

## 3.9 自動要約の誤り分析

西川 仁 (東京工業大学)

### 自動要約

新聞の見出しや小説のあらすじ、論文のアブストラクトなど、何らかの要約はありふれたものである。これらはそれぞれ、新聞、小説、論文の要約であり、それらの媒体に含まれている情報を簡便に入手するために用意されている。さまざまな媒体に対してあらかじめ要約を用意することができれば、何らかの情報を探そうとする際の利便性は大幅に向上すると考えられるため、要約の作成を機械の手にもゆだねようと試みることは自然である。自動要約はまさに、機械によって、要約の対象となるテキストの内容を、より短いテキストで簡潔にまとめる処理のことを言う<sup>2)</sup>。

### 自動要約の技術

自動要約は、入力されたテキストを意味表現に変換する自然言語理解の技術、入力されたテキストに含まれる重要な情報を同定しその意味表現を要約としての意味表現に変換する技術、要約としての意味表現を自然言語に変換する自然言語生成の技術の3点からなる<sup>1)</sup>。一方、高精度の自然言語理解や、意味表現からの自然言語の生成は依然として容易ではない。そのため、現在の自動要約システムの多くは、意味表現からの自然言語生成の代わりに、入力されたテキストに含まれる、重要な情報を含む表現を抽出することで要約を作成している。具体的には、入力テキストを構成する文から、要約としてふさわしい文を抽出し要約を生成しており、この処理を重要文抽出という。また、単に文を抽出するだけではなく、文に含まれる余分な節や句を除去する書き換えを行うことで、より端的に重要な情報のみを含む文を生成する、文短縮という処理も行われる。さら

に、要約の出力は文章であるため、文章全体を論理的に一貫したものとする処理も必要となる。

### 自動要約の課題

自動要約には上に述べた処理が必要であり、以下のような課題が存在する。

1. 要約としてふさわしい情報を含む文の特定。
2. 文法性を維持した文の書き換え。
3. 要約を論理的に一貫したものとする、接続表現などの適切な追加。

これらはそれぞれ異なる性質を持った課題であり、自動要約の大きな課題は、これらさまざまな要素を総合的に勘案し、妥当な要約を出力できるシステムを設計するという点にある。

### 自動要約の誤り分析

上に述べたように、自動要約の出力にはさまざまな要素が関与しており、誤りの分析が容易ではない。このことを踏まえ、Project Next NLP 要約課題グループでは、不完全な要約を不完全たらしめる要因と、不完全な要約が生じる原因とを分類し、何が原因となってどのような誤りが要約に生じるのか整理した。その結果を表-1に示す。

この枠組みでは、不完全な要約が生じる場合は3つにわけられる。1つめは要約に含まれる文が非文法的(非文)であったり、文同士の論理的関係が明らかでなく(便宜的にそのような文章を非文章と呼ぶ)文章から適切に文意を読み取れない場合である。2つめは入力テキストと要約とで読み取れることが異なり、要約処理によって文意が歪曲されている場合である。3つめは要約としてふさわしくない情報が要約に含まれている場合である。

### 3.9 自動要約の誤り分析

		非文章の出力	文意の歪曲	重要部同定の失敗
操作の不足		文を書き換える機構などが不十分であるため、非文や非文章が生成される場合	文を書き換える機構などが不十分であるため、要約を読んだ結果、読者が入力テキストとは異なる結論に到達する場合	文を書き換える機構などが不十分であるため、要約の長さの制約などから重要な情報を要約に含めることができなかった場合
特微量の不足	特微量の設定不足	文の書き換え規則に関する特微量が得られず、非文を出力してしまった場合	省略を含む文にそのことを表す特微量が付与されておらず、これを考慮せずに出力した要約が結果として文意を歪曲する場合	入力テキストに固有表現などの情報が付与されておらず、機械が重要文の同定に失敗した場合
	言語解析の失敗	自然言語解析器が解析に失敗し、それが原因となって文の書き換えに失敗した場合	自然言語解析器が解析に失敗し、本来の文意とは異なる解析結果を出力した場合	自然言語解析の失敗によって適切な特微量を機械が取得できず、重要文の同定に失敗した場合
パラメタの誤り		文の書き換え規則の適用順序が正しくなく、誤って必須格の格要素を削除してしまった場合	文意の歪曲の原因となる文のパラメタが大きく見積もられている場合	ある特微量が適切な重みを得ておらず、重要文として認定されるべき文が重要文として認定されなかった場合
探索の誤り		パラメタは問題がないが、最適解が得られなかったために文の書き換えに失敗した場合など	左に同じ	左に同じ
情報の不足		入力テキストにおいて、言及されている情報が十分に説明されておらず、出力の要約を読んだ読み手が要約を適切に解釈できない場合	入力テキストが曖昧性を含んでおり、外部の情報なしには入力を正しく解釈できない場合など	新聞記事の見出しなど、要約のために必要な情報がそもそも要約システムに与えられていない場合

表-1 自動要約の誤り分析の枠組み

このような誤りが生じる原因は5つにわけられる。それぞれ、要約システムの、文を書き換える操作が不足している場合、利用できる特微量が不足している場合、パラメタが正しくない場合、要約を生成する際に行う探索に問題がある場合、要約システムに入力された情報が不十分でありそもそものような工夫によっても適切な要約を生成できない場合である。

Project Next NLP 要約課題グループではこの分析の枠組みに基づき各自の要約システムの出力の分析を行った。分析の結果、現在の要約システムの出力に含まれる誤りは主として、文の書き換えに関する操作が不十分であることによる非文章の出力、および、要約対象のテキストの論理的な構造を把握する機構が不足していることによる重要な情報の同定の失敗の2点にあることが分かった。

### 自動要約の今後

今後の自動要約には、より柔軟な文の書き換えや、元のテキストに含まれない表現の生成を可能にする機構が重要になる。また、テキストの論理的な構造に基づいて重要な情報を特定する機構も重要になる。これらのより高度な機構に基づきよりよい要約を生成することによって、人々が情報を入手しようとする際の利便性を向上させることが自動要約の目標である。

#### 参考文献

- 1) Jones, S. K. : Automatic Summarising : The State of the Art, Information Processing & Management, 43, pp.1449-1481 (2007).
- 2) 奥村 学, 難波英嗣: テキスト自動要約, オーム社 (2005). (2015年9月24日受付)

西川 仁 (正会員) hitoshi@cs.titech.ac.jp

東京工業大学大学院情報理工学専攻助教, 博士 (工学). 自動要約の研究に従事.