自然言語処理 116-20
(1996. 11. 19)

言い替えを使用した要約の手法

近藤恵子, 奥村 学
e-mail:{k-kondo,oku}@jaist.ac.jp
北陸先端科学技術大学院大学 情報科学研究科

[概要]

本稿では言い替えにより要約を生成する手法を提案する。要約の手法は過去に幾つか提案されているが、それらの方法ではキーワード、要約的表現などの表層的な情報を利用した抄録の方法が主である。しかし、物語文を対象とした要約システムにおいては、ストーリーを損なうことなくまとめることが重要であり、重要箇所を抽出する手法は適さない。そこでエピソードを言い替える事により物語文を要約する手法を提案する。言い替えのためにはEDR 電子化辞書の日本語単語辞書の語釈文と概念辞書を使用する。語釈文からは時間経過による動作を抽出したテンプレートを作成する。概念辞書は前後の動作が概念的に似ている場合に一語へ言い替えを行なうのに使用する。

Summarization by Paraphrasing

KONDO Keiko,OKUMURA Manabu
School of Information Science, Japan Advanced Institute of Science and Technology
(Tatsunokuchi Ishikawa 923-12 Japan)

Abstract

In this paper, we present a summarization method which uses paraphrase for generating an abstract. In the past, several various methods were proposed, and most of them use surface information to extract important parts for the whole text. However, in the process of summarization for tales, it is important to keep the continuity of the story. As the most previous approaches did, the extraction of important parts from the text is not fit for summarizing tales. We propose a framework of summary production based on paraphrasing episodes. The definition sentences in Japanese Words Dictionary and Concept Dictionary of EDR electric dictionary are used for paraphrase. The templates for paraphrasing are created by extracting sequence of action from definition sentences in Word Dictionary. When some actions in the text conceptually resemble each other, Concept Dictionary is used combining these actions into a single action.

1 はじめに

要約生成の手法は、過去にいくつか提案されている。表層的には単語の頻度、要約的表現、キーワード、テキスト内の位置などの情報を利用する手法がある。また、スクリプトや物語文法を使用してテキストを解析し、重要箇所を抽出する方法、意味ネットワークなどを用いて意味表現を行い、重要箇所を決定する方法などもある。しかしいずれにせよ、これらの手法はその目的と対象とするテキストにより使い分けなければならない。

当然、物語文を対象とした要約では、論説文の要約とは違ったアプローチが必要とされる。論説文は一つの結論を導くための明確な構造を有しているため、重要箇所や事項を抽出する手法の構築も比較的容易であると考えられる。しかし物語文においてはトピック間の関係は複雑であり、その構造も論説文のように一様ではないため、キーセンテンスのみを抽出する方法は適さない。また、人間が物語文の要約をキーセンテンスだけで生成するとは考えにくく、不自然である。

過去に提案された物語文を対象としたシステムに、スクリプトから作り出された予期に基づき処理を行うSAM[4]、登場人物のゴールからそれに続く動作を解釈することにより、プランから物語を理解するPAM[5]がある。Plot unitの解析に基づき要約を生成する手法も発表されている[2]。また、最近では意味構造生成システムELF[6]、物語の筋のエピソードをネットワークで表現するシステムJstory[7]などがある。これらの研究では意味構造を解析し「理解」することが試みられている。言い替えにより要約を行なう手法も[3]において提案されている。

本稿では物語を抽象度による階層構造を持つものとして捉え、言い替えにより語の抽象度を上げていくことで物語の要約（ストーリー）を生成する手法を提案する。なお、要約の目的は「物語本文を読者に類推させる手掛かりとする」と設定した。

言い替えにはEDR電子化辞書の日本語単語辞書と概念辞書を利用する。日本語単語辞書はその語釈文のうち利用可能なものを抽出し、時間経過に依存する動作のテンプレートを作成し、言い替えに利用する。また、単語は概念辞書により上位-下位関係を与えられており、上位概念を辿ることにより前後の単語との概念的な相似を見付け、言い替えることが出来る。

