文書データベース検索支援のためのtf.idf法による要約文の自動抽出法の提案

室工大紀要第50号(2000)167～174

文書データベース検索支援のためのtf.idf法による要約文の自動抽出法の提案

蓮井　洋志[*1]

The System for Automatic Extraction of Abstract by tf.idf Method to Assist in Searching the Document Database

Hiroshi HASUI

(原稿受付日　平成12年4月28日　　論文受理日　平成12年8月31日)

Abstract

I have studied the tf.idf method as the automatic extraction of abstract in order to assist in retrieving the document database. I defined sentence importance as the sum of the word importance of tf.idf value, and the system extracted the several sentences which are the most important. In this paper, I propose the extended tf.idf method in order to get more comprehending abstract. It is the method which adds three ideas to the tf.idf method, the case weight, the eliminataion of demonstrative words and the conjunctions, and the decision of the region for the extraction. Result of the experiment shows that the abstract of the extended method is more comprehensible than one of the tf.idf method.

Keywords: tf.idf method, retrieval system, extraction of abstract, case weight, query keyword

1　はじめに

ワードプロセッサやパーソナルコンピュータなどの電子機器の普及によって，現代社会には膨大な量の電子化文書のコレクションが存在する．文書データベースシステムはこれらの文書コレクションを管理して，欲しい情報をそれらから得るためのシステムである．大規模な文書データベースシステムでは，欲しい文を取り出すのに，すべての文書を読むことは不可能に近い．そこで，必要な文書を取り出すためにキーワードによる分類，検索という技術が開発されてきた．しかし，分類や検索を利用しても結果として得られる文書の数が多く，人手でそれらの中から欲しい文書を選ぶのには大変な手間がかかる．文書を選び出す時には，要約文があると便利である．要約文を見れば本文の内容を読まなくても欲しい文書が分かる．学術論文のデータベースは著者が用意するが，他の多くのデータベースには要約文が存在しない．

[*1] 情報工学科

著者はこれまでに文書データベースシステムにおける検索支援のために，要約文を自動的に抽出するtf.idf法[1]を提案した．文の重要度を文中の単語のtf.idf値[2, 3]の合計であると定義し，重要度の大きい文を抽出する．本論文では，よりわかりやすい要約文を作るために，tf.idf法に対して以下の3つの工夫をした方式を提案する．

(1) 格助詞の主辞である単語の重要度には重みを加える．(格の重みづけ)

(2) 指示語，接続詞などを除去する．(指示語，接続詞の除去)

(3) 文書中の重要な部分を特定し，その部分だけを要約する．(要約部分の限定)

本研究ではtf.idf法とそれを拡張した手法の精度を評価するために，拡張法を実現したExtraSummaryと呼ぶ要約文抽出システムを開発した．朝日新聞の「社説」[4]を集めた文書コレクションを活用して，文書データベースを作った．その検索システムにExtraSummaryを組み込んだ．検索システムおよびExtraSummaryはインターネットのホームページ上でCGIを利用して作成した．

本論文では，2章ではtf.idf法について，3章ではそれを拡張した方法について述べる．次に，4章でExtraSummaryの構成について述べる．5章ではこのシステムが抽出した要約文の精度および検索支援効果の評価実験とその結果について考察する．7章で内容をまとめる．

2　tf.idf法による要約文の抽出

2.1　要約文抽出法

要約文の作成方法には抽出法と生成法がある．生成法は次のような手順をとる．まず，自然言語解析の結果から文の意味情報をとりだし，冗長な文や文節を取り除くことで要約文を生成する．意味情報をとりだすために，形態素解析，統語解析，意味解析，文脈解析などが必要である．これら4つの自然言語解析は不正確さが故に実用的でない．また，これらの解析には辞書が必要であるが原文の中の語が未登録の場合は処理できない．

これに対して，抽出法は，まず出現する単語の統計量あるいは接続詞，助詞などの単語の言語的手がかりを利用して文の重要度を決定する．抽出法は，形態素解析だけを利用する．現在，最新の形態素解析システムの精度は95%前後である．この数値は信頼に足るものである．

