社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2004－NL－164 (27)
2004／11／6

重要文抽出を利用したテキストからのストーリー抽出

相良 直樹† 砂山 渡‡ 谷内田 正彦†

†大阪大学大学院基礎工学研究科 〒560-8531 豊中市待兼山町 1-3
‡広島市立大学情報科学部 〒731-3194 広島市安佐南区大塚東 3-4-1

E-mail: †n-sagara@yachi-lab.sys.es.osaka-u.ac.jp, ‡sunayama@sys.im.hiroshima-cu.ac.jp

あらまし 情報化社会の発展に伴い，文書の電子化や大量管理，共有が可能となった．これにより，人が処理しなければならない情報量も増加している．現在，この問題に対処すべくテキスト要約において主題抽出の研究は盛んに行われているが，ストーリー抽出に関しては人手に頼らざるを得ないのが現状である．本研究においては，従来の重要文抽出を利用したテキストからのストーリー抽出手法を提案し，この有効性について考察する．

キーワード ストーリー抽出，メインストーリー，サブストーリー，重要文抽出

Story Extraction from Text Using Key Sentences Extraction Method

Naoki SAGARA† Wataru SUNAYAMA‡ and Masahiko YACHIDA†

†Graduate School of Engineering Science, Osaka University 1-3 Machikaneyama-cho, Toyonaka-shi, Osaka, 560-8531 Japan
‡Faculty of Information Science, Hiroshima City University 3-4-1 Ozuka-Higashi, Asa-Minami-ku, Hiroshima-shi, Hiroshima, 731-3194 Japan

E-mail: †n-sagara@yachi-lab.sys.es.osaka-u.ac.jp, ‡sunayama@sys.im.hiroshima-cu.ac.jp

Abstract Development of information society caused computerizing of documents, and made a lot of managements and share of documents possible. It caused the increase in the amount of information which people have to process. Currently although research of extracting subjects is active in text summarization to deal with this problem, story extraction cannot be performed without manpower. In this paper, we propose a method of story extraction from text using key sentences, then discuss the proposing method.

Keyword story extraction, main-story, sub-story, key sentence extraction

1. はじめに

近年のパソコンや WWW(World Wide Web)の急速な成長に伴い，インターネットは世界規模でその利用人口が増え続けている．これに伴い大量の情報を誰もが容易に入手可能となってきた．そのような状況下で現在我々は情報過多の状態に陥っており，情報量の増加に処理能力が追いついてない．このような状況では，情報の奔流に埋もれている本当に必要な知識を効率よく獲得することが難しくなってしまう．そこで，文書の内容をまとめてストーリーとして簡潔に提示するシステムがあれば，非常に有益であると考えられる．

また文章作成においても，自分の作成した文章のストーリーを確認することができれば，自分の意図したストーリーと作成した文章のストーリーの差異を知ることが可能となり有用である．

本研究においては，テキストのストーリーモデルについて述べ，ストーリーモデルを構成するメインストーリーとサブストーリーを表す単語を評価する方法の提案と，その評価実験について述べる．

2. 研究背景

文書要約の手段としては，文章中の各文に評価値を与え，その評価値の高い文を文章の主題を表す文として抽出する手法がよく用いられている．文章中の各文を評価する手法としては，以下のものがある．

1. 文に含まれる単語を評価する
2. 文と他の文との関わりを評価する
3. 文を表層的な手がかりによって評価する

まず 1. については，文章のキーワードを取り出して評価することに相当し，高頻度語を取り出す方法[1]，他の文にない単語を取り出す tfidf 法[2]の２つが広く用いられている．2. は文章中の重要語群を抽出する語彙的連鎖[3]を用いる手法が多く用いられており，文章中の多くの文と関係をもつ文を重要文として抽出する方法がある[4]．3. については，接続詞など特定の表記を手がかりとする方法[5]などが挙げられる．しかし，これらの手法によって文章の主題のみを抽出するだけでは，筆者が主張したい重要な部分は理解できても文章のストーリーを理解することはできない．

一方で，文章の流れを意識した要約の研究[6]もある．この研究においては，文章のメインテーマのみに沿った要約を作成することを目的としており，文章のメインストーリーだけでなくサブストーリーにも焦点を当てる本研究とは異なっている．

