# 新聞記事の自動要約によるニュース速報配信*

1D－1

畑山満美子　松尾義博　大山芳史
NTTコミュニケーション科学基礎研究所†

白井諭
ATR音声翻訳通信研究所‡

## 1　はじめに

情報通信処理技術，通信技術の飛躍的な進歩により，社会生活や産業活動のあらゆる面で情報化が進んでいる．特に近年，インターネットを始めとするネットワークの拡大や，携帯電話などのモバイル通信が普及し，情報流通の速度は増すばかりである．一方，量的・質的に増大する情報から自分に必要な正しい情報を選択することが必要になってくる．

本論文では，情報をいち早くニュース速報として配信するシステムとして，新聞記事を自動要約し，日本語・英語の見出しを自動生成するシステムについて論じ，このシステムを用いたニュース配信，情報検索についてのサービスイメージを提案する．

従来，ヘッドラインを自動生成する研究は行なわれていない．しかし，ヘッドラインを生成することは，記事から最も重要な一文を選定する要約，または，重要要素を抽出し一文に構成する要約であるとも考えられる．このような観点から従来の要約研究との比較を考えると，テキスト中の出現頻度によって単語に重み付けを行ない重要文を選定する手法[1, 2]，文間関係を利用した手法[3, 4]，心理実験を利用する手法[5]などがあるが，これらはいずれも重要選定文をそのまま抜き出すことが考えられている．また，要約の評価として基準となるものが人間の直感によるところが大きく不安定である．本手法では，評価基準として人間の直感の他，元となるヘッドラインとの整合性を基準にすることができ，実験者の主観が入らないという利点がある．

## 2　概要

### 2.1　ヘッドライン自動生成システムの概要

本論文で研究開発を行なっているヘッドライン自動生成システムは，日本語の新聞記事本文から短く一文にまとめた要約文を生成し，それを日本語要約文・英文ヘッドラインとして出力する．

例えば，図1の新聞記事を入力すると，下線の部分の情報が抽出され図2の日本語要約文と，それをヘッドラインスタイルに翻訳した英語ヘッドラインを生成する．

この記事の原文につけられている新聞見出しは，「公的

社会保障制度審議会（首相の諮問機関、会長・隅谷三喜男東大名誉教授）は八日、社会保障の将来像についての報告書を発表、老齢者の介護サービスを保障する公的介護保険制度の導入を提言した。厚生省はこの報告を踏まえたたちに本格的な検討に入るが、早ければX年度の導入をめざし（1）六十五歳以上を保険給付の対象とし（2）二十歳以上のすべての国民から、月収の一%弱相当の保険料を徴収する一などを考えている。年内にも具体案を提示するが、大幅な負担増に強い反発も予想され、実現までには曲折が予想される。

図1: 元になる日本文新聞記事

【日本語要約文】
社会保障制度審議会は公的介護保険制度の導入を提言した。

【英語ヘッドライン】
Panel proposes creation of public nursing insurance

図2: システムで自動生成した見出し

三菱信託銀行と住友海上火災保険は、兵庫銀行系ノンバンク十社向けの金利減免債権を金融期間から分離管理するための特別目的会社「ポートアイランド・アクセプタンス」に対して、新規に出資することを決めた。日中に出資金を払い込む。金利減免債権を処理するための特別目的会社の第一号であるポート社に出資することで、特別会社方式の手法を取得することが狙いだ。

図3: 元になる日本文新聞記事(2)

【日本語要約文】
三菱信託銀行と住友海上火災保険は特別目的会社ポートアイランド・アクセプタンスに出資する。

【英語ヘッドライン】
Mitsubishi Trust Bank and Sumitomo Insurance to invest at special purpose company Port Island Acceptance

図4: システムで自動生成した見出し(2)

*News flashes using automatic article summarization

†Mamiko HATAYAMA, Yoshihiro MATSUO, Yoshifumi OOYAMA, NTT Communication Science Laboratories.
‡Satoshi SHIRAI, ATR.

介護保険導入を提言」「社会保障制度審報告」「厚生省検討」「６５歳以上に給付」の４文であった.

同様に，図3の新聞記事では，下線の部分の情報が抽出され図4の日本語要約文と，英語ヘッドラインが生成される.

この場合，原文の新聞見出しは「三菱・住友」「特別目的会社に出資」「兵銀系向け」「ノウハウ取得狙う」の４文であった。

このように，ヘッドライン自動生成システムは日本文新聞記事の本文を入力すると，それを１文に要約して日本語要約文と英文ヘッドラインを出力する.

