重要文抽出，自由作成要約に対応した新聞記事要約システム YELLOW

大竹清敬[†] 岡本大吾[††]
児玉充[††] 増山繁[††]

日本語の新聞記事を対象とした新聞記事要約システム YELLOW について報告する．YELLOW は，「重要な情報を洩れなく抽出する」ことに重点をおいて作成した．本システムは，二重修飾に着目した削除を中心とした文内要約と，重要度付与による文選択の 2 つの部分より構成される．文内要約では，構文解析結果を積極的に利用する．ある名詞に対し，複数の修飾部がある場合，名詞を限定する働きが弱い修飾部を削除する新たな手法を提案する．また，換言処理，例示の削除などの要約手法も用いる．重要度付与では，主要語，高頻度の名詞，位置情報，見解文であるか否かなど，従来，文の重要度を決定するにあたって重要であるといわれてきた種々の情報を，複合的に用いる．情報検索と自動要約の評価のためのワークショップ，NTCIR-2 の要約タスク TSC-1 に参加した結果，YELLOW は平均値で良好な結果を得た．

A Summarization System YELLOW for Japanese Newspaper Articles

Kiyonori Ohtake,[†] Daigo Okamoto,[††] Mitsuru Kodama[††]
and Shigeru Masuyama[††]

We propose a new automatic summarization system, YELLOW, for Japanese newspaper articles. YELLOW is designed to avoid omission of important information. The system was composed of two components, an abstract-type summarizer and an extract-type summarizer. The abstract-type summarizer summarizes sentences by deleting one of multiple modifiers for nouns and illustrations and by paraphrasing, etc. In the extract-type summarizer, features such as main terms, high frequency words, location information in a paragraph, are used to decide the weight of each sentence. We participated in tasks A-1 and A-2 of TSC-1 in NTCIR-2 and the evaluation results showed that YELLOW outperformed all other participants in average precision.

1. はじめに

計算機およびネットワークの発展によって，膨大な量の文書が分散されて蓄積されるようになった．我々は自らがかかえる問題を解決するために，これらの文書から必要な情報を探しだし，利用しなければならない．一方で，有史以来生物としての人間の情報処理能力はほとんど変化していない．そのため，自動要約技術などにより，読み手が読む文書の量を制御できることが求められている．

単一の文書に対する要約研究は，長い歴史を持っており[1]，重要な文を選ぶ重要文抽出型の要約や，1 文ごとに要約を行う文内要約などがある．

山本らによって開発された GREEN[2] では，論説文章を対象に，文書内の談話構造の利用による重要文抽出を行い，連体修飾部の削除などの文内要約によって文章要約を試みている．しかしながら，GREEN は報道記事など，論説以外の特徴を持つ文書への対応はしていない．

三上ら[3] は，TV ニュース原稿を対象として構文構造を用いて文内要約を行っている．しかしながら，三上らは，連体修飾節の削除，固有名詞へ係る修飾語句の削除のいずれにおいても，重要部の認定が困難であると報告している．つまり，重要な情報が削除される点が問題である．

そこで，我々は，山本ら，三上らの問題点を考慮し，「重要な情報を洩れなく抽出する」こと，すなわち報知的（informative）な要約を生成することに重点をおき，論説，報道記事の両方に対応した要約システ

† ATR 音声言語コミュニケーション研究所
ATR Spoken Language Translation Research Laboratories
†† 豊橋技術科学大学知識情報工学系
Department of Knowledge-based Information Engineering, Toyohashi University of Technology

ム YELLOW☆1を開発した．また，国立情報学研究所主催の情報検索と自動要約評価のためのワークショップ NTCIR-2☆2の要約タスク TSC-1 における重要文抽出サブタスク，自由要約作成サブタスクに参加し，YELLOW の評価を行った．

YELLOW は，各文の文内要約と各文に対する重要度付与をもとに要約を行う．

文内要約では，二重修飾に着目した削除を中心に，5 つの手法によって要約を行う．二重修飾に着目した連体修飾要素の削除では，KNP☆3による構文解析結果を積極的に利用する．そして，「重要な情報の削除は極力行わない」という方針に基づいて連体修飾要素を削除する．また，ヒューリスティクスによって直接引用表現内の冒頭文の削除なども行う．さらに，山崎ら[4]，および若尾ら[5] の換言手法も採り入れ，冗長な文頭や文末表現を簡潔に換言する．

文への重要度付与には，主要語，高頻度の名詞，位置情報，見解文であるか否かなど，従来，文の重要度を決定するにあたって重要であるといわれてきた種々の情報[6] を複合的に用いている．文の重要度は，各情報の重要度と諸条件から導き出される得点の総和により決定する．得点の算出方法は，NTCIR-2 の DRYRUN（事前に行われた評価の試行）に基づき，人手で決定した．なお，記事の種類（報道か論説）によって主要語の抽出箇所および得点の算出方法が異なる．

以下，YELLOW の仕様，NTCIR-2 における評価を報告し，文内要約に使用した各手法の比較を行う．なお，今後の研究の参考となるよう，作成した規則などのデータをすべて開示する．なお，分量が多く論文中に収録できないデータは，WWW☆4にて公開する．

2. YELLOW の概要

YELLOW は，Linux 上で，Perl を使用して作成した．システムは，記事の各文に対し，文内要約を行う部分と各文へ重要度付与を行う部分の 2 つから構成される．

YELLOW は，まず，記事中の各文に対し，KNP による構文解析を行う．そして二重修飾に着目した連体修飾要素の削除，直接引用表現の処理などの文内要約を行う．また，記事の原文に対し，主要語や語の頻度，段落内構造などの表層情報をもとに各文に重みを付け，各文の優先順位を決定する．最後に，文内要約された各文の中から，各文の優先順位をもとに文を選択し，要約結果を出力する．

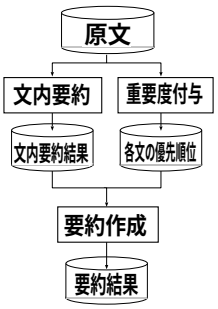


図 1 YELLOW の概要
Fig. 1 Overview of YELLOW.

