小説における本文とあらすじ文の対応付け

立見 英士 † 笹野 遼平 ‡ 高村 大也 ‡

† 東京工業大学 総合理工学研究科，‡ 東京工業大学 精密工学研究所

†tatsumi@lr.pi.titech.ac.jp，‡{sasano，takamura}@pi.titech.ac.jp

1　はじめに

あらすじとは，小説などの内容を簡略化し，大まかな流れを容易に把握できるようにした文章のことである．小説を全て読むことは長い時間を要するが，あらすじを読むことによって短時間で概要をつかむことができる．そのため，あらすじはその小説が読者の嗜好に即しているかを判断するために非常に有用である．電子書籍で溢れる現代においては，その有用性はますます高くなってきている．しかし，あらすじを作成するためには，本文を少なくとも一度全て読み，内容を全て把握してからまとめる作業になるため，人手で作成することは非常に高コストである．そのため，あらすじの自動生成技術の実現が望まれている．

我々は，小説に対するあらすじの自動生成を実現することを将来的な目標と位置づける．そのうえで，一部の小説にはあらすじが人手で既に作成されていることに着目し，これをあらすじ自動生成システムの訓練データとして用いることを考える．そのためには，あらすじの各文が本文のどの部分に対応するかという情報が付与されていると便利である．よって本研究では，あらすじの各文と本文の部分との対応付けを自動的に行う手法を提案する．

2　関連研究

文書とその要約のペアに対し，要約中の各文と対応している箇所を元テキストから同定する研究は，これまでにも行われてきた [1, 2, 3, 4]．Marcu は要約中の各文と本文の各文との余弦類似度を求め，要約文との類似度が最大となる節や文を本文中から同定する手法を提案している [2]．Jing らは，隠れマルコフモデルを用いて，要約中の単語と本文中の単語との対応付けを行っている [1]．Bamman らは，隠れマルコフモデルを用いて，あらすじ中の文と本文中の文を対応付ける手法を提案している [3]．Bamman らはさらに，あらすじに対応付けされた文と，それ以外の文を分類する二値分類問題としてあらすじ生成課題を定式化した．

上記の先行研究では，要約あるいはあらすじ中の各文と対応する文を本文中から同定しているが，本研究では，あらすじ中の各文が対応する本文中の範囲を同定することを目的としている．すなわち，先行研究ではあらすじ中の各文は本文中の重要文から生成されるという考えに基づいているのに対し，本研究では，あらすじ中の各文は，本文中のある範囲の内容をまとめることにより生成されるという考えに基づいている．

3　提案手法

本研究では，あらすじ中の各文 (以下，あらすじ文) が対応している本文の範囲を同定する手法を提案する．あらすじは，本文の内容を順番を変えずに表わしており，本文の分割を与えるものと仮定する．より詳細には，次のような仮定を置く：

(a) あらすじと本文との対応が交差することはない，

(b) あらすじ文は必ず 1 文以上の本文と対応する，

(c) 本文中の各文は必ずあるあらすじ文と対応する，

(d) 複数のあらすじ文が本文 1 文と対応することはない．

図 1 に本研究で考える対応付けの例を示す．

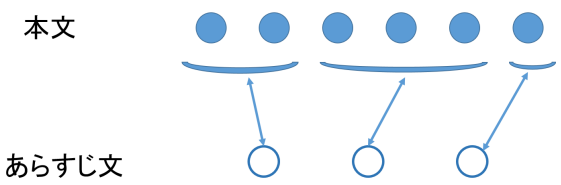


図 1: 対応付けの例

3.1 提案モデル

上記の仮定の下，n 文から成るあらすじと m 文から成る本文の対応付けの良さを表す目的関数を次のように設定する：

$$\max_{S} \quad \sum_{i=1}^{n} score(i, S_i). \tag{1}$$

ここで，S_i はあらすじ文 i と対応させる本文中の文番号の集合である．また，$S = (S_1, \cdots, S_n)$ である．$score(i, S_i)$ はあらすじ文 i を本文中の文番号集合 S_i と対応させたときの良さを表す．したがって，あらすじと本文の対応の良さ (式 (1)) は，各あらすじ文 i とそれが対応する S_i の対応の良さの総和で表されるとモデル化したことになる．これが最大となるように各あらすじ文が対応する本文の範囲 $S = (S_1, \cdots, S_n)$ を同定することが目的となる．