## 2.2 サービスイメージ

本システムは記事内容を短く端的に表すことができるため，この情報を携帯電話，ポケベル，メールなど携帯端末へニュース速報として配信するサービスに利用できると考えられる．また，英文のヘッドラインを生成できるため，日本から世界への情報発信，又は国内の外国人ユーザへの情報提供を行なうことができる.

配信する側だけではなく，ユーザーリクエストでの利用も考えられる．インターネット等で情報が氾濫している現在，大量のデータから自分に必要な記事を探し出すことは容易ではない．そこで，記事データの検索，ブラウジングなどに適用し，記事内容を要約し端的に表示することで情報検索の効率化をはかることが可能である．同様に，外国人ユーザのための情報検索の効率化といったサービスが考えられる.

# 3 ヘッドライン自動生成システムの実装

## 3.1 システムの流れ

システムは大きく２つの部分に分かれている．１つは，日本文記事から１文の日本語要約文を生成する和文要約部，もう１つは，日本語要約文をヘッドライン特有のスタイルに英語翻訳して英文ヘッドラインを生成するヘッドライン加工部である．システムの日英翻訳部は，NTTの日英機械翻訳システムALT-J/Eを利用している.

システム（図5）に日本語新聞記事本文を入力すると，和文要約部では各文に対して(1)形態素解析・係り受け解析[6]を行なった後，(2)位置情報や手係り語，文の長さといった情報から文の重要度を判定し得点をつける．更に(3)一文の中から重要情報を抽出し，(4)一文に再構成して日本語要約文を「日本語見出し」として出力する．ヘッドライン加工部では，(5)4で得られた日本語要約文に対して，英語ヘッドラインの特徴をふまえた英訳を行ない「英語ヘッドライン」として出力を行なう.

## 3.2 和文要約部

1) 重要文選定

要約部では，形態素,係り受け解析を行ったあと，重要文を選定する．各文に対して，位置情報や手がかり語，文の長さといった情報から文の重要度を判定し得点をつけ，最も得点の高い文を最重要文とする.

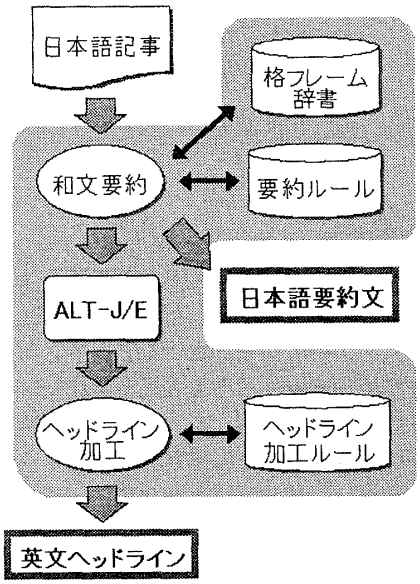


図 5: ヘッドライン生成の流れ

2) 主動詞の特定

一文の中から更に重要情報を抽出するが，まず，主動詞を特定する．通常１文の中には複数の動詞が記述されており，複文で形成されている文も少なくない．その中で最も重要な意味を持つ動詞を特定し，抽出しなければならない.

新聞記事によく見られる表現として，以下のようなものがある.

「～する（した）ことを明らかにした。」
「～する見通しになった（だ）。」
「～する（した）と発表した。」

表層的に見た場合，これらの述語が動詞となるが，文の意味を考えると実質的に意味のある動詞，つまり要約として残したい主動詞は「～する（した）」の部分であることが分かる．本論文ではこのような主動詞にならない述語動詞を広義の意味で様相的表現と呼ぶが，本システムでは，このような様相的表現を判断し，文中から主動詞を特定する.

3) その他の情報の抽出

格フレーム情報を用いてその他の情報を抽出する．ALT-J/Eの格フレーム辞書によって動詞の必須格とその条件が分かる．この情報を利用して，主動詞に対する必須格を文中から探しだし，抽出する.

その他，文選択，主語の抽出，目的語の選定基準などの詳細な分析は[9]による.

## 3.3 ヘッドライン加工部

ヘッドラインには時制の使い方に特別なルールがある．ヘッドラインの動詞の時制は，現在形(49%)，to不定詞(47.5%)に大別され,

- 和文動詞が現在形の場合，内容が“予定, 推定, 未実現”の場合，英文動詞の時制は“to 不定詞”.
- 和文動詞が過去形の場合，内容が“確定した事実, 過去”の場合，英文動詞は“現在形”となる.
- 様相表現の場合，主動詞によって変化する.