3. ストーリーモデル

本章では，テキスト中のメインストーリーとサブストーリーの定義を述べた上で，テキストのストーリーモデルを定義する．テキストには全体を通して筆者が述べたい主な話題（メイントピック）と，このメイントピックに関連してテキストの一部で述べられるサブトピックとがある．そこで，各トピックを代表するキーワードを以下のように定める．

1. メインキーワード：テキストのメイントピックを表すキーワード

2. サブキーワード：テキストのメイントピックと関連したサブトピックを表すキーワード

また，メインキーワードに関する話をメインストーリー，サブキーワードに関する話をサブストーリーと呼ぶ．これらで定義されるストーリーモデルを図1に示す．ただし，図中の「メイン」と「サブ」は，それぞれメイントピックとサブトピックを表し，矢印は各トピックに関するストーリーを表す．すなわちテキストには，主題に関わる大きな話の流を表すメインストーリーと共に，メインストーリーと関連を持ったテキストの各部分での話題に伴う局所的な話の流れを表すサブストーリーが存在する．本モデルにおいては，テキストのストーリーとは上記のメインストーリーとサブストーリーとが織り成す話の流れであると定義する．

4. システム概要

前章で述べたストーリーモデルに基づく本システムの処理の流れを図2に示す．システムはテキストを入力として，入力されたテキストからメインキーワードを抽出し，メインキーワードに関するストーリー抽出を行う．その後，サブキーワードを抽出し，サブキーワードに関するストーリー抽出を行う．最後に，メインストーリーとサブストーリーの組み合わせを行ったものをテキストのストーリーとして出力する．以下本章では，システムの各モジュールについて述べる．

4.1. メインキーワード抽出モジュール

本研究においては，メインキーワードをテキストの広範囲に出現し，テキストにおいて最も重要な意味を持った名詞であると定義する．この定義に基づくメインキーワード抽出アルゴリズムを以下に記す．

メインキーワード抽出アルゴリズム：

[STEP1] 以下の条件1を満たす名詞 n の集合 *Set*1 と条件2-1，2-2を満たす名詞 n の集合 *Set*2 を抽出する．

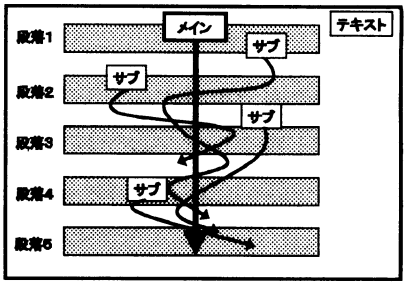

図1 ストーリーモデル

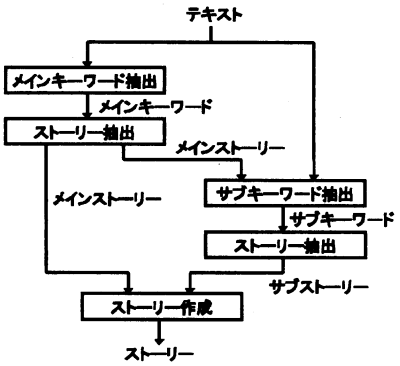

図2 システムの処理の流れ

条件 1. テキストの2段落以上に主語として用いられている文が存在する．

条件 2-1. テキストの1段落以上において高頻度で出現する，すなわちテキスト中で重要な役割を果たしていることを表す式(1)を満たす．

$$freq_{seg}(n) \geq mean_{seg} + var_{seg} \quad (1)$$

$freq_{seg}(n)$: 段落における名詞 n の出現頻度

$mean_{seg}$: 段落の名詞の平均出現頻度

var_{seg} : 段落の名詞の出現頻度の分散

条件 2-2. テキストの2段落以上において，その名詞が段落中で偶然出現したのではなく，一定の意味をもって用いられていることを表す式(2)を満たす．

$$freq_{seg}(n) \geq mean_{seg} \quad (2)$$

[STEP2] *Set*1 の各名詞 n にテキスト中で主語となる回数に基づいて式(3)により(0,1]の評価値 $E_{Set1}(n)$ を付与する．*Set*2 の各名詞 n にテキスト中の出現頻度に基づいて式(4)により(0,1]の評価値 $E_{Set2}(n)$ を付与する．

$$E_{Set1}(n) = \frac{subjectFreq_{text}(n)}{max\_SubjectFreq_{Set1}} \quad (n \in Set1) \quad (3)$$

