

小説を対象とした知的ナビ機能の付いた電子書籍リーダの提案

西岡 渉[†] 沼尾 雅之^{††}

[†] 電気通信大学電気通信学部情報工学科 〒182-8585 東京都調布市調布ヶ丘 1-5-1

^{††} 電気通信大学情報理工学部情報・通信工学科 〒182-8585 東京都調布市調布ヶ丘 1-5-1

E-mail: [†]n0611077@edu.cc.uec.ac.jp, ^{††}numao@cs.uec.ac.jp

あらまし 電子書籍が普及し、従来と異なる方法で読書を楽しめるようになった。そのひとつに電子書籍リーダの読書支援機能がある。機能の例に、小説の登場人物などのキーワードの出現場所の可視化や、キーワードに対する初出箇所付近の文の提示などがある。しかし、電子書籍リーダは使えるコマンドに制限がある。電子書籍リーダに場面の自動分割や場面ごとの意味抽出の機能があるとより便利である。そこで、本稿では小説を対象として、テキストの場面解析を用いた知的ナビ機能の付いた電子書籍リーダを提案する。

キーワード 固有表現抽出, 情報要約, 小説

1. 背景と目的

電子書籍が広く定着し利用されるようになり、従来と異なる方法で読書を楽しめるようになった。小説が電子化されたことで計算機による処理を行えるようになり、ジャンルの推定やストーリーの抽出などを自動で行えるようになった。その応用のひとつに電子書籍リーダの読書支援機能がある。たとえば Amazon Kindle の X-ray には小説の登場人物などの重要な語の出現箇所の可視化や、重要な語に対して初出箇所付近の文を表示する機能などがある。しかし、電子書籍リーダには機能の制限がある。キーワードの出現箇所を検索する場合でも対象は単一のキーワードに限られ、複数のキーワードを組み合わせた出現箇所の検索はできない。ことに小説を読む場合には、意味的なまとまりである場面を単位にした複数の登場人物の出現を検索できると便利である。そこで本稿では小説を対象として、場面分割および登場人物を対象とした場面要約を行う場面解析システムおよびテキストの場面解析手法を提案し、用いた知的ナビ機能の付いた電子書籍リーダ作成する。

2. 関連技術

2.1 形態素解析

2.1.1 形態素

計算機で自然言語を扱うことを自然言語処理という。自然言語処理ではテキストをなんらかの単位に分解したうえで、その並びなどを扱う。

その単位のひとつが形態素である。形態素は意味を持つ最小の音素のまとまりである。音素とは人間の意思伝達に用いられる際の音の単位のことである。

文を形態素の並びとみたとき、形態素はその前後で影響を及ぼしあい音韻的なつながりによって変形した状態で表れている。この文中の現われている状態のものを形態と呼ぶ。[13]

2.1.2 形態素解析

形態素解析とは、与えられたテキストを形態素に分解し、その品詞などを推定することという。

自然言語には英語のように語がわかち書きされているものと、日本語のようにわかち書きされていないものがある。前者の場合、形態素解析ではテキストを形態素に分解しその品詞などを決定する。後者の場合、これらの前に、形態素解析でテキストの語への分解も行なう。

とど 分割例をここに示す

2.1.3 形態素解析器 JUMAN

形態素解析をおこなうシステムを形態素解析器という。

形態素解析器のひとつに JUMAN^(注1)がある。JUMAN は辞書や品詞体系の定義ファイルが解析システムと独立している。このため、解析方法についてユーザが自由に定義をすることができる。形態素の品詞のほかにカテゴリやドメインなどの意味情報を持つことが特徴のひとつである。

2.2 固有表現抽出

2.2.1 固有表現

固有表現とは人名、地名、日時などの情報をもつ表現のことをいう。IREX [6] は、固有表現とは固有名詞的表現、時間表現、数値表現のことであると定義している。この固有表現は 8 種類の細目からなる。

