自然言語処理 129－15
（1999. 1. 21）

句表現要約の句合成手法

岡 満美子，小山剛弘，上田良寛

E-mail: {mamiko, koyama, ueda}@rsl.crl.fujixerox.co.jp

富士ゼロックス(株) 総合研究所

検索結果のふるい分けを、ユーザに負担を与えずに短時間で正確に行うことを目的として、「読む」のではなく「一目で分かる」要約、At-a-glance 要約を目指している。その一実現形態として、句表現要約手法を考案した。これは、重要概念を含む短い「句」を列挙することにより、文書の概要を示すものである。これらの句は、①テキスト中の単語と単語の関係を解析し、②重要概念を表す関係をコアとして選択し、③意味の特定度を上げ、句に意味的なまとまりを持たせるために必要な関係を補完する、というプロセスによって合成される。ここでは、②でどのような関係を選択し、③でどのような関係を補完するかを中心に述べる。現在、この手法を用いたプロトタイプを開発中であり、生成される要約のチューニングアップを行っている。

Phrase Construction Method for Phrase-represented Summarization

OKA Mamiko, KOYAMA Takahiro, UEDA Yoshihiro

Corporate Research Laboratory, Fuji Xerox Co., Ltd.

Summaries are used as a clue for sifting information retrieval results. "At-a-glance" summarization is the goal we set to reduce stress to read sentences and serve fast and accurate sifting. We have developed the phrase-representation summarizing method as a realization of "at-a-glance" summarization. Summaries are presented as a pile of rather short phrases that contain important concepts. Each phrase is constructed by (1) analyzing the original text to a collection of relations between two words, (2) selecting an important relation as a core, and (3) attaching auxiliary relations to the core to bring a minimal sense. Here we describe (2) what relation is selected and (3) what relations must be attached. A prototype based on this method is now under development and the phrase construction rules are tuned up.

1　はじめに

要約の目的には、indicative なものと informative なものの二種類がある。Web の検索結果のふるい分けなどに使われる indicative な要約に求められる要件は、できるだけ短時間で、できるだけユーザに負担を与えずに、要約を処理し、タスクを成し遂げ得ることである。

現状の自動要約技術は、単語頻度や出現位置を用いてピックアップした重要文から抄録を作成する手法が中心であり、本文中の文がそのまま用いられることが多い。このようにして作成された要約文を用いて文書のふるい分けを行う場合、自分の要求に合った文書かどうかを判断するのには、要約文を頭から「読む」必要がある。長くて複雑な構造を持つ文や、離れたところからピックアップされ、形態的にも意味的にもつながりが悪い文を読解することは、ユーザに負担を与え、判断に時間を要する。

このように、従来の「読む」要約は indicative な目的には適していない。我々は、ユーザに負担をかけない「一目で分かる」要約を目指して、「At-a-glance 要約」の研究を開始した。

本論文では、At-a-glance 要約のひとつの実現手法として「句表現要約」を提案し、その句合成手法について述べる。

2　句表現要約とは

At-a-glance 要約のひとつの目標となるのは、電車の中吊りで見られる雑誌広告である。中吊り広告は、記事本体を読むかどうかを判断する indicative な情報であり、各記事の内容が見出し的に書かれている。これらの見出し文は、長さが短く、構造が単純なのが特徴である。

我々は、この短さ、単純さを、通常の「文」とは区別して、「句」という言葉で表している。句表現要約とは、重要概念を含む「句」の並びによって文書の概要を表現するものである。