2.2　tf.idf法

2.2.1　文の重要度

tf.idf法における文の重要度は，文を構成する単語の重要度の和である．要約文抽出システムは対象とした文書内の文の重要度の大きいものからユーザが指定した数だけ抽出する．以下に文の重要度を表す式を示す．文を構成する単語の数をnとし，各々の単語に番号をつけた．$Word_i$は文の中のi番めの単語の重要度を表す．

$$Sentence = \sum_{i=1}^{n} Word_i \quad (i = 1, 2, \ldots, n)$$

$Word_i \leq 0$であるために，多くの単語を含む程，文の重要度がより大きいということを示している．短い文は情報量が少ないために，要約文として抽出されても前後の文脈から，意味が理解できない場合が多い．長い文を選ぶことで，理解しやすい要約文を作る狙いがある．

2.2.2　単語の重要度

単語の重要度はtf.idf式で定める．tf.idf式は文書データベースにおける語の重要度を表す．対象としている文書内での単語の出現回数をtf，文書データベース内においてその語を含む文書の個数をdfとする．dfを文書データベースに登録された文書数で割った値に対して，負の対数をとったものをidfとする．この時の対数はeを底とする．tfとidfを掛け合わせた数がtf.idf式の出力である．

重要度は名詞だけに与える．日本語の文書は，話題を主に名詞で表現する．助詞や助動詞などは，直接的にしか話題に関連しない．

単語の重要度$Word_i$を以下のように定義する．

$$Word_i = \begin{cases} tf \times idf \\ \quad (forNoun) \\ QueryKeyword_i \\ \quad (forKeyword) \\ 0 \quad (Otherwise) \end{cases}$$

$$QueryKeyword_i = MaxTermFrequency \times log_e AllDocumentNumber$$

$$tf = TermFrequency$$

$$idf = -log_e\left(\frac{DocumentFrequency}{AllDocumentFrequency}\right)$$

TermFrequency: 対象とする文書中のi番目の単語の出現回数のこと

DocumentFrequency: 文書データベースの内単語が出現した文書数のこと

AllDocumentFrequency: 文書データベースに登録されたすべての文書数のこと

QueryKeyword: データベース検索時にユーザが指定した単語のこと

MaxTermFrequency: 対象とする文書中で一番頻出した単語の出現回数のこと

文書データベースで多くの文書に出現する名詞の *idf* は小さい．また，少数の文書にしか現れない名詞は *idf* は大きい．つまり，文書データベースの中で少数の文書にしか現れない名詞の重要度が大きい．文書データベース検索システムでは検索結果から欲しい文書を選択する必要がある．検索を支援するためには，要約文が他の文書と異なった部分を示していることが大切である．そのために，少数の文書にしか現れない名詞が重要だとされる tf.idf 式を利用した．

検索で指定されたキーワードはユーザの興味を端的に表している．要約文の中には興味のある情報についての詳しい記述が必要である．キーワードを含んだ文を多く抽出するために，キーワードの重要度は，その文書の最大の出現回数と，総文書数に対して底 e で対数をとった値の積で，同一文書中の他の単語の重要度より大きい値である計算になる．

本論文では，この方法で単語の重要度を計算する方法を原法と呼ぶ．

3 tf.idf法の拡張

原法で抽出される要約文は，一般に直接話題を表した文が少ない．精度の向上のために，これから述べる3つの改良を加えた．これらの改良を加えた方法を拡張法と呼ぶ．

3.1 格の重みづけ

拡張法は原法と単語の重要度の与え方が違う．単語の重要度を格の種類によって重みを加えた．このことを格の重みづけとよぶ．以下に，重みづけを行なった場合の単語の重要度の計算式を示す．

$$Word_i = \begin{cases} tf \times idf \times CaseWeight\ (forNoun) \\ QueryKeyword_i \times CaseWeight\ (forKeyword) \\ 0 \quad (Otherwise) \end{cases}$$

$$QueryKeyword_i = MaxTermFrequency \times log_e AllDocumentNumber$$

$$CaseWeight = (1.0 + CF \times PN)$$

$$PN = PhraseNumber$$

$$CF = \frac{CaseFrequency}{AllCaseFrequency}$$

$$tf = TermFrequency$$

$$idf = -log_e\left(\frac{DocumentFrequency}{AllDocumentFrequency}\right)$$

CaseWeight: 格の重み

PhraseNumber: その単語が含まれる文の中の文節数

CaseFrequency: 文書データベース中の文書のタイトルに含まれる単語の格の数

AllCaseFrequency: 文書データベース中のすべての格の数

名詞は接続する助詞の種類によって，その名詞の文の中での役割がわかる．特に，格助詞を含んだ文節は文の意味を決定するうえで重要な役割を持っている．接続する助詞によって名詞が文の話題に関連する度合が異なる．重要な単語が話題となる文は重要である．

