文脈ベクトルと細分化した単語ベクトルを用いた語彙的換言

芦原 和樹†，梶原 智之‡，荒瀬 由紀†，内田 諭*

†大阪大学大学院情報科学研究科，‡大阪大学データビリティフロンティア機構，*九州大学大学院言語文化研究院

語彙的換言タスク

語彙的換言タスク [1, 2, 3]

... and you are required to listen **hard**.

One event in particular hits the platoon **hard** ...

carefully

badly

- 同じ単語でも異なる意味を持つ場合がある
- 文脈中の語義に合わせたベクトルが必要

アプローチ1 [4, 5] ：静的な単語ベクトル生成モデル

SoTA：DMSE [5]

✓ 文脈化した単語ベクトル

✗ 文脈中の1単語しか考慮しない

アプローチ2 [6, 7] ：動的なベクトル生成モデル

SoTA：context2vec [6]

✗ 単語ベクトルは文脈化されていない

✓ 1文全体の情報を考慮したベクトル

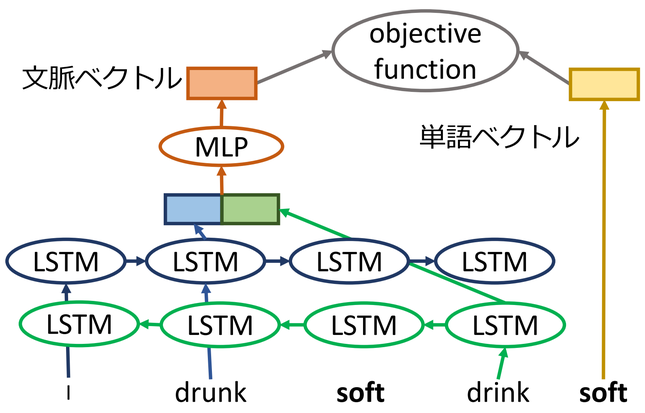
手法

アプローチ1とアプローチ2を組み合わせる

✓ 文脈化した単語ベクトル

✓ 1文全体を考慮した文脈ベクトル

事前学習

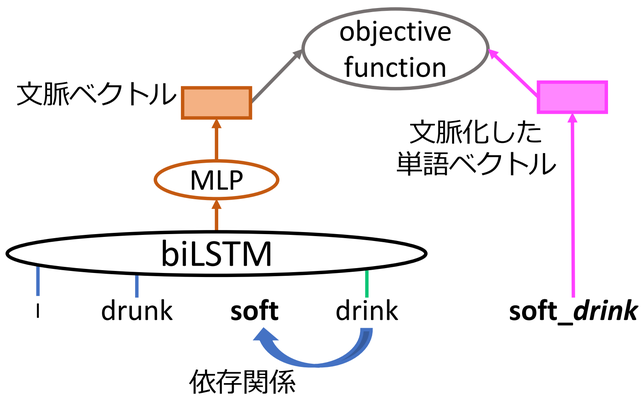


- context2vec [6]
- 単語ベクトルの学習
- LSTMのパラメータ学習

✓ 1文全体を考慮した文脈ベクトル

✗ 単語ベクトルは文脈化されていない

事後学習



- DMSE [5]
- 依存関係にある単語(dependency-word)ごとに単語ベクトルを割り当て
- LSTMのパラメータと事前学習した単語ベクトルは固定