YELLOW の概要を，**図 1** に示す．

3. 文 内 要 約

1 文内に存在する修飾関係には様々な関係がある．日本語においては，名詞の比率が高い[7] ために連体修飾が他の修飾関係より頻繁に用いられる．そのため連体修飾要素の削除による要約が効果的である．山本らの GREEN[2] では，ヒューリスティックスによって連体修飾要素を削除している．三上ら[3] は，構文解析器を利用し，連体修飾要素を削除する手法を用いてニュース文を対象とした要約実験を行った．その結果は，削除後の文が不自然になることが多く，連体修飾要素の削除による要約作成が容易ではないことを示している．

YELLOW における文内要約では，新聞記事の各文に対し KNP による構文解析を行い，その結果を利用して，二重修飾に着目した文内要約を行う．また，二重修飾に着目した削除だけでは，冗長さが残存した要約となるため，「直接引用表現の処理」，「補足情報の削除」，「例示の削除」，「換言処理」の 4 つの手法による要約もあわせて行う．

YELLOW で行われる文内要約は，以下の 2 つの基本方針に基づく．

- 要約は，文ごとに独立に冗長な部分を削除（または換言）することで行う．文と文のまとめあげ，文間の冗長性の認定による削除など，複数の文を対象とした処理は行わない．
- 重要な情報が損なわれないことを重視する．

3.1 二重修飾に着目した削除

3.1.1 定　　義

1 つの名詞に対して 2 つの修飾要素が修飾している

☆1 YEt another summarization system with two moduLes using empiricaL knOWledge
☆2 http://research.nii.ac.jp/ntcir/workshop/work-ja.html
☆3 http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html
☆4 http://www.smlab.tutkie.tut.ac.jp/research/NL/DATA/yellow_data.html

状態を二重修飾，複数の修飾要素が修飾している状態を多重修飾と定義する[2]．ただし，3 以上の修飾要素で構成される多重修飾は，それほど用いられない．これは，多重修飾の一部が他の修飾要素への連用修飾として表現されることによる．たとえば,「白い長い大きな手」は 3 重修飾であるが，この表現よりも「白い」を「白くて」とした「白くて長い大きな手」の方がより自然に感じる．この場合,「白くて」は「長い」を修飾する連用修飾であり,「手」を修飾している要素は「長い」と「大きな」の 2 つである．あるいは,「白くて長くて大きな手」になると，そこには多重修飾は存在しない．そこで，YELLOW では，二重修飾に着目する．

3.1.2 二重修飾に着目した削除手法

YELLOW では，構文解析器を積極的に利用することで，二重修飾の修飾要素のどちらか一方，あるいは両方を削除する要約を行う．

三上らは文献 3) の中で，簡易構文解析を用いて連体修飾要素の削除による要約を試みた．つまり，これは一重修飾を中心とした修飾要素の削除である．三上らは修飾要素内の重要語や，被修飾要素である名詞の種類によって重要な情報が削除されることを回避しようとした．しかし，その対処は不十分であったと報告している．山本らの GREEN では，多重修飾の処理に対して被修飾名詞に最も近い修飾要素を残し，他を削除するヒューリスティックスを導入している．しかし，このヒューリスティックスには問題がある．たとえば,「ここに規制緩和の大きな役割がある」は「規制緩和の」と「大きな」がともに「役割」を修飾している．このとき，GREEN による要約では「ここに大きな役割がある」となり不自然である．

そこで，YELLOW では，三上らの報告に基づき一重修飾の削除には問題がともなうことから，一重修飾の修飾要素の削除を行わず二重修飾に着目する．また，山本らの手法の問題点から，構文解析を積極的に利用し，二重修飾の削除に関する条件を網羅的に調査する．

我々は構文解析器として，一般に広く，自由に用いられている KNP を使用する．KNP は，ルールベースの比較的高精度な構文解析器であるが，その解析結果には誤りが存在する．さらに，要約の入力として与えられる文が非文である可能性もある．そこで，我々は，KNP の解析結果に対し，人手で作成した規則を照らし合わせることにより，解析誤りやその他の原因によって起こる不自然な要約，ならびに重要な情報の欠落を回避する．

このようなアプローチをとる場合，構文解析結果に人手で作成した規則を照らし合わせるほかに，2 つ以上の構文解析器による結果を比較し，解析誤りを回避することも考えられる（たとえば，文献 8) など）．しかし，我々は，実用的なシステムを構築するためには，構文解析結果に対して人手で作成した規則を適用するアプローチの方が，より短時間で精度の高いシステムを構築できると考えた．

YELLOW は，二重修飾に着目し，いずれかの修飾要素を削除することにより要約を行う．問題は，どちらの修飾要素を削除するか一意に決めることができないことである．山本らは名詞に最も近い修飾要素を残す処理をしている．これは英語の場合には適切な処理である．なぜならば，英語において形容詞がいくつか重なる場合，修飾する名詞の本質に関係の深いものほど名詞の近くに置かれる傾向がある[9] ためである．

YELLOW の基本的方針としては，名詞を限定する働きが同程度と考えられる場合は，山本らの GREEN と同様に後方の修飾要素を残し，前方を削除する．

例：注目されている，農作業の 機械が発売．
→ 農作業の機械が発売．

それ以外の状況については，KNP によってコーパスを構文解析し，具体的な二重修飾を調査し，削除条件を決定する．構文解析器を用いて二重修飾の処理を行う場合，誤って二重修飾と解析される場合がある．YELLOW では，そのような誤りに起因する不自然な要約を回避するために，KNP が誤って二重修飾と解析している可能性がある場合を規則として記述し，削除を行わない．一方，構文的には間違いなく二重修飾であるが，いずれかの修飾要素を削除しても重要な情報が欠落すると考えることができる場合も同様に規則化し，削除を回避する．

