文書要約のための一貫性モデル

金澤 尚史 † 高村 大也 ‡§ 奥村 学 ‡
東京工業大学工学院 †， 東京工業大学科学技術創成研究院 ‡，産業技術総合研究所 §
†kanazawa@lr.pi.titech.ac.jp， ‡{takamura， oku}@pi.titech.ac.jp

1 はじめに

要約文書の質は，文書自体の読みやすさで測ることができ，情報を正確に伝えるためには，文書自体の読みやすさを考慮することが重要となる．

文書自体の読みやすさには，文の文法性と文章の一貫性の側面が存在する．文の文法性とは，文が適格かどうかを表し，文章の一貫性とは，文間の因果関係などつながりのよさを表す．これらは人が正確に文書の内容を理解するために必要な要素となる．また，一貫性は局所的な一貫性と大域的な一貫性に分けることができる．局所的な一貫性とは連続する 2 文間の一貫性であり，大域的な一貫性とは文書におけるトピックの遷移に関する一貫性である．局所的一貫性は大域的一貫性にとっても重要な要素であり，文書の局所的一貫性の向上は大域的一貫性の向上にもつながる．一貫性を正確に捉えることは，複数文書要約において重要箇所抽出後，抽出した箇所を適切な順序に並べる際，より良い並びを選択することに繋がるため，文書要約においても重要となる．

このような背景から，本論文では一貫性に着目し，意見記事集合を対象とした複数文書要約において，文書から抽出した重要箇所を適切な順序に並べるための一貫性モデルを提案する．本研究の提案手法の流れを図 1 に示す．複数文書要約において抽出した重要箇所は，1 つの重要箇所が複数文から構成されている場合も考えられるので，複数文から構成されるパラグラフを基本単位とし，パラグラフ集合に対して一貫性モデルを適用することにより，一貫性を測る．

提案手法では，一貫性モデルとして多層パーセプトロンを使用する．モデルの学習方法は，局所的一貫性を捉えるため，同一文書の特定のパラグラフの分散表現から次のパラグラフの分散表現を予測するように学習を行う．一貫性モデルは，抽出されたパラグラフ集合に対して各並びの一貫性を測り，最も一貫性が高い並びを出力するが，評価実験では，その並びと人手要約の並びとの相関を元に評価を行う．

2 関連研究

本研究と関連する研究分野として，要約における重要箇所抽出後の抽出箇所の並び順決定に関する研究が

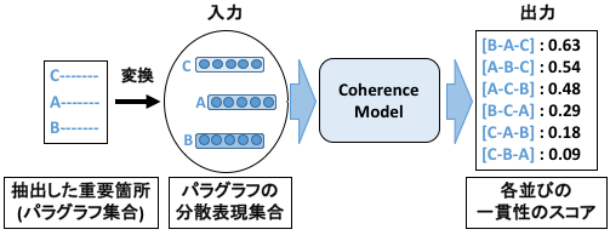

図 1: 文書要約に沿った一貫性モデル

挙げられる．単一文書要約における，並び順の決定方法としては，元文書中の並び順をそのまま利用する方法が用いられてきた．しかし，複数文書要約においては，元文書間の関係も考慮しなければならず，元文書中の並び順をそのまま利用することはできない．

岡崎ら [9] は複数文書要約における並び順の決定方法として，元文書が書かれた時系列と各抽出文の類似度を考慮して並べる方法を提案している．元文書が書かれた時系列は災害等の報道記事を対象とした複数文書要約において有効であるが，本研究で扱う意見記事のように特定のトピックに関して賛否等の議論を行う場合，時系列が有効な素性であるとは限らない．また，本研究は，あくまで文書要約に沿った一貫性モデルの提案であるため，並び順の決定に関する研究とは枠組みが異なる．

一貫性モデルに関する研究として， Barzilay ら [1] は局所的一貫性モデル entity grid を提案している．entity grid は，一貫性がある文章では文中の要素の遷移に規則性があるというセンタリング理論 [2] に基づいており，文中に出現する主語，目的語等の談話要素の遷移に着目し，一貫性を捉えている．

