小説から演劇台本への書き換え過程のアノテーション

金子 遥渚 † 松吉 俊 ‡ 内海 彰 ‡
† 電気通信大学 情報理工学域, ‡ 電気通信大学大学院 情報理工学研究科
{h.kaneko, matuyosi, utsumi}@uec.ac.jp

1 はじめに

一般に、一から物語を創作することは難しいため、物語の小説を原本としてそのテキストを書き換えることにより演劇台本を作成することがある。この書き換え作業は、小説のストーリーから重要部分を切り出し、それらを演劇台本の文法に則って記述するというものであり、専門性を伴うとても難しいタスクである。この書き換え作業を自然言語処理技術により自動化もしくは支援することができれば、新しい演劇台本作成のコストを大幅に下げることができる。

物語小説から重要部分を切り出すタスクは、自然言語処理の自動要約や情報抽出の技術が応用できると思われるが、残念ながら、< 原本, 演劇台本 > というペアに関して利用可能な電子データは存在しない。

そこで、本研究では、小説から演劇台本への自動書き換えシステム構築の基盤として、< 原本, 演劇台本 > の対応付けを書き換えの各過程ごとにアノテーションする方法を提案する。本稿では、5 作品の小説を対象としたアノテーションの現状についても報告する。

2 関連研究

演劇台本を一から書く方法や演劇台本に書き換える方法に関する文献は数多く発行されている (例えば、[3, 4, 1])。演劇台本作成において、全体の上演時間や場面転換の回数を考慮した結果、採用する場面の数を絞らなければならないことがある。平田 [3] は、そのような場合に、ストーリーの抜けを補足するため、想像力を喚起させるような対話 (**エピソード**) を適所に挿入することでストーリーの破綻を防ぐ工夫を紹介している。本研究でもこの工夫を採用する。

自然言語処理の分野において、小説のストーリーから重要部分を切り出す先行研究として、相良ら [8] の研究がある。この研究では、重要度の高い名詞をキーワードとして重要文のリストを抽出することでストーリーを抽出する方法を提案している。テキストを構成単位に自動分割する先行研究として、TextTiling 法 [2] がある。この手法では、近傍段落との語句の重なり度合いを見ることで分割すべきかどうか判断する。しかしながら、物語小説では、場所や時間の変化が生じた際にそれに関連する語句が何度も現れることは少ないため、TextTiling 法は小説ストーリーの場面分割には有効でないことが知られている [7]。

通常の形式の物語テキストを演劇台本の形式に自動変換する先行研究として、今ら [6] の研究がある。この研究では、日本の昔話を原本として、台詞とト書きを抽出し、台詞の話者を特定する方法を提案している。一般に昔話はストーリーが短いため、今らの研究では重要部分抽出は一切行われていない。

一般の演劇からは少し外れるが、テキストを原本として漫才台本を自動生成する先行研究 [9, 5] がある。生成対象が漫才であるので、上演時間や場面転換等は考慮されない。

3 小説から演劇台本への書き換え

演劇台本は、1. 台詞とその話者、2. ト書きと呼ばれる役者の動作や出入りの記述、3. 場面転換の情報から構成される[1]。演劇台本の例を図 1 に示す。1 章でも述べたとおり、物語の小説を書き換えることで演劇台本を作成する際には、全体の上演時間や場面転換の回数等を考慮しながら、ストーリーの重要部分を切り出し、それらを上記の台本形式で記述する必要がある。

自動書き換えシステムの開発を見通しの良いものとするために、我々は、次の 2 つの方針を取った。

- 表 1 のように、小説から演劇台本への書き換えに 6 つの手順を導入し、書き換え過程を明確に定義した。

[1] 一方、**脚本**には、台本の情報に加え、使用機材や音響関連、裏方スタッフの動きなども指定される。劇団の人員や舞台設備を把握しないと脚本を完成させることはできないため、本研究では、台本を対象とする。