拡張法の単語の重要度は，原法の単語の重要度を $1.0 + CF \times PN$ 倍した値である．CF は単語が構成する文節の助詞の格の重みで，データベース内の文書の格の総文節数に対する，タイトルの中の単語を含む文節の格の割合である．各々の格に対して，データベース内の文書から CF を統計的に定める．

PN は文節数を表す．文節を隣接する自立語連鎖と付属語連鎖を結合した文字列と定める．自立語連鎖とは自立語の連続した単語列で，付属語連鎖とは付属語の連続した列である．文節数 PN が大きい文は多くの情報を持っているために重要である．

3.2 格の重み CF の計測

朝日新聞「社説」4926件を対象として CF 値の計測を行なった．これらの「社説」は1985年から1988年にかけて朝日新聞に掲載されたものである．

格助詞が提示助詞「は」と「も」と連結した文節は別の格とみなした．例えば，「に」と「には」は別とした．また，格助詞が後続しない「その他」という文節も一つの格とみなした．

表1に計測した CF を示す．全体的に見て，「とは」，「には」などの提示助詞「は」がついた助詞の重要度が大きい．これは，「は」は話題を提示する助詞であるためだ．その他には，「が」，「を」などを含んだ文の CF が大きい．これは，文書を構成する上で，要となる主格，目的格を表すからだ．

また，格助詞を含まない文節は，CF は0.04で「は」や「が」の半分くらいである．「まで」や「から」と比較すると，かえって格助詞を含まない文節の CF の方が大きい．先行する単語が主題に関連している度合は助詞によって異なる．

表 1:「社説」データベースにおける CF

格	CF	格	CF
とは	0.10	には	0.08
では	0.06	からは	0.04
までは	0.01	へは	0.00
と	0.06	に	0.06
が	0.07	は	0.07
で	0.05	も	0.03
を	0.07	から	0.04
まで	0.02	へ	0.12
とも	0.04	にも	0.05
でも	0.05	からも	0.04
までも	0.00	へも	0.00
その他	0.04		

3.3　指示語，接続詞の除去

拡張法は，要約文として抽出した文から指示語，接続詞を除去する．指示語は隣接した文に先行詞がある場合が多い．また，接続詞は隣接した文の間の関係を表す場合が多い．抽出法によって作られた要約文は文書中の文の数より要約した文の数の方が少ないから，抽出した文と隣接した文が要約文の中にない場合が大半である．その結果，取り出された要約文は，指示語の先行詞がわからないし，接続詞による文間の関係が理解できない．指示語，接続詞はない方が理解しやすい．

3.4　要約部分の決定

拡張法は，文書の最初 10 文と最後 10 文の部分だけを抽出対象とする．もし，12 文目の文が重要度が大きくても抽出されない．説明文は，一般に話題が 1 つの文書は，序文，本文，結論で構成されている．内容を提示する問題提起があり，その問題に対する議論があり，最後には内容をまとめ，意見を述べる結論がある．要約文はこの中の問題提起と結論を明確に伝えればよい．

4　文書データベース検索システム

4.1　検索システムの構成

本研究では，要約文抽出システムをサブシステムとする文書データベース検索システムを開発した．その検索システムの構成を図1に示す．この文書データベースは朝日新聞「社説」4926件を保持する．このシステムはインターネットホームページブラウザと検索システムと要約文抽出システムからなる．検索システムはブラウザの CGI として動き，UNIX 上で Perl 言語をもとに実現した．要約文は ExtraSummary に

