CEFRレベルを付与した語彙的換言データセットの構築

芦原和樹 †，荒瀬由紀 †，内田諭 ‡
† 大阪大学大学院情報科学研究科，‡ 九州大学大学院言語文化研究院
{ashihara.kazuki，arase}@ist.osaka-u.ac.jp ， uchida@flc.kyushu-u.ac.jp

1 はじめに

言語学習・教育支援，ノンネイティブ話者の文章作成支援を目的として，語彙的換言技術が研究されている．SemEval-2007 における語彙的換言タスク [1] では，言い換え対象であるターゲット単語の文脈中での意味を考慮し，与えられた言い換え候補群を言い換えの可能性に準じてランク付けする．ランク付けの精度は，複数のアノテータが実施したアノテーションから得られたスコアを元に評価される．しかし，一般的に利用されるデータセット [1, 2] では，言い換え候補はアノテータが列挙しているため網羅性が低く，また単語の難易度を考慮していないため，言語学習・教育支援への応用が難しいという問題があった．

著者らの研究グループでは，言い換え候補となりうる全ての同義語について言い換え可能性をアノテーションしたデータセットとして，CEFR-LS [3] を作成している．CEFR-LS は言語教育支援を目的とし，平易な語に換言する語彙平易化のためのデータセットである．ターゲット単語及び言い換え候補単語に CEFR [4] レベルを付与し，言い換え対象の単語より平易な同義語を言い換え候補としてアノテーションを行っている．信頼性の高いデータセットを構築するため，英語教育に携わっているネイティブ話者によるアノテーションを実施している．以上の性質上，言い換えによって難易度を向上させる場合も含む一般的な語彙的換言タスクには CEFR-LS は適用できない．またアノテータは信頼度は高いが 1 名のみであり，換言可能かどうかの 2 値のアノテーションとなっているため，換言可能性を連続値として表現できない．

そこで，本稿では CEFR-LS を拡張することで，語彙的換言タスクのためのデータセットである CEFR-LP (CEFR basd Lexical Paraphrase) を作成した．CEFR-LP では，言い換え対象の同義語全てに CEFR レベルの付与と言い換え可能かどうかのアノテーションを行っている．アノテーションは Amazon Mechanical Turk[*1] を用い，複数人で行った上で集約することで，アノテーション品質の保証をしながら言い換え可能性を連続値として表現することが可能となった．

本稿では，ターゲット単語総数 263，言い換え候補総数 4,339 に対して 3 段階でアノテーションを行い，25,222 件のアノテーション結果を得た．構築した CEFR-LP は語彙的換言タスクの評価セットとして利用できるよう，公開している[*2]．

2 関連研究

語彙的換言タスクでは，ターゲット単語とそれを含む文脈，そして言い換え候補群が与えられる．様々な手法 [5, 6] が提案されており，LS-SE [1] や LS-CIC [2] が評価セットとして広く用いられている．LS-SE は SemEval-2007 の語彙的換言タスクで用いられたデータセットである．201 種類のターゲット単語について，それぞれ 10 種類の文脈が与えられている．各ターゲット単語には，5 人のアノテータが最大 3 種類ずつの言い換えを文脈を考慮して生成している．LS-CIC は語彙的換言タスクのための大規模なデータセットであり，15,629 語のターゲット単語について，6 人のアノテータが最大 5 種類ずつの言い換え候補を文脈を考慮して生成している．これらのデータセットでは，生成したアノテータの人数を言い換え候補群のスコアとして利用している．また，同一のターゲット単語に対して異なる文脈で生成された言い換えを不正解（スコア 0）の言い換え候補群として用いる．そのため，本来存在するであろう言い換え候補について網羅性が低いという問題がある．

LexMTurk [7] および BenchLS [8] は語彙平易化タスク [9] のためのデータセットである．これらのデータセットも，アノテータが生成した言い換えのみを正