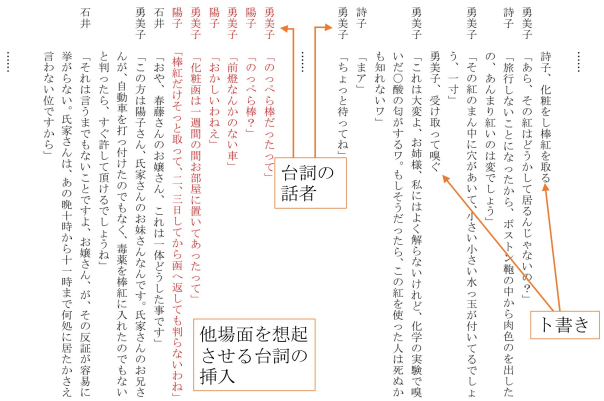

図 1: 『身代わりの花嫁』(野村胡堂) の演劇台本

表 1: 小説から演劇台本への書き換え手順

手順	内容	関連研究
1	台詞・話者抽出	[6]
2	ト書き抽出	[6]
3	場面分割	[2, 7]
4	場面選択	[8]
5	エピソード変換	–
6	エピソード挿入	–

- ＜原本, 演劇台本＞の対応関係を見やすくするために、図 2 のように、原本のテキストに XML タグを挿入するアノテーション形式を採用した。

前者の方針を取ることで、手順ごとにモジュールを開発することができ、モジュールごとに評価を実行することも可能となる。小説から演劇台本を一気に作成し、毎回人手でその質を評価することは現実的ではないため、スムーズなシステム開発のためにこの方針をとった。後者の方針を取ることで、XML タグをすべて除去すればオリジナルのテキストとなり、逆に、関係する XML タグのみを残せば、演劇台本が再現されるというデータ構造になっている。この方針には、表 1 の各手順のモジュールが手がかりとすべきテキスト中の文字列を明示的にアノテーションできるという利点もある。

以下、この章では、表 1 に挙げた 6 つの手順の内容とそのアノテーション方法について説明する。

3.1 台詞・話者抽出

台本の基礎となる、台詞とその話者を特定するタスクである。このタスクの前提として、テキスト中に記述された登場人物の集合を自動獲得する必要がある。

```
...
<utterance speaker="広 田">「 ま だ 出 そ う も な い の で す か ね 」</utterance>と 言 い な が ら 、今 行 き 過 ぎ た<action actor="広 田">西 洋 の 夫 婦を<edit after="">ちょい　と</edit>見<edit after="る">て</edit></action>、
<utterance speaker="広 田">「 あ あ 美 し い 」</utterance>と 小 声 に 言って、すぐに生欠伸をした。
<actor><ne type="person">三　四　郎</ne></actor>は<actor id="三四郎">自分</actor>がいかにもいなか者らしいのに気がついて、さっそく首を引き込めて、<action>着座した。</action><actor id="広田">男</actor>もつづいて<action>席に返った。</action>そうして、
<utterance speaker="広田">「どうも西洋人は美しいですね」</utterance>と言った。
  <actor><ne type="person">三四郎</ne></actor>はべつだんの答も出ないのでただ<utterance speaker="三四郎" parentheses="no">はあ</utterance>と受けて笑っていた。
...
```

図 2: 『三四郎』(夏目漱石) に対するアノテーション

テキスト中の台詞を<utterance>タグで囲み、このタグの@speaker 属性に話者を記述する。手がかりとして、テキスト中に記述された固有名詞を<ne type="person">タグで、一般名詞か固有名詞か問わず、登場人物を<actor>タグで囲む。台本において登場人物名は名寄せする必要があるので、代表表記以外の場合には<actor>タグの@id 属性にその代表表記を記述する。

3.2 ト書き抽出

台本のト書きに採用すべき動詞句を抽出し、その動作主を特定するタスクである。

テキスト中のそのような動詞句を<action>タグで囲み、このタグの@actor 属性に動作主の登場人物の代表表記を記述する。台本のト書きの末尾は、助動詞等が付かない動詞終止形で記述されるため、動詞句末に必要となる編集処理を図 2 に例示されるように<edit>タグを用いて記述する。

3.3 場面分割

小説内の場面の区切りを認識するタスクである。本稿では、時間的・空間的に一続きである内容のテキスト断片を**場面**と呼ぶ。演劇では、場面が変われば、暗転とともに大道具や照明の変更が伴うことが多いため、場面は重要な概念である。