$$E_{Set2}(n)=\frac{freq_{text}(n)}{max\_Freq_{Set2}} \quad (n \in Set2) \ (4)$$

$subjectFreq_{text}(n)$: テキスト中でnが主語となる頻度

$max\_SubjectFreq_{Set1}=\max\{subjectFreq_{text}(n) \mid n \in Set1\}$

$freq_{text}(n)$: テキスト中におけるnの出現頻度

$max\_Freq_{Set2}=\max\{freq_{text}(n) \mid n \in Set2\}$

[STEP3] テキストから$E_{Set1}(n)+E_{Set2}(n)$で上位の名詞nをメインキーワードとして抽出する．

本モジュールの評価実験については，サブキーワード抽出モジュールとあわせて 5.1.節で述べる．

4.2. ストーリー抽出モジュール

本モジュールの構成を図 3 に示す．本モジュールはキーワード，およびテキスト全体と各段落を入力とし，テキスト中の各文に入力キーワードに対する評価値を付与し，その評価値で上位の文を入力キーワードに関するストーリーの重要文として抽出する．最後に抽出された文の意味的接続を行い，入力キーワードに関するストーリーとして出力する．

本研究では，あるキーワードに関するストーリーの重要な文とは，以下の２点を満たすものと定義する．

1. テキストのキーワードに関する主張と大きな関わりを持ち，かつそのテキストを特徴づける文
2. その文が属する段落のキーワードに関する主張と大きな関わりを持ち，かつその段落を特徴づける文

一方で，展望台システム[7]と呼ばれる重要文抽出システムは，情報を探す人間の観点に基づいた重要文抽出が可能である．また展望台システムは，各単語の頻度（文章の主張）と各文内における単語の共起頻度に基づく条件付確率（文章の特徴付け）を用いて重要文抽出を行うシステムである．このように展望台システムは，入力キーワードに関する重要文抽出が可能であり，文の重要度評価が本研究におけるストーリーの重要度の定義と一致するため，本研究においては重要文抽出システムとして展望台システムを採用した．

本モジュールの入力キーワードに関するストーリーの重要文抽出までの処理について以下に述べる．

まず，テキスト中の各文S_nに対して 1. の観点から入力キーワードkey_mに関するテキスト全体における重要度順位（テキスト順位：$R_{text}(S_n,key_m)$）と，2. の観点からS_nが属する段落における重要度順位（段落順位：$R_{seg}(S_n,key_m)$）を展望台システムにより与える．ただし，段落中にキーワードkey_mが存在しない場合は，S_nに対して$R_{seg}(S_n,key_m)$を与える処理は行わない．次

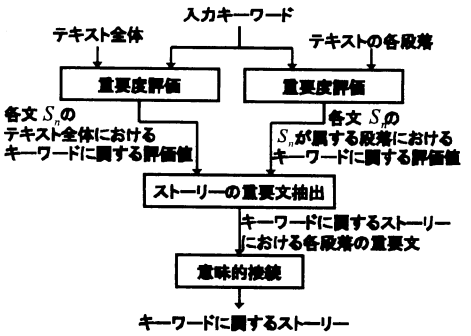

図 3　ストーリー抽出モジュール

に，テキスト順位と段落順位の組み合わせを式(5)～(9)によって行い，テキスト中の各文S_nに対して入力キーワードkey_mに関するストーリーにおける重要度を表す評価値$story(S_n,key_m)$を付与する．上記で述べた段落中にキーワードkey_mが存在しない場合には，式(7),(8)の$\overline{R_{seg}}(S_n,key_m)$の値は 0 とする．式(5)～(9)ではテキスト順位と段落順位をkey_mに関するストーリーにおける重要度を測る評価基準として同じ重みで扱うために，各々を[0,1]の範囲にうつした後，線形和を取っている．このようにテキスト中の各文S_nに$story(S_n,key_m)$を付与した後，各段落から$story(S_n,key_m)$で上位N文をキーワードkey_mに関するストーリーの重要文として抽出する．ただし，Nは段落文数の 1/10 とする．

if $R_{text}(S_n,key_m) \geq Ave_{text}(key_m)$

$$\overline{R_{text}}(S_n,key_m)=0.5-\frac{R_{text}(S_n,key_m)-Ave_{text}(key_m)}{Worst_{text}(key_m)-Ave_{text}(key_m)}\times 0.5 \ (5)$$