YELLOW で用いた二重修飾処理のための規則は，毎日新聞 CD-ROM 95 年版の 1，2 月の記事を観察し作成した．規則を作成する際に注意した点は，重要な情報の欠落を避けることである．

規則の作成方法は，まず構文解析結果の観測に基づき規則を作成する．そして，規則を作成するために参照したコーパスに対して，その規則を適用し，要約を行う．もし，その規則によって重要な情報が欠落する文が存在する場合は，規則の条件を細分化し，規則を書き直す．このようにして最終的に 36 個の規則を得た．その一部を**表 1** に示す☆．ただし，規則のいずれにも該当しなかった場合には，2 つの修飾要素は名詞を限定

☆ http://www.smlab.tutkie.tut.ac.jp/research/NL/DATA/yellow_data.html にて，全規則を公開．

表 1 二重修飾処理のための規則の一部
Table 1 Examples of rules for double modifier.

前方	後方	被修飾要素	動作
＜動詞連体＞	～い，＜形判連体＞	—	後方を削除
＜動詞連体＞	～な，＜形判連体＞	—	後方を削除
～の	＜数量修飾＞	＜数量＞	後方を削除
—	—	＜数量ノ＞	両方を削除
～という	～の	—	削除しない
～のは	～の	—	削除しない
—	—	～との	削除しない

する働きがともに同程度であるとし，基本方針に従って前方の修飾要素を削除する．

KNP を用いた二重修飾処理において，被修飾要素となるのは＜体言＞属性を持つ文節である．しかし，この条件に該当する文節でも，さらに＜用言＞属性をあわせて持ち，体言止めや，文末と判断できる場合は，その文節を被修飾要素として認めない．表 1 に示してある被修飾要素の条件は，この条件に加えてさらに適用するものを示している．また，表 1 においては，KNP の出力結果における文節の属性を＜＞で囲み表示している．その他の表現は文字列照合による制約を示している．また “—” は特に制約が存在しないことを示している．

なお，YELLOW が持つ二重修飾処理のための規則には，KNP の解析誤りに起因する重要な情報の欠落を回避するためのものも多く含まれる．そのため，直観的に連体修飾とは思えない修飾要素も条件として記述されることがある．

3.1.3 二重修飾の特例

二重修飾であるにもかかわらず，KNP により二重修飾と解析されない，あるいはその逆の場合がある．ここでは，その代表的，かつ頻繁に見られる例として「連体修飾節+名詞の+名詞」に着目する．たとえば，「私が聞いた作家の話」と「私がインタビューした作家の話」は，いずれも品詞の並びは同じであるが，その依存構造は**図 2** に示すとおり異なる．このような文における連体修飾節の係り先を正しく決定することは構文解析器にとって困難である．

山本らの GREEN では，いずれの構造においても，「名詞の名詞」という依存構造は同一であることから，この場合には連体修飾節を削除している．YELLOW でも基本的に同一の方針をとり，連体修飾節を削除する．しかし，それは連体修飾節に重要な情報が含まれていないと見ることができる場合に限る．ところが，そのような判断を客観的かつ機械的に行うことは困難である．そこで，残された「名詞の」のみで十分に最後の「名詞」が限定されており，その文において不自

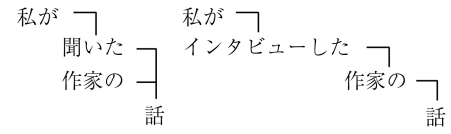


図 2 依存構造が異なる 2 文の例
Fig. 2 Example of two sentences from which dependency structure differs.

然ではないと見なせる場合に削除する．限定の強さの近似として「名詞の」に含まれる名詞の抽象性を用いる．名詞の抽象性の判断は，日本語語彙大系[10] の各カテゴリを代表する語を抽象的な語と定義し☆，それに該当するか否かで行う．

例：カテゴリ名 **茶**に属する語：
甘茶 烏龍茶 麦茶 レモンティー
→ 「茶」のみを抽象的な語と見なす．

3.1.4 主題部内から外への係り受け

「X は」の形で文の陳述の対象を表す要素を,「主題」と呼ぶ[11]．主題を提示する働きをする助詞を提題助詞といい，一般に主題は名詞と提題助詞で構成される．提題助詞には,「は，なら，って，ったら」がある[11]．それに対して，同類の他の事項を背景にして，ある事項を取り上げる働きをする助詞を,「取り立て助詞」と呼ぶ．取り立て助詞には,「は，も，さえ，で，すら，だって，まで，だけ，ばかり，のみ，しか，こそ，など，なんか，なんて，くらい」がある．

文の係り受け構造に基づいた二重修飾の削除では，係り先が削除の対象となった場合には，その係り元も削除する．しかし，主題の一部が係り元であるような場合は，主題の一部を削除することになる．その結果，その文が何について述べようとしているのか分かりにくくなってしまう．そこで，本手法では主題部を定義し，この主題部中から主題部の外へ係る依存関係は無視する．そのことにより，上記の削除を回避する．主題はすでに述べたように，名詞と提題助詞で構成され

☆ 日本語語彙大系から得られた名詞に，さらに 20 ほど追加し，2490 の名詞からなる．

るが，本手法では提題助詞に準ずる働きをする取り立て助詞も主題を構成する助詞と考える．また，提題助詞のうち「なら，って，ったら」は口語にみられる表現であり，新聞記事にはほとんど見られないため，本手法では取り扱わない．結果的に本手法における提題助詞は，上で述べた取り立て助詞と等しくなる．主題部を以下に定義する．

定義 文頭から提題助詞と同一の文字列を末尾に持ち，かつ KNP の解析結果として助詞があると認定された文節までを「主題部」とする．ただし，提題助詞のあとに格助詞がある場合もこれに該当する．

3.2 直接引用表現の処理

新聞記事内の直接引用表現は，基本的に重要度が低く，それが存在しない場合でも前後の文により理解できることが多い．また，第 1 文より，第 2 文以降の意見が重要視される傾向がある．そこで，直接引用表現が複数文で構成されるとき，第 1 文を削除する．