表 1 固有表現種類一覧

	固有表現種	対応する開始タグ	対応する終了タグ
固有名詞的表現	組織名	<ORGANIZATION>	</ORGANIZATION>
	人名	<PERSON>	</PERSON>
	地名	<LOCATION>	</LOCATION>
	固有物名	<ARTIFACT>	</ARTIFACT>
時間表現	日付表現	<DATE>	</DATE>
	時間表現	<TIME>	</TIME>
数値表現	金額表現	<MONEY>	</MONEY>
	割合表現	<PERCENT>	</PERCENT>

(注1): 黒橋・河原研究室 日本語形態素解析システム JUMAN
<http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN>

2.2.2 固有表現抽出

テキストに対してそれぞれの固有表現文字列のタグ付けを行うことを固有表現抽出とよぶ。固有表現のタグを用い入れ子にならず、重複もしないようにタグ付けを行う。表の8種のほかに、タグ付けが困難なものやタグ種を一意に決定できないものを示す <OPTIONAL></OPTIONAL> を用いる。

2.3 セグメンテーション

テキストを意味的なまとまりで分割することをセグメンテーションという。

手法には併合と分割の二つがある[2]。併合は、テキストを文などの単位で小分けにしたものを結束性によってマージしてゆく手法である。分割は、区切りのないテキストに対して分割点を探す手法である。

分割の代表的な手法は Text Tiling 法[1]がある。

2.4 小説テキストの特徴

2.4.1 作品独自の固有表現

小説テキストの特徴のひとつには作品独自の固有表現がある。小説にはしばしば架空の登場人物名や地名、組織名などが現われる。固有表現を収集した辞書はさまざまなものがある。人名に例をとると、歴史上の人物を中心に収集した人名辞典や存命の人物を中心に収集した紳士録、フィクション中の登場人物をまとめた架空人名辞典などがある。また、形態素解析器 JUMAN は新聞記事をもとに語彙辞書を作成しており、固有名詞に対して人名や地名などの情報を付与してある。

しかし、作品独自の固有表現は既存の知識やテキストから生成した辞書には含まれないことがある。解析の対象となる小説においてのみ独自に表われる固有表現については、既存のテキストから作成した辞書を利用できない。

とど 例を示すたとえば、山月記の人名にりちょう、えんさん、地名にろうさい文字禍 では 人名にナブアヘエリバ、地名にアッシリアなどなど がある

2.4.2 場面

もうひとつの特徴として場面概念があげられる。セグメンテーションの単位として場面を用いることができる。

小説テキストはその物語内容の変遷によって、いくつかのまとまりに区切ることができる。小説において意味的につながりのある連続した文のまとまりを場面とよぶ。計算機によって小説テキストに対して場面の区切りを推定することを場面分割とよぶ。

場面分割のために、場所、時間、登場人物などの変化に着目する。場所を移動したとみられる表現、時間が移り変わったとみられる表現、登場人物が入れ替わったとみられる表現から場面の区切りを推定する。

場面は小説の意味解析の際にネットワークを構成するノードとして用いられる。

3. 関連研究

3.1 場面分割

小林[4]は、シーンの定義およびその分割手法を提案している。自然言語処理における従来の時間と場所によるシーンの定

義に加え、構成主義の人物に着目した物語分析を参照し、シーンの境界要件として場所、時間、登場人物を導入する。これら3種の場面の構成要素に変化に注目し、場面分割を行う。ここでは場面を構成する最小単位を文とした場面分割の手法を提案している。テキストを形態素解析器 JUMAN を用いて形態素解析を行い、形態素の共起などから場所、時間、登場人物を表す語を推定する。それぞれの文間の異なり数の線形和から求めたペナルティ値が閾値を超えたか否かで、文境界がシーン境界であるかを判定する。

渡邊[5]では形式段落を最小単位にした場面分割を行なっている。場所、時間、人物を表わす語によってそれぞれの形式段落間の単語距離を求め、距離が閾値を越えたものをそれぞれの区切り候補とする。あるいは時間と人物の区切り候補、あるいは時間と場所の区切り候補である形式段落境界を場面の境界と判定する。このときの閾値の条件による分割精度を評価を行なっている。