ということが分かった.
また，英文ヘッドラインには次のような特徴がある [7].

- 現在形を使う
  S will see → S to see
- be 動詞の省略
  S be to V → S to V
- 短い単語に置換される（略語化）
  The Ministry of International Trade and Industry → MITI
  S will approve → S to OK

このような英文ヘッドライン特有のスタイルをふまえた翻訳を行なう.

### 3.4 実例による説明

例えば図1の記事の場合，形態素解析，構文解析を行なったあと，重要文判定により第1文目を重要文に選定する. 次に第1文目から，様相的表現はないので主動詞を述部の「提言する」に決定する. 主動詞「提言する」の格フレーム情報 (図6) を見ると「を格」が必要要素であるから，文

| [主体 \| 文書] が　[抽象] を　提言する |
|---|

図 6:「提言する」の格フレーム情報

中から「を格」を探し，抽出する. この場合，「導入を」だけでは意味が通らないため，その直前の「の格」を同時に抽出する.

ヘッドライン加工部では，和文動詞の時制が過去形なので英文動詞の英文動詞の時制は現在形になる. また，ヘッドラインの特徴として冠詞を省略する. また，「社会保障制度審議会」は“Panel”1単語に省略される.

図3の記事の場合は，形態素解析，構文解析を行なったあと，重要文判定により第1文目を重要文に選定する. 次に第1文目から主動詞を決定するが，述部の「することを決めた。」は様相的表現なので，主動詞はその直前の動詞「出資する」であると判定する. 主動詞「出資する」の格フレーム情報は図7のようになっているが，この場合，2段

| [主体] が [制度 (経済)] を [主体 \| 人間活動] に出資する<br>[組織] が [組織] に 出資する |
|---|

図 7:「出資する」の格フレーム情報

目の情報に相当するので，主語と「に格」を抽出する.

和文動詞は現在形で内容がまだ未実現の出来事であるから，英文動詞の時制は to 不定詞になる.

## 4 ヘッドライン自動生成結果の評価・実験

### 4.1 実験方法と入力データ

実験に用いるデータは，対応づけ可能な日英記事として，日本経済新聞社の新聞記事に着目し，日経テレコンデータベースから，日本語記事はテレコン BIZ，英語記事は Japan News & Retrieval を用いた. テレコン BIZ と Japan News & Retrieval はある程度の記事対応付けが可能 [8] であるため，日本語記事に対応した英語記事に付与されているヘッドラインを正解の評価基準にすることができるからである.

実験では，上記データから無作為抽出した新聞記事 100 記事について日本語要約文と英文ヘッドラインの自動生成を行ない，被験者による，正解データのない評価と，英語記事原文のヘッドラインを正解データとした評価の2種類を行なった.

### 4.2 実験結果

図1，図3の日本語記事を入力とした場合，図2，図4の結果が得られるが，Japan News & Retrieval の同内容の英語記事に付与されているヘッドライン (以下，正解 HL) と比較すると，次のようになる.

【システムが自動生成したヘッドライン】
Panel offers creation of public nursing insurance
【正解 HL】
Panel proposes creation of public nursing insurance

【システムが自動生成したヘッドライン】
Mitsubishi Trust Bank and Sumitomo Insurance to invest at special purpose company Port Island Acceptance
【正解 HL】
Mitsubishi, Sumitomo to buy into Port Island Acceptance

図 8: 生成結果と正解 HL

### 4.3 評価と考察

実験で得られた日本語要約文と英文ヘッドラインについて評価を行った. 評価の対象は，要約文の生成とヘッドラインスタイル加工の2つの観点から行う. 前者は，日本語要約文について，どれだけ要約・情報抽出が出来ているかの観点で評価を行った. 後者は，理想的な日本語要約ができたと仮定したとき，どれだけ英文ヘッドラインスタイルに適した翻訳がされたか，という観点で評価を行った. なお，どちらも日経ヘッドライン (正解 HL) を正解基準とした.