[*1] https://www.mturk.com/mturk/welcome
[*2] http://www-bigdata.ist.osaka-u.ac.jp/arase/pj/CEFR-LP.zip

解としており，可能な言い換え候補を網羅しているとはいえない．

3 データセットの作成

この章では CEFR-LP の作成法について述べる．CEFR-LP は CEFR-LS [3] を拡張することで作成する．CEFR-LS は，Rice 大学が使用する OpenStax website[*3]で公開されている教科書データを元に作られている．各単語の難易度は Common European Framework of Reference for Languages (CEFR) [4] に準拠している．CEFR は世界中で学習者の言語運用能力を客観的に評価するために使われている指標であり，平易な順に $A1$，$A2$，$B1$，$B2$，$C1$，$C2$ の 6 種類のレベルから構成されている．CEFR-LS ではターゲット単語の CEFR レベルは $B2$，$C1$，$C2$ に限定し，言い換え候補群の CEFR レベルは $A1$，$A2$，$B1$ に限定している．本データセットでは同様に単語の難易度として CEFR を使用し，一般的な語彙的換言タスクに利用できるようにするため，言い換え候補群は全ての CEFR レベルを対象とする．

3.1 アノテーションセットの作成

本アノテーションでは，CEFR-LS 同様 CEFR レベルが $B2$，$C1$，$C2$ の単語をターゲット単語とし，それぞれに言い換え候補群を抽出したものをアノテーション対象とする．言い換え候補群については，全ての CEFR レベルの単語をアノテーション対象する．言い換え候補群は CEFR-LS 同様，類義語辞書[*4]を利用して獲得した．各単語に CEFR レベルが付与された辞書[*5] [*6]を利用し，単語と Stanford Parser [10] から得られる品詞情報を元に単語の難易度を付与した．難易度が付与できない類義語は言い換え候補群には加えていない．アノテータの負荷を考慮し，ターゲット単語は言い換え候補数が 30 未満のものとした．表 1 にアノテーション対象となるターゲット単語及び言い換え候補群の数と CEFR レベル別の内訳を示す．アノテーション対象となる言い換え候補群の総数は 4,339 である．

過負荷とならないよう，アノテーションデータはターゲット単語が平均 6.3 個，言い換え候補群は平均 103.3 個（最大 105 個最少 94 個）となるよう分割し，アノテータに配布した．

[*3] https://cnx.org/
[*4] http://www.thesaurus.com/
[*5] http://www.englishprofile.org/wordlists
[*6] http://www.cefr-j.org/download.html

表 1: アノテーション対象の内訳

CEFR レベル	ターゲット単語	言い換え候補群
計	263	4,339
$A1$	0	650
$A2$	0	946
$B1$	0	1403
$B2$	186	951
$C1$	30	178
$C2$	47	211

3.2 アノテーションの実施

アノテーションは Amazon Mechanical Turk を通して行った．アノテータの条件としては，アメリカでの学位を取得していることを必須条件とした．加えて MTurk マスターである，もしくは MTurk 上での acceptance rate が 98% 以上であることを条件とした．

図 1 に示すのが実際にアノテータが作業する画面の一部である．アノテータには Text，Target，Candidate が与えられる．Text ではターゲット単語を含む 1 文とその前後 2 文ずつを表示した．これは，ターゲット単語を含む 1 文のみでは判定が難しい場合も考えられるためである．赤文字となっている単語がターゲット単語である．Target にはターゲット単語を含む 1 文及びターゲット単語を記載した．Candidate には，言い換え候補群を記載した．与えられた言い換え候補とターゲット単語の言い換えの可否を Sure，maybe，not possible の 3 段階からアノテータが判定する．判定には [3] で設計した下記のアノテーション基準を用いる．

Grammatical Reformation Stage ターゲット単語と言い換え候補の単語を言い換えた際，品詞や前置詞とのつながり等，文法的な正確性が維持されること．ただし，時制および人称については原文と同一のものが自動的に適用されるものとする．