例：そのうえ「明日の公式試合には出なくてええ．背番号も返せ」と言われたという． →
そのうえ「背番号も返せ」と言われたという．

ただし，第 2 文以降に第 1 文を参照する可能性のある指示詞が存在する場合は，削除を行わない．このような削除を行った場合，記事が理解しにくくなり，削除後の自然さも損なわれることを経験的に確認している．

また，新聞記事では，鉤括弧でくくられた直接引用表現の直後の節が，その引用表現の要約である場合がある[12)]．YELLOW では，直接引用表現を含む文の陳述表現が文献 12) に示されるパターンにマッチする場合，発言部分を削除しても意味内容を保持できると仮定し，直接引用表現部分をすべて削除する．

例：検察側は「捜査段階で事実を認めていた」と主張して，タイミングを計って証拠申請をする構え．→ 検察側はタイミングを計って証拠申請をする構え．

3.3 補足情報の削除

新聞記事において，丸括弧で囲まれた情報はフリガナや略称など，補足的な説明が多い．そこで，丸括弧で囲まれた情報および他の記号（=，<>など）で判断できる補足説明は削除する．

3.4 例示の削除

「〜などで」「〜などの」のような例示は，広い意味での修飾と考えられ，削除しても意味的に変化が生じないと近似的に仮定する．YELLOW では，「〜などの+名詞」という表現において「〜などの」を削除する．ただし，「〜などの」が修飾する名詞が，3.1.3 項で定義した抽象的な語である場合は削除しない．また，「〜などで」が用言に単独で係る場合，「〜などで」を削除する．

例：経済や外交戦略などの 専門知識はもとより， → 専門知識はもとより，

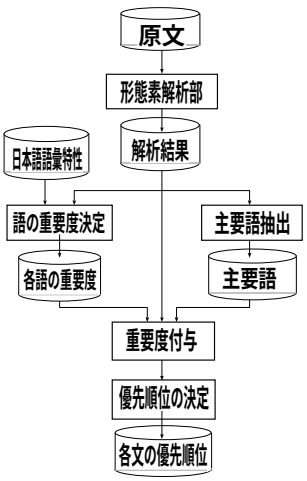


図 3 重要度付与プロセスの概要
Fig. 3 Outline of extract-type summarizer.

3.5 換 言 処 理

新聞記事に見られる語尾のまわりくどい表現は，体言止めにするなどの換言により要約できる．また，前文との結束性を表す冒頭の接続詞は，前文が抽出されなければ意味を失う危険がある，さらに，接続詞を削除しても文の大意が変わらない場合が多いことから，冒頭の接続詞はすべて削除する．

YELLOW では，このような処理を文字列照合に基づいて，ある文字列を別の文字列へ変換する規則（1 対 1 換言）によって実現する．このような規則を，文末表現の簡潔化を中心に，人手で 106 用意した☆．

例 1：…決まらないようだ． → …決まらない．
例 2：そんな中， …． → …．

4. 重要度付与

記事中の各文に対し，5 つのプロセスから重要度を付与する．その関係を**図 3** に示す．

4.1 形態素解析部

形態素解析には，JUMAN☆☆を用いる．解析した結果は，語の重要度決定，主要語抽出，重要度付与に使われる．

☆ http://www.smlab.tutkie.tut.ac.jp/research/NL/DATA/yellow_data.html にて，全規則を公開．
☆☆ http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html

4.2 語の重要度決定

新聞記事において，記事中に多く出現する語は筆者が伝えたい情報であり，重要と考えられる．また，他の記事において多く出現する語は，一般的な語であり，重要ではないと考えられる．

そこで語の重要度決定部では，記事 a における語 w の重要度 $W(a,w)$ を式 (1) により決定する．

$$W(a,w) = \frac{tf(a,w)}{\log(lp(w))} \quad (1)$$

ここで，$tf(a,w)$：記事 a での語 w の頻度，$lp(w)$：日本語の語彙特性[13] に記載されている w の頻度である．

日本語の語彙特性に記載されている単語の出現頻度は，朝日新聞の 1985 年から 1998 年の記事から計算されており，同じく新聞記事を対象としている本システムに適していたため，これを用いた．

4.3 主要語抽出

見出しは，記事の「究極的な要約」であり，その記事の主題を表すことが多い[2]．したがって，見出しに含まれる名詞は，少なからず主題に関係していると考える．また，記事中の重要な文も，見出しと同様に，主題について述べられている可能性が高い．報道記事では，冒頭段落に全体のまとめが書かれる傾向がある．そのため主題に関係する語の多くは冒頭段落に含まれていると考える．論説記事では，主要な結論が述べられる最終見解文と，前提をまったく持っていない読者に初めて著者の持つ情報を提供する冒頭文に，主題に関係する語が多く含まれていると考える[2]．なお，見解文に関しては，4.4 節で定義する．

YELLOW では，見出し，および報道記事における冒頭段落中の文と，論説における冒頭文と最終見解文に含まれる名詞を主要語とする．なお，JUMAN によってカタカナ，アルファベットが未定義語と解析されるが，本研究ではこれらを名詞とする．また，それぞれの主要語の主題への関連度は，語の重要度に比例すると仮定している．

4.4 重要度付与

各文の重要度は，主要語，高頻度語，記事構造，段落内構造，不要文，見解文と現象文の計 6 つの要素に着目して決定する．最終的な重要度は，それぞれの要素と諸条件から導き出される得点の合計値によって与えられる．

ただし，記事の種類により重要視される要素が異なるため，YELLOW では，論説と報道記事に応じてそれぞれ異なった得点算出を行う．以下，6 つの要素について説明する．

主要語 主要語は，その記事の主題など，重要な情報に関連していると見なした語であり，それを含む文は重要であると考える．また，主要語がその文において主語（主格を表す格助詞を持つ語）である場合，その文は主題について述べていることが多い．そのため，より大きな重みをその文に付加する．