2 言い替えによる要約の手法

人間は要約の過程においてストーリーを保持するために、エピソードを削ることではなく、一つのエピソードに費やす字数を減らすことを考えると思われる。そのためには、具体的な記述ではなく抽象度の高い語を使用した記述を行うことが考えられる。

物語は意味的には階層構造で表現できる。最下層は単位文と呼び、テキストを意味レベルの最小単位まで解析したものである。単位文の意味的な集合は高次の概念を持つ語へと言い替えられ、トピックを形成する。このトピックの意味的な集合がさらに高次の概念を持つ語へと言い替えられ、もう一段、高次のトピックの階層を形成する。このように言い替えは抽象度の低い語から抽象度の高い、高次の概念を持つ語へと行われる。この言い替えの最上層をエピソードとする。

以下にこれらの提案する手法について詳細に述べる。

2.1 人間による要約方法

人間が過去に読んだ（書いた）物語の要約を生成する場合、まず想起されるのは特定のエピソードやキャラクターであるかもしれない。しかし、要約文を生成するためには頭の中で複数のエピソードを連結させ、その全体の構成を構築する必要がある。要約を生成するためには、この二つが重要となる。本を実際に読み返しながら要約文を生成できるとすれば、文章の引用が増えることが考えられる。しかし、もし引用のみをつなげたとしたら、それは文章として不自然なものになることは明らかである。

要約には、多くの場合、文字数の制限が付与される。要約する側はこの制限により、エピソードを収縮していく。制限文字数が多ければ要約文に使用されるエピソードは増加し、少なければ減少する（図1）。しかし、その関係は比例しない。最長制限文字数とはその本文文字数そのものであると考えられるが、制限文字数をその半分にしたときに拾われるエピソードも半分ということはありえない。それではストーリーが成立しない。エピソードの取捨選択は、テキストの構成、即ちストーリーの展開により決定される。

そして、選択したエピソードに対して制限文字数が少ないときに通常、人間が対処する方法として、高次の概念を持つ語への言い替えによる短縮が考えられる。一つ一つのエピソードを表現している文字数を、エピソードの内容を変えずに短縮するのである。この言い替えにより、ストーリーの欠落なく、指定の文字数に要約することが可能となる。また、物語文の要約は物語内での時系列順ではなく、エピソードの物語への出現順に生成される。

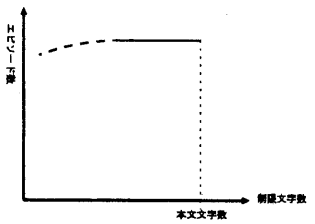

図 1: エピソード数と制限文字数の関係のグラフ化

2.2 物語の構造

物語文のテキストは、複数の文を一列に連続的に並べた形で表記される。しかし、その内容は意味的には図 2のような階層的構造を有していると思われる。

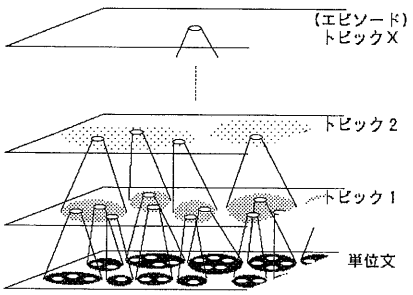

図 2: 階層構造

階層のレベルはその抽象度による。最下層は単位文であり、最上層はエピソードである。その間の階層はテキストの記述の詳細の程度により決まる。単純なテキストならば、中間にトピックの階層を持たない場合もありうる。

最下層の「単位文」とはテキストの内容を最小限の単位まで解析した意味表現である。一つの単位文は一つの状態もしくは動作を表し、テキストである物語文の一文とは異なる。複数の単位文を意味的にまとめ、高次の概念を持つ語に言い替えたものがトピックとなり、トピック１の階層を形成する。トピック１の階層において同じように意味的なまとまりを収集し、言い替えたものがトピック２の階層を形成し、同様の方法でテキストの深さによって幾つかのトピックの階層を形成する。そして、トピックの階層の最上層がエピソードの階層となる。

