# 自動要約のための文重要度の比較

内山　将夫[†]　井佐原　均[†]

本稿では，重要文抽出によるテキスト自動要約のための各種重要度を比較した．特に，タイトルとの類似度の高い文から抽出するという要約方法を想定し，各種の類似度を比較した．類似度としては，共起関係を利用する方法と利用しない方法とを試みた．その結果，共起関係を利用する方法の方が高精度な要約が作成できた．また，要約の手法としては，他に，本文の先頭数文を抽出する方法と，単語の重要度の総和を文の重要度とする方法も試みたが，これらの方法よりも，タイトルとの類似度に基づく方法の方が高精度であった．これらの結果は，共起関係を利用した類似度が自動要約に有効であることを示している．

キーワード：　テキスト自動要約，重要度，類似度，共起関係

# Empirical Comparison of Sentence Importance Measures for Automatic Text Summarization

Masao Utiyama[†] and Hitoshi Isahara[†]

The effectiveness of various statistical measures of sentence importance was compared for automatic text summarization done by extracting important sentences. We focused on comparing various measures of sentence similarity on the assumption that important sentences in an article are similar to the title. Two types of similarity measures were compared: one uses word co-occurrence statistics and the other does not. The former proved superior to the latter. Other automatic text summarization methods, such as extracting the leading part of an article, or extracting sentences with important words, proved inferior to the similarity-based method. These results show that similarity measurement using word co-occurrence statistics is effective for automatic text summarization.

**KeyWords:** *automatic text summarization, importance, similarity, co-occurrence statistics*

## 1　はじめに

テキスト自動要約は，自然言語処理の重要な研究分野である．自動要約の方法には様々なものがあるが，現在の主流は，テキスト中から重要文を抽出して，それらを連結することにより要約を生成する方法である (奥村 難波 1999)．

重要文を選ぶための文の重要度は，一般に，

- 位置情報 (例：先頭部分の文は重要)
- 単語の重要度 (例：重要単語を含む文は重要)

† 郵政省通信総合研究所, Communications Research Laboratory, Ministry of Posts and Telecommunications

- 文間の類似関係 (例：タイトルと類似している文は重要)
- 文間の修辞関係 (例：結論を述べている文は重要)
- 手がかり表現 (例：「要するに〜」などで始まる文は重要)

などのテキスト中の各種特徴に基づいて決める (奥村・難波 1999).

これらの特徴の組合せは，人手で決める (Edmundson 1969) ことも，機械学習により決める (Kupiec, Pedersen, and Chen 1995) こともできるが，いずれの方法で決めるとしても，それぞれの特徴を精度良く自動的に求めることが重要である．

そのため，我々は，これらの特徴を個別に調査し，それぞれの自動要約への寄与を調べることを試みた．特に，本稿では，文間の類似度の各種尺度を，新聞記事要約を対象として比較した．類似度の良さは，要約の良さにより比較した．すなわち，精度の高い要約ができるような類似度ほど，高精度の類似度であると解釈した．ここで，文間の類似度を求める方法としては，単語間の共起関係を利用する方法と利用しない方法とを試みた．その結果は，共起関係を利用する方法の方が高精度であった．

なお，各種の類似度を比較するための要約方法としては，タイトルとの類似度が高い文から重要文として抽出するという方法を利用した．この要約方法を利用して類似度を比較した理由は，タイトルは本文中で最も重要であるので，それとの類似度が文の重要度として利用できると考えたからである．なお，タイトルが重要であるという考えに基づく要約には，(Yoshimi, Okunishi, Yamaji, and Fukumochi 1998; 奥西, 吉見, 山路, 福持 1998, など) がある．

また，要約の手法としては，他に，本文の先頭数文を抽出する方法 (Brandow, Mitze, and Rau 1995) と，単語の重要度の総和を文の重要度とする方法 (Zechner 1996) も試みたが，これらの方法よりも，タイトルとの類似度に基づく方法の方が高精度であった．