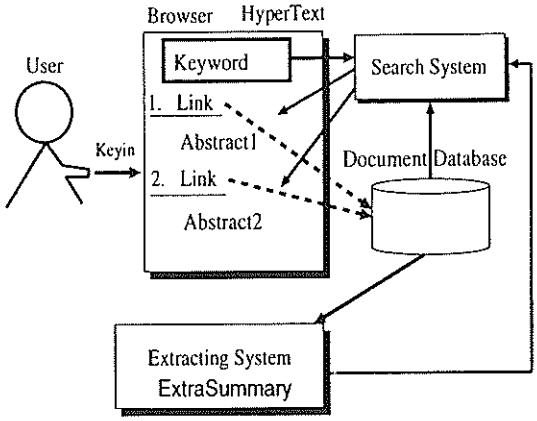

図 1: 検索システムの構成

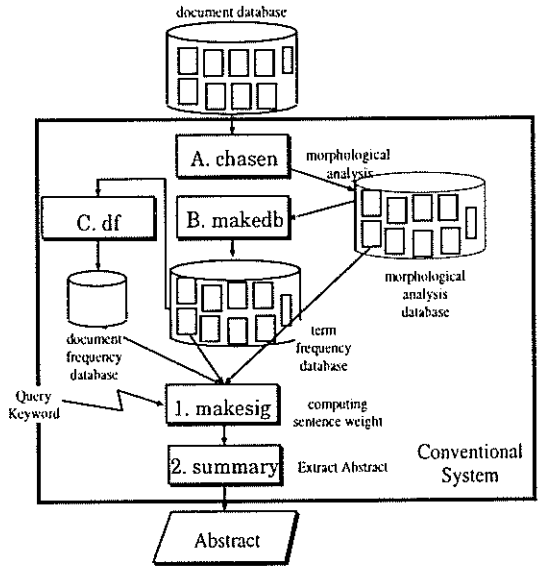

図 2: 原法のシステム構成

よって，あらかじめ抽出し，要約文データベースに登録する．

検索処理の流れでは，まず，ユーザがブラウザに検索キーワードを入力する．ブラウザはそのキーワードを検索システムに伝える．検索システムはキーワードが含まれる文書を取り出し，ExtraSummary が要約文を抽出する．要約文の長さは 4 文とした．付録 A にその例を示す．検索結果の文書における検索キーワードの tf.idf 値を計算し，大きいものから順位づけする．その順番に従ってブラウザで表示する．

4.2　要約文抽出システムの実現

(1) 原法 (Conventional System)

原法の構成を図 2に示す．

原法に必要な処理プログラムは次の 3 つである．

A. chasen[5][1] (形態素解析システム)

B. makedb (tfデータベースを作る)

C. df (dfデータベースを作る)

1. makesig (文の重要度の計算)

2. summary (重要度の大きい文の抽出)

原法で要約文を作るために必要なデータは次の2つである.

- Query Keyword (検索キーワード)
- document database (文書データベース)

処理途中で必要になるデータは, 次の3つである.

- tfデータベース
- dfデータベース
- 形態素解析データベース

文書データベースの各々の文書をA.で形態素解析する. その結果を用いて, 単語とその出現回数を対応づけたtfデータベースを作る. これは, B.が行なう. そのtfデータベースを利用してC.が値を数える. その結果をdfデータベースに保存する. 1.が*tf.idf*値, 形態素解析の結果から文の重要度を計算する. 最後に2.がその結果から重要度の大きい文を抽出し, それを文書の先頭から順番に並べたものを要約文として出力する.

プログラムA, B, Cは, 検索前に起動し, dfデータベース, 形態素解析データベース, tfデータベースを作っておく. 検索の要求があると, 検索システムは, 要約文を作る段階で1., 2.が抽出処理をする. 本手法は名詞のみに重要度を与える. 本システムでは, A.で使用された品詞体系の中の普通名詞, サ変名詞, 固有名詞, 未定義語を名詞とみなす.

(2) 拡張法 (ExtraSummary)

拡張法のシステム構成を図3に示す.

原法に加えて, 処理プログラムとしてはD. case, 処理途中で必要となるデータとしてはcase tableを追加する. D.は格の重みを計測し, case tableに登録する.

拡張法は, データとしては文書のタイトルと, 処理プログラムのcaseを持つ. 文書データベースからcaseによって, 格の重みを表すcase tableを作成する. 拡張法のmakesigはそれを用いて単語の重要度を計算する.