つまりここでは、原文のテキストを解析し、単位文により表現したものとエピソードの間は、０以上の階層に分けられると考える。この中間の階層をトピックの階層と呼ぶことにする。

2.3 区切りの決定

言い替えに際しては、どの階層のレベルで行うにしても、区切りをどこでつけるかという問題が生じる。まず、図 3にスクリプトを使用した言い替えの区切りの例を示す。一つ一つはトピックである。これは「電車に乗る」というスクリプトであり、「電車に乗る」という一階層上のトピックになる。

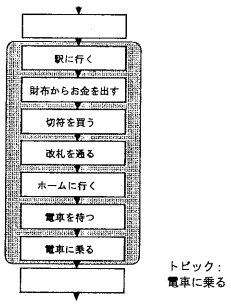

図 3: スクリプトを利用した区切りの決定

さらにネットワーク表現を用いると、図 4に示すように関係の低い単位文、もしくはトピック間のリンクは当然少なくなる。リンクの少ない部分は区切りであると推測でき、ネットワークのまとまりは時間経過に依存しない、一階層上のトピックのまとまりに当たると考えられる。さらに一階層上のトピックにまとめる

ときには、再びネットワークを生成し、区切りの決定を行う。

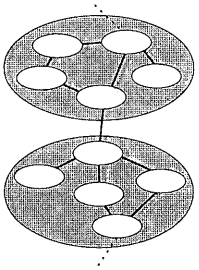

図 4: ネットワークを利用した区切りの決定

3 テキストからの動作の抽出

まず、物語文の冒頭には多くの場合、物語世界の設定の説明がある。「むかしむかしあるところに」や「これは私が小さい時に聞いたお話です」等の世界説明は以降の物語のストーリーと容易に切り離すことが出来る。この物語世界の説明は設定の説明であり、本文のストーリー展開には直接には関係しないためである。また、ストーリーの始まりは「ある日のことでした」といった定式を持っていることからも、区切りは容易に発見され得る。この物語世界の設定の説明についてはストーリーを保持するという本研究とは直接関係しないため、削除した。

次にテキストから文章の流れにそった動詞を抽出する。動詞は文章において動作を表現しており、動作を抽出することによりストーリーの展開を推測できると考えたためである。これを階層構造における単位文とみなす。しかし、文章中の全ての動詞を抽出することは適切でない。主節に対する従属節は文脈を形成しない [1]。そのため、一文中に複数の動詞が含まれる文の場合は主節の動詞だけを採用し、文中の名詞を修飾しているような連体節の動詞、入れ子になっている節などの従属節の動詞は採用しない。

具体的には以下のような形で動作を抽出した。

- 私は学校に行って、お弁当を食べた。
  ⟶「行く」＋「食べる」
- 私は拾ったボタンを捨てた。
  ⟶「捨てる」
- 私は勉強をするために学校に行く。
  ⟶「行く」

実装にあたってはjuman3.0Betaにより形態素解析を行ない、動詞とそれに続く助詞の情報より必要な動詞を選択した。

4 EDR電子化辞書を利用した実装

言い替えをEDR電子化辞書の中の日本語単語辞書と概念辞書を用いて実装する。日本語単語辞書は単語に対してレコード番号、見出し情報、文法情報、意味情報、運用、その他の情報、及び管理情報を与えている。本手法では語釈文として与えられている日本語概念説明を利用し、時間に依存する言い替えのためのテンプレートを作成する。一方、概念辞書は単語に与えられた概念識別子により上位-下位の関係を位置付けている。これを利用して前後する単語の概念的関係を認定し、同一の概念を持つものを上位の語により言い替える。