高頻度語 記事中に多く出現する語は筆者が伝えたい情報に関連しており，高頻度の語を含む文は重要であると仮定する．そこで，高頻度の語を含む文に重み付けをする．なお，YELLOW では，1 記事に 2 回以上出現した名詞を，高頻度語と定義する．

記事構造 新聞記事において，冒頭文が重要であることは先に述べた．さらに，報道記事だけに着目すると，著者はより重要な情報ほど冒頭に書く傾向がある[6]．したがって，冒頭に近い段落中の文ほど重要視する必要がある．また，報道記事の最終段落には，テーマに関する今後の展開など読者が関心を持つ情報が書かれる場合がある[2]．本手法では，このような情報も重要視し，最終段落中の文も重み付けを行う．

段落内構造 各段落の冒頭文は，新しい主題に関する前提を持っていない読者に，著者の持つ情報を初めて伝達する役割を持つ．また，各段落の最終文は，著者がその段落の話を締めくくるという特別な意図を持って書いた文と考えることができる．したがって，各段落の冒頭文および最終文は，その段落において重要な文と考えられる．

不要文 新聞記事には，段落のタイトルや補足説明など，不要な文が存在する．このような文は，記事特有の記号（◇，＝ など）を含む傾向がある．また，括弧で囲まれた発言のみの文は，前後の文脈により情報を把握できることが多い．YELLOW では，記事特有の記号を含む文，および発言内容のみを示す文は，不要文と定義し，重要視しない．

見解文と現象文 記事中の文は，著者の主張，意見，希望などを述べた文と，出来事，事実，現象を述べた文の 2 種類に大別することができる．以下，前者を「見解文」，後者を「現象文」と呼ぶ[2]．論説記事においては，見解文が特に重要視される傾向がある．見解文を抽出するために，「〜が必要である」「〜すべきである」などの文末表現に注目する．GREEN に基づいて作成した 55 個のテンプレートと各文の文末表現を照合することにより近似的に見解文の抽出を行う☆．

各文の重要度は，**表 2** に示す条件を独立に判定し，

☆ http://www.smlab.tutkie.tut.ac.jp/research/NL/DATA/yellow_data.html にて，全テンプレートを公開．

表 2 重要度付与条件とその得点
Table 2 Factors and points for weighting.

条件	得点	
主語が主要語である文	主要語の重要度× 10	
主要語を含む文	主要語の重要度× 2	
高頻度語を含む文	高頻度語の重要度× 1	
各段落の冒頭文	20	
各段落の最終文	10	
不要文	文の重要度を 1/10	
条件	報道記事	論説記事
第 1 段落	100	20
第 2 段落	50	0
第 3 段落	20	0
最終 (まとめ) 段落	文の重要度× 10	0
見解文	0	10
それ以外	0	

得点を総和することで計算する．ただし，計算の対象とする文が不要文である場合は，最終的に求めた文の重要度の 1/10 をその文の重要度とする．また計算対象の文が報道記事の最終段落に存在する場合は，最終的に求めた文の重要度を 10 倍した値をその文の重要度とする．なお，主要度，高頻度語に関する得点は，文中に含まれる各語ごとに判定し，得点を総和する．

各条件の元での得点の算出方法は，DRYRUN（事前に行われた評価の試行）での評価結果に基づき，人手による試行錯誤で決定した．したがって，NTCIR-2 の評価方法に特化した得点の与え方といえる．

4.5 優先順位の決定

各文の重要度をもとに，重要度の高い文から順番に，各文の優先順位を決定する．ただし，記事の冒頭文（冒頭文が不要文の場合は次の文）は，重要度にかかわらず，強制的に優先順位 1 番とする．

5. 要 約 作 成

YELLOW は，文内要約で要約された各文と，重要度付与による各文の優先順位をもとに，使用する文を選択し，それを要約とする．このとき，規定文字数を入力することで，YELLOW は，規定文字数を超えないように要約を行う．

要約結果を出力する際のアルゴリズムを以下に示す．

要約作成アルゴリズム

1. 各文に対して換言処理を除く文内要約を行う．
2. 規定された文字数を超えるまで，優先順位が大きい順に文を選ぶ．
3. すべての文に対してもう一度，換言処理を除く文内要約を行う．
4. 選択した文すべての長さが，規定された文字数以下ならば要約結果を出力して終了．
5. 選択した文に対して換言処理を行う．
6. 選択した文すべての長さが，規定された文字数以下ならば要約結果を出力して終了．
7. 選択した文の中で最も優先順位の低い文を除き，(規定文字数 − 現在の文字数) 以下の長さの文をまだ選択されていない文の中から重要度が高い順に探す．
8. 該当する文があれば，その文を換言処理を行ってから選択し，要約結果を出力して終了．
9. 該当する文が存在しない場合，8. の処理を終えた段階で選択されている文を要約として出力する．もし，選択されている文が 1 つもなければ，規定文字数が小さすぎると判断し，出力を行わない．

Step 3. において，文内要約を繰り返している理由は次のとおりである．まず，YELLOW ではいくつかの文内要約手法が用いられている．また，KNP の解析結果には誤りが含まれる．そのため，いずれかの手法によって，ある部分が削除された場合，それをもう一度解析することによって，一度目の解析とは依存構造が変化した結果が得られ，二重修飾の削除などを行える可能性がある．そこで，YELLOW では，文内要約を繰り返し行うことでより短い要約の作成を試みる．

6. 評 価

YELLOW による要約の有効性，および，重要度付与の精度の評価を行うため，情報検索，自動要約評価のためのワークショップ NTCIR-2 に参加した．NTCIR-2 は，システム間の比較や，情報検索やテキスト処理技術の評価手法，および，繰り返し利用できる大規模データセット構築法について研究を行うことを目的としたワークショップである．我々は，要約に関するタスクのうち，自由要約作成サブタスク，重要文抽出サブタスクに参加した．

また，YELLOW の文内要約手法を代表する，二重修飾に着目した削除，および，直接引用表現の処理，それぞれの妥当性を評価するために評価実験を行った．

6.1 自由作成要約サブタスクの評価

自由作成要約サブタスクでは，システムが作成した文字単位の自由な要約に対する評価を行う[14]．

毎日新聞 CD-ROM 94 年版，98 年版から指定された 30 記事（報道 15 記事，論説 15 記事）を対象に，記事全体の 20%，40%に相当する文字数が規定文字数として与えられ，各参加システムは規定文字数をもとに要約を行う．

自由作成要約タスクでは，主観評価と content-based

表 3 主観評価の結果
Table 3 Subjectivity evaluation.