テキスト中の場面を<scene>タグで囲む。小説内の章・節の始まり等、そこに場面区切りが確実に存在する

表 2: 現在構築中のデータセットに関する統計情報

	三四	道順	上海	王室	王冠
文字	171,897	182,963	146,344	71,195	180,939
地の文の句点	4,934	2,878	2,849	965	3,020
ト書き	552	449	560	173	509
段落	2,049	1,774	1,129	667	1,893
場面	106(100%)	68(100%)	86(100%)	52(100%)	122(100%)
場面 (確定分割)	13(12%)	14(21%)	46(53%)	16(31%)	51(42%)
場面 (段落末が区切り)	95(90%)	52(76%)	82(95%)	47(90%)	119(98%)
場面 (それ以外)	11(10%)	16(24%)	4(5%)	5(10%)	3(2%)
場面 (採用)	25(24%)	11(16%)	26(30%)	16(30%)	37(30%)
場面 (不採用)	69(65%)	47(69%)	52(60%)	31(60%)	68(56%)
場面 (エピソード変換)	12(11%)	10(15%)	8(10%)	5(10%)	17(14%)

箇所がある。このような場合、<scene>タグの@block 属性に “new” を指定することでこれを明示する。本稿では、このような箇所を**確定分割**箇所と呼ぶ。

3.4 場面選択

実際に台本に採用する場面を選択するタスクである。2 章の冒頭で述べたように、我々は、想像力を喚起させるような対話形式のエピソードを適所に挿入することで、採用する場面の数を抑える工夫を用いる。すなわち、前節で分割した場面を、(a) 採用、(b) 不採用、(c) エピソード変換 のいずれかに分類する。

<scene>タグの@use 属性に “yes”、“no”、“episode” のいずれかを記述する。

3.5 エピソード変換

前節で「エピソード変換」に分類された場面のテキストから、図 1 の赤字で示されるような対話を生成する難しいタスクである。この図の赤字は、採用する場面の数を減らすために、別場面における「前燈のない車」と「化粧函」に関する内容を人手で対話形式に編集したものである。

<episode>タグを用意し、その中に対話形式のエピソードを記述する。@source 属性に生成元の場面の ID を記述する。

3.6 エピソード挿入

前節で生成したエピソードを適所に挿入し、それぞれの台詞の話者を決定する難しいタスクである。

テキスト内の適所に前節の<episode>要素を挿入する。対話内の各台詞の@speaker 属性に話者を記述する。

4 データセット構築

本章では、現在進めているアノテーション作業の現状について報告する。XML タグのアノテーションは、普段から演劇に携わっており、演劇台本に関する知識がある大学生 1 名が行った。

4.1 対象の小説

現在の対象は、次の 5 作品である (一番左に本稿での略称を示す。)。

三四 『三四郎』(夏目漱石)

道順 『道順は彼女に訊く』(片岡義男)

上海 『上海』(横光利一)

王室 『グリュックスブルグ王室異聞』(橘外男)

王冠 『王冠の重み』(ホワイトフレッド・M、訳 奥増夫)