else

$$\overline{R_{text}}(S_n,key_m)=0.5+\frac{Ave_{text}(key_m)-R_{text}(S_n,key_m)}{Ave_{text}(key_m)-Best_{text}(S_n,key_m)}\times 0.5 \ (6)$$

if $R_{seg}(S_n,key_m) \geq Ave_{seg}(key_m)$

$$\overline{R_{seg}}(S_n,key_m)=0.5-\frac{R_{seg}(S_n,key_m)-Ave_{seg}(key_m)}{Worst_{seg}(key_m)-Ave_{seg}(key_m)}\times 0.5 \ (7)$$

else

$$\overline{R_{seg}}(S_n,key_m)=0.5+\frac{Ave_{seg}(key_m)-R_{seg}(S_n,key_m)}{Ave_{seg}(key_m)-Best_{seg}(S_n,key_m)}\times 0.5 \ (8)$$

$$story(S_n,key_m)=\overline{R_{text}}(S_n,key_m)+\overline{R_{seg}}(S_n,key_m) \quad (9)$$

$Ave_{text}(key_m)$: テキスト中の各文S_kの$R_{text}(S_k,key_m)$の平均

$Ave_{seg}(key_m)$: 段落中の各文S_kの$R_{seg}(S_k,key_m)$の平均

$Worst_{text}(key_m)$: テキスト中の各文S_kの$R_{text}(S_k,key_m)$の最大値

$Worst_{seg}(key_m)$: 段落中の各文S_kの$R_{seg}(S_k,key_m)$の最大値

$Best_{text}(key_m)$: テキスト中の各文S_kの$R_{text}(S_k,key_m)$の最小値

$Best_{seg}(key_m)$: 段落中の各文S_kの$R_{seg}(S_k,key_m)$の最小値

しかし，このようにして抽出したストーリーの重要文を並べただけでは，必ずしもそれらの文間に意味的な繋がりが存在しないため意味的接続を行う．

この意味的接続では，連続する文が同一の名詞を含むと，人はその文間に意味的繋がりを感じやすい[8]という性質を利用して，入力として受け取った連続する文に少なくとも1つの同一の名詞を含むように文を挿入する処理を行う．すなわち，図4の上段のテキストから抽出されたあるキーワードに関するストーリーの重要文が中段であった場合には，上段のテキスト中の3文目を文間に挿入することにより，下段のように連続する文に同一の名詞を含むようにする．本モジュールについての評価実験については，5.2.節で述べる．

4.3. サブキーワード抽出モジュール

本研究においては,サブキーワードとはテキストの広範囲に出現し，メインストーリーと関わりを持つ重要な意味を持った名詞であると定義する．この定義に基づくサブキーワード抽出アルゴリズムを以下に示す．

サブキーワード抽出アルゴリズム：

[STEP1] 以下の条件 1，3 を満たす名詞の集合 *Set1* と条件 2-1, 2-2, 3 を満たす名詞の集合 *Set2* を抽出する．

条件 1. テキストの 2 段落以上に主語として用いられている文が存在する．

条件 2-1. テキストの1段落以上において 4.1.節のメインキーワード抽出アルゴリズムの式(1), すなわち段落内において高頻度で出現し，テキスト中で重要な役割を果たしているという条件を満たす．

条件 2-2. テキストの2段落以上において 4.1.節のメインキーワード抽出アルゴリズムの式(2), すなわち段落中で偶然出現したのではなく，一定の意味をもって用いられているという条件を満たす．

条件 3. メインストーリーの重要文中に出現する．

[STEP2] *Set1* の各名詞 n にテキスト中で主語となる回数に基づいて，4.1.節のメインキーワード抽出アルゴリズムの式(3)により(0,1]の評価値 $E_{Set1}(n)$ を付与する．

Set2 の各名詞 n にテキスト中の出現頻度に基づいて 4.1.節のメインキーワード抽出アルゴリズムの式(4)により(0,1]の評価値 $E_{Set2}(n)$ を付与する．

[STEP3] テキストから $E_{Set1}(n)+E_{Set2}(n)$ で上位の名詞 n をサブキーワードとして抽出する．

上記のように,サブキーワード抽出アルゴリズムは 4.1.節のメインキーワード抽出アルゴリズム[STEP1] に，サブキーワードはメインストーリーと関わりを持つことを表す条件 3 を追加したものとなっている．