このモデルの長所の一つは，i と S_i の間の類似性を，文同士の類似性に帰着させることなく様々な方法で測れる点である．例えば，S_i 内での話題の一貫性などもこのモデルに組み込むことが可能である．

ここでは，$score(i, S_i)$ として，あらすじ文 i の bag-of-words ベクトルと対応する本文の部分 S_i 全体の bag-of-words ベクトルの間の余弦類似度を用いた．S_i 内の各文との類似性に帰着させていないことに注意してほしい．余弦類似度を求めるにあたり，文書中に出現した単語 w がどのくらい特徴的であるかを識別するための指標である tf-idf を用いた．類似度計算で考慮する単語は，名詞，動詞，形容詞に限定した．

3.2 動的計画法

本研究では 3.1 節で述べたモデルを動的計画法を用いることによって実装した．我々の目的は，最初のあらすじ文 1 から最後のあらすじ文 n まで順次対応する本文の文番号集合の組 $(S_1, \cdots, S_n)$ を同定することである．一般的な動的計画法によるマッチングアルゴリズムでは，各あらすじ文と本文の各文とが対応したときのスコアを算出し，スコアの合計が最大となるような経路を探索する．しかし我々が提案するアルゴリズム (Algorithm 1) では，各あらすじ文と本文の複数の文との間の類似度を足し合わせるので，少し一般的な動的計画法と違いがある．具体的には，あらすじ文 i と本文の文 j が対応したときの最大スコアを考えるのでなく，i が j までと対応したときの最大スコア $g(i, j)$ を考える．$g(i, j)$ を計算するときには，その定義から i が対応する最後の文は j と決まっているので，直前のあらすじ文 $i-1$ が $j-k-1$ までと対応しているならば，i に対応するのは $j-k$ から j までの文となる．つまり $S_i = \{j-k, \cdots, j\}$ である．よって $g(i, j)$ は $g(i-1, j-k-1)$ に $score(i, \{j-k, \cdots, j\})$ を加えることで計算できる．このスコアが最大となるように k を決定すればよい．これが Algorithm 1 の max で行っている計算である．ただし，あらすじ文 1 文目は必ず本文の 1 文目と対応付いている必要があることから，$l = 0$ の場合は $g(0, l) = 0$，それ以外の場合は $g(0, l) = -\infty$ とし，また max は $0 \leq k \leq j - i$ の範囲で探索を行う．アルゴリズムを実行した後，$g(n, m)$ から最大値を与えたパスをバックトラックすることにより S を求めることができる．

Algorithm 1 スコア算出アルゴリズム（動的計画法）

```
for i = 1 to n do
  for j = i to m − (n − i) + 1 do
    g(i, j) ← max_k(g(i − 1, j − k − 1)
                    + score(i, {j − k, ··· , j}))
  end for
end for
```

$score(i, \{j-k, \cdots, j\})$ は前節で説明したように，あらすじ文 i と対応付けされた本文集合 $\{j-k, \cdots, j\}$ との余弦類似度である．

4 実験

4.1 データセット

小説の本文は青空文庫 [5] から収集した．青空文庫には，新字新仮名，新字旧仮名，旧字新仮名，旧字旧仮名という 4 種類の文字遣い種別があるが，本研究では現在一般的に使用されている表記である新字新仮名で書かれている小説を扱う．そして青空文庫の書籍の中でも，分野別リストの「文学」の中の「日本文学」に分類されている「小説，物語」に属する作品を収集した．その結果，121 個の小説を収集することができた．

あらすじは，それらの小説の Wikipedia のページの「あらすじ」項目から収集した．収集した青空文庫の小説の中で Wikipedia にあらすじが掲載されている小説は 60 個存在した．現在存在している小説の中でこの数の小説にしか Wikipedia にあらすじが存在していないことを考えると，あらすじの自動生成の有用性は明らかである．本研究ではこの 60 個の小説の本文とあらすじのペアを使用する．対象の小説の中から 18

個の小説に対して，各あらすじ文と本文の各文との対応付けを人手でアノテーションした．

4.2 ベースライン手法

提案手法との比較のため，2 種類のベースライン手法を考える．本研究では，全てのあらすじ文を本文に対応付けている．そこで，本文をあらすじ文数に等分し，それぞれを各あらすじ文に対応付けたものを 1 つ目のベースライン手法 (ベースライン 1) とする．