3.2 固有表現抽出

固有表現抽出の手法として、人手によって抽出規則を作成する手法がある。竹元ら[11]では固有表現辞書とパターンマッチルールを用いた固有表現抽出手法を提案している。ここでは辞書の強化や人手によるルールの整備により精度を改善している。

一般的な固有表現抽出の手法として機械学習を用いるものがある。山田ら[7]では SVM を用いて学習を行う。固有表現をタグ付けしたテキストをサンプルとして、固有表現の抽出規則を学習する。

馬場ら[9]は人名辞書を利用した小説テキストを対象とした人名抽出の手法を提案している。人名抽出の際に、形態素解析器 JUMAN の人名辞書に加え、既存の人名辞典を利用して候補語を選別する。また、候補語の出現頻度に基づいて人名と判別するか否かの推定を行う。

荒井ら[8]は造語を対象にした固有表現抽出手法を提案している。これは Web 検索の検索数から候補となる語を選別し、候補語の素性を Amazon.com や Wikipedia で検索した結果から判定するものである。このように判定した固有表現を既存の手法で抽出した固有表現と合併する。

馬場[10]では小説テキストにおける人名の抽出と、抽出した人名を用いた人物の関係情報の体系化を行なっている。人名抽出には、形態素解析器 ChaSen の解析情報を利用している。また、「8万人西洋人名よみ方綴り方辞典」に収録された人名情報を ChaSen の辞書に追加している。

3.3 課題

小説テキストを対象とした自然言語処理の課題として作品独自の固有表現の扱いがあげられる。

既存研究ではタグ付けしたテキストによる機械学習、辞書や Web などの外部のデータベースの利用などによって固有表現抽出を行っていた。いずれも既存のテキストをもとにしたデータベースを参照するものであり、処理の対象である小説テキストにおいて初めて現れるような作品独自の固有表現を扱うことができない。

本稿では解析対象である小説に表われる作品に特有の固有表

現を独自の固有表現と呼ぶことにする。独自の固有表現うち、とくに人名を表わすものを扱った。

4. 予備実験

提案に先立ち、二つの予備実験を行なった。

ひとつは小説テキストの場面分割の実験、もうひとつは固有表現辞書作成についての実験である。

場面分割の実験では既存研究の手法を参考に簡易的なシステムを実装し、その解析精度を F 値に基づいて評価した。

固有表現辞書作成の実験では人力で作成した抽出規則を用いた固有表現抽出を行ない、抽出規則を固有表現の再現率によって評価した。

予備実験の結果をもとに提案を行う。

4.1 予備実験 (1): 小説テキストの場面分割

渡邊 [5] の手法を参考にし、形式段落を最小単位とした小説テキストの場面分割を行うシステムを作成した。実験を行いこのシステムの解析精度を評価した。

4.1.1 構成

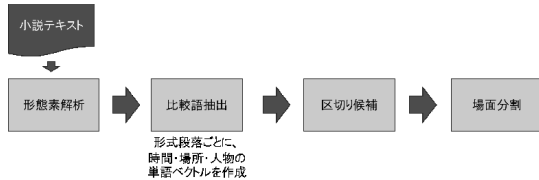
システムの構成を図 4.1.1 に示す。

テキストを形態素解析し、JUMAN の意味情報を利用して、人物・時間・場所を表す語を抽出する。

形式段落ごとに、抽出した語からなる単語ベクトルを作成する。

この単語ベクトルを用いて、ウィンドウサイズ 1 の Text Tiling 法によって人物・時間・場所それぞれの区切り候補を求める。2 種類以上の区切り候補である箇所を場面の区切りと判定する。

