E-002

文から単語への言い換え手法の提案

Proposal of Paraphrasing Methods from Sentences to Words

榎並 龍大† 杉本 徹†
Tatsuhiro Enami Toru Sugimoto

1. はじめに

コミュニケーションにおいて，言い表したい概念を簡潔に表す言葉が思いつかないことがある．例えば，「自然言語処理」という概念を言い表したいが，この単語を忘れてしまった場合に「人が使う言葉をコンピュータで処理する技術」といった文を用いることで概念を表すことができる．しかし，このような表現は冗長であり理解も難しくなることが懸念される．そのため，与えられた文の内容に合う意味をもつ単語を自動的に発見し，文から単語への言い換えを提案するシステムが望まれる．

自然言語処理において，言い換えに関する様々な研究が行われているが，文から単語への言い換えを目的とした研究は少ない．

このような背景を踏まえて本研究では，入力文と意味的に近い単語を推定することで文から単語への言い換えを行う手法を提案する．

2. データセットの作成

2.1 言い換え候補単語リスト

本研究では，言い換え候補となる単語の意味を表す文として岩波国語辞典第五版タグ付きコーパス 2004[1]に含まれる語義説明文を用いる．岩波国語辞典に含まれる約 5 万単語のうち次の 3 つの条件をすべて満たす 4,142 単語を本研究における言い換え候補単語リストとする．

(1) 一文で終わる
(2) 文の文字数は 15 文字以上
(3) 文末単語の品詞が名詞

2.2 言い換えデータセット

言い換えの実験に使う入力文は，前節で述べた岩波国語辞典とは異なる 2 つの辞書（新明解国語辞典[2]，例解新国語辞典[3]）における語義説明文を用いる．すなわち，単語 w の辞書[2]または[3]における語義説明文を S_w とするとき，入力文として S_w を与えたときの言い換え先の正解単語を w と見なす．ただし w は言い換え候補単語リストに含まれる単語であるとする．このような S_w と w の組をモデル学習用に 100 組，評価実験用にそれとは異なる 50 組，それぞれ用意した．言い換えデータの例を表 1 に示す．

表 1 言い換えデータの例

入力文	正解単語
竹や木の骨組みをスギ・ヒノキなどの青葉で包んだ門	アーチ
秋、穂状の赤い小花をつける、たで科の一年生植物	あい 藍

†芝浦工業大学 Shibaura Institute of Technology

3. 提案手法

入力文 S に対して，言い換え候補単語リストに含まれる各単語 w の岩波国語辞典における語義説明文 T_w と S との文間類似度をすべて計算し，類似度が大きい順に言い換え候補単語リストの単語をソートして出力する．2 つの文 S, T の類似度を計算する方法として，本研究では次の 6 つを提案する．

3.1 Sentence-BERT

2 つの文 S, T をそれぞれ Sentence-BERT[4]に入力し，分散表現を獲得する．そして分散表現同士のコサイン類似度を計算する．

3.2 Jaccard 係数

ストップワードを除いた上で，文 S の単語集合と文 T の単語集合における共通要素が占める割合を式(1)により計算する．

$$Jaccard(S,T) = \frac{\text{文}S\text{と}T\text{の両方に含まれる単語の数}}{\text{文}S\text{と}T\text{のいずれかに含まれる単語の数}} \quad (1)$$

3.3 単語分散表現の平均

2 つの文 S, T に対して，それぞれの文に含まれるストップワード以外の単語の Word2Vec[5]による分散表現の平均を求めて，それらの間のコサイン類似度を計算する．

3.4 文末単語の分散表現

2 つの文 S, T に対して，それぞれの文の末尾の単語の Word2Vec による分散表現間のコサイン類似度を計算する．

3.5 Word Mover's Distance

2 つの文 S, T の距離を Word Mover's Distance（WMD）[6]を用いてスコアとして求める．そして求めた距離を 1 から引いた値を類似度とする．

3.6 アンサンブル手法

3.1 節から 3.5 節で述べた 5 つの手法を組み合わせてよい高い精度の類似度を得るために回帰モデルの学習を行う．言い換えデータセットにおける（S_w, w）の形をしたモデル学習用データ 100 組を用いて，入力文 S_w と言い換え候補単語リストに含まれる各単語 w' の語義説明文 $T_{w'}$ の組を作り，S_w と $T_{w'}$ の類似度を 5 つの手法で求めた値を説明変数とする．また，単語 w と w' の分散表現のコサイン類似度を求めて，これを目的変数とする．単語の分散表現は未知語に対応可能である fasttext[7]を用いて獲得した．この説明変数と目的変数を用いて，重回帰分析とサポートベクトル回帰（SVR）モデルの学習を行う．SVR のカーネル関数には，線形カーネルと RBF カーネルの 2 つを用いた．

得られた 3 つの回帰モデルのうち決定係数が最大のモデルを用いて，2 つの文 S, T に対する 5 つの説明変数の値から推測した値をアンサンブル手法による S と T の類似度とする．

