PPDB) および w2v_m (PPDB) では 0.864 の Target ベース適合率を達成した．これは，提示する Candidate の中に適切な言い換え候補が高い確率で含まれていることを示しており，目的である英語教育者支援において望ましい特性を示している．

表 3 より，Candidate ベースの適合率では，$n = 3$ の場合，w2v_m (PPDB) が最も高い性能を示した．Target ベースの適合率も考慮すると，w2v_m (PPDB) では w2v (PPDB) と同数の Target について正解となる Candidate を出力し，かつ各 Target の出力に含まれる正解 Candidate 数を増加できていることがわかる．一方で，Candidate 数の n を増加させるほど適合率が低下している．これは，一つの Target 当たりの正解 Candidate 数が平均 2.1 個のため，n を増加させるほど不正解の Candidate を出力してしまうからである．今後，Target ごとに適切な数の言い換え候補を出力するよう，手法を改善する予定である．

次に Retrofitting の効果について，n が 2 以上の場合では Target ベース及び Candidate ベースの両方において元の手法よりも高い性能を示している．Retrofitting を適用することにより，正解 Candidate が上位にランク付けされる割合が高まることが分かる．教育者支援においては複数の Candidate を提示し，妥当な候補を教育者が選択することを想定しているため，目的に対して望ましい特性といえる．

w2v と w2v_m 比較すると，w2v_m の方が Target ベース，Candidate ベースともに高い結果となった．これは，単純な単語ベクトルの足し合わせのみでは複単語表現を表現するベクトルとしては不十分であることを示唆している．表 4 に $n = 1$ の場合の w2v，w2v_m の Target ベースの適合率について，compositionality の影響を示す．現状アノテーションできている non-compositional な複単語表現の数は小さいが，non-compositional と判定された Target については特に，単語ベクトルの和によって複単語表現ベクトルとするアプローチが不向きである傾向が伺える．図 2 にその例を示す．下線が引かれている単語が正解 Candidate である．Target の course of action に対して，w2v では全構成単語の意味を足し合わせてベクトルを求めたため，まったく意味の異なる不正解の course が出力されている．一方，w2v_m では正解の approach を出力できている．今後アノテーションデータを増やし，compositionality の影響について詳細に分析する予定である．

また，w2v では図 3 に示すように入力文の文脈を考慮しないため，誤った Candidate を言い換え可能と判定してしまう Target が存在する．come down は fail の意味を含んでいるが，入力文中では inflation rates が主語となっているため，fail への言い換えはできない．文脈によって意味が異なる複単語表現が存在することから，Target だけでなく入力文の文脈情報も考慮する必要がある．

4 まとめ

本研究では，英語教育者支援を目的とした複単語表現の平易化手法を提案した．実験の結果，複単語表現のベクトル化は構成単語全てのベクトルの単純な和ではなく，構成単語を連結し一つのベクトルを生成することが有効であることが明らかとなった．特に，non-compositional な複単語表現についてはその傾向が強いことが分かった．

今後，深層学習等を利用して，入力文の文脈を考慮に入れた平易化ができるように手法を拡張する予定である．

5 謝辞

本研究は KDDI 財団による助成を受けたものである．

参考文献

[1] Batia Laufer. What Percentage of Text-Lexis is Essential for Comprehension? *Special Language: From Humans to Thinking Machines*, pp. 316–323, 1989.

[2] Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword Expressions: A Pain in the Neck for NLP. *In Proc. of CICLing*, pp. 1–15, 2002.

[3] Gustavo Henrique Paetzold and Lucia Specia. Lexical Simplification with Neural Ranking. *In Proc. of EACL*, Vol. 2, pp. 34–40, 2017.

[4] Siva Reddy, Diana McCarthy, and Suresh Manandhar. An Empirical Study on Compositionality in Compound Nouns. *In Proc. of IJCNLP*, pp. 210–218, 2011.

[5] A.J.Charles. The CEFR and the Need for More Research. *The Modern Language Journal*, pp. 659–663, 2007.

[6] Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. Retrofitting Word Vectors to Semantic Lexicons. *In Proc. of ACL*, pp. 1606–1615, 2015.

[7] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB : The Paraphrase Database. *In Proc. of NAACL*, pp. 758–764, 2013.

[8] Radim Rehurek and Petr Sojka. Software Framework for Fopic Modelling with Large Corpora. *In Proc. of LREC*, pp. 45–50, 2010.