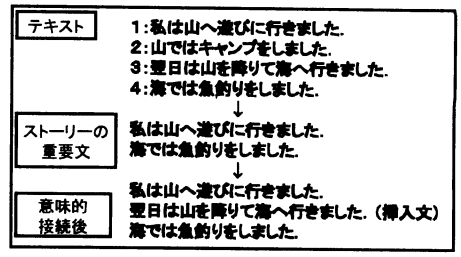

図4 意味的接続処理

表1 キーワードの正解セット

テキスト	メイン	サブ
浦島太郎	浦島	亀，乙姫，竜宮，玉手箱
桃太郎	桃太郎 鬼	お爺さん，お婆さん，桃，犬，猿，きじ，島，黍団子
猿蟹合戦	猿 蟹	おにぎり，柿，栗，蜂，昆布，臼，子
舌切り雀	お爺さん お婆さん 雀，葛篭	舌，宿，金銀珊瑚
母	母，死， 子供	老人，夜，コスモス，湖水，神様，老婆，花
親指姫	マイア	燕，嫁，お婆さん，穴，蛙，虫，人，王妃，王様，もぐら，お母さん，花，チューリップ

5. システムの評価実験

5.1. キーワード抽出モジュールの評価実験

4.1.節，4.3.節で述べたメインキーワード抽出モジュールとサブキーワード抽出モジュールの有効性を確認するために行った評価実験について以下で述べる．

実験テキストとしては「浦島太郎」,「桃太郎」,「猿蟹合戦」,「舌切り雀」,「母」,「親指姫」を採用した．各テキストのメインキーワード，サブキーワードの正解セットの作成は，まず被験者7名に各テキストの中で重要と考える名詞を選択してもらう．次に，その集合の中からテキストの主題と深く関係する名詞を選択してもらい，選択された集合の中で被験者の過半数に選択された名詞をメインキーワードの正解セットとする．その後，最初に選択された集合からメインキーワードの正解セットを除いた集合の中で被験者の過半数に選択された名詞をサブキーワードの正解セットとして採用した．表1に各テキストの正解セットを示す．提案システムとの比較システムとしては以下のSystemA～Fを採用した．

[比較システム]

SystemA： テキスト中の名詞を出現頻度で降順にソーティング

SystemB: 4.1.節のメインキーワード抽出アルゴリズム[STEP1]の条件 1 を満たす名詞をテキスト中で主語と

して出現した頻度で降順にソーティング

SystemC: 4.1.節[STEP1]の条件 2-1 を満たす名詞をテキスト中での出現頻度で降順にソーティング

SystemD: 4.1.節[STEP1]の条件 2-2 を満たす名詞を条件を満たした段落数で降順にソーティング

SystemE: 4.1.節[STEP1]の条件 2-1, 2-2 を共に満たす名詞をテキスト中での出現頻度で降順にソーティング

SystemF: 4.1.節のアルゴリズムによって抽出された名詞を $E_{Set1}(n)+E_{Set2}(n)$ によって降順にソーティング

各システムの比較は，正解セットの名詞と各システムによる評価値で上位の正解セットと同数の名詞を比較することにより行った．ただし，テキスト中のすべての名詞に異なった評価値が付与されるとは限らないため，システムの出力個数と正解セットの総数が常に等しくなるとは限らない．このようにして，各システムの適合率，再現率を算出した結果を表 2 に示す．ただし，表中の Proposal は提案システムを表す．また，提案システムによる各キーワードの抽出結果を表 3 に示す．この表 3 から，提案システムによりメインキーワード，サブキーワードとして適切な名詞が精度よく抽出されていることがわかる．しかし，表 1 と表 3 を比較すると「舌切り雀」，「母」における精度が低いことがわかる．この原因を考察する．「舌切り雀」の金銀珊瑚という名詞はテキスト中に 1 度しか出現しないが，物語の結論に関係する語であるため被験者により正解セットに選ばれている．また「母」の老人，コスモス，湖水という名詞は局所的に集中して出現するのみで，サブストーリーを構成するほど重要な語であるとは考えにくいが，テキスト中で物語の状況変化を引き起こす語であるため，被験者により重要な名詞として抽出されたと考えられる．これらの名詞は読み手には印象的な語ではあるが，正解セットであるサブストーリーを構成する語として適切かどうか疑問が残る．このような語を取得する手法をシステムに組み込むことは可能かもしれないが，結果としてノイズを多く拾ってしまいシステム全体としての精度の低下を招くと考えられるため有用ではないと考えられる．