日本語の文書への entity grid の適用は英語と比較し，談話要素の省略が多いため難しいと考えられるが，横野ら [8] は格助詞による分類や接続詞等の結束性を考慮した entity grid モデルを提案している．しかし，本研究では文ではなく，複数文から構成されるパラグラフを扱っているため，entity grid を利用してパラグラフ間の談話要素の遷移から一貫性を捉えることは，難しいと考えられる．

また，Barzilay らや横野らは，枠組みとして，一貫性の高い文書と低い文書それぞれの談話要素の遷移情

報を ranking SVM に学習させ，文書の一貫性の度合いを 2 値分類するタスクに取り組んでいる．しかし，本研究では，パラグラフ集合の各並びに対して一貫性を測り，最も一貫性が高いと判断した並びを出力するタスクに取り組む．

関連する枠組みとして，Li ら [4] は文書再構築タスクに取り組んでいる．文書再構築タスクは，Barzilay らや横野らの 2 値分類タスクより難しいタスクとされており，文書を構成する文集合が与えられ，文集合を元の文書の並びに再構築するタスクである．Li らは一貫性モデルに sequence-to-sequence を利用した．連続する 2 文に対して，一方の文をモデルの入力として利用した時，もう一方の文が生成される確率を文間の局所的一貫性とみなすモデルを提案している．しかし，本研究では，文集合ではなくパラグラフ集合を使用しており，複数文から構成されるパラグラフは平均単語数が多いため，sequence-to-sequence モデルを利用し，生成確率から一貫性を捉えることは難しいと考えられる．文書再構築タスクとの違いとしては，本研究で扱うパラグラフ集合は，もともと 1 文書から構成されていない．また，Li らのタスクでは先頭の文が与えられているが，本研究では要約時の利用を想定しているため，先頭のパラグラフを与えない設定である．

3 提案手法

3.1 一貫性モデル

本研究では，連続する 2 パラグラフ間の局所的一貫性を捉えるため，一貫性モデルとして多層パーセプトロン (multilayer perceptron; MLP) を利用する．図 2 は局所的一貫性の計算の流れを示したものである．提案手法では，入力したパラグラフの分散表現に対し，次に来るべき一貫性が高いパラグラフの分散表現を予測する一貫性モデルの構築を行う．

3.2 パラグラフの分散表現

本節では，一貫性評価に用いるパラグラフの分散表現の獲得方法について述べる．パラグラフの分散表現の獲得方法には様々な手法が存在する．Skip-thought[3] 等の文の分散表現を直接獲得する手法も存在するが，パラグラフは文と比較し構成単語数が多く，学習に多大な時間がかかるという問題が存在する．そこで，本研究では，パラグラフの全構成単語の分散表現を平均したものをパラグラフの分散表現として利用する．

しかし，パラグラフの構成単語数が多いと，助詞や助動詞等の機能語の出現回数の増加につながり，全てのパラグラフの分散表現が似てしまうという問題点が存在する．そこで提案手法 (Co) として，パラグラフの構成単語のうち，内容語（名詞・動詞・形容詞・副詞・接続詞・感動詞・接頭詞）に限定して単語の分散表現を平均したものをパラグラフの分散表現として利用する．これらの単語の分散表現の獲得には新聞記事を用いて訓練を行った word2vec[5] を利用する．

3.3 一貫性モデルの訓練

本節では，一貫性モデルの訓練方法について述べる．人手によって書かれた文書は一貫性があると仮定し，訓練データとして，同一文書内の連続するパラグラフの分散表現 (p_i，p_{i+1}) のペアを与える．p_i を MLP の入力として利用し，出力を y_i とする．このとき，MLP の活性化関数に双曲線正接関数を利用する：

$$y_i \quad = \quad MLP(p_i). \tag{1}$$

パラグラフの分散表現 p_i を入力とした時，一貫性のあるパラグラフの分散表現 p_{i+1} を予測するために，出力 y_i と正解パラグラフの分散表現 p_{i+1} の平均二乗誤差を取り，学習を行う．N は訓練事例数とする：