Definition Stage ターゲット単語と言い換え候補の単語とが共通した意味を持つこと．

Context Stage 言い換えた際に言い換え候補の単語が表す意味や他の単語との整合性が維持されること

以上の条件を全て満たす場合に sure を選択する．1 つでも満たさない条件があった場合は not possible を，

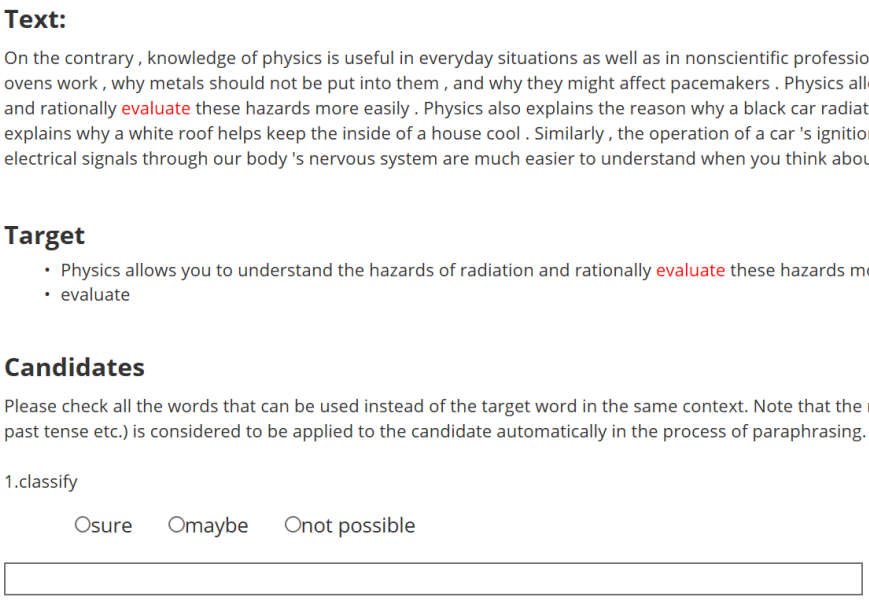

図 1: アノテーション画面

表 2: アノテーション結果の一例

言い換え候補	P1	P2	P3	P4	P5
block	0	0	0	0	0
development	2	0	2	2	2
advancement	2	2	2	2	2
break	0	0	1	0	1
breakthrough	2	2	2	2	1
failure	0	0	0	1	0

判断が難しい場合は maybe を選択する．特筆事項があった際は枠線で囲まれたボックスの中に記載する．以上のアノテーションをアノテーションセットあたり 5 人以上が行うものとした．

4 CEFR-LP の分析

表 2 に示すのは，*From alchemy came the historical **progressions** that led to modern chemistry : ...* という文脈中のターゲット単語 *progressions* に対するアノテーション結果の一部である．それぞれの言い換え候補に対して P1〜P5 の 5 名がアノテーションを行っている．アノテーションは 0 (not possible)，1 (maybe)，2 (sure) の 3 段階で行われている．各アノテーションセットには最低 5 名，平均 6.1 名がアノテーションを行い，延べ 26330 件のアノテーションを得た．アノテーションの内訳は 0 が 15457 件，1 が 5477 件，2 が

表 3: データセットの一例

Sentence	From alchemy came the historical **progressions** that led to modern chemistry : ...
Target	progression
Index	5
CEFR	$C1$
POS	noun

Candidate	CEFR	Score	アノテーション
block	$B1$	0.00	0 0 0 0 0
development	$B1$	2.00	2 2 2 2
advancement	$B2$	2.00	2 2 2 2 2
break	$A2$	0.40	0 0 1 0 1
breakthrough	$B1$	1.80	2 2 2 2 1
failure	$B1$	0.20	0 0 0 1 0

