

要約支援システムCOGITOにおける
文章の重要性評価について

4D-7

日比 孝 小松 英二 安原 宏
(沖電気工業株式会社 総合システム研究所)

1. はじめに

要約支援システムCOGITOでは、文脈理解に基づいた要約をめざしている。

COGITOでは、文書の構造を解析して重要な部分を選ぶことにより要約を行う。このために、構文解析のあと照応、省略、接続関係の解析等の文脈処理を行い、さらに、その結果を用いて要約処理を行う。本稿では要約処理における重要性の評価について述べる。

2. 重要性評価ルール

一般に、重要性の評価には文脈情報を用いた深層的な情報を用いた方法と、接続語やキーワード等による、表層的な情報を用いた方法がある。表層的な情報だけを用いた場合、意味を使わないことから不要なものが重要と判断されてしまうことが多い。そこで、COGITOでは重要性の評価には双方の情報を用いている。

要約の生成には、重要性の評価ルールを用いて文や概念に点数を与え、評価の高いものを要約とする。

2.1 文脈情報を用いた評価

ある文(あるいは概念)に対し、説明文や補足のための文が多いとき重要であるとする。そのために文の間の接続関係のネットワークを参照する。ルールとしては次のようになっている。

- (1) 多くの文がかかっている文を重要とする。
このとき、間接的にかかっている文に対しても考慮する。

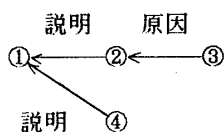


図1 文章構造の模式(①が重要)

- (2) 照応の対象となる回数の多い概念を重要とする。
- (3) 事実と意見の区別

報告文、説明文等では事実が重要であり、論説等では意見、見解が重要である。

2.2 表層的な情報を用いた評価

以下の情報を用いて重要性の評価を行う。

- (1) 語の頻度
特に頻度の高い語およびそれを含む文。
- (2) 重要性を示している語
「要するに」「結局」など、明らかに重要性を示している接続語を含む文。
- (3) 強調表現
強調された表現に対しては重要と評価する。
- (4) 分野ごとの重要語

文書の分野、種類によって異なる重要語が存在する。そこで、そういう種類の重要語を調べるため、実際に新聞記事で検討してみた。3ヵ月分の記事(朝日新聞社提供)から、キーワードに「コンピュータ」を含んでいるもの(記事数101)を検索し、その中での語の頻度を調べた。そして、頻度の高い語の中から重要性を示す語となりそうなものを選んでみた(図2に示す)。これらの語は、いわゆるキーワードとしては適切でないが、その種類の文書の特徴をとらえているといえるものであり、重要性の評価には有用であると思われる(特に動詞)。

定型的文章では、決まった語を含む文が重要になってくる。このような語を見つけるためには、同種の文書を集めて統計をとれば、その特徴が明らかになるであろう。

開発	115	研究	27
導入	39	発売	26
発表	34	輸出	25

図2 重要語の例(数字は頻度)

3. まとめ

要約支援システムCOGITOにおける重要性の評価について述べた。

文脈情報と表層的な情報の両方を利用することにより、より正確な要約が可能になる。今後はより多くの文書に対してシステムの評価を行い、改良していく予定である。

参考文献

小松他, 要約支援システムCOGITO-文書の構造解析, 自然言語処理研究会64-11(1987)

本研究は第5世代コンピュータプロジェクトの一環としてICOTからの委託で行われたものである。