$$loss_1 = \frac{1}{N}\sum_{i}^{N}(y_i - p_{i+1})^2. \tag{2}$$

しかし，同一トピックについて述べられているパラグラフは出現単語が似てしまい，出力される分散表現自体も似てしまうという問題点が存在する．そこで提案手法 (Neg) として，一貫性モデルの訓練時，正解パラグラフ以外の同一文書内のパラグラフに似ないように学習を行う．訓練方法としては，一貫性モデルへの入力を p_i とした時，基本モデルでは正解パラグラフの分散表現である p_{i+1} のみと誤差をとるが，提案手法 (Neg) では，同一文書内の p_i，p_{i+1} 以外のパラグラフ集合 P_i を負例として利用し，誤差計算を行う．正解パラグラフ p_{i+1} と負例パラグラフ p_j との距離を $\mathrm{L}(p_{i+1}, p_j)=|(i+1)-j|$ と表し，以下のように正解のパラグラフと距離が離れるほど負の重み付けを大きくし，負例パラグラフに似ないように学習を行う：

$$loss_2 = -\frac{1}{N}\sum_{i}^{N}\Bigg(\sum_{p\in P_i}\frac{L(p_{i+1}, p)}{\sum_{p\in P_i}L(p_{i+1}, p)}(y_i - p)^2\Bigg). \tag{3}$$

最終的な，提案手法 (Neg) の損失関数は $loss_1$ と $loss_2$ を足し合わせたものを利用し，学習を行う．

3.4 局所的一貫性の計算

本節では，図 2 に示す局所的一貫性の計算方法について述べる．パラグラフの分散表現 p_i-p_k 間の局所的一貫性の計算を行う場合，訓練済みの一貫性モデルに対してパラグラフの分散表現 p_i を入力として使用する．出力された分散表現 y_i と p_k の cos 類似度の値を p_i-p_k 間の局所的一貫性の度合いとみなす：

$$coherence_1 = cos(y_i, p_k). \tag{4}$$

しかし，式 (4) では，p_i を一貫性モデルの入力に利用し，p_k との一貫性を判断しているが，これでは一貫性

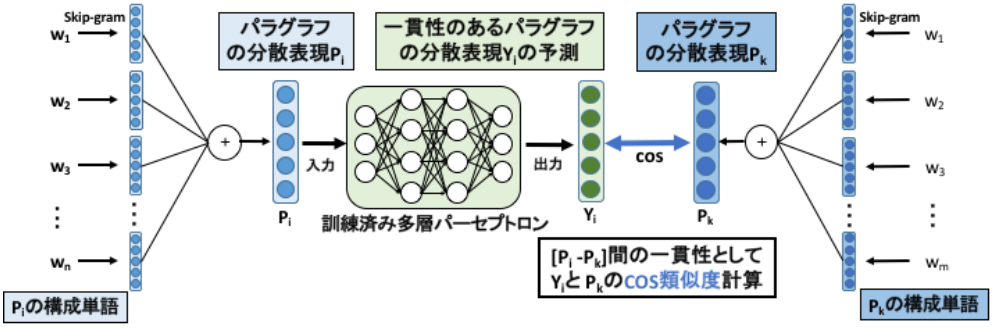

図 2: 一貫性モデルによる局所的一貫性の計算

モデルに入力しない p_k に含まれる接続詞等の情報を一貫性モデルが上手く捉えられない可能性が考えられる．そこで提案手法 (Bi) では，連続するパラグラフ $(p_i,\ p_{i+1})$ が与えられた時，p_{i+1} から p_i の予測を行う逆方向の多層パーセプトロンモデルを新たに一貫性モデルとして用いる．一貫性の度合いとして cos 類似度を計算する際には，以下のように従来の順方向モデル MLP_F と新たに作成した逆方向モデル MLP_B の両方向の cos 類似度の値を足し合わせ，一貫性の度合いとみなす：

$$\begin{aligned} coherence_2 =& cos(MLP_F(p_i), p_k) \\ &+ cos(p_i, MLP_B(p_k)). \end{aligned} \tag{5}$$

3.5 <start> タグの導入

本研究では，要約時の利用を想定しているため，先頭パラグラフを与えない設定で，最も一貫性の高いパラグラフの順序の推定を行う．しかし，先頭パラグラフを推定するためには，一貫性モデルへの入力として初期状態で何かしらの分散表現が必要となる．そこで，全ての要素が 0 の分散表現を，推定のため，最初に一貫性モデルへの入力として利用することが考えられる．