図 1 システムの構成



4.1.2 特徴語抽出

JUMAN の意味情報より、人物・時間・場所を表わす語からなる 3 種類の単語ベクトルを作成する。

- 人物を表わす語は、「カテゴリ:人名」あるいは「カテゴリ:人物」と解析されたものとした。
- 時間を表わす語は、「カテゴリ:時間」と解析されたものとした。
- 場所を表わす語は、「カテゴリ:場所」「カテゴリ:地名」と解析されたものとした。

単語ベクトルの次数は小説テキストの全文に表われる比較語とし、各成分は形式段落における特徴語の出現回数を表わすものとする。

表 2 単語ベクトルの例: 人物を表わす語

私	人	太郎
(4,	0,	1)

4.1.3 区切り候補

単語ベクトルからそれぞれ区切り候補を求める。

テキストの分割には Text Tiling 法を用いる。Text Tiling 法は、基準点から前後に同じだけ広げたウィンドウを作成し、となりあう基準点におけるウィンドウの類似度を求め、ウィンドウの類似度が低い基準点間でテキストを分割するものである。ここでは形式段落の境界における、形式段落を単位としたウィンドウサイズ $n = 1$ の分割を行った。

すべての形式段落 S_i を場面分割候補として、以下の操作を行なう。予め定めたウィンドウサイズ n により、任意の形式段落 S_i に対応するウィンドウ W_i を作成する。ウィンドウ W_i は形式段落 S_i およびその前後 n 個の形式段落のまとまりである。すべてのウィンドウ W_i について、ウィンドウに含まれる形態素からなる 3 種類のベクトル、人物ベクトル C_i 、時間ベクトル T_i 、場所ベクトル P_i を作成する。

任意の形式段落 S_i のベクトル種類ごとの距離 $D_{c,i}$ 、 $D_{t,i}$ 、 $D_{p,i}$ を、隣り合う形式段落に対応するウィンドウの特徴ベクトルとのユークリッド距離とする。すなわち、

$$D_{c,i} = C_{i+1} - C_i = \sqrt{\sum_{k=1}^m (c_{i+1,k} - c_{i,k})^2}$$

$$D_{t,i} = T_{i+1} - T_i = \sqrt{\sum_{k=1}^m (t_{i+1,k} - t_{i,k})^2}$$

$$D_{p,i} = P_{i+1} - P_i = \sqrt{\sum_{k=1}^m (p_{i+1,k} - p_{i,k})^2}$$

とする。ただし、 m は特徴語の数とする。

人物、時間、場所についてそれぞれ距離が閾値を超えたとき、形式段落 S_i と S_{i+1} との境を場面の分割候補点とする。

4.1.4 場面分割

時間と人物の分割候補点、または時間と場所の分割候補点である形式段落の境を場面の分割点と判定する。

4.1.5 実験

システムによって中島敦の「山月記」「文字禍」「狐憑」の場面分割を行なった。

さらに、人手の場面分割を正例としてシステムの判定結果を評価した。

ただし、場面分割の閾値は中島敦の「木乃伊」の分割結果が最良となった 10 とした。

4.1.6 結果

作品名	適合率	再現率	F 値
山月記	4/9	4/8	0.47
文字禍	5/7	5/9	0.62
狐憑	3/8	3/6	0.42

表 3 場面分割の結果

3 作品の F 尺度の平均値は.50 であった。これは [5] の結果と同程度であり、[4] の結果より.03 低かった。

4.1.7 考 察

形態素解析の結果を見たところ人物を表すと見做されるべき語の見落としがあった。

ひとつに、日本語では常用しない漢字からなる中国の人名は複数の形態素として分割されていた。例えば、『山月記』における「李徴」や「袁□（人偏に参）」が該当する。

また、カタカナからなる古代の神の名前や外国人名は未定義語と解析されていた。例えば、『文字禍』における「死者の霊エティム」や「ナブ・アヘ・エリバ」などがこれにあたる。

