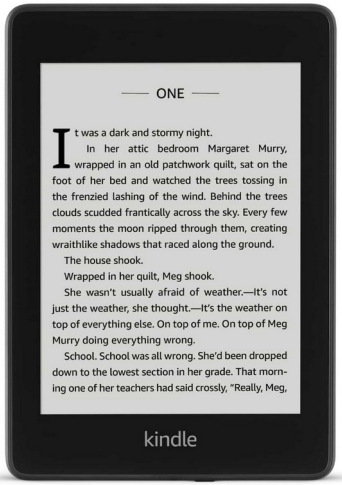
物語内の人物と場所情報の時系列可視化による読書支援

立命館大学情報理工研究科　立命館大学情報理工学部
MA JIAXIU　西原 陽子　山西 良典

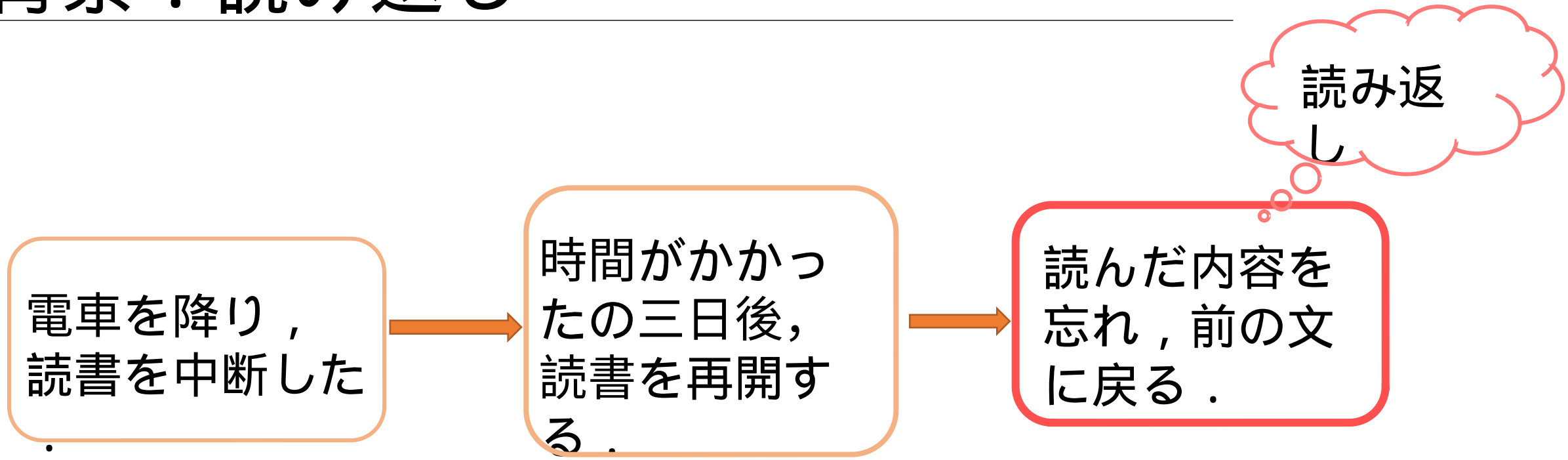
背景：電子書籍

電子書籍の利用率が高い

- ★ 「いつも読むことができる」
- ★ 「場所を取れない」
- ★ 「一つの端末に複数の書籍を入れることができる」

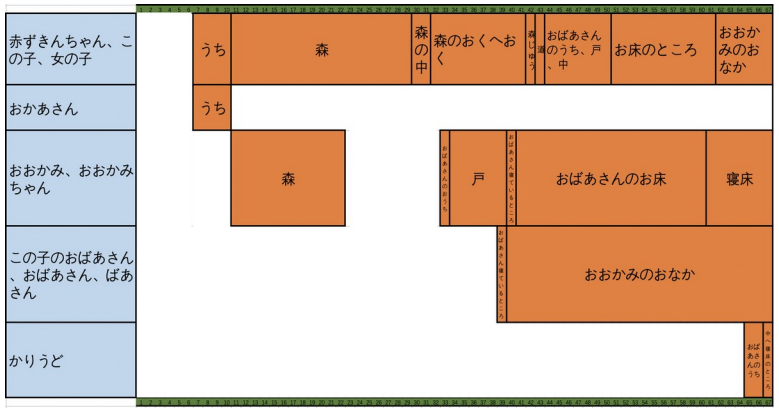


背景：読み返し



読み返すことは時間が無駄になる．
忘れたまま読み進めると，書籍の内容を把握しにくくなる

研究目的：読書支援



- 既読部分の人物と場所の情報が可視化され，どの人物が誰とどこに居る情報が分かる．
- 「人物をしたこと」と「同じ場面で共起したこと」を提示することもある．

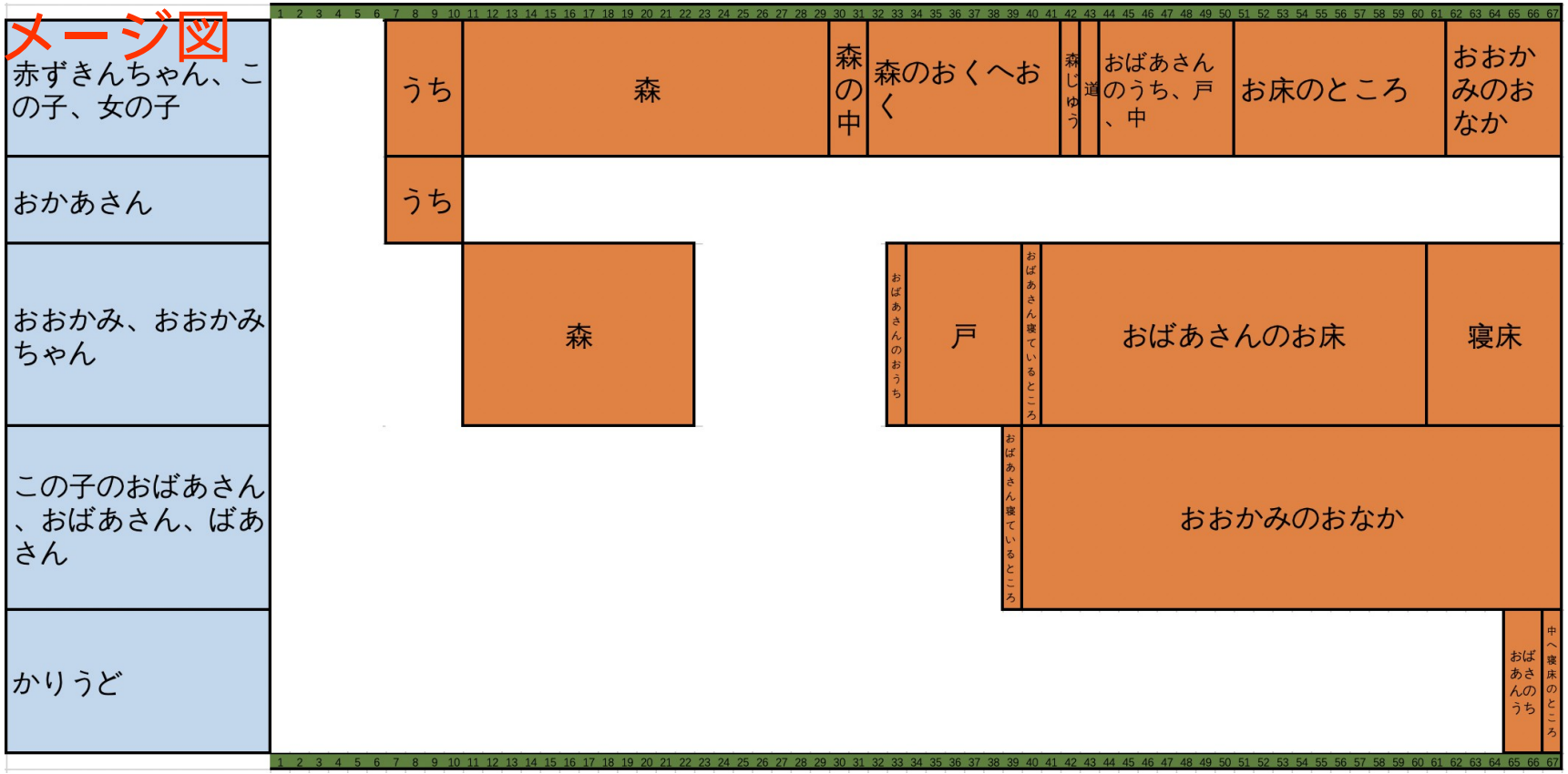
読書再開前にながめると，読んだ内容を思い出すことが容易になる．

読み返しは不要になり，時間が有効に使える．

物語の内容を把握しやすくなる．

可視化（イメージ図）

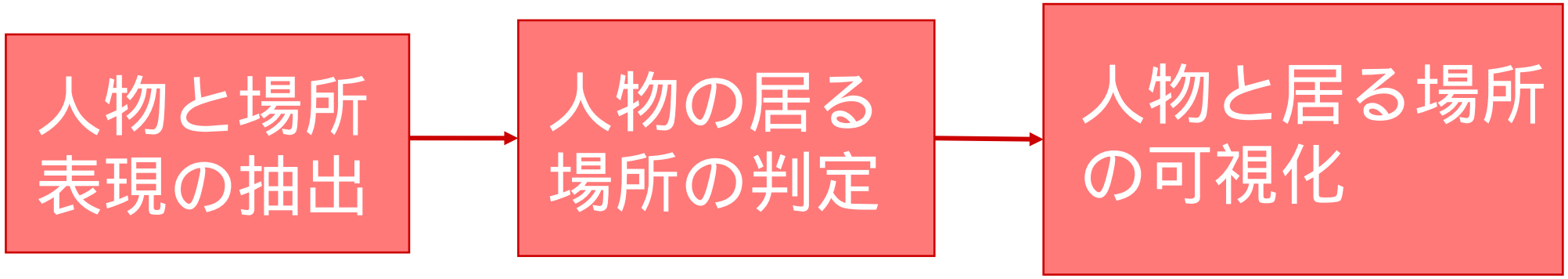
「赤ずきんちゃん」冒頭から 67 文目までの可視化イメージ図



- 横軸（緑色の長方形）
  「文の番号」
- 縦軸（水色の長方形）
  「人物」