また，提案手法ではあらすじ文 i と文集合 S_i とが対応したときの良さをあらすじ文 i の bag-of-words ベクトルと対応付けされた本文集合 S_i 全体の bag-of-words ベクトルの余弦類似度で表わしている．本文集合 S_i 全体を用いたことの妥当性を確かめるために，上記余弦類似度の代わりに，あらすじ文 i の bag-of-words ベクトルと本文集合 S_i の各文の bag-of-words ベクトルとの余弦類似度の平均を用いて i と S_i との対応の良さを測るモデルをもう 1 つのベースライン手法 (ベースライン 2) とした．

4.3 実験結果

4.1 節で述べた人手によるアノテーション結果と，提案モデルによって得られた対応付けとを比較し，その一致度によって提案モデルの評価する．ベースライン手法についても同様に評価する．表 1 に結果を示す [1]．

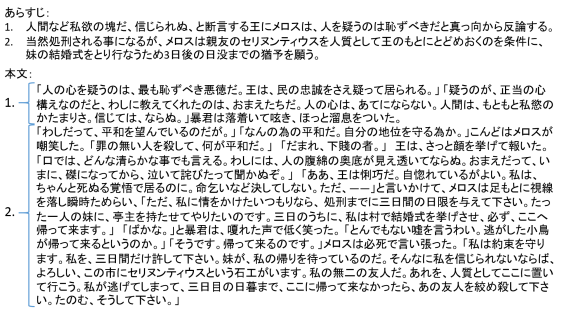
表 1: 実験結果

手法	F 値
ベースライン 1	0.138
ベースライン 2	0.450
提案手法	0.633

表 1 から，提案手法が 2 つのベースライン手法に比べて優れていることがわかる．特にベースライン 2 よりも良い結果だったことから，あらすじ文 i と S_i 全体との類似度をモデルに組み込んだことが効果的だったことがわかる．図 2，3 に提案モデルによって得られた対応付けの一部を示す．

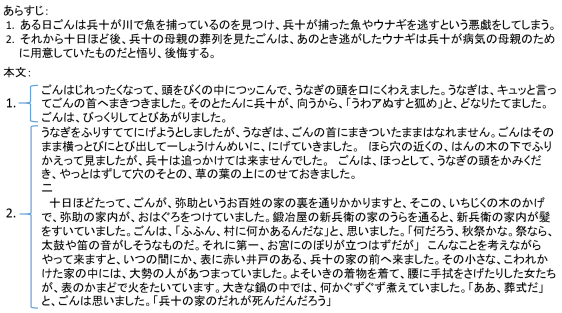
図 2 より，あらすじ文と最も似ている文だけではなく，そのあらすじ文がカバーしている範囲との対応付

[1]対応付けの研究では多くの場合，適合率，再現率，F 値で評価される．本研究での実験設定では，適合率と再現率が同じ値になり，結果として F 値も同じ値となる．


あらすじ:
1. 人間など私欲の塊だ、信じられぬ、と断言する王にメロスは、人を疑うのは恥ずべきだと真っ向から反論する。
2. 当然処刑される事になるが、メロスは親友のセリヌンティウスを人質として王のもとにとどめおくのを条件に、妹の結婚式をとり行なうため3日後の日没までの猶予を願う。

本文:
1. 「人の心を疑うのは、最も恥ずべき悪徳だ。王は、民の忠誠をさえ疑って居られる。」「疑うのが、正当の心構えなのだと、わしに教えてくれたのは、おまえたちだ。人の心は、あてにならない。人間は、もともと私慾のかたまりさ。信じては、ならぬ。」暴君は落着いて呟き、ほっと溜息をついた。
2. 「わしだって、平和を望んでいるのだが。」「なんの為の平和だ。自分の地位を守る為か。」こんどはメロスが嘲笑した。「罪の無い人を殺して、何が平和だ。」「だまれ、下賤の者。」 王は、さっと顔を挙げて報いた。「口では、どんな清らかな事でも言える。わしには、人の腹綿の奥底が見え透いてならぬ。おまえだって、いまに、磔になってから、泣いて詫びたって聞かぬぞ。」「ああ、王は悧巧だ。自惚れているがよい。私は、ちゃんと死ぬる覚悟で居るのに。命乞いなど決してしない。ただ、――」と言いかけて、メロスは足もとに視線を落し瞬時ためらい、「ただ、私に情をかけたいつもりなら、処刑までに三日間の日限を与えて下さい。たった一人の妹に、亭主を持たせてやりたいのです。三日のうちに、私は村で結婚式を挙げさせ、必ず、ここへ帰って来ます。」「ばかな。」と暴君は、嘆れた声で低く笑った。「とんでもない嘘を言うわい。逃がした小鳥が帰って来るというのか。」「そうです。帰って来るのです。」メロスは必死で言い張った。「私は約束を守ります。私を、三日間だけ許して下さい。妹が、私の帰りを待っているのだ。そんなに私を信じられないならば、よろしい、この市にセリヌンティウスという石工がいます。私の無二の友人だ。あれを、人質としてここに置いて行こう。私が逃げてしまって、三日目の日暮まで、ここに帰って来なかったら、あの友人を絞め殺して下さい。たのむ、そうして下さい。」