これらは、青空文庫[2]に存在する以下の 2 つの条件を満たす作品の中からランダムに選出した。

- 確定分割箇所が 10 以上ある
- 「一連の台詞の合計文字数が 2,000 字以上」が 6 箇所以上ある

[2] https://www.aozora.gr.jp

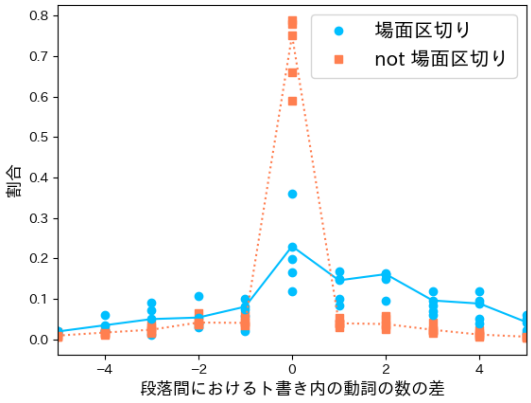

図 3: 連接する 2 つの段落間の「ト書き内の累計動詞数」の差の分布

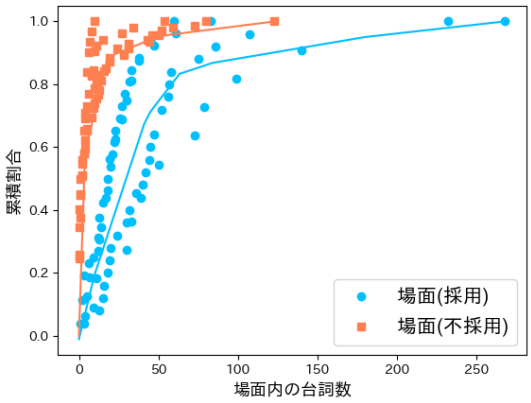

図 4: 1 場面内の台詞数に関する累積割合

4.2 統計情報

対象の小説に関する基礎統計値およびアノテーション結果の統計値を表 2 に示す。

「地の文の句点」の数は、いわゆる「文数」を指す。手順 2 のト書き抽出では、この「文数」の候補の中から、ト書きとなる動詞句を抽出することになる。この表から、およそ 1/10~1/5 の適切な動詞句を選択するタスクであることが分かる。

「場面 (···)」には、「場面」数を分母とした割合も併記した。「場面 (確定分割)」の割合は、作品により大きく異なることが見て取れる。今回使用した作品の著者はすべて異なり、ジャンルも同じというわけではないので、これは当然の結果かもしれない。「場面 (段落末が区切り)」には、段落の途中ではなく段落間に場面区切りがあった事例の数を示す。5 作品のうち 2 つに関しては段落間のみ考慮すればよいことが分かるが、残り 3 つに関しては段落途中にある時間的・空間的な切れ目を適切に認識する必要があることが分かる。「場面 (採用)」に、採用した場面の数を示す。20~30%の場面が採用される傾向にあることがこの表から見て取れる。また、「場面 (エピソード変換)」の行から、10%強の場面は、対話形式のエピソードに変換することにより、上演場面から省略できることが分かる。

手順 3 に関して、『王冠』を除く 4 作品を対象として、連接する 2 つの段落間の「ト書き内の累計動詞数」の差を調査した。調査結果を図 3 に示す。データ点は 4 つの作品それぞれに対するものであり、折れ線はそれらの平均値を示す。作品に依存せず、場面冒頭の段落においてト書き内の動詞が相対的に多いことが見て取れる。

手順 4 に関して、『王冠』を除く 4 作品を 1 つにまとめて、場面内の台詞数の分布を調査した。「場面 (採用)」と「場面 (不採用)」に対するこの累積割合を図 4 に示す。台詞が少ない場面は演劇台本に採用されにくいことがこの図から分かる。

5 おわりに

本論文では、小説から演劇台本への書き換え過程のアノテーション方法を提案した。今後は、まずは表 1 の手順 3 と 4 のシステムを実装する予定である。

参考文献

[1] 原田佳夏. 脚本を書こう！青弓社, 2004.

[2] Marti A. Hearst. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, Vol. 23, No. 1, pp. 33–64, 1997.

[3] 平田オリザ. 演劇入門. 講談社, 1998.

[4] 北村想. 高校生のための実践劇作入門—劇作家からの十二の手紙. 白水社, 2000.

[5] 真下遼, 灘本明代. 対立語抽出に基づく web ニュースからの漫才ロボット台本自動生成手法の提案. In *DEIM Forum 2014*, 第 8 巻, 2014.

[6] 今誠一, 吉田文彦, 菊池浩明, 中西祥八郎. 昔話の自動シナリオ化システムの構築. 言語処理学会年次大会発表論文集, pp. 317–320, 2005.

[7] 小林聡. 場・時・人に着目した物語のシーン分割手法. 情報処理学会研究報告 NL179, pp. 25–30, 2007.

[8] 相良直樹, 砂山渡, 谷内田正彦. 重要文抽出を利用したテキストからのストーリー抽出. 情報処理学会研究報告 NL164, pp. 159–164, 2004.

[9] 吉田裕介, 萩原将文. 漫才形式の対話文自動生成システム. 日本感性工学会論文誌, Vol. 11, No. 2, pp. 265–272, 2012.