句表現要約では、単語と単語の関係を概念の基本単位とし、その組合せによって句を合成する。単語と単語の関係を基本関係と呼び、次の三つ組で表す。

< 関係名　構成要素 1　構成要素 2 >

関係としては、キーリレーションに基づく検索(岡他 1994; Miyauchi, et al. 1995)でも用いた係り受け関係を用いる。

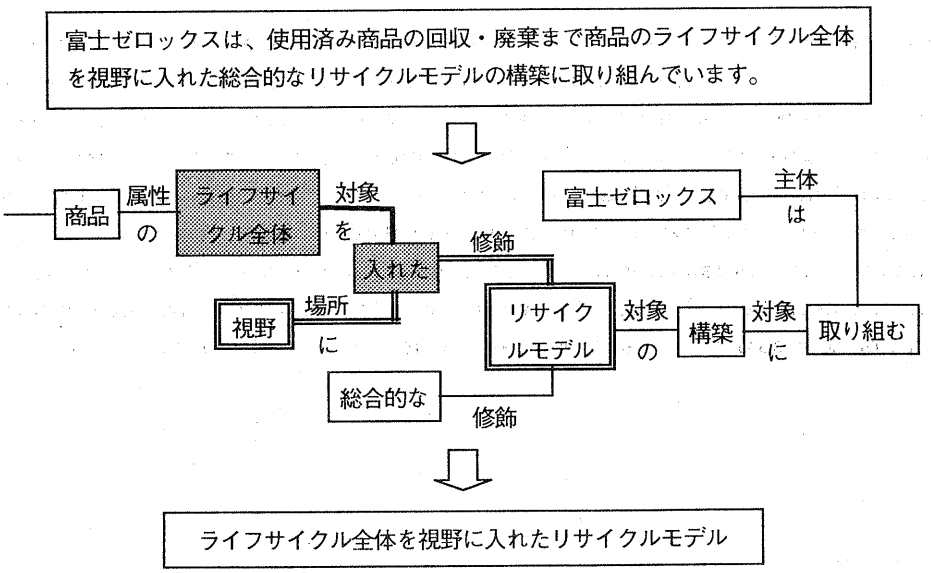


図 2.1　要約句生成の概略

図2.1は、文が基本関係のネットワークで構成されている様子を示すものである。グレーの部分は、重要概念を表す基本関係である。これが句の核になる部分で、コア関係と呼ぶ。二重線で囲んだ部分は、句の表す意味の特定度を上げるために、コア関係に補完される関係である。これらの基本関係から、次のような要約句が合成される。

《ライフサイクル全体を視野に入れた
リサイクルモデル》

このようにして合成された短い句を複数並べることにより文書の概要を把握させることが、句表現要約の基本的な考えである。

3 アルゴリズム

句表現要約の基本アルゴリズムを図3.1に示す。

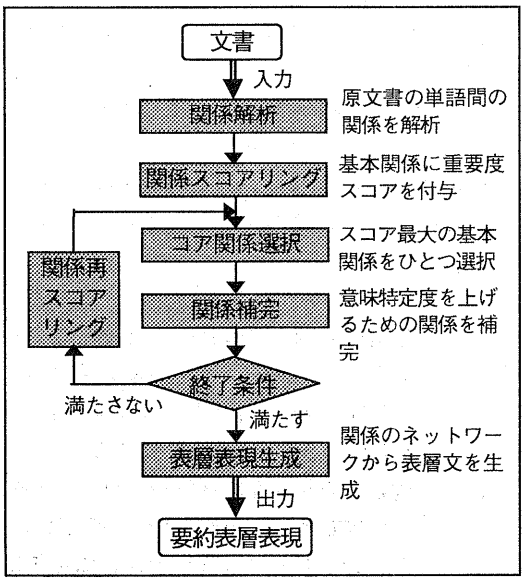


図3.1 アルゴリズム基本構造

以下、各ステップについて簡単に説明する。

関係解析

原文書の単語間の係り受け関係を解析する。まず原文を形態素解析し、結果の単語列に対してパターンマッチを行うことにより関係を抽出する(Miyauchi, et al. 1995)。

基本関係のスコアリング

コア関係を選択するために、すべての基本関係に重要度スコアを与えておく。スコアは、構成要素の重要度と関係の重要度に基づいて計算される。詳細は第4章で述べる。

コア関係選択

スコアの最も大きい基本関係を、コア関係として選択する。ここでコア関係を複数選んでおくのではなく、要約句をひとつ合成するごとに重要度を計算し直し、次のコア関係を選ぶようにする。

関係補完

コア関係に対して、意味の特定度を上げ、要約句に意味的なまとまりをもたせるのに必要な関係を補完する。具体的にどのような関係を補完するかについては、第5章で詳述する。

終了条件判定

これまでに作った要約の量が十分かどうかを判断する。現在のところ、要約全体が含む基本関係の数を判定基準としている。

関係再スコアリング

終了条件を満たさない場合には、再びコア関係選択に戻って新たな要約句を合成する。その際に、要約に同じ単語が繰り返し登場するのを避けるため、前回の要約句に含まれる単語のスコアを落とし、基本関係重要度の再計算を行う。