このように、JUMAN の辞書にはない語について形態素解析に失敗していた。

また、人物の表現と比べて時間や場所の表現は抽出数が少なく、距離が閾値を超えにくい傾向にあった。

4.2 予備実験 (2): 単一の形態素からなる独自の固有表現辞書の作成

4.2.1 手 法

テキストから独自の固有表現を抽出し辞書を作成するシステムの実験を行った。

テキスト全文を形態素解析する。助詞「は」の直前にある形態素のうち未定義語であるものを独自の固有表現の候補とする。

さらに、候補のうちテキスト全文において複数回出現したものを独自の固有表現として判定する。

4.2.2 実 験

中島敦「文字禍」「山月記」を対象に、独自の固有表現辞書を作成する実験を行った。

さらに人手で抽出した人名と独自の固有表現辞書とを比較してシステムの評価を行なった。

4.2.3 結 果

システムの再現率は 5 % であった。ただし、人名が複数の形態素からなる場合はその末尾の形態素のみを独自の固有表現と判定した。

4.2.4 考 察

漢字からなる人名についてはその末尾の一字が抽出された。例えば「李徴」のうちの「徴」などが抽出された。

カタカナからなる人名についてはその末尾の形態素ひとつが抽出された。例えば「ナブ・アヘ・エリバ」のうち「エリバ」などが抽出された。

これらから、求めるべき独自の固有表現を得るためには、複数の形態素にまたがった抽出を行なう必要があることがわかった。

5. 提 案

5.1 概 要

以上の予備実験や前述の課題をふまえて、独自の固有表現を扱う場面解析システムを提案する。これは次のような特徴がある。

a) 複数の形態素からなる独自の固有表現辞書を作成

予備実験 4.2 では単数の形態素からなる独自の固有表現のみを抽出したため、複数の形態素からなる固有表現を取り零した。そこで、独自の固有表現を正確に扱うために複数の形態素からなる独自の固有表現辞書を作成する。

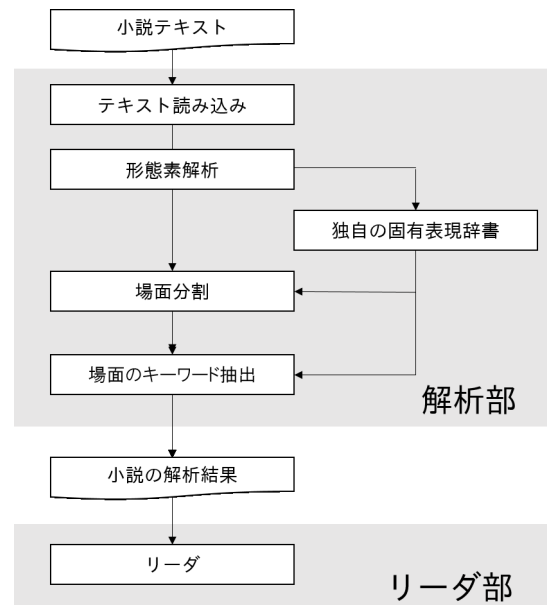


図 2 システムの構成

「文字の精霊」(中島敦「文字禍」)や「私」(夏目漱石「こころ」)などを固有表現と見做したいがために、IREX [6] の固有表現の定義を逸脱し接続助詞の「の」を含むことや、代名詞を許すような判定条件を用いた。

予備実験と同様に、登場人物は八格になるという考えから助詞「は」のひとつ前の形態素をなめて候補を抽出する。

次に、さらに前に遡った形態素を加えても人名として妥当である場合は候補の形態素の列を拡張する。これ以上拡張できなくなるか、列の長さが上限値に到達するまでこの拡張を繰り返す。

このようにして、複数の形態素からなる独自の固有表現辞書を作成する。

b) 場面分割の改善

場面分割は各比較語の単語ベクトルが正確になれば精度を改善できると考えた。予備実験では、とくに人物について取り零しが目立った。そこで、JUMAN の意味情報に加え、独自の固有表現辞書を用いて抽出した比較語をもとに場面分割をする。