	評価値 (参加システムの平均)
読みやすさ 20%	2.53 (3.16)
内容評価 20%	2.93 (3.24)
読みやすさ 40%	2.73 (3.05)
内容評価 40%	2.77 (3.12)

表 4 content-based での評価結果
Table 4 Content-based evaluation.

	評価値 (参加システムの平均)
FREE 20%	0.4727 (0.4418)
FREE 40%	0.6483 (0.6065)
PART 20%	0.5137 (0.4740)
PART 40%	0.6608 (0.6342)

表 5 文内要約における手法の比較
Table 5 Comparison by methods.

手法	削除文字数 (全削除文字数に占める割合 [%])	使用回数	平均削除文字数	content-based の評価	
				FREE 20%	FREE 40%
補足情報の削除	662 (31.1)	335	1.97	0.4631	0.6067
直接引用表現の処理	298 (14.0)	8	37.25	0.4679	0.6198
二重修飾に着目した削除	729 (34.2)	61	11.95	0.4489	0.6199
例示の削除	126 (5.9)	6	21.00	0.4627	0.6326
換言処理	314 (17.1)	89	3.53	0.4652	0.6206

での評価という 2 種類の評価が行われた☆.

主観評価では，記事ごとに「テキストとして読みやすいかどうか」と「元テキストの重要な内容を不足なく記述しているかどうか」の 2 つの観点から，要約筆記の専門家が要約を評価した．なお，評価値は，原文および，人間の要約（自由作成要約，重要個所抽出要約），システムの要約，ベースラインシステムの要約の 4 つを提示し，原文の重要な内容をどの程度要約がカバーしているか，要約の読みやすさの 2 つの評価基準で，要約を順序付けた．したがって，読みやすいものから，1，2，3，4，同様に内容の点で見て，良いものから，1，2，3，4 となり，評価値の小さい方が良い結果となる．

content-based での評価は，まず，人間の作成した要約およびシステムの作成した要約をともに形態素解析し，内容語のみを抽出する．そして，人間の作成した正解要約の単語頻度ベクトルとシステムの要約の単語頻度ベクトルの間のコサイン距離を計算し，どの程度内容が単語ベースで類似しているかという値を求める．また，正解要約は，人間が自由作成した要約（以下，FREE）と，人間が重要個所抽出により作成した要約（以下，PART）の 2 種類がある．主観評価の結果を**表 3**，content-based での評価を**表 4** に示す．

6.1.1 考　　察

今回，文への重要度付与によって採用した重要文（22,812 文字）に対して，文内要約で行った冗長部の削除（2,129 文字）の要約率は約 91%である．YELLOW では，重要な情報の欠落を防ぐため，大胆な削除は行わず慎重な要約を行った．その方針は比較的成功し，良好な評価にも結び付いている．

6.1.2 各手法の比較

YELLOW では，5 つの手法を併用することにより文内要約を行っている．それぞれの手法単独による削除文字数，content-based での評価を調査した結果を**表 5** に示す．

各手法のうち最も要約に貢献した手法は，一度に削除する文字数が多く，使われる場面も多い，二重修飾に着目した削除であった．自由作成要約サブタスクで，良好な評価が得られていることから，二重修飾に着目した削除は，文内要約において，比較的有効な手法であったといえる．一方で，content-based での評価は，他の手法に比べ，やや精度が落ちている．これは，必要な情報の削りすぎがあるためであり，特に「連体修飾節＋名詞の＋名詞」という構成の連体修飾節を削除した場合に，不自然に感じられる場合が多いとの印象を受けた．この点に関して，厳密に評価し，改良することで，さらなる精度の向上が期待できる．

6.2 重要文抽出サブタスクの評価

重要文抽出サブタスクでは，人間が選択した重要文とシステムが選択した重要文との間の一致度を評価する[14]．このタスクでは，自由作成要約サブタスクと同一の毎日新聞 30 記事を対象として，記事全体の 10%，30%，50%に相当する規定文数が与えられる．

我々は，YELLOW の重要度付与による文選択部分のみを用いてこのタスクに参加した．最終的な文の選択には，重要度付与による優先順位の高い文から順に，規定文数だけ文を選択する単純なアルゴリズムを用いた．

重要文抽出サブタスクでは，評価尺度として再現率，精度，F 値の 3 つを用いている．

☆ http://oku-gw.pi.titech.ac.jp/tsc/

$$再現率 = \frac{システムが選んだ正解文の数}{正解文の総数}$$
$$精度 = \frac{システムが選んだ正解文の数}{システムが選んだ文の総数}$$
$$F\ 値 = \frac{2 \times 再現率 \times 精度}{(再現率 + 精度)}$$

要約率 10%，30%，50%ごとに，30 記事（報道 15 記事，論説 15 記事）の平均を求めたときの F 値，および，記事を報道限定，論説限定にしたときの結果を**表 6** に示す．

6.2.1 考　　察

結果より，報道記事に比べ，論説記事の重要文抽出の精度が全体的に低いということができる．これは，論説が報道のように定型的な構成を持たないことや，報道に比べ記事中の文数が多いなどの特徴が影響しているためである．また，10%への要約のように，少数の文を選択する場合，論説においては，冒頭文強制採用の誤りによる影響が大きく現れた．これは，冒頭文が必ずしも全体の概要にはならず，筆者によって書き出しが大きく変わるためであり，冒頭文が全体の概要を示しているかを判定したうえで，採用するなどの改良が必要である．