表層表現生成

終了条件が満たされると、基本関係の組合せから要約の表層表現を生成する。まず、各要約句の構成要素を原文の語順に基づいて線形化し、活用語尾、助動詞、助詞等を追加する。さらに、要約句を原文での出現順にしたがって並べる。

4 基本関係のスコアリング

基本関係の重要度スコアは、次式によって計算される。

$$Score = S_{rel} * (W_1 * S_1 + W_2 * S_2)$$

ここで、S_{rel} は関係の重要度スコア、S_1、S_2 は係り側、受け側の各構成要素の重要度スコア、W_1、W_2 は係り側、受け側それぞれの重みである。なお、現在は $W_1=W_2=1$ であり、重みづけは行っていな

い。以下、構成要素の重要度と関係の重要度について説明する。

4.1 構成要素の重要度

単語の重要度は、一般的な方法である tf*IDF 積(Salton 1989)を用いる。構成要素は複合語を含むが、この場合、複合語の tf*IDF 積と構成単語の tf*IDF 積の和を比較し、大きい方を複合語の重要度スコアとする。

IDF は、文書集合全体から計算する必要があるが、Web 文書を対象とする場合、全体を規定することは困難である。現在のところ、新聞記事[1]1万文書から作ったものと Web 文書 100 万 URL から作ったものの2種類の df データを用意している。

4.2 関係の重要度

関係の重要度は、次の原則に従って決めている。

① 重要な意味を担いやすい関係の重要度を大きくする。

② 表層的に異なる関係でも、深層的に同じ関係を表し得るものは同じ重要度にする。

文の中心は動詞概念であり、文の骨格をなすのは、動詞の格関係である。したがって、動詞の格関係、およびそれと同等の関係を表す埋め込みなどを最も高いスコアに設定した。逆に、並列のように関係そのものにあまり意味がないものはスコアを低く設定した。詳細を表 4.1 に示す。

表 4.1 関係のスコア

スコア	関係	例(表層表現)
3	○動詞の格関係 ○動詞が名詞を修飾する関係 ○動詞がサ変名詞化したもの	アイデアを募集する 募集したアイデア アイデアの募集
2	○形容詞・形容動詞を含む関係 ○名詞間関係(並列以外)	新しいアイデア アイデアが新しい 騒音のレベル
1	○並列	鉄と銅
0	○副詞を含む関係 ○複合語・複合語相当語	すでに乗り込んだ 環境推進体制

5 関係の補完

コア関係に対して関係を補完する目的は、次の二つである。

① 句の意味の特定度を上げる。

② ひとつの句としての意味的なまとまりをもたせる。

まず、情報の必要性の判断が可能な程度に限定された意味を持つために、要約句が含むべき関係や要素として、次の三つを定めた。

1) 動詞的概念を含んでいる。

2) 用言は、必須格要素によって意味が限定されている。

3) 意味の抽象度が高い名詞は、修飾語などにより限定されている。

さらに、意味的なまとまりを持つために、要約句が含むべき関係や要素として、次のふたつを定めた。

i) 関係の中心となる語を含む。すなわち、用言の名詞への修飾や、必須格要素と用言といった結びつきの強い関係の、係り側の要素が要約句に含まれる場合、受け側の要素も含むようにする。

ii) 現状の表層的な関係解析では、係り先が特定できない場合がある。例えば「AのBのC」という表現のAの係り先は意味的な情報なしにはわからないため、現状では隣接するものを基本関係としている。このような表現の一部が要約句に含まれる場合、曖昧性を持つ部分全体(「AのBのC」)をまとめて取り込むことにより、不正確な要約句の生成を防ぐ。

一方、要約句の長さは「一目でわかる」程度に保っておく必要があり、補完される関係は必要最小限でなければならない。以上を考慮して、コア

[1] CD-毎日新聞 95 年版を使用。

関係の種類に応じて六種類の補完規則を作った。以下、それぞれについて簡単に説明し、表層表現の形で例を示す。例では、アンダーラインがコア関係に相当する部分であり、ボールドで示す部分は関係補完によって追加される要素である。

(a) 用言の必須格要素の補完

コア関係に用言を含む場合、その用言の必須格関係を補完する。日本語では用言ごとに異なる必須格をもつと思われるが、ここでは、[が][を][に]格を必須格とみなしている。また、副助詞が格助詞を置き換えて用いられる場合や無形格の場合、格の特定を行っていないので、[が][を][に]格に該当する関係を表し得るものとして、[は][も][の][無形]格も必須格とみなしている。