表 2 から，メインキーワード抽出精度については SystemB 以外については差異がない．これは，メインキーワードは非常に重要な語であるため，各システムで採用しているすべての評価基準で高評価を得るためである．SystemB がメインキーワード抽出精度が低いのは，SystemB はテキスト中で主語となった語のみを評価対象としているが，テキスト中で重要なメインキーワードであっても，必ずしも主語として用いられるとは限らないためである．サブキーワード抽出精度を比較すると提案システムが適合率，再現率友に高い値を示しており，もっとも優れた手法であるといえる．以下本節では，4.4.節のサブキーワード抽出アルゴリズム[STEP1]の各条件の働きを表 2 より考察する．

表 2　各システムの適合率と再現率

	適合率 (メイン)	再現率 (メイン)	適合率 (サブ)	再現率 (サブ)
Proposal	92.3% (12/13)	92.3% (12/13)	73.7% (28/38)	66.7% (28/42)
SystemA	92.3% (12/13)	92.3% (12/13)	55.8% (29/52)	69.0% (29/42)
SystemB	76.9% (10/13)	76.9% (10/13)	57.1% (20/35)	47.6% (20/42)
SystemC	92.3% (12/13)	92.3% (12/13)	57.1% (24/42)	57.1% (24/42)
SystemD	92.3% (12/13)	92.3% (12/13)	58.3% (21/36)	50.0% (21/42)
SystemE	92.3% (12/13)	92.3% (12/13)	64.5% (20/31)	47.6% (20/42)
SystemF	—	—	61.9% (26/42)	61.9% (26/42)

表 3　提案システムによる各テキストからの抽出結果

テキスト	メイン	サブ
浦島太郎	浦島	亀，乙姫，竜宮，海，顔
桃太郎	桃太郎 鬼	お爺さん，お婆さん，舟， 犬，猿，きじ，島，黍団子
猿蟹合戦	猿，蟹	柿，栗，蜂，昆布，臼，子
舌切り雀	お爺さん お婆さん 雀，葛篭	宿
母	母，死，花	神様，老婆，木，子供，目
親指姫	マイア	燕，嫁，お婆さん，穴，蛙， 人，もぐら，お母さん，花， お家，葉，子，水

5.1.1. 条件 1 に関する考察

SystemE と SystemF の比較を行うと，SystemF が適合率では 2.6%低いが，再現率では 14.3%高くなっている．このように条件 1 を組み込むことにより再現率が大きく上昇している．これは，条件 2-1 かつ 2-2 を採用している SystemE では出現頻度のみを考慮しているため，出現頻度が低い名詞は評価対象とならないが，SystemF では条件 1 により出現頻度が低くても，文の主語となっている名詞は重要であると見なして評価対象とするためである．このように，条件 1 により頻度情報のみでは取れないサブキーワードが取れるようになる．

5.1.2. 条件 2-1 に関する考察

SystemD と SystemE の比較を行うと，SystemE が適合率では 6.2%高く，再現率では 2.4%低くなっている．このように条件 2-1 を組み込むことによって適合率が上昇するのは，条件 2-2 のみを採用している SystemD では，ある程度広い範囲に出現している名詞はすべてサブキーワードの候補としているため，重要でなくとも広い範囲に出現する一般的な語がノイズとして取ら