- 内部（オレンジ色の長方形）
  「人物が居る場所」

一目でどの人物が誰とどこに居たのかが分かることを目的とする．

可視化手法



書籍の中の物語を研究の対象とする

- 物語の既読部分から文ごとに人物と場所情報を抽出する
- 各文において人物と人物が居る場所をひもづける
- 1文ずつに人物と場所情報を可視化する

可視化に必要な技術要素

(1) 物語のテキストから、人物と場所を表す単語の抽出

(2) 異なる表記であるが、同一人物を表す単語をまとめる
例「赤ずきんちゃん」「この子」「女の子」は同一人物を指す

(3) 人物がいる場所の推定
「赤ずきんちゃんはおばあさんの家に出かけました。
おばあさんは森の奥に住んでいます。
赤ずきんちゃんが森の入り口に着きました。」
赤ずきんちゃんは森の入り口にいる
おばあさんは森の奥にいる

本発表では既存技術の利用により、(1)を達成可能なことを報告

人物と場所表現の抽出

全体で最適な固有表現のためのタグ付けを行う手法 CRF を用いた，
人物と場所表現を抽出する．

CRF を用いた固有表現抽出器の作成手順．

1．物語のテキストを文ごとに分割する．
2．各文に対して形態素解析を行い，単語と品詞情報を得る．
3. CRF で学習を行うために，単語に対し固有表現抽出のためのタグを付与する．
4. CRF を用いて学習を行う．１つの単語に対し，自分自身と前後３つの単語，および品詞情報を学習し，固有表現抽出器を得る．

文ごとに分割し、形態素解析をする

- 文末の句読点（ ．や。 ）