これらのことから，共起関係を利用した方法によりタイトルとの類似度を求め，その類似度が高い方から重要文として抽出する方法が，自動要約に有効であることが分かった．

以下では，まず，2章で，各種の文の重要度の定義を述べ，次に，3章で，各種重要度を比較した実験について述べる．4章は結論である．

## 2　文の重要度の定義

重要文を抽出するためには，文に重要度を付与する必要がある．そのための各種の重要度を以下に定義する．なお，以下では，重要度が数値的に高い文ほど重要な文であるとする．

### 2.1　先頭の数文を抽出する手法

文章 (特に新聞記事) の先頭の数文 (冒頭部) は重要と考えられる (Brandow et al. 1995)．その先頭の数文を抽出するためには，文 $S$ の文章中での位置を $p(>=1)$ とすると，$S$ の重要度

$lead(S)$ を，たとえば，

$$lead(S) = \frac{1}{p} \tag{1}$$

と定義すれば良い．なお，$lead(S)$ としては，その他には，$lead(S) = -p$ などを採用することもできる．

## 2.2　単語重要度の和による文重要度

重要単語を多く含む文は重要と考えられるので，単語の重要度の和を文の重要度とすれば良いと考えられる (Zechner 1996)．文の重要度を単語重要度の和として求めるためには，文 $S$ を構成する単語集合を $W(S)$ とすると，単語 $w$ の重要度を $f(w)$ とし，$w$ の $S$ 中での頻度を $n(w, S)$ とするとき，文 $S$ の重要度 $sum(S, f)$ を，

$$sum(S, f) = \sum_{w \in W(S)} n(w, S) \times f(w) \tag{2}$$

と定義すれば良い．

単語 $w$ の重要度 $f(w)$ としては，以下のものを定義する．

$$\begin{aligned} one(w) &= 1 && (3)\\ tf(w) &= \text{要約対象文書中での } w \text{ の頻度} && (4)\\ idf(w) &= \log \frac{\text{全文書数}}{w \text{ を含む文書数}} && (5)\\ tfidf(w) &= tf(w) \times idf(w) && (6) \end{aligned}$$

なお，(2) 式における $f(w)$ を $tfidf(w)$ としたときの重要度が，(Zechner 1996) で採用された文の重要度に相当する．

## 2.3　タイトルとの類似度による文重要度

タイトルは本文中で最も重要と考えられる (Yoshimi et al. 1998) ので，それと類似度が高い文は重要と考えられる．そのタイトルとの類似度を文の重要度とするために，文 $S$ とタイトル $T$ の類似度を以下に定義する．まず，共起頻度を利用しない方法 (以下の $common$, $ip$, $cos$) を定義する．これらは，類似度の定義として良く知られたものである．

### 共通単語の重みの和

$$common(S, T, f) = \sum_{w \in W(S) \cap W(T)} n(w, S) f(w) \tag{7}$$

内積

$$ip(S,T,f) = \sum_{w \in W(S) \cap W(T)} n(w,S)f(w) \times n(w,T)f(w) \tag{8}$$

コサイン

$$cos(S,T,f) = \frac{ip(S,T,f)}{\sqrt{ip(S,S,f)}\sqrt{ip(T,T,f)}} \tag{9}$$

ここで，$f$ は，(3) 式から (6) 式で定義された関数のいずれかである．

次に共起頻度を利用する方法 (以下の $coProb$, $coIDF$) を定義する．共起頻度は条件付き確率として式中に現れる．なお，以下で述べる，比較的簡単な式である $coProb$ と，IDF(Inverse Document Frequency) の拡張としての $coIDF$ とは，類似度の尺度として，本稿で新たに提案するものである．

共通単語の条件付き確率の和

$$coProb(S,T) = \sum_{w \in W(S)} n(w,S) \Pr(w|T) \tag{10}$$

**IDF の拡張**

$$coIDF(S,T) = \sum_{w \in W(S)} n(w,S) \times \Pr(w|T) \log \frac{\Pr(w|T)}{\Pr(w)} \tag{11}$$

