ネットワークを利用した要約のための話題の抽出

近藤恵子　荒井良徳　上里福美
東京工芸大学　工学研究科

1　はじめに

文章は文の単純な羅列ではなく、全体として意味的な構造を有している。複数の話題について順番、並列、あるいは交差するように述べられている。記述についても、深く綿密に述べられている話題もあれば、簡潔な解説で済まされている話題もある。文章を要約するためには、この文章全体の構成を把握する必要がある。文章は最小の単位である中間言語的な表現におとされているものとする。それらを概念的に高次のものへと階層的に言い替えていくことにより、最終的に幾つかの話題に収束させ、文章全体の構造を把握する。本研究では物語文を対象として、話題を収束させるためネットワークを利用し、意味的に関連のある話題を集める方法を試みた。

2　テキストの階層構造化

物語は意味的には階層構造で表現できる[1]。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。

2.1　人間による要約方法

人間が過去に読んだ物語の要約を生成する場合、複数のエピソードを連結させ、その全体の構成を思い出す必要がある。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だろう。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。

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束していく。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。この関係を図1に示した。

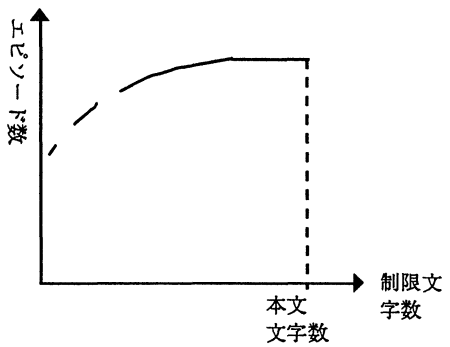

図1　エピソード数と制限文字数の関係

この二つ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。

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。

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。

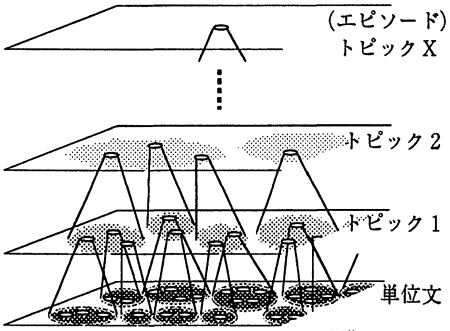

図2　物語の意味的階層構造

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。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。この中間の全階層をトピックの階層と呼ぶことにする。

3　言い替えのための区切りの決定

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。このときに重要な課題となるのが、時間の取り扱いである。言い替えには時間経過を重視する語と、時間経過に関係しない語が考えられる。この二種類について次に説明する。

3.1　時間経過に依存する言い替え

時間経過を重視する言い替えの例として、単位文の列を動作「拭く」に言い替える模式図を図3に示す。

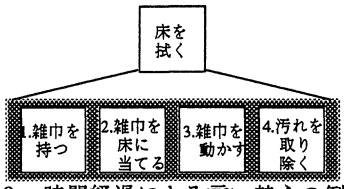

図3　時間経過による言い替えの例1

図3において1,2,3,4の動作のすべてが時間順に起こることにより、この言い替えは成立する。言い替えのための語の概念辞書において「拭く」という語は図3の単位文1,2,3,4に当たる動作の連続により記されている。単位文の列がこの辞書に照応することにより、「拭く」への言い替えは成功する。この言い替えのための語の概念辞書はスクリプト[2]であるとも考えられる。これらの単位文中に欠落がある場合や、順番の違う場合、例えば、1,4,3,2の場合などは、言い替えは成立しない。

3.2　時間経過に依存しない言い替え

言い替えは常に時間経過に依存するとは限らない。語の概念辞書によるスクリプトのような形を用いる言い替えは、単位文の列が常に時間的に一直線上になければならず、このような言い替えは不可能な例もある。時間経過に依存しない言い替えの例を図4に示す。