- 発話終了の鍵括弧など記号があれば文末と判定し，物語テキストの文への分割を行う

- 形態素解析器は MeCab 、辞書は NEologd を用いる
- 得られる形態素と品詞の細分類を CRF への入力に用いる

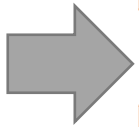
CRF とは

Conditional Random Field という、系列ラベリング問題を解くための手法
入力としてデータの列を与えると、データに対しラベルが付与される

例
入力：私　は　あかずきん　です（データの列）
出力：名詞　助詞　名詞　助詞（ラベルの列）

固有表現抽出のためのタグ

表現タグ	説明
B-CHAR	人物表現の先頭の形態素
I-CHAR	人物表現内の先頭以外の形態素
B-POS	場所表現の先頭の形態素
I-POS	場所表現内の先頭以外の形態素
O	人物と場所表現以外



赤ずきん	B-CHAR
ちゃん	I-CHAR
、	O
おばあさん	B-POS
の	I-POS
ところ	I-POS

「赤ずきんちゃん」の BIO2 タグ結果例

ＣＲＦを用いた人物名場所名のラベルの学習

位置	入力単語	品詞細分類	表現タグ
i-3	これ	名詞-代名詞　--一般	O
i-2	を	助詞-格助詞--一般	O
i-1	赤ずきん	名詞-固有名詞--一般	B-CHAR
i	ちゃん	名詞-接尾-人名	I-CHAR
i+1	、	記号-読点	O
i+2	ここ	名詞-代名詞--一般	O
i+3	に	助詞-格助詞--一般	O