[1]奈良先端科学技術大学松本研究室で開発されたフリーウェア

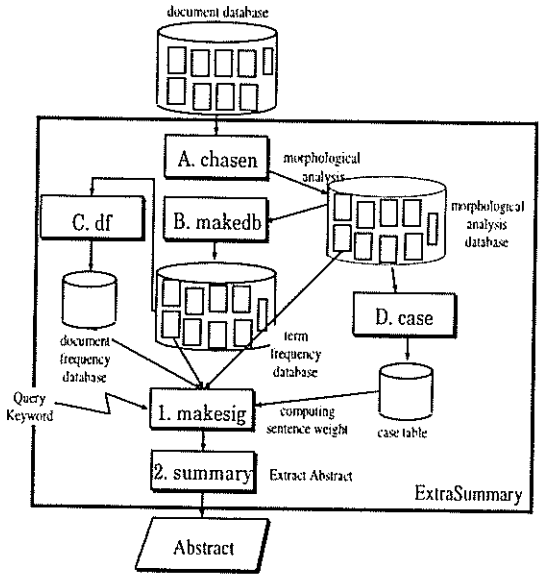

図 3: ExtraSummary のシステム構成

拡張法では, 検索の前に, 原法の処理に加えてcase tableも作る.

5 評価実験

5.1 評価基準

このシステムで抽出した要約文の読みやすさ, あるいは原文の内容に対する忠実さを評価した. 評価に用いた基準は次の6つである. なお, 要約文を評価する被験者は3人である.

(1) 要約文の内容が把握できるか (読みやすさ)

- a. 話題が掴める
- b. 話の流れが掴める
- c. 何の話題か第1文目から理解できる
- d. 結論があり, 文章の終りであることが理解できる

(2) 原文に忠実か (原文に対する忠実さ)

- a. 原文の問題提起と同じか
- b. 原文の結論と同じか

(1)の4項目は, 要約文の内容の理解しやすさをはかる. (1) b., d.は要約文の話の流れや文脈の理解

表 2: 要約文の精度

抽出法の種類	(1)				(2)	
	a.	b.	c.	d.	a.	b.
被験者 1						
拡張法	15	13	9	15	5	3
原法	15	13	8	14	6	1
被験者 2						
拡張法	13	13	9	15	8	5
原法	14	11	10	13	10	10
被験者 3						
拡張法	15	14	10	14	6	6
原法	14	7	10	9	6	2
合計						
拡張法	43	40	28	44	19	14
原法	43	31	28	36	22	13

しやすさを判定する．理解しやすい要約文を作るには文脈が理解できなければいけない．

(2) の 2 項目は要約文が原文の内容を反映している度合を評価する．朝日新聞の「社説」は問題提起，議論，結論で構成される．この中で，原文と要約文の問題提起と結論の内容が一致すれば，原文の内容を忠実に表している．

5.2　評価結果

朝日新聞の「社説」4926 件をもとにした文書データベースにおいて，15 個の文書に対して原法，拡張法の 2 種類で要約した．この 15 個の文書は「関西国際空港」をキーワードとして検索された文書である．

これら 2 種類の要約文に対して 3 人の被験者が精度を評価した．結果を表 2に示す．

全般として，読みやすいが，原文に対する忠実度は低い．文脈に関わる (1) b. d. は拡張法が原法より低い結果がでた．つまり，拡張法は原法と比較して文脈のとりやすい要約文を抽出する．

5.3　検索支援の効果

次に，拡張法を利用して作った要約文がユーザの欲しい文書の選択を支援する効果について評価する．20 個の文書が検索結果となる 5 種類のキーワードを用意する．これらのキーワードで検索した結果，得られる文書群の中から 1 つの文書をユーザに提示する．その後，検索結果となった 20 個の要約文を被験者に見せ，被験者が最初に見た文書がどの要約文と対応しているか判断させる．

これは，ユーザがキーワードで検索した結果，得られた 20 個の要約文の中から 1 つ欲しい文書を選び出す状態をシミュレーションしている．選択に成功した文書数を表 3に示す．結果は 3 人とも正解率が 100%であった．明らかに検索支援の効果がある．