ここで，

$$\Pr(w) = \frac{w \text{ を含む文書数}}{\text{全文書数}} \tag{12}$$

$$\Pr(w|T) = \max_{v \in W(T)} \Pr(w|v) \tag{13}$$

$$\Pr(w|v) = \frac{w \text{ と } v \text{ を含む文書数}}{v \text{ を含む文書数}} \tag{14}$$

である．

これらの確率の定義によると，$\Pr(x|x) = 1$ である．そのため，(11) 式は，

$$\begin{aligned} coIDF(S,T) &= \sum_{w \in W(S) \cap W(T)} n(w,S) \times idf(w) \\ &+ \sum_{w \in W(S) - W(S) \cap W(T)} n(w,S) \times \Pr(w|T) \log \frac{\Pr(w|T)}{\Pr(w)} \end{aligned}$$

と変形できる．つまり，$coIDF(S,T)$ は，共通単語 $w$ については，$idf(w) = 1 \times \log \frac{1}{\Pr(w)}$ の和を求め，それ以外については，共起確率を考慮した値 $\Pr(w|T) \log \frac{\Pr(w|T)}{\Pr(w)}$ の和を求めていると言える．これから分かるように，$\Pr(w|T) \log \frac{\Pr(w|T)}{\Pr(w)}$ は，$idf(w)$ の拡張と言える．

また，(10) 式は，

$$\begin{aligned} coProb(S,T) &= \sum_{w \in W(S) \cap W(T)} n(w,S) \times one(w) \\ &+ \sum_{w \in W(S) - W(S) \cap W(T)} n(w,S) \times \Pr(w|T) \end{aligned}$$

であるので，共通単語数を確率的に求めたものとも言える．

# 3　実験

本章では，各種重要度により重要文を抽出し，その抽出精度を求めた．

## 3.1　実験材料

### コーパス等

IDF 等を求めるときのコーパスとしては，「CD-毎日新聞」の 91-98 年版 (8 年間分) の約 86 万記事を茶筌 version2.02(松本, 北内, 山下, 平野 1999) により形態素解析した結果を用いた．なお，IDF 等の計算においては，1 記事を 1 文書とした．

また，各種の重要度を求めるときに，各文の単語を必要とするが，このときの単語としては，10 記事以上に出現した単語で，かつ，茶筌の品詞体系における，「名詞」「未知語」「記号-アルファベット」に該当するもののみを選んだ．ただし，名詞のうちで，その下位分類が「数」「代名詞」「非自立」「特殊」「接尾」「接続詞的」「動詞非自立的」に該当するものは除いた．

### 正解データ

「CD-毎日新聞」94 年版から抽出した 56 記事についての，被験者による重要文抽出結果を正解データとした[1]．これらの記事は以下の分布である．

- 14 記事からなるセットが 4 セットで合計 56 記事
- 各セットの 14 記事は、記事の長さを 100 文字単位で区切って，各文字数の範囲から 1 記事を無作為に選択．つまり，

1 正解データは筑波大学山本幹雄助教授に提供していただいた．

0〜99 文字　1 記事
100〜199 文字　1 記事
200〜299 文字　1 記事
...
1300〜1399 文字　1 記事

各セットは，3 または 5 人の被験者により要約された (set1,3 が 5 人，set2,4 が 3 人)．

被験者は，各記事から重要文を，抽出結果の分量が，元の記事の約 30%になるように抽出した[2]．その抽出結果についての諸元を表 1 に示す．これらの結果は，各記事を，抽出された文の数によりクラス分けした場合の統計である．なお，ある文が抽出されたとは，その文が，過半数の被験者 (5 人の場合は 3 人，3 人の場合は 2 人) により抽出されたことであると定義する．また，全記事とは，56 記事全てについての結果である．

表 1　被験者による重要文抽出結果の諸元

| 抽出文数 | 記事数 | 平均抽出数 | 平均記事長 | 抽出率 |
|---|---|---|---|---|
| 1 | 10 | 1.0 | 2.7 | 0.37 |
| 2 | 10 | 2.0 | 5.9 | 0.34 |
| 3〜4 | 13 | 3.6 | 12.6 | 0.28 |
| 5〜6 | 11 | 5.2 | 16.9 | 0.31 |
| 7〜11 | 12 | 8.6 | 27.6 | 0.31 |
| 全記事 | 56 | 4.2 | 13.8 | 0.31 |

