文献紹介

自然言語処理研究室
後藤　大明

出典

- 整合性を考慮した物語要約システムの構築
- 横野　光,自然言語処理，Vol.15，No.5，pp. 45-71（2008）

はじめに

- 要約･･･ユーザが自分の望む情報を手早く手に入れるための要素技術

- 文書構造を利用した要約手法が提案され、一定の成果が上げられている。

背景

- 青空文庫のように著作権の切れた文学作品
- 原文書の代わりとして機能する要約（報知的要約）が必要

- 論説文・・・著者の主張
- 物語・・・全体の流れや他の箇所との関係

- 新聞記事、論説文を対象とした要約は物語の要約には適さない

目的

- 物語の要約は重要な要素を含むだけでなく要約中の整合性まで考慮しなければ十分に機能しない

- 話題のつながりに焦点を置いた物語要約システムを構築

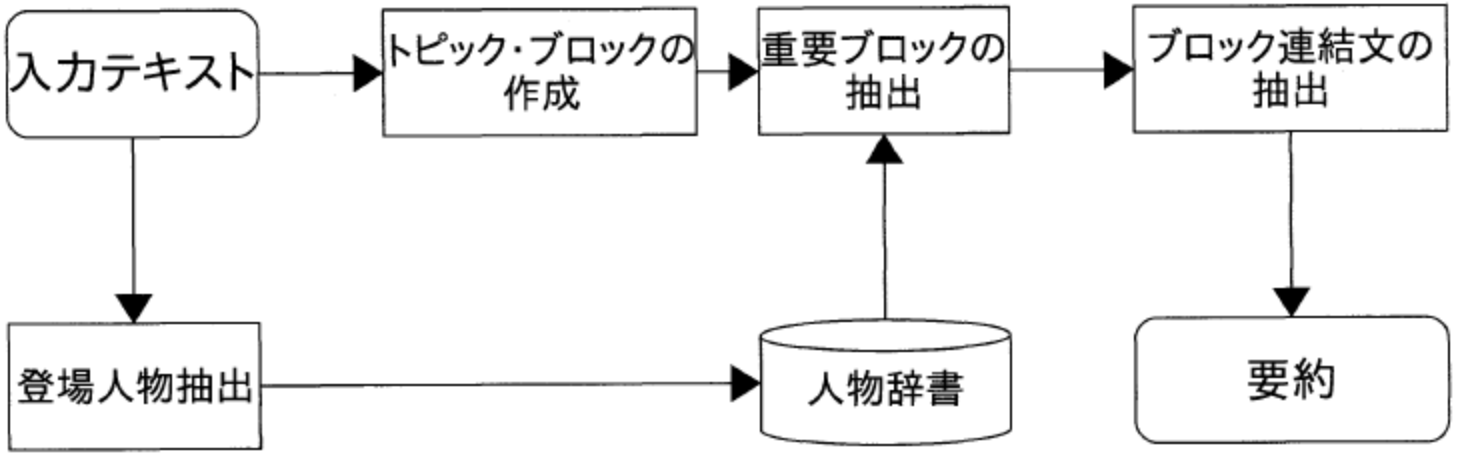
整合性を考慮した要約文の抽出

- 整合性を保つためには話題間のつながりを扱うことが重要

- 話題が一貫した箇所（トピック・ブロック）を単位として抽出
- ブロック間の整合性を補完する文（ブロック連結文）

整合性を考慮した要約文の抽出

- トピック・ブロックと登場人物を抽出
- 人物を軸としたブロックの重要度を計算
- ブロック連結文の抽出



登場人物抽出

- 登場人物抽出タスクの問題点
- 1）固有表現のものが多く、一般の辞書に登録されていない

- 2）人名でないものが人物化することがある

  「さるかに合戦」に登場する「うす」など

登場人物抽出

- 文字n-gramを利用して頻度の高い部分文字列を獲得
- 部分文字列に対してMeCabによる形態素解析を行い以下のフィルタリングを行う
  - 平仮名、カタカナの小文字から始まっていない
  - 句読点、カギ括弧、記号を含まない
  - 接続詞、連体詞、代名詞を含まない
  - 複数の形態素で構成されている
  - 文章中に“は”を接続した形で出現する