4.1 日本語単語辞書の利用

日本語単語辞書レコードの構造は表1のような体系により表され、エントリは約25万語である。ここで与えられている日本語概念説明の文を語釈文であるとみなす。語釈文は語の意義を文で表しており、これから時間経過を含む動詞及び述語句を抜き出し、テンプレートを作成した。

表1 日本語単語辞書の仕様

レコード番号	単語見出し
不変化部-連接属性対	かな表記
発音	品詞
構文木	活用情報
表層格情報	相情報
機能語情報概念識別子	概念見出し
概念説明	用法
頻度	管理情報

まずEDR日本語単語辞書から動詞及び述語句の語釈文を抜き出した（86954語）。以下に例をあげる。

- 走る [ハシ・ル] 捕らえられていた場所から逃げ出す
- 説得する [セットク・スル] よく話して納得させる
- 足場を固める [アシバヲカタメ・ル] 物事を行う場合の拠り所を確立する

次にこれらを解析し、主節の動詞だけを抜き出した。その結果、複数の動詞が抽出されたものを時間経過に依存するものとみなし、約1万のテンプレートを採用した。テンプレートの例を図5に示す。

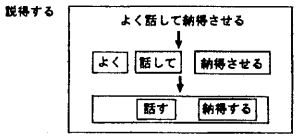

図5: テンプレートの作成

このテンプレートを用いて文章から抜き出した動作の列とマッチングを計り、言い替えを行なう。この処理は時間経過に依存しており、テンプレートは一種のスクリプトと言うことが出来る。図6はマッチングの例を示している。言い替えは複数回に渡り階層的に行なうことが出来る。まず「聞き入れる」という言葉が『承認する』の語釈文「相手の言い分を聞き入れる」にマッチし、『承認する』に言い替えられる。次に「理解する」と「承認する」が『納得する』の語釈文「物事を理解して承認する」にマッチし、「納得する」に言い替えらえる。そして「話す」と「納得する」が『説得する』の語釈文「よく話して納得させる」にマッチし「説得する」に言い替えられる。このような手順により動作の列、「話す」「理解する」「聞き入れる」は「説得する」に言い替えられる。

このように動作を語釈文と対応させ、その語釈文の語に言い替えるという作業を繰り返すことにより複数の前後する動作を一つの語に言い替えることが出来、要約が可能になる。

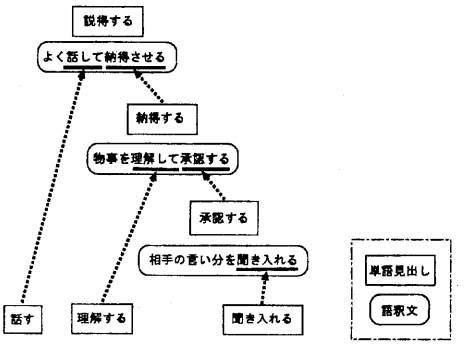

図6: 語釈文による言い替えの例

4.2 概念辞書の利用

各単語にはそれぞれ概念識別子が振られている。概念識別子は16進数で表示されている。語は一語に複数の意味が存在するため、一つの語に複数の概念が存在することもある。また、同様のことを表現するにも様々な言い表し方があるように、複数の語に共通の概念識別子が振られていることもある。更にこの概念は概念辞書において上位-下位関係が与えられ、グラフ構造をなしている。上位概念を辿っていくと対応する語のない概念だけの概念識別子になり、最終的には「概念」「事象」などを意味する幾つかの概念識別子に行き着く。

時間経過に依存しない要約を行なう方法としてこの概念辞書を利用する。前後する単語がその上位概念において同じものを持っていれば、それらはまとめてその共通した上位概念識別子を持つ語に言い替えることが可能である。

図7に例を示す。

文中に「掃き清める」という動作と「拭く」という動作が連続して見られたとする。「掃き清める」という語は [103778] と [103779] という二つの概念識別子を持つ。具体的には