表 1で,「抽出文数」により分かれる記事のクラスにおいて,「記事数」とは，そのクラスに属する記事の数である.「平均抽出数」とは，そのクラスの各記事から抽出された文数の平均値である.「平均記事長」とは，そのクラスの各記事に含まれる文数の平均値である.「抽出率」とは，平均抽出数を平均記事長で割った値である．

**被験者の重要文抽出精度**　5 人の被験者 (a,b,c,d,e) について，それぞれが選んだ文と正解データ (過半数の被験者が選んだ文) との再現率と適合率を表 2に示す．ただし，被験者 $x$ が選んだ文の集合を $S(x)$ とし，過半数の被験者に選ばれた文の集合を $M$ とするとき，

$$\text{再現率}\,(x) = \frac{|S(x) \cap M|}{|M|} \tag{15}$$

$$\text{適合率}\,(x) = \frac{|S(x) \cap M|}{|S(x)|} \tag{16}$$

である．

2 被験者は，実際には，重要文から，更に，重要文節も抽出したが，その情報は今回の実験では使用しなかった．また，抽出された重要文についても，特に重要な文と，その他の重要文という 2 通りが被験者により区別されているが，今回の実験では，この区別は無視して，抽出された文は，区別なく全て重要文とした．

表 2　被験者の重要文抽出精度

| 抽出文数 | 再現率 | | | | | 適合率 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | a | b | c | d | e | a | b | c | d | e |
| 1 | 0.86 | 0.80 | 1.00 | 1.00 | 1.00 | 0.75 | 0.67 | 0.88 | 0.80 | 0.78 |
| 2 | 0.88 | 0.75 | 0.88 | 0.92 | 0.94 | 0.88 | 0.67 | 0.82 | 0.85 | 0.94 |
| 3〜4 | 0.75 | 0.71 | 0.81 | 0.82 | 0.98 | 0.64 | 0.61 | 0.69 | 0.68 | 0.83 |
| 5〜6 | 0.83 | 0.63 | 0.86 | 0.79 | 0.97 | 0.79 | 0.56 | 0.78 | 0.71 | 0.90 |
| 7〜11 | 0.78 | 0.76 | 0.82 | 0.80 | 0.86 | 0.78 | 0.74 | 0.83 | 0.76 | 0.87 |
| 全記事 | 0.80 | 0.71 | 0.84 | 0.82 | 0.92 | 0.75 | 0.65 | 0.79 | 0.73 | 0.87 |

表 2の再現率や適合率が高いのは，正解データをこれらの被験者から作成したので，ある程度は，当然であるが，それでも，後掲の表 3に示す，自動抽出の結果に比べるとずいぶんと高い．統計的には，全記事を対象として，再現率と適合率とを考えたとき，もっとも数値の低い被験者 b の適合率 0.65 を除いては，自動抽出で最高精度である $coIDF$ の結果と比べても，比率の差の検定による片側検定で，全てが有意水準 5% で有意に再現率や適合率が高い．

## 3.2　実験方法と実験結果

正解データの与えられた 56 記事を要約の対象とし，茶筅により形態素解析し，その結果について，各種重要度を適用して重要文を抽出した．各記事から抽出する文数は，正解データにおける抽出文数と同じにした．これは，(15) 式と (16) 式において，$|M| = |S(x)|$ であることを意味する．したがって，再現率と適合率とが等しくなる．そのため，本節では，それらを単に精度と呼ぶことにする．

表 3は，2章で定義した各種重要度について，抽出文数によりクラス分けされた記事について，抽出精度を求めたものである．たとえば，まず，$lead(S)$ は，抽出文数 1 の記事に対しては，精度 1.00，つまり，1 文だけを抜き出すなら，先頭文を抜き出すと必ず正解であることを示す．次に，たとえば，$sum(S, one)$ は，(2) 式の関数 $f$ として，(3) 式の関数 $one$ を用いたことを示し，$common(S, T, one)$ は，タイトル $T$ との類似度を，関数 $one$ により，(7) 式を用いて求めたことを示す．