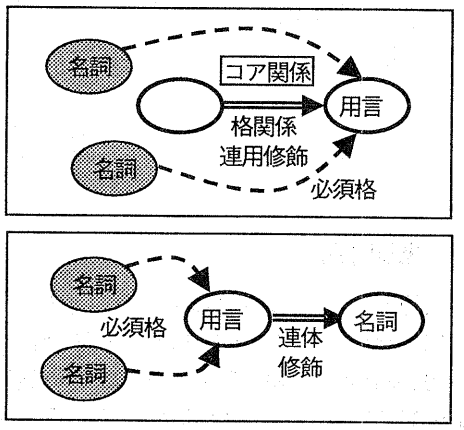


図5.1 必須格要素の追加

例 《Tiger Beer は**1931 年に**生まれた》
《**アラビア文字を**書いた土産物》

(b) 用言の連体修飾先の補完

コア関係の受け側が用言の場合、その用言が名詞を修飾していれば、その名詞を追加する。また、規則(c)(d)で動詞を追加した場合にも、それが名詞を修飾していればその名詞も追加する。

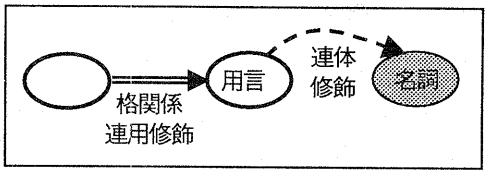


図5.2 用言の連体修飾先の追加

例 《お母さんが入院している**病院**》

(c) 必須格関係の用言の補完

コア関係の受け側の名詞が用言の必須格を占めている場合、その用言を補完する。

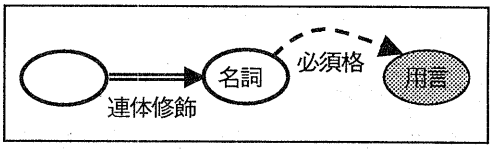


図5.3 必須格要素の用言の追加

例 《洗浄に使用していた**フロン類を削減**》

(d) 用言の補完

コア関係が名詞間関係の場合、受け側の名詞を格要素とする用言を追加する。規則(f)でコア関係の後ろに名詞を追加した場合も同様である。

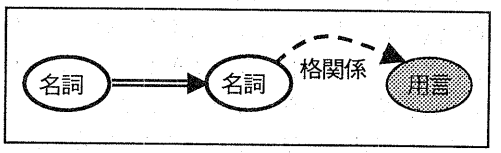


図5.4 用言の追加

例 《砂糖キビと塩**を使う**》

(e) 抽象的な名詞への修飾語の補完

「時代」「人」「場所」といった意味が抽象的な名詞に対して、修飾語を追加する。これは、規則(a)で追加された名詞に対しても同様である。

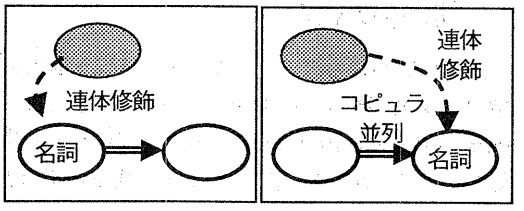


図5.5 抽象的な名詞への修飾語の追加

例　《**激動の**時代に活躍した》

(f)「の」を介した修飾先の補完

コア関係が「の」を介した名詞間関係の場合、受け側の名詞がさらに「の」を介して係る名詞を追加する。

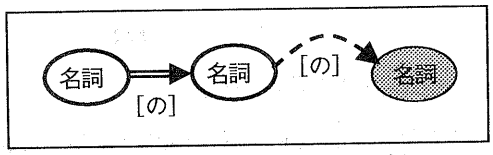


図5.6　「の」を介した名詞の追加

例　《玄関のアーチ型**の庇**》

6　特性評価

予備的な評価として、要約の特性評価を行った。新聞記事[2]をサンプルとし、句表現要約手法と重要文ピックアップとを比較した。評価項目は次の二点である。

① 同じキーワードをカバーするのに必要とする要約量

② 一文あたりの長さ

①の要約量は、本手法は重要文ピックアップの約 40%であった。これは、句表現要約により重要な概念をコンパクトに表現できることを意味し、読む時間を短くすることができるといえる。