表 3: 検索結果の正解文書数

	正解文書数	正解率 (%)
被験者 1	5	100
被験者 2	5	100
被験者 3	5	100

6　考察

要約文抽出法の中には，単語の重要度を用いる方法 [6]，[7] と，助詞，接続詞などの修辞表現を利用する方法 [8] の 2 種類がある．単語の重要度を利用した方法は，普通，単語の重要度を利用して文の重要度を定義し，その重要度を利用して要約文を抽出する．文献 [6] は文の重要度の変化量から文章をセグメントに分割し，セグメントごとに重要度の大きい文を抽出する．これは，文章の話題の区切り目で出現する単語が変化することに着目した方法である．複数の話題を含む文の要約に向いている．それに対して，後者 [8] の抽出法は，文と文の関係，段落と段落の関係を接続詞，助詞から，主題を表していると推測できる文を抽出する．

tf.idf 法は前者のタイプに属する．重要度は名詞のみに与え，tf.idf 式によって計算する．重要度が大きい文は tf.idf 値の大きい名詞を多く含む．tf.idf 値の大きな名詞は少数の文書にしか出現しないために，絞り込みやすい検索キーワードである．要約文の中から，ユーザが再検索に有効である検索キーワードを見つけることもできる．

英語では，要約文抽出法として tf.idf 法 [9] が発表されている．英語は単語の間にスペースがある．しかし，日本語は膠着語であるために単語の分割から行なわなくてはならない．原法，拡張法の両方とも名詞だけに重要語を与えたが，英語の tf.idf 法はテーブルに登録された前置詞など付属語の単語以外の語に重要度を与えている．これは，英語の単語の大半が多品詞語で，単語の品詞を特定する技術が完全でないためである．原法，拡張法の両方とも形態素解析をするために，完全ではないが品詞は特定できる．

7　まとめ

本稿では，文書データベースの検索を支援するための要約文抽出法として，tf.idf法について説明し，それを拡張した方法を提案した．そして，それらを実現した要約文抽出システムを開発し，それを評価した結果について述べた．その結果，以下の2点が明確になった．

- 拡張法は原法より文脈のとりやすい要約文ができた．
- 拡張法は文書の検索を支援する効果がある．

拡張法では「社説」の最初の10文と最後の10文を要約の対象とした．これは，文書の話題を最初に述べ，最後には内容をまとめた文を記述することを利用している．しかし，文章の記述方法は一つではない．「社説」以外の文書の要約では，文書の要約部分を変更する必要がある．今後，文書の中の重要部分を自動的に特定する方法を考えることで，より良い要約文を作る方策を考える．

A　要約文の例 A.1 原文：

【'85．4．14　朝刊　5頁　(全1539字)】

毎年四百万人を超す日本人が海外へ旅行する。南極、北極、エベレストへも行く。ところが辺地でもないのに、外国人が極めて入りにくいのがアルバニアだ。「純粋マルクス・レーニン主義」の旗のもと、資本主義や修正主義という外からの風を入れまいと、突っぱりぬいて四十年になる。

体制のちがいを超えた国際交流や経済協力が世界の常識になっている今、こういう国が存在し続けていることは一種の奇跡とも思える。地中海に面したバルカン半島の一角。九州の四分の三ほどの国土に、二百七十余万の人口。このふしぎな国を支配してきた労働党のホッジャ第一書記が十一日死去した。その死によって、アルバニアにも必ず変化が起こるだろう。しかし鎖国から開放への道は必ずしも容易ではあるまい。

ホッジャ政権の四十年は、三つの修正主義（ユーゴスラビア、ソ連、中国）、米ソ中の三大国との闘争だった。

第二次大戦末期、ナチス占領下のアルバニアで、ホッジャ氏は隣国ユーゴのチトーひきいる共産ゲリラに助けられて、祖国を解放する。しかし、チトーが自主・民族路線を打ち出すと、ホッジャ氏は一九四八年、ソ連とともにこれを修正主義と非難して決別する。続いて六〇年代に中ソが対立すると、毛沢東とともにクレムリンを修正主義と断罪する。その中国が七〇年代に米国と接近すると、これも修正主義に転落したと、手を切った。ではホッジャ氏の純粋社会主義とは何なのか。七六年制定の憲法はいう。

