青山学院大学理工学研究科経営工学専攻

文脈情報も考慮した要約システムの高精度化

原田研究室　　野　口　貴（35502016）

近年，テキストマイニングの一つとして自動要約研究が注目されている．昨今の要約研究を調査したところ，重要文抽出型要約では，重要な文や語句が削除されて主旨が伝わらないことがある．重要語句抽出型では，構文情報，表層情報に基づく研究が中心で，表層格で述語の格を補うため，語意に基づく適切な格が抽出されず要約文が不自然になってしまう．さらに，意味解析していないので文中の語意の精度が低く，必須格の抽出の誤りがあり，要約文が日本語として不適格であった．そこで原田研究室では昨年度，文章を意味解析した結果の格フレーム群から要約文を生成するシステムABISYS を開発した．しかし，従来の ABISYS では，反復語句を重要語句とし，必須格を補うことで要約を生成していたが，反復情報のみを用いているために重要語句選定に関して問題点があった．

本研究では，重要語句選定用に従来の反復得点に，文脈情報である文間の深層格情報，位置情報，意見語情報，主題・焦点情報などの複数の視点からの得点を加えてこれらからマハラノビスの汎距離を用いて体言に対する重要語得点の総合化を行い，そこから正しい日本文として必要な語句を追加して要約文を生成する手法を提案し，これを基に新要約システム ABISYS2003 を開発することにした．

要約処理の具体的な手順は，(１)重要語句の得点化，(２)重要語句得点の総合化，(３) 要約要素語の選定，(４)冗長な表現の削除，の 4 ステップで行う．(１)重要語句の得点化では，体言が文章の内容のポイント（要点）を示すものであると考え，体言を重要語候補とする．重要度合いの得点は反復得点，文脈得点，位置得点，見解得点，主題・焦点得点の 5 種類を用いる．(２)重要語句得点の総合化では，マハラノビスの汎距離を用いた，確率的な重要語句得点の総合化を行った．まず，5 得点で定義された空間において，あらかじめ大量の重要語句候補を正解，不正解に人手で判別しておく．ここで言う正解とはその文章の要約において抽出されるべき語句ということである．その上で新たな重要語句候補を解析する際，それが正解，不正解どちらの群に近いかをマハラノビスの汎距離を用いて判別し，正解重要語句確率という形で算出する．要約率に応じて，その確率の高い順に重要語句とする．(３) 要約要素語の選定は，次の４ステップで行う．(ⅰ)重要語句から主述語までのパスを要約要素語として抽出する．(ⅱ)重要語句と深層格で隣接した用言とこの用言と「reason」「cause」「sequence」格で隣接する用言を抽出する．(iii)重要語句と深層格で下向きに隣接した語句を抽出する．(iv)抽出した要約要素語中に,「こと」,「の」,「という」が含まれていた場合，その which 格と that 格の語を抽出する．(４)冗長な表現の削除では，要約文を短くするために補足語修飾節の削除，引用動詞の削除，換言処理の３つの処理を行う．

本システムによる要約と，人手による３つの要約を主観評価と日本語としての正しさに関して順位付け評価したところ，主観評価では文の読みやすさの点で平均 3.11（他システムの平均 3.11），内容カバーの点では平均 2.65（他システムの平均 3.25）となった．また生成された要約文の日本語としての正しさは昨年度 86%から 95%と，従来の手法より要約品質が向上した．よって本手法の有効性を示すことができたと言える．