②の一文あたりの長さの平均は、重要文ピックアップの 50.1 文字に対して、本手法は 18.4 文字であり、約 37%になっている。ある週の週刊誌の中吊り広告で、見出し文の文字数の平均を調べたところ、A誌 18.4 文字、B誌 15.4 文字、C誌 20.6 文字であった。したがって、目標とした中吊り広告程度の長さがほぼ実現できているといえる。

7　プロトタイプ

上述のアルゴリズムに基づいて、Web 文書を対象とした要約システム X-press(Xerox's Phrase-Represented Summarization System)を開発している。現在は、アルゴリズム検証用プロトタイプを作成中で、開発環境は IBM VisualAge for Java for Windows[3]を用いている。プロトタイプのユーザインタフェースを図7.1 に示す。

また、富士ゼロックスの Web ページに対して要約処理を行った実行結果を付録として添付する。

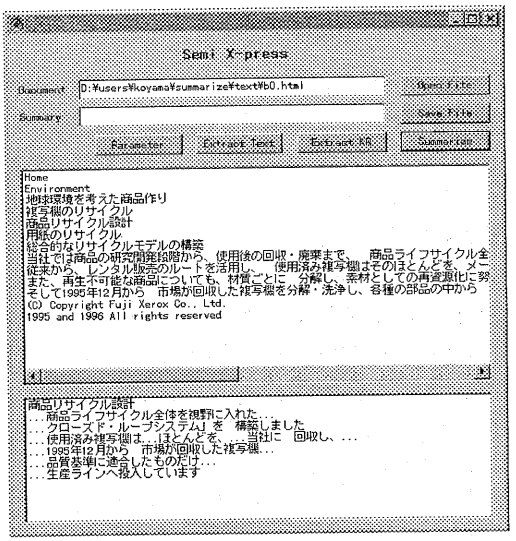


図7.1　プロトタイプのユーザインタフェース

8　関連研究

要約研究は、(Zechner 1996)をはじめとして重要文ピックアップによる方法が中心である。そこでの主要課題は重要文の選択方法である(奥村・難波 1998)。我々は、重要文ピックアップによる要約では読む負担が大きいことを問題にし、関係を組み合わせて句を合成する句表現要約手法を提案した。ここでは、句表現要約との関連から、読む負担が少ないという観点と、関係から文を合成するという観点を中心に関連研究を概観する。

短い文にするという点では、文の言い替えや修飾語の削除による文短縮がある。この方向では、TV ニュースの原稿から字幕を作成することを目的とした研究がある(若尾他 1998; 三上他 1998)。これらは informative な要約であるため情報をなるべく落とさないようにしている。Indicative

[2] CD-毎日新聞 95 年版を使用。

[3] VisualAge は米国 IBM Corp.、Java は米国 Sun Microsystems, Inc.、Windows は Microsoft Corp.の商標である。

な目的でもっと短くすることを考えると、文の中心構造を残して修飾部分を減らすというこの方法では、文の中心から遠い重要フレーズが選択されず、また短縮できる量にも限界がある。

文書読解支援の目的では、文自体を短くするのではなく、重要部分を可視化する研究が行われている。(亀田 1995)では、名詞性キーワードを原文と同じ位置に表示する Screening 支援が indicative な要約に相当し、文の骨格を中心に主要な文節を表示する Skimming 支援が informative な要約に相当する。Screening 支援のキーワードの表示は、キーワード間の関係がわからないため、概要を表すのには不十分である。また Skimming 支援は、上述した文短縮と同様の手法であり、同様の問題点をもつ。

語と語の関係をベースに要約を作るものには、(長尾他 1997)がある。これは、著者があらかじめ GDA(Global Document Annotation)という、関係を表す複雑なタグを付与しておくことにより、要約を含む種々の文書処理を可能にしようという試みである。最も困難な解析部分を人手で行うという点で、我々とは全く前提を異にする。一方、想定されている理想的な要約は句表現要約に近く、我々の研究はこれを自動生成することを目指すものとして位置づけられる。

9 おわりに

本論文では、At-a-glance 要約の一手法である句表現要約を提案し、その句合成手法を示した。

現在の課題は次の二つである。

① 要約プロトタイプのチューンナップ

② プロトタイプシステムを用いた評価

①に関しては、現在、関係解析の精度の向上、関係スコアの重みの調整などを行っている。

②に関しては、プロトタイプの出力を用いて検索結果のふるい分けを行う、タスクベースの評価実験を準備中である。

参考文献

Miyauchi, T., Oka, M. and Ueda, Y. (1995). "Key-relation technology for text retrieval." In *Proceedings of the SDAIR'95*, 469–483.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley Publishing Company, Inc.