しかし，全ての要素が 0 の分散表現では先頭パラグラフを推定するための情報が乏しいため，提案手法 (St) では文書の先頭を表す <start> タグの分散表現を学習し，初期状態の一貫性モデルへの入力として用い，先頭パラグラフの推定を行う．<start> タグの分散表現は，word2vec の訓練時，新聞記事の各記事の先頭に <start> タグを挿入し，1 つの単語として学習を行うことで獲得する．分散表現獲得後，<start> タグを訓練データの文書の先頭に挿入し，一貫性モデルを訓練することで，先頭パラグラフを推定しやすい一貫性モデルの構築が行える．

4 実験

4.1 データセット

本節では実験に使用したデータセットについて述べる．単語の分散表現獲得時に使用した word2vec は，モデルに Skip-gram を利用し，窓幅 5 単語，ネガティブサンプリング 15 単語とした．訓練データは，新聞記事 3 紙 (毎日新聞，読売新聞，日本経済新聞) 合計 43 年分を MeCab[1] で分かち書きし，頻度 5 以上の単語を用いた．

一貫性モデルの訓練データは，意見記事の一貫性を捉えるため，新聞記事 3 紙 (毎日新聞，読売新聞，日本経済新聞) 合計 43 年分から社説記事を抽出し，利用した．抽出した社説記事をパラグラフ毎に分散表現に変換後，連続するパラグラフの分散表現 $(p_i,\ p_{i+1})$ のペアを 362,050 組作成し，訓練に利用した．社説記事の 1 パラグラフあたりの平均単語数は 53.3 単語であった．

評価実験には，TSC4 の意見要約コーパス [7] を利用した．TSC4 は，タバコの増税や大食い競争の賛否など，ある特定トピックに対しての意見記事の複数文書要約コーパスであり，1 トピックに対して平均 10 記事と人手によって書かれた総文字数の 5%，10%の要約から構成される．トピックは計 25 トピック，記事は 4 紙 (毎日新聞，読売新聞，日本経済新聞，朝日新聞)2 年分の新聞記事からのものであり，主に読者の投稿記事や訓練データと被らない社説記事等からなる．また，TSC4 の要約と元文書の対応箇所は人手によりパラグラフ単位で対応関係がアノテーションされている．

4.2 評価手法

図 3 に評価実験の流れを示す．TSC4 における，元文書側の人手要約との対応パラグラフを重要箇所抽出後の抽出したパラグラフ集合とみなし，一貫性モデルの入力に利用した．パラグラフ集合の平均パラグラフ数は 11.9，1 パラグラフあたりの平均単語数は 89.9 である．

評価実験では，パラグラフ集合に対して，提案手法を用いて各パラグラフ間の局所的一貫性の値を求め，貪欲法を用い，最も一貫性の高い並びを決定した．次に，最も一貫性が高いと判断した並びと人手要約の並びがどの程度の相関があるかを計算した．相関の計算には，Li らが文書再構築タスクで評価に用いた Kendall の順位相関係数 τ を使用し，計 49 セットの平均を結果として利用した．

[1] http://taku910.github.io/mecab/

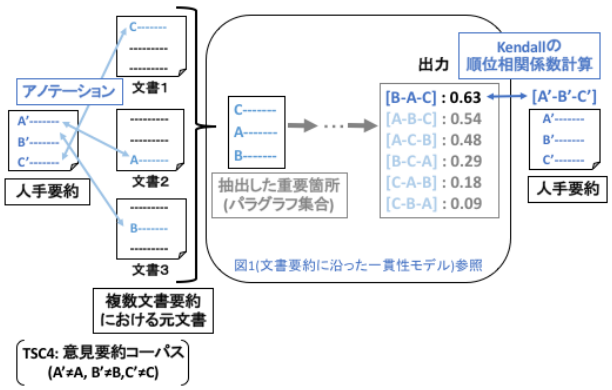

図 3: 評価実験の流れ

4.3 実験設定

提案手法 (Co), (St), (Bi), (Neg) の全てを使用しない基本モデル及び，各提案手法を追加していった計 5 通りの提案手法を用いて，訓練及び評価実験を行った．