図 2: 実行結果の一部（走れメロスより）


あらすじ:
1. ある日ごんは兵十が川で魚を捕っているのを見つけ、兵十が捕った魚やウナギを逃すという悪戯をしてしまう。
2. それから十日ほど後、兵十の母親の葬列を見たごんは、あのとき逃がしたウナギは兵十が病気の母親のために用意していたものだと悟り、後悔する。

本文:
1. ごんはじれったくなって、頭をびくの中につっこんで、うなぎの頭を口にくわえました。うなぎは、キュッと言ってごんの首へまきつきました。そのとたんに兵十が、向うから、「うわアぬすと狐め」と、どなりたてました。ごんは、びっくりしてとびあがりました。
2. うなぎをふりすててにげようとしましたが、うなぎは、ごんの首にまきついたままはなれません。ごんはそのまま横っとびにとび出して一しょうけんめいに、にげていきました。 ほら穴の近くの、はんの木の下でふりかえって見ましたが、兵十は追っかけては来ませんでした。 ごんは、ほっとして、うなぎの頭をかみくだき、やっとはずして穴のそとの、草の葉の上にのせておきました。
二
十日ほどたって、ごんが、弥助というお百姓の家の裏を通りかかりますと、そこの、いちじくの木のかげで、弥助の家内が、おはぐろをつけていました。鍛冶屋の新兵衛の家のうらを通ると、新兵衛の家内が髪をすいていました。ごんは、「ふふん、村に何かあるんだな」と、思いました。「何だろう、秋祭かな。祭なら、太鼓や笛の音がしそうなものだ。それに第一、お宮にのぼりが立つはずだが」 こんなことを考えながらやって来ますと、いつの間にか、表に赤い井戸のある、兵十の家の前へ来ました。その小さな、こわれかけた家の中には、大勢の人があつまっていました。よそいきの着物を着て、腰に手拭をさげたりした女たちが、表のかまどで火をたいています。大きな鍋の中では、何かぐずぐず煮えていました。「ああ、葬式だ」と、ごんは思いました。「兵十の家のだれが死んだんだろう」

図 3: 実行結果の一部（ごん狐より）

けが行えていることがわかる．提案手法により会話の流れの中にある切れ目も捉えることができていることが確認できた．図 3 では，あらすじ文 2 で「それから十日ほど後」とあり，本文では「十日ほどたって」とあるので，理想としてはあらすじ文 2 はその本文から対応付けが開始されることが望ましい．しかしながら，実際の結果はそれ以前の本文からあらすじ文 2 の対応付けが開始されてしまっている．このような本文の話題の区切りは，あらすじ文との対応付けの区切りとなる場合が多いと考えられることから，本文の話題の境界を意識して対応付けをすることによって精度を向上させることができるのではないかと考えられる．

5 境界モデル

前節までで有効性を示したモデルでは，あらすじ文と本文の類似性を利用した．しかし，あるあらすじ文に対応する本文の部分と，次のあらすじ文に対応する部分との境界は，特徴的な性質を持っていると考えられるので，これを利用して対応付けの性能をさらに向上させることを考える．例えば，対応付けの境界は，話題の境界にもなっている可能性もある．また，境界付近では,「あくる日」や「それから十日後」など，場面