Zechner, K. (1996). "Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences." In *Proceedings of the 16th International Conference on Computational Linguistics*, 986–989.

岡, 宮内, 上田 (1994). "キーリレーションに基づくテキスト検索." 情報処理学会研究報告 自然言語処理 103-12, 89–96.

奥村・難波 (1998). "テキスト自動要約技術の現状と課題." 言語処理学会第 4 回年次大会ワークショップ「テキスト要約の現状と将来」, 80–87.

亀田 (1995). "日本語文書読解支援系 QJR の検討." 情報処理学会研究報告 自然言語処理 110-9.

長尾, 橋田, 宮田 (1997). "GDA (Global Document Annotation)タグを用いた文書の要約に関する一考察." シンポジウム「実用的な自然言語処理に向けて」.

三上, 山崎, 増山, 中川(1998). "文中の重要部抽出と言い替えを併用した聴覚障害者用字幕生成のためのニュース文要約." 言語処理学会第 4 回年次大会ワークショップ「テキスト要約の現状と将来」.

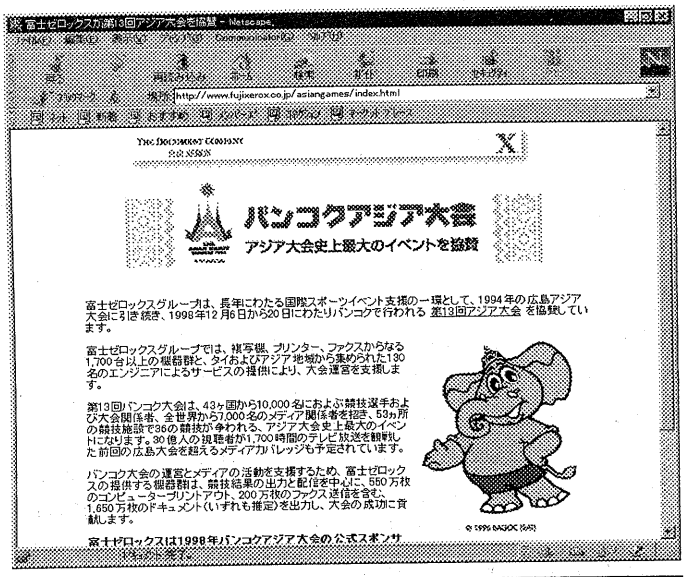
若尾, 江原, 白井 (1998). "テレビニュース字幕のための自動要約." 言語処理学会第 4 回年次大会ワークショップ「テキスト要約の現状と将来」論文集, 7-13.

付録：X-Press 実行結果

■ 例1

◇ URL

http://www.fujixerox.co.jp/asiangames/index.html



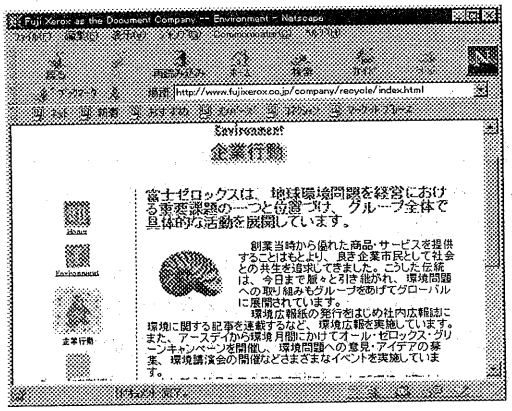
◇ 要約

富士ゼロックスグループは、...1998 年 12 月 6 日から 20 日にわたり...
...長年にわたる国際スポーツイベント支援の一環...
...広島アジア大会に引き続き、...
...13 回バンコク大会は、...競技が争われる、...
...200 万枚のファクス送信を含む、...
...1998 年バンコクアジア大会の公式スポンサーです

■ 例2

◇ URL

http://www.fujixerox.co.jp/company/recycle/c0.html



◇ 要約

...企業市民として...共生を追求してきました
...発行をはじめ社内広報誌に...記事を連載するなど、...
...環境広報を実施しています
...新入社員の集合教育プログラムの中...
...マネジメント層にも環境問題について再認識する場...