実験を行うにあたり，Li らが文書再構築タスクにおいてベースラインに用いた「Foltz et al.(1998) (Glove)」として参照されている手法を一部変更し，本実験のベースラインに利用した．この手法は，2 文間の分散表現の cos 類似度を局所的一貫性とみなす．文の分散表現は，全構成単語の分散表現を平均したものを利用し，単語の分散表現獲得には Glove[6] を利用する．評価実験の際には，先頭文が与えられ，cos 類似度の値を基に，文集合に対して貪欲に探索を行い，最も一貫性が高い並びを決定する．変更点としては，単語の分散表現の獲得に Glove ではなく word2vec を利用した点，文ではなくパラグラフが単位である点，先頭パラグラフは与えずランダム初期化した分散表現から探索を行った点である．

評価用のデータセットは 2 分割し，開発セット，テストセットを作成した．ハイパーパラメータは次元 : 100, 300, 500, 1000, 1500，バッチサイズ : 64, 128, 256，中間層 : 300, 500, 1000, 1500，エポック数 : 100 と設定し，各パラメータで実験を行った．最も開発セットの値が高かったパラメータをテストセットの設定とし，2 分割交差検証を行った．

4.4 実験結果

各提案手法とベースラインの結果を表 1 に示す．各提案手法において次元が上がる毎に精度は良くなり，1000 次元付近で最も良い結果を示した．これは分散表現の次元を大きくすることにより，パラグラフの特徴を上手く捉える事ができたためと考えられる．表 1 に示すように (Co), (St), (Bi), (Neg) の全ての提案手法を加えることで精度向上が見られた．特に，3.2 節の内容語に限定したパラグラフの分散表現 (Co)，及び 3.3 節の負例を用いた一貫性モデルの訓練方法 (Neg) による効果が大きいことがわかった．全ての提案手法を組み合わせた提案手法 (Co+St+Bi+Neg) では 0.3 を超える順位相関係数の値を示し，ベースラインの値を大きく上回った．

表 1: 各一貫性モデルの評価実験結果

提案手法	次元				
	100	300	500	1000	1500
ベースライン	.024	.035	.039	.057	**.099**
基本モデル	.019	.011	.029	.050	**.065**
Co	.089	.110	.153	**.168**	.155
Co+St	.072	.140	.174	**.184**	.143
Co+St+Bi	.121	.159	.116	**.227**	.197
Co+St+Bi+Neg	*.181*	*.195*	*.258*	***.301***	*.237*

bold: 提案手法別最高値, *italic*: 次元別最高値

5 おわりに

本研究では，一貫性モデルとして多層パーセプトロンを使用し，特定のパラグラフの分散表現から一貫性のあるパラグラフの分散表現を予測するモデルを提案した．また，一貫性を捉えるための工夫として，内容語に限定したパラグラフの分散表現 (Co) の使用，先頭パラグラフ推定のための <start> タグ (St) の導入，両方向から局所的一貫性を計算するモデル (Bi) の導入，負例を用いた訓練方法 (Neg) の導入を行った．結果としては，ベースラインの値を大きく上回る結果を示し，一貫性モデルとして有効であることを示した．

今後の展望としては，パラグラフの分散表現として Skip-thought を使用した時との比較，両方向モデルにおいて逆方向モデルが捉える事ができる素性の追加を行いたいと考えている．

参考文献

[1] Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, Vol. 34 (1), pp. 1–34, 2008.

[2] Grosz B. J., Weinstein S., and Joshi A. K. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, Vol. 21 (2), pp. 203–225, 1995.

[3] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. In *Proceedings of NIPS 28*, pp. 3294–3302, 2015.

[4] Jiwei Li and Dan Jurafsky. Neural net models for open-domain discourse coherence. In *Proceedings of the 2017 Conference on EMNLP*, pp. 198–209, 2017.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS 26*, pp. 3111–3119, 2013.

[6] Richard Socher and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on EMNLP*, Vol. 14, pp. 1532–1543, 2014.

[7] 奥村学, 平尾努, 難波英嗣. TSC4: 意見要約コーパスとそれを用いたワークショップ. 言語処理学会第 11 回年次大会, 2005.

[8] 横野光, 奥村学. テキスト結束性を考慮した entity grid に基づく局所的一貫性モデル. 自然言語処理, Vol. 17 (1), pp. 161–182, 2010.

[9] 岡崎直観, 石塚満. 複数の新聞記事から抽出した文の並び順の検討. 人工知能学会第 18 回全国大会, pp. 191–194, 2004.