5396 件であった．それぞれのアノテーションセットごとに対する Fleiss のカッパ値の平均は 0.32 であった．また，各言い換え候補に対するアノテーションの標準偏差の平均は 0.45 であった．アノテーションデータの信頼性を担保するため，各言い換え候補ごとに，中央値とアノテーションが 1 より乖離しているデータを取り除く．該当するアノテーションは全体の 4.21% (1108 件) であり，それらを取り除いた後の標準偏差

表 4: 本データセットの詳細な統計

	CEFR-LP
ターゲット単語総数	263
言い換え候補群総数	4,339
平均言い換え候補群数	16.5
総アノテーション数	25,222
異なり Sentence 数	255
Sentence の平均単語数	25.4

の平均は 0.35 であった．また，アノテーションの内訳は 0 が 15077 件，1 が 5477 件，2 が 4668 件となった．得られたデータセットに対してアノテーションの値をすべて足し合わせ，人数で割った値をその言い換え候補のスコアとする．

最終的に得られたデータの 1 例を表 3 に示す．Sentence はターゲット単語が含まれる 1 文を，Target はターゲット単語を表す．また，Index は Sentence 中のターゲット単語のインデックスを，CEFR は CEFR レベルを，POS はターゲット単語の品詞を表す．Candidate は言い換え候補群を表し，それぞれに対して CEFR レベルと Score が記載されている．Score は先述した言い換え候補のスコアであり，言い換え候補がどの程度ターゲット単語と言い換え可能であるかを $0 \leq Score \leq 2$ で表し，値が大きいほど言い換え可能性が高いことを示す．

最終的に作成された CEFR-LP は，表 4 に示すように 263 のターゲット単語に対して，言い換え候補群である 4,339 個の単語全てにアノテーションが行われたデータセットである．

5 まとめ

本稿では著者らが構築した CEFR-LS を拡張し，語彙的換言タスクのための評価セットである CEFR-LP を構築した．本データセットは言い換え候補群全てに複数人によるアノテーションを行っている．また，単語には難易度が付与されているため語彙平易化タスクの評価セットとしても利用できる．今後はターゲット単語を全ての CEFR レベルに拡張し，より汎用性の高いデータセットを作成する予定である．

6 謝辞

本研究は公益財団法人 KDDI 財団による助成を受けたものである．また，貴重なコメントや議論をいただいた九州大学大学院言語文化研究院の Christopher G. Haswell 准教授に感謝の意を表す．

参考文献

[1] Diana McCarthy and Roberto Navigli. SemEval-2007 Task 10: English Lexical Substitution Task. *In Proc. of SemEval*, pp. 48–53, 2007.

[2] Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. What Substitutes Tell Us - Analysis of an "All-Words" Lexical Substitution Corpus. *In Proc. of EACL*, pp. 540–549, 2014.

[3] Satoru Uchida, Shohei Takada, and Yuki Arase. CEFR-based Lexical Simplification Dataset. *In Proc. of LREC*, pp. 3254–3258, 2018.

[4] Council of Europe. Common European Framework of Reference for Languages: Learning, teaching, assessment. *Cambridge: Cambridge University Press*, 2011.

[5] Oren Melamud, Jacob Goldberger, and Ido Dagan. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. *In Proc. of SIGNLL*, pp. 51–61, 2016.

[6] Kazuki Ashihra, Tomoyuki Kajiwara, Yuki Arase, and Satoru Uchida. Contextualized Word Representations for Multi-Sense Embedding. *In Proc. of PACLIC*, 2018.

[7] Colby Horn, Cathryn Manduca, and David Kauchak. Learning a Lexical Simplifier Using Wikipedia. *In Proc. of ACL*, pp. 458–463, 2014.

[8] Gustavo H Paetzold, Lucia Specia, and Western Bank. Benchmarking Lexical Simplification Systems. *In Proc. of LREC*, pp. 3074–3080, 2016.

[9] Gustavo Henrique Paetzold and Lucia Specia. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, pp. 549–593, 2017.

[10] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. *In Proc. of ACL*, pp. 55–60, 2014.