また、予備実験 4.2 において、人物に比べて時間や場所の表現の種類が少なく、また複数回出現することも少なかった。そのため比較語ベクトルの距離の閾値の超えやすさに差があった。そこで、比較語の種類によって閾値を替えた。

c) 場面のキーワード抽出

独自の固有表現辞書を用いて場面に出現した人物のキーワードを抽出する。

5.2 構 成

システム全体の構成を図 5.2 に示す。以下に流れをおって各部を説明する。

5.2.1 テキスト読み込み

小説テキストのプレーンテキストデータを入力として与える。テキストは本文のみからなるものとする。

青空文庫形式のテキストファイルを扱う場合は、システムに

入力するまえに以下の整形を行う。

- 冒頭部および末尾部の除去
- 入力者註の除去
- ルビの除去

これらにより、メタ情報を除去し本文のみを取り出し、マークアップされたテキストをプレーンテキストにする。

他の場合でも題名、作者名、目次、その他メタ情報などは除去したうえで、本文のみのプレーンテキストを入力とする。

5.2.2 形態素解析

形態素解析器 JUMAN を用いて形態素解析を行う。

5.2.3 独自の固有表現辞書作成

複数の形態素からなる独自の固有表現辞書を作成する。

まず、助詞「は」の直前にある形態素ひとつを始めの候補として、形態素とその出現回数を記録する。ただし、以下の条件を満たすものに限る。

- 句読点や鉤括弧ではない (文を跨がない)
- 動詞ではない
- 判定詞、助詞、接続詞ではない

次に候補の長さをひとつ前に拡張する。上の3つの条件を満たすようなひとつ前の形態素があるならば、その形態素加えた語をあらたに候補とし、形態素の列と出現回数 n を記録するとともに、現在の候補の出現数から n を減ずる。この拡張を候補の長さが最大形態素数 M になるか、これ以上拡張できる候補がなくなるまで行う。

このようにして求めた候補のうち、2回以上出現したものを独自の固有表現として辞書に登録する。

5.2.4 単語ベクトル作成

JUMAN の意味情報と独自の固有表現辞書より、人物・時間・場所を表わす語からなる3種類の単語ベクトルを作成する。

- 人物を表わす語は、「カテゴリ:人名」あるいは「カテゴリ:人物」と解析されたもの。または独自の固有表現辞書にある語句とした。
- 時間を表わす語は、「カテゴリ:時間」と解析されたものとした。
- 場所を表わす語は、「カテゴリ:場所」「カテゴリ:地名」と解析されたものとした。

5.2.5 場面分割

4.1 の予備実験と同様の手法で場面分割を行う。

5.2.6 場面のキーワード抽出

独自の固有表現辞書から、人物を表わす語を抽出する。

また、場面分割およびキーワード抽出による小説の解析結果を次に示すの json 形式で出力し、リーダーに渡す。

結果出力に用いる json フォーマット

```
{
  "meta": {
    "amount": 場面数
    "author": 表示用の著者名
    "title": 表示用の作品名
  }
  "bamen": {
```

```
{ 場面番号: {
  "text": 本文
  "tag": [登場人物のリスト]
}...
}
```

出力は、解析結果として場面ごとの本文と登場人物のリスト、さらに小説テキストのメタ情報として、分割した場面数、リーダーの表示に用いる著者名と作品名からなる。

5.2.7 リーダ

リーダーでは場面単位でページ分けし、小説を閲覧する。

a) インポート

はじめに結果のインポートを行う。解析部から出力された結果の json をパースし、データベースに登録する。

b) 閲覧

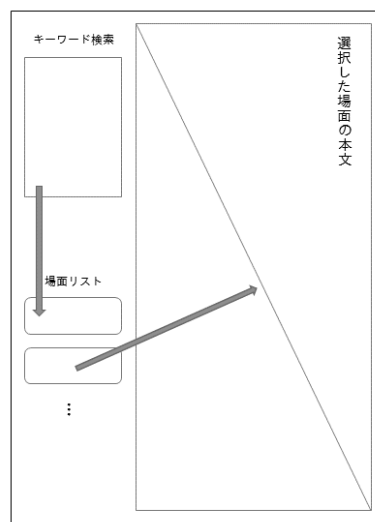
インポートが完了すると、場面一覧が表示される。ここから読みたい場面を選択し、小説を閲覧する。検索メニュー、場面の本文、場面の登場人物、次のページへのリンクが表示される。