✓ 文脈化した単語ベクトル

実験

- 余弦類似度の高い順にランキング

$$S(v_s^d \mid v_t^d, v_c) = (\cos(v_s^d, v_t^d) + 1)(\cos(v_s^d, v_c) + 1)$$

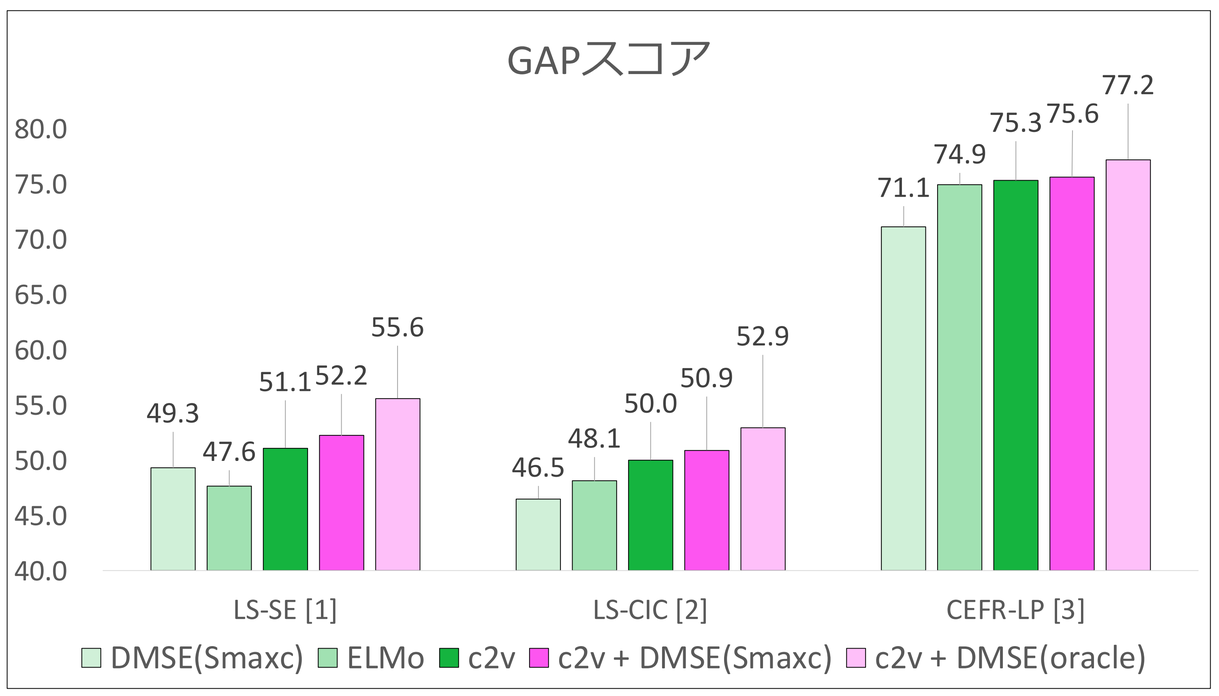
- dependency-wordの選択

S_{maxc}：$d = \underset{d \in D}{argmax}\, S(v_s^d \mid v_t^d,\ v_c)$

S_{oracle}：$d =$ GAPスコアが最も高くなるd

s ：言い換え先単語
t ：言い換え元単語
c ：文脈
d ：*dependency-word*

結果



- DMSE (S_{maxc})よりもc2v + DMSE (S_{maxc})が高い精度
  → 文脈全体を見ることの有効性
- c2vよりもc2v + DMSE (S_{maxc})が高い精度
  → 単語ベクトルを細分化することの有効性
- c2v + DMSE ($oracle$)が最も高い精度
  → 正しいdependency-wordを選択すれば更に精度が向上

実験設定

モデルの訓練

- English Wikipediaの使用
- dependency-wordは内容語のみ（名詞・動詞・形容詞・副詞）

モデル	概要
DMSE [5]	DMSEで学習。dependency-wordごとに単語を文脈化。
c2v [6]	context2vecで学習。提案手法の事前学習モデル。
ELMo [7]	双方向言語モデルによって学習。3層から得られるベクトルを連結*して用いる。

*予備実験の結果最も高い性能を発揮

出力例

Target	**go**
文脈	… , explain the basic concept and purpose and get it **going** with minimal briefing .
DMSE (S_{maxc})	try (0), move (1), proceed (1), leave (0), be (0), …
c2v	proceed (1), run (0), start (4), move (1), take (0), ...
c2v + DMSE (S_{maxc})	start (4), proceed (1), move (1), run (0), take (0), ...

表：DMSE(S_{maxc}) ・c2v・c2v + DMSE (S_{maxc})の出力の比較

- 正しい言い換え候補をより上位に出力
- 文脈ベクトルと細分化した単語ベクトル両方を使うことでより正しい言い換えが可能

[Reference]
[1] McCarthy et al., 2007, “SemEval-2007 Task 10: English Lexical Substitution Task,” In Proc. of SemEval, pages 48-53.
[2] Kremer et al., 2014, “What Substitutes Tell Us - Analysis of an “All-Words” Lexical Substitution Corpus,” In Proc. of EACL, pages 540-549.
[3] 芦原ら, 2019, “CEFRレベルを付与した語彙的換言データセットの構築,” 言語処理学会年次大会, pages 803-806.
[4] Fadaee et al., 2017, “Learning Topic-Sensitive Word Representa-tions,” In Proc. of ACL, pages 441-447.
[5] Ashihara et al., 2018, “Contextualized Word Representations for Multi-Sense Embedding,” In Proc. of PACLIC, pages 28-36.
[6] Melamud et al., 2016, “context2vec: Learning Generic Context Embedding with Bidirectional LSTM,” In Proc. of CoNLL, pages 51-61.
[7] Peters et al., 2018, “Deep Contextualized Word Representations,” In Proc. of NAACL , pages 2227-2237.