Title	長さの異なる要約の自動生成
Author(s)	伊良波，隆
Citation	
Issue Date	1999-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1228
Rights	
Description	Supervisor:佐藤 理史，情報科学研究科，修士

長さの異なる要約の自動生成

伊良波 隆

北陸先端科学技術大学院大学 情報科学研究科

1999 年 2 月 15 日

キーワード： 情報抽出、長さの異なる要約.

現在、インターネットの普及により、ワールドワイドウェブ (World Wide Web)、電子ニュースなどの電子化されたメディアが増加している。電子化されたメディアはネットワークを通じて、地理的な制約や時間的な束縛を受けることなく様々な種類の情報にアクセスできる。そのため、非常に便利なネットワーク社会が構築されつつある。

しかし、利用できる情報が急速に増加することによって、ユーザは必要な情報を的確に捜し出すことが難しくなっている。そのため、有用だが利用されない「眠った」情報を増大させるという状況が引き起こされてきている。

求める情報を得るためにユーザは情報検索を行ない、どの情報が必要であるかどうかを決定しなければならない。その決定を支援する１つの方法として、簡潔で適切な要約を提示することが挙げられる。そうすると、ユーザは全文を読まなくても済むため、必要とする情報を容易に取捨選択することができる。

本研究では、電子ニュースのニュースグループ fj.sys.sun を対象テキストとし、ニュース記事を自動的に要約するシステムを作成する。ニュースグループ fj.sys.sun は、質問応答型のニュースグループで、主に、SUN ワークステーション関連の、ハードウェアやソフトウェアに関する質問記事とそれに対する応答記事から構成される。

佐藤研究室では、自動編集プロジェクトの一環として、質問応答パッケージ SUN QA-Pack を開発し Web 上で公開している (http://www-sato.jaist.ac.jp:8000/faq/)。SUN QA-Pack とは、ニュースグループ fj.sys.sun の質問記事を分類し、短いサマリーをつけて求める記事を捜し出すことを支援するシステムである。

SUN QA-Pack では、それぞれの質問記事に対して１種類の長さの要約を生成している。本研究では、これとは長さの異なる要約を自動生成することを試みる。一つは１文のみからなる短めの要約であり、もう一つは数文で構成される長めの要約である。

長さの異なる要約の生成を実現するため、本研究の研究対象であるニュースグループ fj.sys.sun の調査を行なう。質問記事中でよく現れる特徴的な表層表現を用い、文のタイプを次の１０種類設定した。

1. 挨拶、自己紹介，2. 環境，3.Goal，4.Fail，5.Question，
6.Error または Program の表示，7. 分析，8. 終りの挨拶，9.Signature，10. その他

fj.sys.sun の質問記事において中心となる情報は、「何を行ないたいのか」、「直面した問題は何か」、「質問は何か」ということである。「何を行ないたいのか」を記述した文は、上記のタイプの **Goal** となる。このタイプの文では、意志を表す助動詞「〜たい」がよく用いられる。「直面した問題は何か」を記述した文は、上記のタイプの **Fail** となる。このタイプの文では、否定文であるか、あるいは、文末に否定的な表現を表す動詞が使われることが多い。「質問は何か」を記述した文は、上記のタイプの **Question** となる。このタイプの文では、ほとんどの場合、疑問文となる。これら３つのタイプを fj.sys.sun の重要文として抽出する。

要約の生成は以下の手順で行なう。まず質問記事中の各々の文が、どのタイプ文であるかを、文字列のパターンマッチングによって判定する。その際、意味のある情報を全く含まない文が、抽出すべき重要文（**Goal**,**Fail**,**Question**）と判定される場合がある。それらは重要文として抽出するのは不適切である。以下では、このような情報を全く含まない文を意味なしセンテンスと呼ぶ。

意味なしセンテンスの特徴として、「何が／を／について」という対象を特定する名詞的概念を表す語が、センテンス中に含まれていない。これら名詞的概念を表す語には２種類考えられる。１つはコンピュータ関連の専門用語（固有名詞）である。コンピュータ関連の専門用語は、ほとんどアルファベットかカタカナで記述される語であるため、簡単に判定できる。もう一つは、一般的に使われる名詞（「方法」、「本」など）である。これらの名詞を網羅することは、現実的ではない。そこで、意味なしセンテンスに現れる名詞をリストアップ（意味なし名詞リスト）し、このリストに含まれない名詞を、意味のある名詞とみなす方法を採用する。従って、重要文と判定された文中の名詞が、コンピュータ関連の専門用語を含まず、かつ、意味なし名詞リストに全て含まれる場合、その文を意味なしセンテンスと判定する。

最後に、長さの異なる２種類の要約を生成する。短めの要約は、文が重要文と判定され、かつ、意味なしセンテンスではない文のうち、記事の最初に現れる文を抽出する。長めの要約は、文が重要文と判定され、かつ、意味なしセンテンスではない文を全て抽出する。

要約生成システムの実験を行ない評価した。タイプ判定モジュールでは 91%の精度で文のタイプを判定できた。短めの要約の場合 87%、長めの要約の場合 93% の精度で、適切な要約または許容できる要約を生成することができた。