イベントの参照関係に注目した新聞記事の複数文書要約

○吉岡 真治　　原口 誠
北海道大学大学院 工学研究科

背景と目的

- 背景
  - 複数の新聞記事からの要約作成
    - 対象
      - 特定の事象に関する一連の報道記事
    - 目標
      - 全体の記事の流れから要約を作成
    - 特徴
      - 個別の記事の要約ではない、冗長な表現の存在
      - 重要な事象は、異なる記事で複数回参照される
- 目的
  - イベント（発生した日付を特定した事象）の参照関係に注目した複数文書要約手法の提案

新聞記事におけるイベント

- 対象とする一連の記事
  - ある特定の事象（事件や事故など）に関する続報などを単位とした記事群
  - 一つの記事だけではなく、複数の記事で参照される。
- 新聞報道の特徴：日付の重要性
  - 会社による定例会見、地震の発生
    - 語彙的にはほとんど同じ事象であるが、日付が異なる事象は、必ず異なる事象
    - 事象の細かな属性を取り出すよりも容易に異なるイベントを分離可能
- イベント
  ＝　特定の時期に起きた特定の事象

イベントを構成する情報

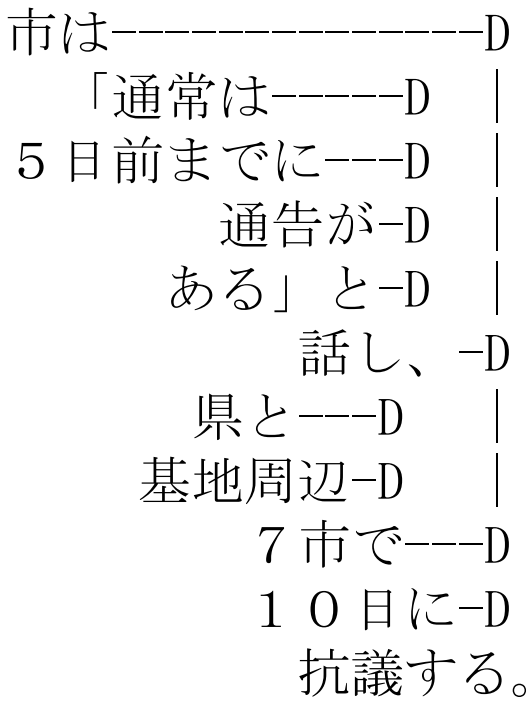
- Root：イベントを表す中心となる語
- 修飾語：イベントの内容を修飾する語
  - 修飾するタイプを、助詞などによって分類
- Date：イベントを特徴づける日付が文から獲得できる場合にはその情報を記載
- ArticleDate：記事が発表された日時
- Negative：否定文か否かをあらわすFlag
- Depth：構文解析木のRootの要素へかかるまでに、どれだけの要素を経由するかを記載する。
- Chunks：イベントに対応する語の文中の位置のリスト

新聞記事からのイベント抽出

1. Cabochaにより各文の構文解析を行い、係り受けの関係を抽出
2. 係り受けの関係がある動詞と名詞を、イベントのRootになる候補として抽出
3. イベントのRootにかかる語について、語の品詞やの後接している助詞の情報に注目して、修飾語をタイプ毎に分類
   - 修飾語には、直接かかっている語だけではなく、そのかかりに直接関係する全ての語を含める。
4. 文章中に日付の情報が入っている場合には、イベントを特徴づける日付に設定

イベント抽出の例

- 市は「通常は５日前までに通告がある」と話し、県と基地周辺７市で１０日に抗議する。(毎日新聞1998年1月10日の記事より)



Root ある
Depth 2
Subject 通常,通告
Date
ArticleDate 980110
Chunks 4,3,2,1
助詞-に ５,日,前

Root 話す
Depth 1
Subject
Date
ArticleDate 980110
Chunks 5,4,2,1
助詞-と 通常,５,日,前,ある

Root 抗議
Depth 0
Subject 市
Date １０日
ArticleDate 980110
Chunks 9,8,7,6,10,0
助詞-で 県,基地,周辺,７,市
助詞-に １０,日

PageRankアルゴリズムに基づく重要文抽出

- 語の参照関係に基づく単一文書要約
  - PageRankアルゴリズムの適用
    - PageRankアルゴリズム
      - Webページの有効性をリンク構造に基づいて求める手法
      - ユーザが文書中のリンクをランダムにクリックした(ランダムウォークをした)と仮定し、収束状態での各々のページへの滞留確率を評価値として利用

$$\vec{r}_{i+1} = M \times \vec{r}_i$$

Mは遷移確率行列、$\vec{r}_i$ は各ページの重要度

  - 語の参照関係に基づく各文の重要性判定
    - 同一の語を含む文書間に参照リンクがあると考える
    - 多くの文で共起している語が多い
      → リンクが多い → 重要な文
    - 高頻度語の影響の排除：DF（Document Frequency)の利用などによるリンクの重み付けの調整

PageRankアルゴリズムに基づく重要文抽出の複数文書への拡張

- 2つのアプローチ
  - 各文書内での文の重要度と文書の重要度の階層的計算
  - 複数の文書をひとまとめとした文書を仮定し、重要度計算
- 新聞記事の特性
  - 続報などにおいて、以前に起きた重要なイベント(事件の発生)についての参照が記事中で複数回あるとは限らない
  - このようなイベントの重要度を表すためには、後者のアプローチが適切