登場人物抽出

- CaboChaによる係り受け解析を行い、副助詞“は”を後接して複数登場する名詞の中で、述部となる動詞に意志性にものがあれば登場人物

- nの範囲は2~10、再現率は0.74、適合率は0.68

トピック･ブロックの抽出

- 1分単位の主題を仮定してそれが連続する文をまとめることでトピック･ブロックを作成

- 小さなトピック･ブロックが多く作られることを防ぐためウインドウ幅を前2文
- 別の主題を間に挟んで同じ主題を持つ文は同じトピック･ブロックに所属するとした

文	s_0	s_1	s_2	s_3	s_4
主題	a	b	a	c	b

$$P_0 = \{s_0, s_1, s_2\}$$
$$P_1 = \{s_1\}$$
$$P_2 = \{s_3\}$$
$$P_3 = \{s_4\}$$

- s=文
- P＝トピック･ブロック

トピック･ブロックの抽出

- 主題は主語が有力な候補、副助詞“は”を後接して登場する名詞、無い場合助詞“が”を主題とする
- 主題は省略される場合があるため、上記の物がなければ全文の主題をその文の主題とする
- 会話文を考慮せず地の文のみを対象とする

トピック･ブロックの抽出

- 代名詞の場合
- 1つ前の文中の名詞を先行詞候補とし、助詞の優先度に従って決定する

- は＞が＞に＞を＞その他

するとどこからやって来たか、突然彼の前へ足を止めた、片目眇の老人があります。
それが夕日の光を浴びて、大きな影を門へ落すと、じっと杜子春の顔を見ながら、「お前は何を考えているのだ」と、横柄に声をかけました。

トピック･ブロックの抽出

- 代名詞に対して出力された先行詞を、人手で判断した結果、精度0.41（134/330）

重要ブロックの計算

- 登場人物が行動を起こし、状況が変化する箇所が重要

$$Score_{sentence}(s) = \sum_{w \in s} (Terms(w) \cdot tw(w,s)) \cdot weight(s)$$

$$Terms(w) = \begin{cases} \text{文書 } D \text{ における単語 } w \text{ の出現頻度} & (w \text{ が登場人物表現}) \\ 0 & (\text{上記以外}) \end{cases}$$

$$tw(w,s) = \begin{cases} 1.2 & (\text{単語 } w \text{ が文 } s \text{ の主題}) \\ 1 & (\text{上記以外}) \end{cases}$$

s=文、w=単語

$$weight(s) = \begin{cases} 1.2 & (\text{文 } s \text{ 中の動詞の意味に“主体の変化”が含まれている}) \\ 1 & (\text{上記以外}) \end{cases}$$

重要ブロックの計算

- トピックブロックの重要度

$$Score_{tb}(P) = \frac{\sum_{s_i \in P} Score_{sentence}(s_i)}{|P|}$$

- Pはトピック・ブロック、|P|はP中の文の数

重要ブロックの抽出

- 登場人物表現が最初に出現する文はその登場人物の説明
- 最後に出現する文はその登場人物の採集状況を説明している

- トピック・ブロックを選択する前に上記を抽出し、残りを重要度順に要約率分だけ抽出

重要ブロックの抽出

- 不要要素（繰り返し同じことを述べている文、補足説明）を除くため接続詞と文末表現から連接関係を推定

種類	説明
展開	前の文の内容を後の文で展開する
反対	前の文の内容に対し，後の文で反対の事柄を述べる
累加	前の文の内容に，後の文で内容を付け加える
同格	前の文と同じ内容を後の文で言い換える
補足	前の文の内容に対し，後の文で説明を補う
対比	前の文の内容に後の文の内容を対比させる
転換	前の文の内容に対し，後の文で話題を変える

- “補足”“同格”のものはその後文を削除

連接関係	接続詞	文末表現
展開	すると，そこで，それで，次に	
反対	だが，だけど，しかし，一方	
同格	すなわち，たとえば，つまり	
補足	もっとも，実は，但し	～のだ

ブロック連結文の抽出

- ブロック間の整合性を保つための状況の変化を示す文を挿入

- 重要ブロック間に存在するトピック・ブロックの登場人物の出現を基にした局所的な重要度を計算

ブロック連結文の抽出

- 隣接する重要ブロックにおいて片方に出現し、もう片方に出現しない登場人物があるときその人物は状況の変化に関わっている

ブロック連結文の抽出

$$Terms_{local}(t, P_i, P_{i+1}) = \sum_{s_j \leq s \leq s_n} freq(t, s) \cdot lw(t, P_i, P_{i+1})$$

$$lw(t, P_i, P_{i+1}) = \left\{ \begin{array}{ll} 1 & \text{単語 t が } IP_i \text{ で出現し，かつ，重要ブロック} \\ & P_i,\ P_{i+1} \text{ のいずれか一方にのみ出現する} \\ 0.5 & \text{単語 t が重要ブロック } P_i,\ P_{i+1} \text{ と } IP_i \text{ の全て} \\ & \text{で出現する} \\ 0 & \text{上記以外} \end{array} \right\}$$