「プロレタリアートの独裁」。「自力による建設」。「土地から工場、銀行、ラジオ、映画までの国有化」。「外国の経済・金融会社、修正主義・資本主義独占家ないし国家と共同する企業を認めず、彼らからの借款を禁ずる」。「税金を支払うことはない」。「教育と医療は無償」。「いかなる宗教も認めない。無神論の宣伝を支持する」。

自給自足をめざしても、この小さな途上国では限界がある。自転車も十分に出回らず、トラクターの役目は人海戦術でまかなう。そのために「生めよ殖やせよ」と人口増加をはかった結果、食糧難があらわれ、生活水準は改善しない。「美しい山河、みじめな服装」と西側からの旅人はいう。

しかしモノの乏しさより、もっと重大なのは「言論・報道・集会・結社の自由」がないことだろう。憲法五三条はこれを保障しているが、「社会主義秩序に反して行使してはならない」。さる八一年、ホッジャ氏の盟友シェーフ首相が「自殺」したが、あとで「処刑」と分かる。党首脳でさえ言論・思想の自由は、命と引きかえなのである。

右であれ左であれ、長い独裁下の国民は、あきらめから慣れに変わり、抑圧が抑圧と感じられなくなる。ましてホッジャ氏は去っても、労働党の一党支配がすぐに揺らごうとは思えない。むしろ危機感から、国内外での引きしめが一時的には強まるかもしれない。ソ連共産党からの弔電を突っ返したとの報道は、それを予感させる。

しかし一方で、中国と衝突した七〇年代後半から、イタリア、ユーゴ、東欧圏などと貿易を広げ始めたのも事実なのだ。八一年にわが日本と国交を結んだのも、そうした微妙な変化の上にある。八三年には中国とさえ貿易協定を結んだ。

ほんらいアルバニアは国連の一員で、米ソなどを除く約百カ国と外交関係をもつ。それなのに鎖国状態を続けるほうが変則なのだ。その結果が明らかになってきた今、開国という新しい挑戦を始めるべきときではないか。

A.2 要約文：

「純粋マルクス・レーニン主義」の旗のもと、資

本主義や修正主義という外からの風を入れまいと、突っぱりぬいて四十年になる。支配してきた労働党のホッジャ第一書記が十一日死去した。ホッジャ政権の四十年は、三つの修正主義（ユーゴスラビア、ソ連、中国）、米ソ中の三大国との闘争だった。ましてホッジャ氏は去っても、労働党の一党支配がすぐに揺らごうとは思えない。

参考文献

[1] 蓮井洋志, 魚住超, 小野功一. 検索要求語を利用した tf.idf 法による要約文抽出. 情報処理学会第57回全国大会講演論文集 (1998),pp 3/245–246.

[2] 大森信行, 岡村潤, 森辰則, 中川裕志. tf.idf 法を用いた関連マニュアル群のハイパーテキスト化. 情報処理学会自然言語処理研究会, No. 16(1997).

[3] 長尾真, 佐藤理史, 黒橋禎夫, 角田達彦. 自然言語処理. 岩波ソフトウェア科学 15. 岩波書店 (1996).

[4] 朝日新聞. 朝日新聞:天声人語・社説:1985-1991. Electric Book. 日外アソシエーツ (1992).

[5] 松本裕治, 北内啓, 山下達雄, 平野善隆, 今一修, 今村友明. 日本語形態素解析システム「茶せん」version1.5 使用説明書. (1997).

[6] 任福継, 定永靖史. 統計情報と文書構造特徴に基づく重要文の自動抽出. 情報処理学会自然言語処理研究報告, Vol. 98, No. 48(1998), pp 71–78.

[7] 平尾努, 木谷強. 単語の重要度に基づくテキストの要約. 情報処理研究会データベースシステム・情報基礎研究報告, Vol. 98, No. 34(1998), pp 41–47.

[8] S. Miike, E. Itoh, K. Ono, and K. Sumita. A full-text retrieval system with a dynamic abstract generation. *SIGIR '94. Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, (1994), pp 152–61.

[9] Zechner, K. Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences. *Proc. COLING-96*, (1996), pp. 986-989.