6.2.2 重要度付与に用いた各要素の比較

YELLOW では，6 つの要素をもとに各文の重要度を付与している．重要度の決定に用いた各要素の効果を，それぞれ単独に無効化することで，比較した結果を**表 7** に示す．表中のそれぞれの値は，10%，30%，50%の平均である．

各要素を比較した結果，全体的には不要文の影響が一番大きいことが分かる．つまり，精度を向上するためには，重要ではない情報をいかに丁寧に取り除くかが重要である．

また，段落内構造による重み付けは，大きな影響があり，特に，段落冒頭文の重み付けのみを無効化した場合，論説のみの F 値は，−7.7%と大きく精度が落ちた．これは，文章をいくつかのまとまりに分けた場合，筆者はその冒頭に重要な情報を書く傾向が強いためである．したがって，段落のみならず，筆者の考えたまとまりと同様に文章を分けることが，より精度をあげるための有効な手段と考える．

さらに，報道記事においては，記事構造による重み付けが大きく影響しており，冒頭に近い段落ほど重要な構造を持っていることが確認された．また，論説においては，見解文や記事中の高頻度語が重要であることが分かった．しかしながら，論説において，見出しなどから抽出した主要語は，効果を発揮しなかった．これは，主要語を主語とする文が必ずしも重要ではなく，むしろ主題に対する意見が述べられる，それ以降の文を重要視すべきことを示唆している．したがって，主要語と各文の重要度との関係を再検討する必要がある．

6.3 二重修飾に着目した削除の評価

本手法の特徴である二重修飾に着目した削除の妥当性を評価するため，二重修飾に着目した削除を単独で適用したときの要約について評価する．また，YELLOW の二重修飾削除規則を作成するために参照したコーパスは限られている．そのため，規則が実際に用いられる二重修飾を網羅しきれていない可能性がある．ここでは，これらの点を明らかにする．

毎日新聞 CD-ROM 94 年版，98 年版から，NTCIR-2 で指定された 30 記事（1,021 文）に対し，KNP による構文解析を行った．構文解析結果から二重修飾部分を自動抽出したところ，109 カ所の二重修飾を得た．これらのうち，直接引用表現内から外への係り受けなど，明らかに構文解析誤りをしている箇所を除いた全 72 カ所を対象に評価を行った．

まず，二重修飾と解析された箇所の前方の修飾要素

表 6 重要文抽出サブタスクの評価結果
Table 6 Evaluation result in extract-type summaries subtask.

	F 値 (全参加システムの平均)	報道の F 値	論説の F 値
10%	38.6 (28.9)	47.8	29.4
30%	46.8 (43.2)	49.2	44.4
50%	62.2 (57.7)	63.4	60.9
ave	49.2 (43.3)	53.4	44.9

表 7 文への重要度付与に用いた各要素の比較
Table 7 Comparison by factors for weighting.

無効化した要素	F 値	報道の F 値	論説の F 値
見解文	48.4 (−1.6%)	53.4 (0%)	43.4 (−3.3%)
記事構造	48.0 (−2.4%)	50.9 (−4.7%)	45.1 (+0.4%)
段落内構造	46.4 (−5.7%)	52.3 (−2.1%)	40.5 (−9.7%)
主要語	49.4 (+0.4%)	53.2 (−0.4%)	45.7 (+1.8%)
記事中の高頻度語	47.8 (−2.8%)	53.3 (−0.2%)	42.3 (−5.8%)
不要文	44.5 (−9.6%)	50.0 (−6.4%)	39.0 (−13.1%)

と後方の修飾要素のそれぞれに対し，重要な情報が含まれているか否かの 2 値選択式のアンケート調査を，被験者 7 名に対して実施した．このアンケート結果をもとに，重要な情報が含まれていないと 4 名以上が判断した修飾要素を削除可能と仮定し，正解データを作成した．

YELLOW が削除を抑制した修飾要素のみを見た場合，その 86.2%が，正解データと一致していた．このことから，二重修飾に着目した削除は,「重要な情報を極力削除しない」という YELLOW の方針を実現している．

さらに，対象箇所全体における YELLOW の出力と正解データを比較した結果，前方，後方，ともに一致した箇所は，全体の 59.7%，片方の修飾要素だけ一致した箇所は，全体の 29.2%，前方，後方，ともに不一致だった箇所は，全体の 11.1%であった．前方，後方，ともに一致した箇所が多いことから，二重修飾に着目した削除手法は，比較的，重要な情報の削除を抑制し，重要ではない箇所の削除を実現していることが分かる．

しかしながら，YELLOW によって実際に削除が行われた箇所が，前方の修飾要素で 18 カ所，後方の修飾要素においても 18 カ所あり，前方の削除が正解データと一致しなかった箇所は 9 カ所（対象箇所全体の 12.5%），後方の削除が正解データと一致しなかった箇所は 4 カ所（対象箇所全体の 5.6%）であった．このように，前方の削除の一致しなかった箇所が後方に比べ，やや多い原因は，削除規則が必ずしも十分に網羅しきれていないためである．したがって，規則を作成するために参照するコーパスをさらに増やし，規則を追加する必要がある．

なお，正解データでは，72 カ所中 26 カ所が削除されておらず，修飾要素の一方だけが削除されたのは 45 カ所（後方を削除：29 カ所，前方を削除：16 カ所），前方，後方，ともに削除された箇所は 1 カ所のみであった．このことから，二重修飾要素の前方，後方，ともに削除するのは情報の欠落を招く危険があり，一方だけ削除することの妥当性を確認できた．ただし，2 つの修飾要素の名詞を限定する働きが同程度の場合にどちらを削除するかについては，今後検討が必要である．なぜならば，正解データ全体の平均的な結果としては前方が削除されたのは後方が削除された場合の約半分と明らかな違いがあるからである．