転換を表すような手がかり語句が出現しやすいだろう．

まず，話題の境界の利用のため，テキスト分割の指標を用いた．本研究では，注目している文間の直前 r 文と直後 r 文との間の余弦類似度を求め [2]，1 からこの余弦類似度を引いた値を，この文間の切れやすさのスコア $score_b$ として定義した．この r を窓幅と呼ぶことにする．本文の j 文目と $j+1$ 文目の間の $score_b$ を提案手法のあらすじと本文との対応付けに反映させるため，(1) 式の関数 $score(i, S_i)$ と $score_b(j, j+1)$ を重み λ で線形和をとった，新たな関数 $score'(i, S_i)$ を定義した：

$$\lambda score(i, S_i) + (1-\lambda) score_b(j, j+1). \qquad (2)$$

実験では，λ を 0 から 1 の範囲で 0.05 刻みで変化させた．また，窓幅を 1 から 20 まで変化させた．窓幅が 1，5，10，15，20 のときに λ を変化させた結果を図 4 に示す．

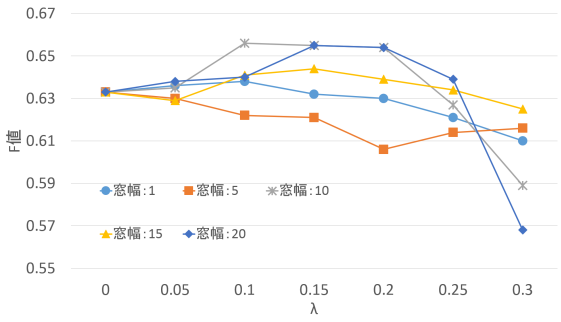


図 4: 境界での余弦類似度を考慮した結果

この図からわかるように，現時点ではパラメータの値によってはF 値が向上しているものの，不安定な結果となっており，境界での余弦類似度を考慮したことの効果を示すには至っていない．各小説での結果を見てみると，多くの小説で向上しているパラメータ値があり，これらをうまく捉えることができれば，全体としての性能向上が期待できるだろう．余弦類似度の代わりに，話題分割のための標準的な手法である Text Tiling[6] の指標も試したが，同様の結果であった．

また，場面転換を表す手がかり語句を捉えるために，教師付きの識別モデルを用いた．具体的には，実験に用いていない小説データから，話題の切れ目であることが章番号などで形式的に示されてる箇所を取り出しこれを正例とし，そうでない箇所を負例とした．そして，周辺に出現する単語を素性としてロジスティック回帰を学習した．この分類器をテストセットの各文間に適用して，その出力値と $score(i, S_i)$ の線形和を新たなスコア関数とした．しかし，良い実験結果は得られなかった．

場面転換を表すような手がかり語句が含まれている小説がいくつか存在していることはデータ分析により明らかになっている．しかし，章番号などにより明確に示されている切れ目に出現していることは少なかった．また，単純に語句の出現を 2 値によって判断しただけなので，「それから 10 日後」の「10 日後」がテストデータに含まれていなければならず，本来であれば，「数字＋時間表現」の出現を捉えるべきである．現状ではまだそのような素性を組み込むことができておらず，良い結果につながらなかった原因の 1 つであると捉えている．

[2] 名詞，形容詞，動詞に限定した．また，テストデータとして用いた 18 個の小説とは異なる 20 個の小説を用い，頻出語上位 30 単語をストップワードとして取り除いた．

6 まとめ

本稿では，あらすじの各文と小説の本文の部分を対応付ける方法を提案した．対応付けの際に，1 文同士ではなく，あらすじ文が対応する本文の部分全体との類似度を用いることによって，F 値 0.633 という結果を得ることができた．さらに，本文における話題の境界を考慮したモデルを作成したが，対応付けの性能の向上は確認できなかった．最後に，本文における話題の境界を考慮しても精度が向上しなかった原因について考察を行い，今後の進むべき方向性を明らかにした．

参考文献

[1] Hongyan Jing and Kathleen R. McKeown. The Decomposition of Human-Written Summary Sentences. *SIGIR'99*, pp. 129–136, (1999).

[2] Daniel Marcu. The automatic construction of large-scale corpora for summarization research. *SIGIR'99*, pp. 137–144, (1999).

[3] David Bamman and Noah A. Smith. New Alignment Method for Discriminative Book Summarization. `http://arxiv.org/abs/1305.1319`, (2013).

[4] 亀田尭宙, 李元, 内山清子, 武田英明, 相澤彰子. 論文における要約記述に対応するパラグラフの同定手法. *JSAI2013*, (2013).

[5] 青空文庫. `http://www.aozora.gr.jp/`.

[6] Marti A. Hearst. TextTiling: Segmenting Text into multi-paragraph Subtopic Passages. *Computational Linguistics*, 23(1), pp.33–64, (1997)