CRF を用いて学習を行う際は，ｉ 番目の形態素を中心として，前後３つの形態素をあわせた合計７つの形態素に対し，入力単語，品詞細分類，表現タグを用いる．

評価実験

使用データ：５つ青空文庫の物語

- 日本人に読まれることが多い（物語タイトル１から４）
- 青空文庫の著者名順の先頭の物語（物語タイトル５）

	物語タイトル
1	赤ずきんちゃん
2	浦島太郎
3	桃太郎
4	猿蟹合戦
5	良夜

適合率と再現率：

- 適合率：正と予測したデータのうち，実際に正であるものの割合
- 再現率：実際に正であるもののうち，正であると予測されたものの割合
- 適合率と再現率は B−CHAR ， I−CHAR ， B−POS ， I−POS ごとに算出した

適合率と再現率

人物と場所表現の抽出の適合率と再現率

	B-CHAR		I-CHAR		B-POS		I-POS	
物語	適合率	再現率	適合率	再現率	適合率	再現率	適合率	再現率
赤ずきんちゃん	0.94	0.90	1.00	0.96	1.00	0.90	0.99	0.97
浦島太郎	0.97	0.98	0.94	0.91	0.98	0.79	0.96	0.97
桃太郎	0.96	0.97	0.95	1.00	1.00	0.89	1.00	1.00
猿蟹合戦	0.98	0.98	1.00	1.00	1.00	0.87	0.97	0.97
良夜	0.55	0.22	1.00	0.68	0.78	0.51	0.78	0.94
平均	0.88	0.81	0.98	0.91	0.95	0.79	0.94	0.97

適合率の平均は 0.88 から 0.98 になった
再現率の平均は 0.79 から 0.97 になった
CRF を用いることで、人物名と場所名の抽出が可能とわかった

実験結果

	B-CHAR		I-CHAR		B-POS		I-POS	
物語	適合率	再現率	適合率	再現率	適合率	再現率	適合率	再現率
赤ずきんちゃん	0.94	0.90	1.00	0.96	1.00	0.90	0.99	0.97
浦島太郎	0.97	0.98	0.94	0.91	0.98	0.79	0.96	0.97
桃太郎	0.96	0.97	0.95	1.00	1.00	0.89	1.00	1.00
猿蟹合戦	0.98	0.98	1.00	1.00	1.00	0.87	0.97	0.97
良夜	0.55	0.22	1.00	0.68	0.78	0.51	0.78	0.94
平均	0.88	0.81	0.98	0.91	0.95	0.79	0.94	0.97

- CHAR タグ，POS タグのいずれにおいても，B タグの方が I タグよりも抽出における適合率と再現率が低かった．
- 実験に用いた物語のテキストのうち 1 から 4 は適合率と再現率は高かったが，5 のテキストは適合率と再現率が低かった

考察

失敗した例

物語	人物・場所	人手で付けたタグ	手法で付けたタグ
1. 赤ずきんちゃん	おばあさん おばあさん｜の｜着物 森｜じゅう｜かけまわっ	B-CHAR O｜O｜O B-POS｜I-POS｜O	O B-CHAR｜O｜O B-POS｜I-POS｜I-POS
2. 浦島太郎	浜｜べ 海 かめの子	B-POS｜I-POS B-POS B-CHAR	O｜O O O
3. 桃太郎	川 帰り 陸	B-POS O B-POS	O B-CHAR O
4. 猿蟹合戦	山道 かに 山｜へ	B-POS B-CHAR B-POS｜O	O O B-POS｜I-POS
5. 良夜	父 新潟｜県｜下 伯父 猿 母 東京	B-CHAR B-POS｜I-POS｜O B-CHAR O B-CHAR B-POS	O B-POS｜I-POS｜I-POS O B-CHAR O O

- 1 から 4 のテキストは子供向けの物語であった．
- 5 のテキストは「良夜」という日本の物語で，対象とする読者の年齢層は低くない．

- 「良夜」の登場人物は「伯父」，「父」が多く，これらの抽出に失敗することが多かった．

人物名が具体的な名称（「赤ずきんちゃん」「浦島太郎」「桃太郎」など）ではないことがあり，人物名と認識がされにくかったと考えられる．

まとめ

- まとめ

　　人物と人物が居る場所の情報に着目し読書支援を実現するための手法を提案した

　　人物と場所表現の抽出部分について評価実験を行い，適合率と再現率により評価を行った

- 結果

　　適合率の平均は 0.88 から 0.98 になった

　　再現率の平均は 0.79 から 0.97 になった

　　CRF を用いることで、人物名と場所名の抽出が可能とわかった

　　子供向けの物語の方が人物，場所表現の抽出が容易であることがわかった．

- 今後の課題

　　同一人物の表記をまとめる

　　人物がいる場所の推定