複数記事要約のためのサマリパッセージの抽出

橋本 力

北陸先端科学技術大学院大学 情報科学研究科

2001 年 2 月 15 日

キーワード: 複数記事要約, サマリパッセージ.

本研究では, 新聞記事コーパスを検索した結果の記事集合を対象に, それらの記事集合の中からサマリパッセージを抽出する方法と, 抽出されたサマリパッセージを用いて記事集合を要約する方法について述べる. 新聞記事の中には, 意見記事, 解説記事, まとめ記事があり, それらの記事の中には, 対象となっている話題の過去の経緯における重要な出来事が, 新聞記者の視点でまとめられている個所が存在する. 本研究では, この個所をサマリパッセージと呼ぶ. ある話題に関する複数記事を要約する際は, その複数記事中の重要個所を同定する必要があるが, サマリパッセージを参照すれば, 対象となっている話題のそれまでの経緯の中で, どの内容が重要なのか判断でき, 複数記事要約の際の重要個所同定に有効である.

検索結果の記事集合からサマリパッセージを抽出するには, まず, 記事集合中から意見記事, 解説記事, まとめ記事を検出するのだが, 記事によっては異なるカテゴリ (意見, 解説, まとめ) に属する内容が 1 記事中に存在する場合がある. このような記事に対応するため, カテゴリ毎のまとまりを単位として記事を処理する. このカテゴリ毎のまとまりをセクションと呼ぶ. 従って, 記事集合からサマリパッセージを抽出する手順は以下のようになる.

1. 記事集合から意見セクション, 解説セクション, まとめセクションを検出し,
2. 検出された各セクションから, そのセクションの形式に応じた方法でサマリパッセージを抽出する.

なお, 意見, 解説, まとめセクションは全てサマリパッセージを含むので, この 3 つを一括してサマリを含むセクションと呼ぶ. また, 報道的なセクションなどの, サマリを含むセクション以外のセクションをその他セクションと呼ぶ. 本研究では, このような枠組で検

索結果の記事集合からサマリパッセージを抽出するシステムを実装した. 全体のシステムは, サマリを含むセクション検出システムとサマリパッセージ抽出システムから構成される. サマリを含むセクションは, 記事をセクションに分割するセクション分割モジュールと, 分割された各セクションに, そのセクションが属するカテゴリを表すラベルを付与するラベル付与モジュールから構成される. さらに, ラベル付与モジュールは 9 つのフィルタから構成されており, 各フィルタは 4 つのカテゴリのいずれかに対応している. 入力されたセクションは以下で示す順にフィルタにかけられ, かかったフィルタに対応するカテゴリのラベルを付与され, その後のフィルタにはかけられず直ちに出力される.

1. まとめ (「週間日誌」)	4. その他 (公文書)	7. その他 (報道)
2. 意見 (「社説」)	5. まとめ (箇条書)	8. 解説 (残り)
3. 解説 (「解説」)	6. 意見 (文末)	9. その他

例えば 2 番目のフィルタにかかったら 3 番目以降のフィルタにはかけられない. フィルタの順番は, まず, 見出しからカテゴリを判断できるセクション用のフィルタ (1〜4) を, 次に, 内容から判断するセクション用のフィルタ (5〜7) を, 最後に, 現在はまだ特徴をつかめていないセクション用のフィルタ (8〜9) を起動するようにしている.

本研究では, サマリを含むセクション検出システムとサマリパッセージ抽出システムをそれぞれ評価した. サマリを含むセクション検出システムの評価では, セクション分割処理は人手で行ない, ラベル付与モジュールのみを定量的に評価した. その結果, 記事集合からのサマリを含むセクションの検出は, Recall,Precision ともに 60 %から 80 %であった. また, 各セクションからのサマリパッセージ抽出もおおむねうまくいくことを示した.

また, 抽出したサマリパッセージから複数記事要約を生成する方法として, 以下のように考えている.

1. 検索結果の記事集合中からサマリパッセージを抽出する.
2. 抽出されたサマリパッセージを, それぞれ, 記事集合中の過去の記事の最も関連の強い個所と対応づける.
3. サマリパッセージと関連の強い個所を重要個所と見なし, それらを元に要約を生成する.

今後は (1) サマリパッセージ抽出の精度向上を目指すとともに (2) 今回は行わなかったセクション分割モジュールの評価と (3) サマリパッセージから要約を生成するまでの処理の詳細の検討が必要である. また今回は一人によって作成されたデータを用いたが,(4) 複数の被験者を募り意見, 解説, まとめの判定実験を行い, その結果を元により客観性の高いデータ, ルール作成を行う必要がある.

なお, 本研究では記事コーパスとして毎日新聞社のものを用いた.