表 3では，$lead(S)$ をベースラインとして，各種重要度を評価した．このとき，もし，ある重要度が $lead(S)$ と比率の差の検定による両側検定により有意に精度が異なるときには，有意水準 5%のときには，‘+/−’，有意水準 1%のときには，‘++ / −−’ を付けてそれを示した．ここで，正の符号は，その重要度が $lead(S)$ よりも精度が高いことを示し，負の符号は，その逆を示している．

表 3から，抽出文数 1, 2, 3〜4, 5〜6(平均記事長は，それぞれ，2.7, 5.9, 12.6, 16.9 文) については，$lead(S)$ の精度が他よりも良いか同等であることが分かる．これは，先頭部に重要なことが書かれているという新聞記事の性質を反映している．しかし，抽出文数 7〜11(平均記事長

表 3　各種重要度による重要文抽出精度 (=適合率, 再現率) の比較

| 重要度 | 抽出文数と精度 | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3〜4 | 5〜6 | 7〜11 | 全記事 |
| $lead(S)$ | 1.00 | 0.65 | 0.68 | 0.49 | 0.42 | 0.53 |
| $sum(S, one)$ | $0.50^{--}$ | 0.75 | $0.43^{-}$ | 0.40 | 0.53 | 0.50 |
| $sum(S, tf)$ | 0.70 | 0.70 | 0.55 | 0.46 | 0.50 | 0.53 |
| $sum(S, idf)$ | $0.50^{--}$ | 0.70 | $0.45^{-}$ | 0.42 | 0.50 | 0.49 |
| $sum(S, tfidf)$ | 0.70 | 0.65 | 0.49 | 0.40 | 0.55 | 0.52 |
| $common(S, T, one)$ | 0.80 | 0.75 | 0.49 | 0.49 | 0.55 | 0.55 |
| $common(S, T, tf)$ | 0.80 | 0.75 | 0.49 | 0.46 | 0.53 | 0.54 |
| $common(S, T, idf)$ | 0.80 | 0.70 | $0.47^{-}$ | 0.49 | $0.56^{+}$ | 0.55 |
| $common(S, T, tfidf)$ | 0.90 | 0.70 | 0.49 | 0.47 | 0.52 | 0.54 |
| $ip(S, T, one)$ | 0.80 | 0.75 | $0.47^{-}$ | 0.49 | 0.55 | 0.55 |
| $ip(S, T, tf)$ | 0.90 | 0.75 | 0.49 | 0.44 | 0.51 | 0.53 |
| $ip(S, T, idf)$ | 0.80 | 0.70 | 0.49 | 0.47 | 0.53 | 0.54 |
| $ip(S, T, tfidf)$ | 0.90 | 0.70 | $0.47^{-}$ | 0.44 | 0.48 | 0.50 |
| $cos(S, T, one)$ | 0.80 | 0.65 | 0.49 | 0.49 | 0.51 | 0.53 |
| $cos(S, T, tf)$ | 0.80 | 0.65 | $0.47^{-}$ | 0.46 | 0.50 | 0.51 |
| $cos(S, T, idf)$ | 0.80 | 0.65 | $0.47^{-}$ | 0.46 | 0.52 | 0.52 |
| $cos(S, T, tfidf)$ | 0.80 | 0.70 | $0.45^{-}$ | 0.44 | 0.49 | 0.50 |
| $coProb(S, T)$ | 0.80 | 0.75 | 0.53 | 0.47 | $0.62^{++}$ | 0.59 |
| $coIDF(S, T)$ | 0.80 | 0.75 | 0.53 | 0.54 | $0.61^{++}$ | 0.60 |

=27.6) になると，先頭部のみでは，カバーできる重要文が少なくなるため，$lead(S)$ は他と比べて有効な方法ではなくなる．

全記事での精度に基づいた結果から，まず，単語重要度の組合せ方の精度を比較すると，

$$coIDF, coProb(0.595) >= common(0.545) >= ip(0.53) >= cos(0.515) >= sum(0.51) \quad (17)$$