#### 4.3.1 和文要約評価

正解 HL の和訳と比較して，必要な情報 (文節単位) がどれだけ抽出されているかを再現率と適合率で判定する. 再現率は，（正解和文に含まれる要約後和文の文節数／正解和文の文節数）で表される. 適合率は，（要約後和文に含

まれる正解和文の文節数／要約後和文の文節数）で表される．以下のような結果が得られた．

| | % |
|---|---|
| 再現率の平均 | 69.0 |
| 適合率の平均 | 84.5 |

正解 HL がある場合，再現率・適合率の評価から，必要情報の大部分が正しく抽出できていると考えられる．

次に，正解データを与えない場合の被験者の主観による文の意味判定，日本語要約文のみを見て，記事の内容が分かるかどうかを評価する．評価基準は以下の通りである．

◎：意味的に正しい要約になっており，正解 HL と語句的にも一致している．

○：意味的に正しい要約になっているが，語句が正解 HL と一致しない．

×：要約になっていない．

また，自動生成された英文ヘッドラインの内容についても同様の評価基準で文判定を行う．（この場合，ヘッドラインスタイルについての評価は行なわない）

以下のような結果が得られた．

| | ○ | × | |
|---|---|---|---|
| 和文意味判定 | 57 | 43 | |
| 英文意味判定 | 54 | 46 | (%) |

正解データを与えない場合，5～6割が意味的に要約として評価することができることが分かった．なお，英文ヘッドラインを生成した後にも，内容の劣化はあまりみられなかった．

#### 4.3.2 英文ヘッドラインスタイル評価

理想的な和文要約が出来た場合，どれだけ英文スタイルに変換できるかを判定する．人手で作成した理想和文要約からヘッドラインを自動生成した結果を評価する．

評価項目と基準は以下の通りである．

1. 略語化されているかどうか
2. 正しい動詞かどうか．
   ◎：正解 HL と一致している，
   ○：一致していないが意味は同じ，
   ×：一致していない．
3. 時制が整っているかどうか
4. 主語の一致（トピックがつかめているかの判定）．
   判定基準は 2 と同様．
5. 文の意味判定．
   判定基準は 4.3.1 節と同様．

以下のような評価結果を得た．

| | ○ | × | |
|---|---|---|---|
| 1. 略語化 | 91 | 9 | |
| 2. 動詞 | 75 | 25 | |
| 3. 時制 | 58 | 42 | |
| 4. 主語 | 78 | 22 | |
| 5. 文の意味 | 73 | 27 | (%) |

ヘッドライン加工ルールとしてみると，略語変換，動詞の選定は，7割以上が正しく訳されていることが分かる．同様に，主語が8割近く取れていること，意味判定が7割以上取れていることから，このルールを用いることによって，トピックを押さえたスタイル加工が可能であることが分かる．ただし，時制の変換については機械翻訳機 ALT-J/E の時制処理との不整合により，正解率が低下していることが分かったため，今後改良を行う．

## 5 おわりに

本稿では，日本語新聞記事本文から内容の要約である日本語見出しと英文ヘッドラインを自動生成する技術を紹介し，また，速報型新聞記事翻訳による情報発信支援と情報検索の効率化に向けたサービスイメージを提案した．見出し・ヘッドラインの生成結果は，情報抽出として7割程度，要約として5～6割程度の結果を得ることができた．これからは，2文目以降をターゲットとした要約を考えていく予定である．

## 参考文献


[1] H.P.Edmundson. New methods in automatic abtracting. Journal of ACM, Vol..16, No.2 (1996).

[2] H.P.Luhn. The automatic creation of literature abstracts. IBM Journal of Research and Development, Vol..2, No.2 (1958).

[3] S.Miike, et al.. A full-text retrieval system with a dynamic abstract generation function. In Proc. of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (1997).

[4] D.Marcu. From discourse structures to text summaries. In Proc. of the ACL Workshop on Intelligent Scalable Text Summarization (1997).

[5] 難波英嗣，奥村学．観点に基づいた新聞記事の重要文抽出に関する心理実験と考察．言語処理学会第4回年次大会 (1998).

[6] Satoru IKEHARA, Satoshi SHIRAI, Akio YOKOO and Hiromi NAKAIWA. Toward an MT system without pre-editing - effects of new methods in ALT-J/E. In Third Machine Translation Summit: MT Summit III, 101-106, Washington DC.(1991).

[7] 藤井章雄．ニュース英語の翻訳プロセス，早稲田大学出版部 (1996).

[8] 高橋大和，他．日英新聞記事の記事対応コーパス自動作成，言語処理学会第3回年次大会 (1997).

[9] 畑山満美子，他．日本文新聞記事からの英文ヘッドライン生成法について，情報処理学会第57回全国大会 (1998).