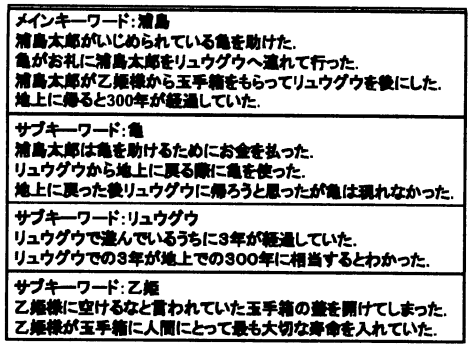

図5 浦島太郎のストーリー

れてしまうためである．したがって，出現範囲だけでなく局所的な出現頻度の高さも考慮に入れて絞り込んでいるSystemEの方が高い適合率を示している．

5.1.3. 条件2-2に関する考察

SystemCとSystemEの比較を行うと，SystemEが適合率では7.4%高く，再現率では9.5%低くなっている．このように条件2-2を加えることにより適合率が上昇している．条件2-1のみからなるSystemCではテキスト中での出現分布は考慮しないため，テキスト中で重要な意味は持っていないが局所的な出現頻度のみが高いノイズを多数取ってしまう．そのため，「複数の段落において複数回出現しなくてはならない」という出現分布を考慮する条件2-2を組み込んだSystemEの適合率が高くなっている．一方で，SystemEでは再現率が10%近く低下している．この低下は5.1.1.項で述べた条件1を組み込むことによって提案システム全体としてはある程度回避できている．すなわち，局所的に頻度が高く他の段落では1回しか出現しないため，条件2-2により広範囲に渡って出現していると判断されない名詞であっても，その1回が文の主語となっていれば条件1により評価対象となるためである．

5.1.4. 条件3に関する考察

ProposalとSystemFの比較を行うと，Proposalが適合率では11.8%高く，再現率でも4.8%高くなっている．このように条件3を組み込むことによって適合率が大きく上昇している．これは，条件3によってサブキーワードとして抽出する名詞を，メインストーリーに関連するものに限定したためである．このことは，本研究の「サブキーワードはテキストのメイントピックと関連がある」という定義の妥当性を示している．

5.2. メインストーリーとサブストーリー抽出実験

4.2.節で記したストーリー抽出モジュールによるメインストーリーとサブストーリー抽出実験について述べる．実験テキストとしては「浦島太郎」を採用した．メインキーワード抽出モジュールにより，浦島太郎のメインキーワードであると判断された「浦島」に関するメインストーリーをストーリー抽出モジュールによって抽出した結果からは，図5に示すように浦島太郎という話の概要を理解することが可能であった．一方，サブキーワード抽出モジュールによってサブキーワードであると判断された「亀」，「リュウグウ」，「乙姫」からは各サブキーワードに関した新たな知見を得ることができた．例えば，浦島が地上に戻ると300年経過していた理由は，「リュウグウ」に関するサブストーリーから「リュウグウでの3年が地上での300年に相当する」ためであることがわかる．このように，テキストのストーリーにサブストーリーを組み込むことにより，メイントピックと大きな関わりを持ったサブキーワードについての理解が促進され，結果としてテキストの内容をより的確に捉えることができる．

6. おわりに

本稿では，テキストのストーリーはメインストーリーとサブストーリーの組み合わせから構成されるという考えに基づく，ストーリーモデルを提案した．また，本モデルの基礎を成すテキストのメインキーワードとサブキーワードの抽出実験を行い，提案システムの有効性を確認した．今後の課題としては，本システムによるメインストーリーとサブストーリー抽出の効果を確認する評価実験を行うこと，メインストーリーとサブストーリーの組み合わせ方の考案が挙げられる．

文　献

[1] H.P. Luhn, “The automatic creation of literature abstracts,” In IBM Journal for Research and Development, Vol.2, No.2, pp.59-165, 1958.

[2] G. Salton and C. Buckey, “Term-weighting approaches in automatic text retrieval,” Readings in Information Retrieval, pp.323-328, 1997.

[3] J.J. Morris and G. Hirst, “Lexical cohesion computed by thesaural relations as an indicator of the structure of text,” Computational Linguistics, Vol.17, No.1, pp.21-48, 1991.

[4] 亀田正之，“段落間および文間関連度を利用した段落シフト法に基づく重要文抽出，” 情報処理学会自然言語処理研究会資料，99-NL-121，vol..97, no.85, pp.119-126, 1997.

[5] 任福継，定永靖史，“統計情報と文章構造に基づく重要文の自動抽出，” 情報処理学会技術研究報告NL125, Vol.98, No.48, pp.71-78, 1998.

[6] 市丸夏樹，飛松宏征，日高達，“話題の流れを保持する自動要約，” 第160回情報処理学会自然言語処理研究会資料，pp.43-48, 2004.

[7] 砂山渡，谷内田正彦，“観点に基づいて重要文を抽出する展望台システムとそのサーチエンジンへの実装，” 人工知能学会論文誌，Vol.17, No.1, pp14-22,2002.

[8] Hearst M., “Text Tiling:Segmenting Text into Multi-paragraph Subtopic Passages,” Association for Conputational Linguistics, Vol.23, No.1, pp33-64, 1997.