c) 場面検索

人物キーワードによって場面の絞りこみ検索を行なう。検索キーの登場人物が現れる場面を検索する。キーワードは組み合わせ、複数の登場人物が同時に登場する場面を検索することができる。

キーの選択は、検索メニューおよび登場人物一覧から行う。

図3 リーダの概観



6. 実験と評価

6.1 目的

本提案では場面分割、辞書作成、場面要約などを行ない、小説テキストを解析する。

それぞれの部分システムの性能を評価するために実験を行った。解析の精度のついて主に人手で作成した正例をもとに計測した。

6.2 場面分割

6.2.1 実験 1

JUMAN の意味情報のみからなる場面分割と、独自の固有表現辞書を加えた場面分割との分割精度を比較した。

予備実験と同様に中島敦の「山月記」「文字禍」をサンプルに、人手の場面分割を正例として判定精度を算出した。また、場面分割の閾値は中島敦の「木乃伊」の分割結果が最良となった 10 を用いた。

6.2.2 実験 1 の結果

結果は作品ごとに以下ようになった。

表 4 実験 1 「山月記」の場面分割の比較

山月記	F 値
独自の固有表現辞書なし	0.47
独自の固有表現辞書あり	0.47

表 5 実験 1 「文字禍」の場面分割の比較

文字禍	F 値
独自の固有表現辞書なし	0.62
独自の固有表現辞書あり	0.62

両手法について共に 3 作品の F 尺度の平均値は .50 であった。また、「文字禍」について分割箇所を比較したところ違いがなかった。

6.2.3 実験 2

閾値を比較語種別に設定して場面分割を行い、共通の閾値による場面分割と分割精度を F 値に基づいて比較した。

予備実験と同様に中島敦の「山月記」「文字禍」をサンプルに、人手の場面分割を正例として判定精度を算出した。また、場面分割の際に、人物の閾値には中島敦の「木乃伊」の分割結果が最良となった 10 を用い、時間と場所の閾値については 3 を用いた。

6.2.4 実験 2 の結果

表 6 実験 2 「山月記」の場面分割の比較

山月記	F 値
共通の閾値	0.47
比較語種別の閾値	0.56

表 7 実験 2 「文字禍」の場面分割精度の比較

文字禍	F 値
共通の閾値	0.62
比較語種別の閾値	0.71

結果は表 6, 7 のようになった。

共通の閾値による場面分割に比べ、10 % 程度精度が改善された。

6.2.5 結果と考察

独自の固有表現辞書の導入によって場面分割の精度は変化しなかった。これは比較に用いた単語ベクトルに含まれる語のうち独自の固有表現の占める割合が小さく、閾値を跨ぐほどの影響を及ぼさなかったためであると考えられる。

また、比較語種別の閾値を設定することで場面分割の精度が向上した。

6.3 独自の固有表現辞書作成

6.3.1 実験

複数の形態素からなる語句をもって特定の人物を指すような作品独自の表現を抽出することが目的である。この実験では中島敦「文字禍」を対象に独自の固有表現を抽出した。

提案手法である複数の形態素からなる独自の固有表現辞書と単一の形態素からなる手法によって独自の固有表現辞書を作成した。

さらに人手で抽出した人名リストを作成した。ここでは JUMAN の解析によって人名と判定されないもののうち人物を表わす語句と、繰り返し表われて特定の人物を指す一般的な語句とを人名として抽出した。