6.4 直接引用表現の処理の評価

ヒューリスティクスを用いた文内要約手法の 1 つである，直接引用表現における第 1 文の削除の妥当性を工学部学生の被験者 3 名によって評価した．ここでは，NTCIR-2 で指定された毎日新聞 30 記事中の 1,021 文に対し，直接引用表現の処理が適用された 22 文を調査した．直接引用表現内の第 1 文の削除が妥当か否かの 2 段階で評価を行った．評価結果の平均を求めた結果，調査対象全体の 84.8%の文が，直接引用表現内の第 1 文の削除が妥当であると評価された．したがって，比較的，良好な値が得られたことから，直接引用表現の処理の妥当性を確認できた．

7. ま と め

報道，論説の両方に対応した要約システム YELLOW を紹介した．このシステムは，構文解析器を利用した文内要約と各文の重要度付与により，要約を行う．また，文内要約では，新たに二重修飾に着目した削除手法を考案し，さらに，GREEN をはじめ様々な先行研究[6] で提案された手法を改良して取り入れた．

NTCIR-2 の要約に関するタスクへ参加することによって，YELLOW を評価した．その結果，重要な情報の欠落を防ぐため，慎重な要約を行う方針に基づいた YELLOW の文内要約の結果は良好であった．一方，文への重要度付与では，報道記事に比べ論説に対する精度が低いことが分かった．報道記事は，冒頭に近いほど重要視されるような定型的な構成が多いのに比べ，筆者により大きく構成の変わる論説は，対応しにくく抽出が難しいのが現状である．

今後の課題としては，YELLOW では考慮しなかった談話構造による文の結束性をどう取り入れるか，または，複数文のまとめあげなどによる，自然さの向上が望まれる．

謝辞 本研究で使用したコーパスは，毎日新聞 CD-ROM'94〜98 版から得ており，その使用許可をいただいた同社に深謝する．また，日本語語彙大系から意味分類を取得するために用いた形態素解析システム ALTJAWS ver.2.0. の使用許可をいただいた（株）日本電信電話に深謝する．

参 考 文 献

1) Mani, I. and Maybury, M.T. (Eds.): *Advances in Automatic Text Summarization*, MIT Press (1999).
2) 山本和英，増山 繁，内藤昭三：文章内構造を複合的に利用した論説文要約システム GREEN，自然言語処理，Vol.2, No.1, pp.39–55 (1995).
3) 三上 真，増山 繁，中川聖一：ニュース番組における字幕生成のための文内短縮による要約，自然言語処理，Vol.6, No.6, pp.65–81 (1999).

4) 山崎邦子，三上 真，増山 繁，中川聖一：聴覚障害者用字幕生成のための言い換えによるニュース文要約，言語処理学会第 4 回年次大会発表論文集，pp.646–649 (1998).
5) 若尾孝博，江原暉将，白井克彦：テレビニュース番組の字幕に見られる要約の手法，情報処理学会研究報告書 NL-122-13, pp.83–89 (1997).
6) 奥村 学，難波英嗣：テキスト自動要約に関する研究動向，自然言語処理，Vol.6, No.6, pp.1–26 (1999).
7) 樺島忠夫：日本語のスタイルブック新装版，大修館書店 (1990).
8) 乾 孝司，乾健太郎：複数のパーザを利用した統計的部分係り受け解析，情報処理学会論文誌，Vol.42, No.12, pp.3160–3172 (2001).
9) 小池清治：形容詞の語順，言語，Vol.29, No.9, pp.28–34 (2000).
10) 池原 悟，宮崎正弘，白井 諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林 良彦（編）：日本語語彙大系 CD-ROM 版，岩波書店 (1999).
11) 益岡隆志，田窪行則：基礎日本語文法—改訂版，くろしお出版 (1992).
12) 児玉 充，片岡 明，増山 繁，山本和英：直接引用表現を利用した要約知識の自動抽出の試み，言語処理学会第 6 回年次大会発表論文集，pp.241–244 (2000).
13) 天野成昭，近藤公久（編）：日本語の語彙特性，三省堂 (2000).
14) Fukushima, T. and Okumura, M.: Text Summarization Challenge: Text summarization evaluation at NTCIR Workshop 2, *Proc. 2nd NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, pp.45–50 (2001).

(平成 13 年 9 月 25 日受付)
(平成 13 年 12 月 28 日採録)

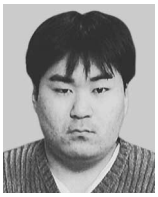
（担当編集委員 仲尾 由雄）

大竹 清敬（正会員）

2001 年豊橋技術科学大学大学院博士後期課程電子・情報工学専攻修了．博士（工学）．同年より ATR 音声言語通信研究所客員研究員，現在に至る．自然言語処理，特に換言処理，要約処理，機械翻訳の研究に従事．言語処理学会，人工知能学会各会員．

岡本 大吾

2000 年豊橋技術科学大学工学部知識情報工学科卒業．同年同大学大学院知識情報工学専攻修士課程入学．現在に至る．テキスト自動要約の研究に従事．



児玉 充

2000 年豊橋技術科学大学工学部知識情報工学科卒業．同年同大学大学院知識情報工学専攻修士課程入学．現在に至る．テキスト自動要約の研究に従事．

増山 繁（正会員）

1977 年京都大学工学部数理工学科卒業．1979 年同大学大学院修士課程修了．1982 年同大学院博士後期課程単位取得退学．1983 年同修了（工学博士）．1982 年日本学術振興会奨励研究員．1984 年京都大学工学部数理工学科助手．1989 年豊橋技術科学大学知識情報工学系講師．1990 年同助教授．1997 年同教授，現在に至る．アルゴリズム工学，特に，グラフ・ネットワーク，組合せ最適化のアルゴリズム，並列アルゴリズム，および自然言語処理，特に，テキスト自動要約等の研究に従事．