である．この順位は，たとえば，$common$ については，$common(S,T,one)$, $common(S,T,tf)$, $common(S,T,idf)$, $common(S,T,tfidf)$ の全記事についての精度の平均を求めると，$(0.55 + 0.54 + 0.55 + 0.54)/4 = 0.545$ であり，$coIDF$ と $coProb$ では，$(0.59 + 0.60)/2 = 0.595$ であることなどから順位付けた．なお，括弧内の数字は，求めた平均値である．

この結果から，$coIDF$ と $coProb$ が他よりも重要文選択に適した重要度であることが分かるが，この結果は統計的には有意ではない．統計的に有意であることを示すには，より規模の大きい実験が必要である．ただし，$coIDF$ と $coProb$ は，表 3でも，抽出文数 7〜11 の場合には，$lead(S)$ と比べて，有意水準 1% で高精度に重要文を抽出できるので，長い記事については，$lead(S)$ を使うよりも，これらの共起情報を利用した重要度を使った方が良いと言える．また，短い記事についても，共起情報を利用した重要度は，$lead(S)$ と比べて，統計的には同等であるので，共起情報を利用した重要度は，自動要約に適していると言える．

## 4　おわりに

重要文抽出によるテキスト自動要約のために，各種の重要度を比較した．本稿では，特に，文間の類似度の各種尺度を，新聞記事要約を対象として比較した．このとき，文の重要度は，タイトルとの類似度により定義した．

文間の類似度を求める方法としては，単語間の共起関係を利用する方法と利用しない方法とを試みた．実験の結果，共起関係を利用した類似度の方が，高精度な要約ができた．この結果から，共起関係を利用した類似度が自動要約に有効であると言える．

我々は，今後，本稿での知見に基づいて，各種情報を統合した自動要約システムを作ることを考えている．また，本稿で提案した，IDF の拡張としての類似度を，自動要約だけでなく，情報検索にも応用して，その有効性を確かめたいと考えている．

## 参考文献

Brandow, R., Mitze, K., and Rau, L. F. (1995). "Automatic Condensation of Electronic Publications." *Information Processing and Management*, **31** (5), 675–686.

Edmundson, H. P. (1969). "New Methods in Automatic Extracting." *Journal of the ACM*, **16** (2), 264–285.

Kupiec, J., Pedersen, J., and Chen, F. (1995). "A Trainable Document Summarizer." In *Proc. of the 18th ACM-SIGIR Conference*, pp. 68–73.

松本裕治, 北内啓, 山下達雄, 平野善隆 (1999). "日本語形態素解析システム『茶筅』 version 2.0 使用説明書第二版." 奈良先端科学技術大学院大学.

奥村学 難波英嗣 (1999). "テキスト自動要約に関する研究動向 (巻頭言に代えて)." 自然言語処理, **6** (6), 1–26.

Yoshimi, T., Okunishi, T., Yamaji, T., and Fukumochi, Y. (1998). "Evaluation of Importance of Sentences based on Connectivity to Title." In *COLING-ACL'98*, pp. 1443–1447.

Zechner, K. (1996). "Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant sentences." In *COLING-96*, pp. 986–989.

奥西稔幸, 吉見毅彦, 山路孝浩, 福持陽士 (1998). "ウェブ英文ページの速読支援." 言語処理学会第 4 回年次大会発表論文集, pp. 572–575.

### 略歴

内山 将夫: 筑波大学第三学群情報学類卒業 (1992). 筑波大学大学院工学研究科博士課程修了 (1997). 博士 (工学). 信州大学工学部電気電子工学科助手 (1997). 郵政省通信総合研究所非常勤職員 (1999). 言語処理学会，情報処理学会，ACL，人工知能学会，日本音響学会，各会員.

井佐原 均： 1978 年京都大学工学部電気工学第二学科卒業．1980 年同大学院修士課程修了．博士（工学）．同年通商産業省電子技術総合研究所入所．1995 年郵政省通信総合研究所関西支所知的機能研究室室長．自然言語処理，機械翻訳の研究に従事．言語処理学会，情報処理学会，人工知能学会，日本認知科学会，ACL，各会員．