4. 実験

4.1 回帰モデルの選択

回帰モデルの学習はモデル学習用に用意した 100 組の言い換えデータをすべて用いる方法の他に学習データを 10, 20, 30, 40 組に限定する方法を試みた．その結果，20 組のデータを用いる場合に決定係数が全体的に大きくなった．その際の 3 つの回帰手法の決定係数を表 2 に示す．表 2 から RBF カーネルを用いた SVR が最も良い結果であったため，次節で述べる実験では SVR(RBF) を用いる．

表 2　回帰モデルの決定係数

	重回帰分析	SVR(Liner)	SVR(RBF)
決定係数	0.035	0.032	0.054

4.2 評価方法

言い換えデータセットの評価実験用データ 50 組に対して，3 章で提案した 6 つの手法それぞれで実験を行い，正解単語の順位が上位に来るかどうか評価する．まず，正解単語の順位 rank を次の式(2)で定義する．

$$rank = p + \frac{q}{2} + 1 \tag{2}$$

ただし，p は言い換え候補単語リストのうち正解単語よりも入力文との類似度が大きい単語の数，q は言い換え候補単語リストのうち正解単語以外で入力文との類似度が正解単語と等しい単語の数とする．

順位の評価は Mean Reciprocal Rank（MRR）と Top-k accuracy（k= 1, 3, 10）を用いる．MRR の計算式を式 (3) に示す．ここで N はクエリの総数（今回は 50）を示す．

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i} \tag{3}$$

4.3 実験結果

6 つの提案手法で得られた実験結果を MRR と Top-k accuracy（k=1,3,10）で評価した結果を表 3 に示す．

表 3　6 つの手法の評価結果

手法	MRR	Top-k accuracy		
	(N=50)	(k=1)	(k=3)	(k=10)
Sentence-BERT	0.635	0.540	**0.720**	**0.760**
Jaccard	0.364	0.300	0.400	0.520
W2V Mean	0.380	0.300	0.420	0.540
W2V Last	0.179	0.060	0.220	0.320
WMD	0.456	0.380	0.500	0.580
Ensemble	**0.638**	**0.600**	0.640	0.720

また，6 つの手法における正解単語の順位の分布を図 1 に示す．

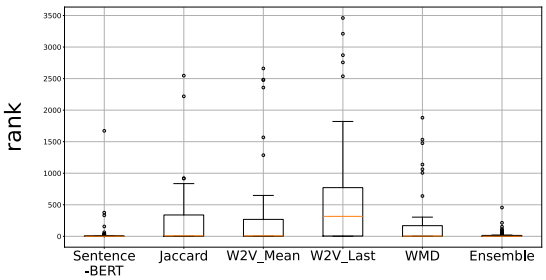

図 1　6 つの手法による正解単語の順位の分布

4.4 考察

実験の結果，MRR と Top-1 accuracy に関してはアンサンブル手法が最も良く，Top-3 と Top-10 accuracy に関しては Sentence-BERT が最も良い結果であった．一方，文末単語の分散表現を用いる手法は全体的に精度が低かったが，表 4 に示す「感触」という単語に関しては，Sentence-BERT 手法の場合に正解単語の順位が 1,672 位だったのに対して 2.5 位と高い値が得られた．その結果，同単語のアンサンブル手法による順位が 15 位と，Sentence-BERT 単独よりも良い結果が得られた．このことから本研究で提案したアンサンブル手法は一定の効果があると言える．

表 4　「感触」の語義説明文

辞書名	語義説明文
岩波国語辞典	広く，相手の談話などから受ける，ぼんやりした感じ
新明解国語辞典	確証は無いが，状況証拠から得られる感じ

5. まとめ

本研究では，文から単語への言い換えを行う手法を 6 つ提案し，評価を行った．その結果 Sentence-BERT とアンサンブル手法が良い結果を示した．

今後は，言い換え先の単語候補を 4,142 単語から拡張した上で類似度計算手法の追加や改良によって更なる精度の向上を目指したい．

参考文献

[1] 西尾 実, 岩淵 悦太郎, 水谷 静夫, “岩波国語辞典 第五版”, 岩波書店 (1994).
[2] 山田 忠雄, 倉持 保男, 上野 善道, 山田 明雄, 井島 正博, 笹原 宏之, “新明解国語辞典 第八版”, 三省堂 (2020).
[3] 林 四郎, 篠崎 晃一, 相澤 正夫, “例解新国語辞典 第十版”, 三省堂 (2021).
[4] Reimers, N., Gurevych, I., “Sentence-BERT:Sentence Embeddings using Siamese BERT-Networks”, Proceedings of EMNLP, pp.3980-3990 (2019).
[5] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J, ” Distributed Representations of Words and Phrases and their Compositionality”, Proceedings of NIPS, pp.3111-3119 (2013).
[6] Kusner, M. J., Sun, Y., kolkin, N. I., Weinberger, K., “From Word Embeddings To Document Distances”, Proceedings of ICML, Vol. 37, pp.957-966 (2015).
[7] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., ”Enriching Word Vectors with Subword Information”, Proceedings of TACL, Vol.5, pp.135-146 (2017).