この人名リストを正例として、予備実験 4.2 で行なった単一の形態素からなる手法と比較し結果を評価した。

また、複数の形態素からなる独自の固有表現辞書の作成における最大形態素数 M は 10 とした。これは人手で抽出した人名をすべて扱うために十分な長さである。

6.3.2 結果

作成した独自の固有表現辞書の語彙数は 17 項目であった。

複数の形態素からなる独自の固有表現辞書では、人名リストのうち 15 % が独自の固有表現と判定できた。予備実験の単一の形態素からなる場合に比べ、3 倍の語を判定することができた。

さらに辞書に含まれる語の内容について、JUMAN が複数の形態素に分析したものに対してひとつの固有表現として判定することができた。

6.3.3 考察

単一の形態素からなる手法に比べ、複数の形態素からなる独自の固有表現辞書にすることで網羅性を改善することができた。

提案手法では助詞「は」の前にある語句のみを見たため、八格として表われなかった人名は抽出できなかった。また、テキストに複数回表われることで重要度を測り選別を行うため、人名のうち一度しか表われない語句についても抽出できなかった。

図 4 リーダの実行例



7. ま と め

本稿では、小説の電子書籍リーダの作成を目的に、小説テキストを対象とした固有表現抽出、場面分割、場面要約を行うシステムの提案し、検討を行なった。さらにこれらの成果からキーワードによる検索機能を備えた電子書籍リーダを作成した。

文 献

- [1] M.A.Hearst.:TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages, Computational Linguistics, Vol.23, No.1, pp.33-64, 1997.
- [2] 平尾努, 北内啓, 木谷強:語彙的結束性と単語重要度に基づくテキストセグメンテーション, 情報処理学会論文誌データベース, Vol.41, No.SIG03, pp.24-36, 2000.
- [3] 福田 雅志, 延澤 志保, 太原 育夫: 語彙的結束性に基づく話し言葉のテキストセグメンテーション, 言語処理学会第 11 回年次大会発表論文集, pp.620-623, 2005.
- [4] 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自然言語処理, Vol.47, No.NL179, pp.25-30, 2007
- [5] 渡邊祥太:電子書籍を対象とした視覚的スタイル自動付与システムの提案, 2013
- [6] IREX:固有表現抽出課題 (version 990214), 入手先 (<http://nlp.cs.nyu.edu/irex/NE/df990214.txt>)(参照 2016-01-01)
- [7] 山田寛康, 工藤拓, 松本裕治:Support Vector Machine を用いた日本語固有表現抽出, 情報処理学会 Vol.43, No.1, pp.44-53, 2002
- [8] 荒井 徹, 大和田 勇人:造語に特化した固有表現抽出, 情報科学技術フォーラム講演論文集 Vol.9, No.2, pp.397-398, 2010-08-20
- [9] 馬場こずえ, 藤井敦, 石川徹也:小説テキスト自動分類のためのジャンル推定と人物抽出, 情報科学技術フォーラム一般講演論文集 Vol.4, No.2, pp.67-70, 2005
- [10] 馬場こずえ:小説テキストを対象とした人物情報の抽出と体系化, 2006
- [11] 竹元 義美, 福島 俊一, 山田 洋志:辞書およびパターンマッチルールの増強と品質強化に基づく日本語固有表現抽出, 情報処理学会論文誌, Vol.42, No.6, pp.1580-1591, 2001-06-15
- [12] 岩倉友哉:固有表現抽出, 情報処理 Vol.57, No.1, pp.16-17, 2016.
- [13] 吉村 賢治:自然言語処理の基礎 新訂版, サイエンス社 (2012).
- [14] 長尾 真, 佐藤 理史, 黒橋 禎夫, 角田 龍彦:岩波講座ソフトウェア科学 自然言語処理, 岩波書店 (1996).
- [15] ジェラルド・プリンス (著). 遠藤健一 (訳):改訂 物語論辞典, 松柏社 (2015).
- [16] 森 辰則:文書からの情報抽出と固有表現抽出— 文書から特定情報を切り出す —, 画像ラボ 1999 年 12 月号.