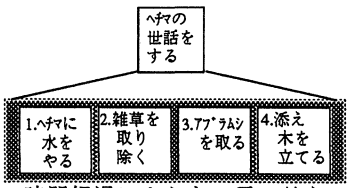

図4　時間経過によらない言い替えの例1

図4における単位文1,2,3,4はどの順に発生してもよく、なおかつすべてが発生する必要はない。単位文3,4,2,1の順に発生しても言い替えは可能であるし、単位文2,4のみでもこの言い替えは成功する。

図4の例の言い替えの条件は、必要なものを与えることと、嫌うものを取り除くことである。条件は言い替えのための語の概念辞書により与えら

れる。必要なもの、嫌うものという一般知識は前もって与えられているものとする。この知識を利用し、概念辞書による条件を満たしてさえいれば、図5に示すように単位文が増えてもこの言い替えは可能になる。

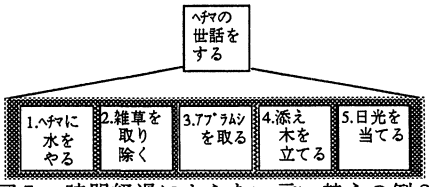

図5　時間経過によらない言い替えの例2

4　言い替えのためのネットワークの使用

3で示したように言い替えは常に時間経過に依存するとは限らない。また、単位文の列は常に時間的な直線上にあるとも言えない。事象は並列に発生、進行することもありうるし、複数の事象の時間関係が明確でないこともある。そこで単位文同士のつながりを時間的な直線によらないネットワーク構造で表現することを試みた。

4.1　ネットワーク構成要素間の関係

ネットワークを構成する単位文同士の関係には幾つかのパターンが考えられる。パターンは詳細に検証すれば単位文間の数だけ存在する。しかし、そのようなネットワークは処理の上で意味をなさない。ここでは関係表現を極力簡素化することを考え、文と文との関係を表わす表現として接続詞の接続の種類[4]を参照し、以下の三パターンを設定した。

順接：時間経過を含んだ関係
　　（過去から未来への矢印で表現する）
並列：時間関係によらず複数の情報が同一レベルである関係
　　（双方向の矢印で表現する）
添加：他の特定の情報に対して何らかの情報を付与する関係
　　（付与する側から付与される側への点線による矢印で表現する）

4.2　話題の抽出

単位文同士の情報が近似しているとき、それらは一つのまとまりとして捉えられ、一つの話題として抽出することが可能である。ネットワーク構造において、近似した情報同士はネットワークが密となり、話題が変わるときにはネットワークが疎となる。添加はネットワークが疎であっても情報を付与しているという特徴から、話題からは外されない。複数の単位文を一つの話題として抽出、言い替えを行う。これがトピックとなる。トピック抽出の例を図6に示す。

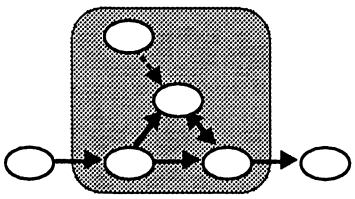
図6　トピックの抽出

抽出されたトピックは、トピック1の階層を形成する。ここでも再び言い替えのために新たにネットワークを構成し、図2に示したようにトピックの抽出を行う。

4.3　ネットワークの再形成

新たなネットワークの形成には下層部でのネットワークの情報がそのまま利用される。単位文からトピック1の階層への言い替えであれば、図7に示すようにトピックAとトピックBとの関係は、トピックAを形成した単位文とトピックBを形成した単位文の関係である。トピックAを形成した単位文とトピックBを形成した単位文はネットワークが疎であることから区切られている。そこに存在する単位文間の関係はそのまま上の階層のトピックの階層にも利用できると考えられる。