イベント間の参照関係の取り扱い

- 語とイベントの対応関係
  - 単一文書：共通する語の存在　→　リンク
  - 複数文書：共通するイベントの存在　→　リンク
- イベント判定の問題
  - 異なる表記（受動態と能動態）の問題
  - 共通する語の存在によるリンクも考慮
- 文書中の文の出現位置に関する情報の利用
  - 重要度の初期ベクトル $\vec{v}$ をバイアスとして利用 (Topic-sensitive PageRank)
  - 重要な文に関しては、重要度を補償し、その重要度をリンクにより分配（α：初期ベクトルのバイアスの強さを決めるパラメータ）

$$\vec{r}_{i+1} = (1-\alpha) \times M \times \vec{r}_i + \alpha \times \vec{v}$$

重要文抽出における抽出文の並べ替え

- 基本スタンス
  - 時系列に基づく文書の並び替え
    - 古い記事に含まれる文から新しい記事に含まれる文へと並び替え
  - 類似文の発見に基づく抽出文の並べ替え
    - 各記事中での文の順序に基づく文の並び替え
      - 同一記事から文を選んだ場合には、その記事中の順序を保存
      - 記事中で選択された文よりも前に存在する文に類似している文がある場合には、その類似している文の後に、選択された文を配置
      - 文の前後関係を決める情報がない場合や、たすき掛けのようになって順序が決まらない場合には、最初に文書を並べた順序で文を配置

イベントの同一性に基づく要約文の圧縮

- PageRankアルゴリズムに基づく重要文抽出
  - 同じようなイベントを記述した文は同じような重要度を持つ
    → 冗長な文の削除による要約文の圧縮が必要
  - 同一イベントについての冗長な記載を削減
    - 語彙重なりではなく、イベント重なりで冗長な文の判定
  - 文中から冗長な説明の除去
    - 名詞を修飾するようなイベント記述が冗長な場合には、文中から記述を削除し、文の長さを圧縮

要約文作成における文の圧縮と並べ替え

- 基本スタンス
  - 時系列に基づく文書の並び替え
    - 基本的な文の並べ方、重要文抽出と同じ
    - 冗長な記述は2回目以降削除
  - 各記事中での文の順序に基づく文の並び替え
    - 基本的な文の並べ方、重要文抽出と同じ
    - 追加する文に対する操作
      - 追加を検討する文に含まれるイベントの内、その文より前に追加する文中に含まれているイベントに対応する文中の要素を、次の基準に基づき削除
        - » 残すことが決まっているイベントにかかっている文中の要素については、一つ以上の内容語(名詞など)を残す
        - » 削除した要素に依存関係を持つ語の要素を削除

要約実験(TSC-3)

- 複数新聞記事(毎日新聞と読売新聞の記事)からの要約作成
  - 30セットのタイトル、記事群、想定質問から重要文抽出、要約文作成のタスクを実行
- イベントの利用に関する有効性を検証
  - イベントの情報を利用した重み付けと語の情報を利用した重み付けの比較
    - イベントの情報のみによるリンク
    - 語の情報のみによるリンク
    - イベントと語の情報を組み合わせたリンク

重要文抽出の評価

- 類似文を含む要約に対する評価
  - 正解データ
    - 全体として選びたい類似している文の集合の集合
  - Coverage：類似していない重要な文をどれくらい網羅的に選んでいるか？（冗長性を考慮）
  - Precision：重要な文をどれくらい選んでいるか（冗長性の考慮なし）

	イベント	語とイベント	語
Coverage	0.309	0.325	0.328
Precision	0.523	0.570	0.557

要約文作成の評価

- 主観評価
  - 読みやすさの評価（おおむね良好の成績）
    - 平均よりも良い項目
      - 時系列の関係が矛盾してないか？
      - 同一の、あるいはほぼ重複する文はいくつあるか？
      - 同一事物を参照する表現の一貫性という観点から修正すべき表現はいくつあるか
      - 先行詞のない指示表現はいくつあるか？
    - 平均よりも悪い項目
      - (ゼロ)代名詞化、指示表現化すべき箇所はいくつあるか？
      - 不適切な格要素の重複はいくつあるか？

	Long	Short
主観評価（内容のCoverage)	0.247	0.207

実験結果の考察

- 全体としての性能は、参加者グループ中、中の上
- 本システムの特徴
  - イベントの情報を使った重複文の判定や前後関係の判定はうまく機能している。
  - アルゴリズムの性質上、複数のイベント、複数の語を含む長い文に良い評価を与えやすく、要約文作成のShortなどの時にCoverageが下がっている。
  - 照応関係を扱っていないために、全般に指示語が少ない。
- リンク作成時には、イベント情報の利用が、現時点では、それほど効果的ではない。
  - 理由
    - イベント同定の問題
    - 語によるリンクで十分な場合が多い
  - イベント情報のリンクの利用
    - スパースなリンク構造をうまく利用するマクロな重要度決定のアルゴリズムが必要

結論と展望

- 結論
  - イベントの参照関係に基づく複数文書要約の方法を提案
    - 基本的な性能としては、悪くないと思われるので、より一層の洗練化が必要
- 展望
  - パラメータチューニングによる限界の確認
  - スパースなイベント情報のリンクをうまく利用するマクロな重要度決定のアルゴリズムの導入