- 103778–掃いて清らかにする
- 103779–ある地域の敵を討ちつくして平穏にする

の二つを意味する。しかし「掃き清める」単体ではどちらが正しいのか決定不可能であるため、まず両

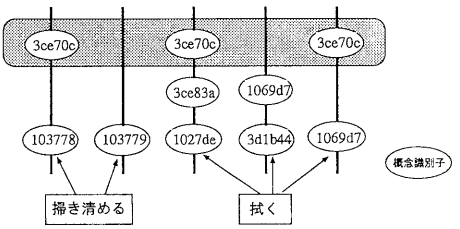

図 7: 上位概念への言い替えの例

方を採用しそれぞれの上位概念を見る。「拭く」には [1027de] [3d1b44] [1069d7] という三つの概念が存在し、「掃き清める」同様にそれぞれ図のように上位概念を辿ることが出来る。これらは更に上位へと辿ることが可能である。この例では上位概念として概念識別子 [3ce70c] を共通の概念として獲得し、言い替えることが可能になる。概念識別子 [3ce70c] は「清」「清め」「浄める」「清める」「清掃する」「掃除する」という単語見出しを持つエントリに照応し、これらのいずれかに言い替えが可能である。実際には言い替えは動作から動作へと行なわれるため、この場合は「浄める」「清める」「清掃する」「掃除する」のいずれかということになる。言い替える語の選択についてはまだ考慮する必要がある。

5 おわりに

物語文要約の枠組みの提案と、EDR 電子化辞書を利用した具体的な言い替えの手法を提案した。枠組みとしては人間の要約処理の過程を参考としたが、これには個人差が考えられること、リサーチの困難さなどから一般化は難しい。しかし、物語文は論説文と比較して定式化された構造を持たないこと、表現方法の違いなどから、表層的な要約には限界がある。より自然な要約文を生成するためには意味解析、文脈解析にまで踏み込んだ解析が必要であり、人間の要約処理過程をモデル化することは有効である。実装には EDR 電子化辞書を使用した。これにより、概念的なものも含めた言い替えが可能となることを示した。

今後の課題としては、日本語単語辞書の日本語概念説明の中には必ずしもテンプレートとして有効であるとは言えないものも含まれていることから、テンプレートの絞り込みを行なう必要がある。概念辞書については高次の概念になると概念が存在してもそれに対応する日本語見出しが存在せず、事実上、言い替えが不可能になるという問題がある。また、一つの概念識別子に対して複数の見出し語が発見されることから、言い替えの対象としてどれを選択するのが適切であるかということも問題となる。テンプレートや上位概念の照応は、前後に幾つかの関係しない動作を挟んでも言い替えは可能であると考えられるが、それをどの程度の幅で許すのかということも課題として残される。更に、今後の処理を進めるにあたり、この二つの手法を如何に組み合わせるかということを検討したい。

参考文献

[1] James Allen. Chapter14,Local Discourse Context and Reference. *Natural Language Understanding*, 1995.

[2] Wendy G. Lehnert. Plot Units and Narrative Summarization. *Cognitive Science 4*, 1981.

[3] Alain Pilguere Lidija Iordanskaja, Richard Kittredge. Lexical Selection and Paraphrase in a Meaning-Text Generation Model. *Natural Language Generation in Artificial Intelligence and Computional Linguistics*, 1991.

[4] and Yale AI Project. Schank, R. SAM-A story understanding. *Research Rep.43,Dept.of Computer Science,Yale Universoty*, 1975.

[5] R. Wilensky. Understanding goal-based stories. *Research Rep.No.140,Dept.of Computer Science,Yale University*, 1978.

[6] 小谷善行上田世志. 昔話「桃太郎」を対象とする自然言語文の意味構造自動生成. 自然言語処理 84-4, 1991.

[7] 重永実中沢俊哉. エピソードネットワークを用いた物語のあらすじ生成. 情報処理学会論文誌 Vol.32 No.10, 1991.