- freq(t,s)は文sにおける登場人物表現tの出現頻度、IPはP_iからP_{i+1}の間の文の集合

ブロック連結文の抽出

- P_iからP_{i+1}の間の文sの局所的重要度

$$Score_{local}(s_i, P_i, P_{i+1}) = \sum_{t \in s_i} Terms_{local}(t, P_i, P_{i+1}) \cdot weight(s_i)$$

$$weight(s) = \begin{cases} 1.2 & (\text{文 } s \text{ 中の動詞の意味に“主体の変化”が含まれている}) \\ 1 & (\text{上記以外}) \end{cases}$$

ブロック連結文抽出

- ブロック連結文の量CS（IP_k）

$$CS(IP_k) = \frac{(1-\alpha)x}{1-\alpha x} \cdot |IP_k|$$

- 要約率xは

$$\text{要約率} = \frac{\text{要約文書の文字数}}{\text{原文書の文字数}}$$

- で与えられ、αを大きくすると限られた要約率の中で重要ブロックを多く抽出し、小さくすると整合性を重視して連結文を多くとりだすことになる。

評価

- 被験者10人に提案手法か比較手法（tf･idf法）にどちらかの要約を知らせずに読ませあらすじを自由筆記させる

- 原文書を読んであらすじを自分で修正する
- 修正は以下の3種類
  - 訂正（1文中の一部を追加削除）
  - 追加（1文を追加）
  - 削除（1分を削除）

評価

- 整合性の評価を行うためどちらの文が読みやすかったか判定

作品名	提案手法	tf·idf 法
杜子春	4	**6**
鼻	**8**	2
風の又三郎	1	**9**
セロ弾きのゴーシュ	5	5
名人伝	**8**	2
雪の夜	5	5
桜の森の満開の下	3	**7**
山椒大夫	**9**	1
俊寛	**7**	3
平均	**5.6**	4.4

考察

- 比較手法は分単位で抽出するため細かな欠損が多く被験者に違和感を与える事が多いため、提案手法がよい評価となった

- 「風の又三郎」の評価が低い原因として会話文の中に物語の展開に関わる表現が入っている

評価

- 内容理解の評価

$$\text{評価値} = \frac{(\text{訂正数} \cdot 1 + \text{追加数} \cdot 2 + \text{削除数} \cdot 0.5)}{\text{被験者が作業1で答えたあらすじの文数}}$$

作品名	あらすじの文数（平均）		評価値（平均）	
	提案手法	tf·idf 法	提案手法	tf·idf 法
杜子春	**42.25**	22.63	**0.93**	1.04
鼻	**21.13**	14.5	**0.57**	1.17
風の又三郎	**58.25**	31.13	1.37	**0.75**
セロ弾きのゴーシュ	**48.38**	26.75	**0.87**	1.38
名人伝	**29.75**	21.43	1.05	**0.48**
雪の夜	**48.88**	19.5	0.77	**0.69**
桜の森の満開の下	**53.88**	27.25	**0.75**	1.00
山椒大夫	**56.75**	23.50	**0.84**	1.42
俊寛	**51.25**	22.38	0.38	**0.28**
平均			**0.84**	0.91
分散			**0.07**	0.14

考察

- 提案手法は重要個所で正しく重要ブロックと抽出できなかった場合、多くの修正が必要となる
- 比較手法は全ては抽出できなくともある程度重要な文を抽出できるため提案手法に比べ修正が少なくなる

- 一般的な語が重要な単語であるとき比較手法では重要語と見なせないため不安定になる
- 分散を見ると比較手法のほうが大きい

おわりに

- 本手法は特に登場人物の移り変わりに着目した
- 抽出した箇所間に移り変わりに関する文を挿入することで整合性のある読みやすい要約を獲得した

- 課題として会話文、より詳細な話題間のつながりを計算するため場所の移動や時間経過などの場面転換も考慮にいれる必要がある