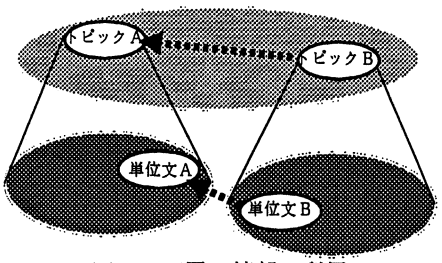

図7　下層の情報の利用

単位文Ａは他の単位文とまとまりトピックＡを形成する。単位文Ｂは他の単位文とまとまりトピックＢを形成する。単位文Ｂは単位文Ａに情報を添加している。二つのまとまりはこの情報だけにより接続している。そのため、単位文Ｂを意味として持つトピックＢは単位文Ａを意味として持つトピックＡに添加の関係を持つ。このように単位文間の関係をトピック間の関係に持ち上げることによりさらにネットワークを形成し、再び話題の抽出を行う。

4.4　複数の関係情報

4.3においてはトピックを形成した単位文間の関係が一種類であると限定して考えた。話題を抽出するためには前述のように、ネットワークが疎である部分を切る。しかし、疎であるという決定基準をどのように定めるかという問題がある。関係情報が一つである状態のみを疎であると見なし区切るならば、その関係情報をそのまま利用することが可能である。しかし、物語を形成するネットワークは様々な形態が考えられる。複雑なネットワークであれば情報関係が一つである部分が存在しないことも考えられるし、簡素なネットワークであればほとんどが順接の関係で一直線上に並んでいることも考えられる。このようなパターンを考える時、情報関係が一種類の場合だけを区切るというのは合理的ではない。疎密を判定する閾値を各ネットワークにおいて設定する必要がある。

さらに情報関係が二種類以上の部分で区切った場合、その関係をいかに上の階層に使用するかという問題が生じる（図８）。

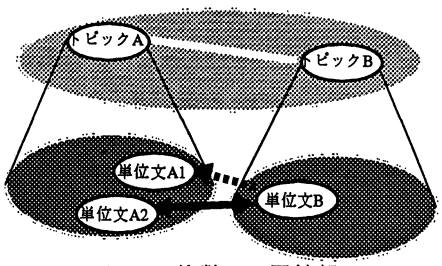

図８　複数の下層情報

単位文A1と単位文A2はネットワーク上である閾値以上の関係である。しかしこれらと単位文Bは一定の閾値以下の関係である。そこでこの部分で区切りが行われるが、この部分には並列と添加の二つの情報が存在している。トピック１の階層においてトピックAとトピックBの関係は下層における二種類の情報を考慮する必要がある。

5　まとめ

物語を最小限の表現に落としたレベルから意味のまとまりにより階層的な言い替えを行い、文章構造を理解する手法を考えた。その意味のまとまりを抽出するためにネットワークを利用する手法を提案した。

ネットワークの特徴として時間的に直線的でない内容に対応できる。物語文において時間経過は常に一様でなく事象の時間的重複、曖昧性などが散見される。このような時間に依存しない表現に対し、ネットワークは有効である。

階層構造であるため上の階層を形成する際に新たなネットワークをどのように形成するかという問題が生ずる。この問題については下層の情報をそのまま利用することとした。

具体的には

1　単位文間の関係を付ける
2　グルーピングを行う
3　トピック１の階層を形成する
4　新たなネットワークを形成する
　（関係の情報は持ち上げられる）

最終的なエピソードの階層に到達するまで、この処理が繰り返される。

課題としては階層を重ねていくうちに情報がまとめられていき、初期に与えた単位文間の関係をそのままネットワーク構造として利用することに困難が生じていく可能性があること、複数の関係情報のより有効な活用などが上げられる。

[1] 近藤恵子,荒井良徳,上里福美:
要約のための文章の部分的言い替えの手法,
NLC95-62(1995).
[2] Schank,R.C. and Abelson,R.P.:
Scripts,Plan,Goals and Understanding,
Lawrence Erlbaum Associates(1977).
[3] 中沢俊哉,重永実:
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).
[4] 松本羊一:国文法,金羊社.