自然言語処理 120−11
(1997. 7. 25)

人間の重要文判定に基づいた自動要約の試み

野本 忠司
(株) 日立製作所 基礎研究所
〒 350-03 埼玉県比企郡鳩山町赤沼 2520
email: nomoto@harl.hitachi.co.jp

松本 裕治
奈良先端科学技術大学院大学
〒 630-01 奈良県生駒市高山 8916-5
email: matsu@is.aist-nara.ac.jp

あらまし
本稿では大学生を中心とした被験者 112 名について要約文指摘能力に関する調査を行い、そのデータをもとにした自動要約手法について述べる。要約問題は日本経済新聞 95 年の記事から随想 (春秋)、社説、一面報道の各分野別に粒度の揃った記事を選び作成した。調査結果は Kappa 統計と呼ばれる尺度で評価し、さらに調査データから要約文判定の被験者間一致度を調べ、その高低に応じていくつかのデータセットを作成した。一方、自動要約の手法として、C4.5 学習アルゴリズムを使い、上記データセットに対する要約モデルの生成とテスト実験をおこなった。実験の結果、K 値と自動要約モデルの性能との間に相関傾向があることが認められた。

キーワード　自然言語処理、日本語、自動要約、コーパス

The Reliability of Human Coding and Effects on Automatic Abstracting

Tadashi Nomoto
Advanced Research Laboratory, Hitachi Ltd.
2520 Hatoyama Saitama 350-03 Japan

Yuji Matsumoto
Nara Institute of Science and Technology
8916-5 Takayama Ikoma Nara, 630-01 Japan

Abstract
We discuss a particular approach to automatic abstracting, where an abstract is created by extracting important sentences from a text. A primary purpose of the paper is to demonstrate that the reliability of human supplied annotations on corpora has crucial effects on how well an automatic abstracting system performs. The corpus is developed through human judgements on possible summary sentences in a text. The reliability of human judgements is evaluated by the kappa statistic, a reliability metric standardly used in behavioral sciences. The C4.5 decision tree method (Quinlan, 1993) is used to build a extraction model. We demonstrate that there is a positive correlation of data reliability with a performance of automatic abstracting, and show results indicating that the reliability of human provided data is crucial for improving the performance of automatic abstracting.

key words　NATURAL LANGUAGE PROCESSING, JAPANESE, AUTOMATIC ABSTRACTING, SUMMARIZATION, CORPUS

1 はじめに

従来の自動要約の研究は大きく生成派と抽出派に分れる。 生成派は、一般に文章内から要約に役立ちそうな手がかり表現を拾いだし、それらをもとに読解可能な文表現を生成すること目的としている。(McKeown and Radev, 1995; Hovy, 1993) 特に生成派では生成された要約の流暢さが問題にされる。一方、抽出派は要約を「切りぬき」と考える立場で、客観的評価が可能になるように問題を設定しようという立場である。(Kupiec et al., 1995; Zechner, 1996; Miike et al., 1994; Edmundson, 1969) 要約の可読性、流暢さは問題にされない。本稿では評価優先の立場から、後者の立場を取り、文章の内容をよく表わすと思われる文を抽出し、それ(ら)をもって要約と定義することにする。

抽出派は、すでに何らかの方法で要約文をマークアップしたデータを評価用あるいは訓練用として使うのが一般的であるが、マークアップの正当性については問題視されてこなかったように思われる。従来研究では、マークアップは実験者が自分で行うか (Watanabe, 1996)、他人に任せてしまうか (McKeown and Radev, 1995)、あるいは Kupiec et al. (1995) のように要約付きの技術論文について、要約から 「逆算」して本文の要約文を見付けるという、少々回りくどいやり方が取られてきた。

このような状況のもと、本稿では要約文マークアップとその評価をより客観的かつ直接的に行うため、複数の被験者の判定に基づいたマークアップ方式を採用することにした。要約文か否かは被験者間の判定の一致度を参考にして決定する。また、値の定性的な意味付けのしやすさなどから、一致度の測り方として行動科学・心理学などで使われている K(kappa) 統計と呼ばれる手法を導入することにした。(Jean Carletta, 1996; Sidney Siegel and N. John Castellan Jr., 1988)

さらに、本稿では自動要約を人手による要約文マークアップを学習データとした C4.5 決定木 (Quinlan, 1993) による文分類という視点で捉える。

2 方法

2.1 要約文指摘に関する調査

今回、関東及び関西の大学(院)生を被験者にして、実際に文章中から要約文を指摘してもらった。総勢 112 名の被験者の協力が得られた。文章は日本経済新聞 95 年 (日本経済新聞社, 1995) から分野を限定して無作為に選択したものを被験者各人に読んでもらい、重要だと思う文を文章中から選択してもらった。選択数は文章中の文の総数の 10% 相当分とした。選択数の範囲は 1 文章あたり大体 2 ～ 4 文程度であった。 また、被験者の年齢の幅は下は 18 才から上は 45 才であった。文章は 3 つの分野 (随想、社説、一面報道) に限定して、日本経済新聞 95 年 1 年分から、分野間で粒度が揃うようほぼ同じ文字数・段落数の記事を選択したが、随想 (春秋) については記事の文字数がほぼ一定であるため、選択の余地がなく、記事あたり平均して 640 文字の長さになった。(表 1) さらに、上記の手順で得られたもののなかから 1 分野 25 本合計 75 本の記事を無作為に抽出し最終的な調査用文章とした。 1 テストはそれぞれの分野の調査用材料から一つづつ無作為に選びだした合計 3 つの要約課題より成る。

表 1: 調査用記事の分野とサイズ

分野	文字数	段落数	記事数
随想 (春秋)	約 640	4-5	352
社説	900-1100	6-9	131
一面報道	800-1000	6-9	147

被験者のうち 85 名は、テストを一つ受けてもらったが、残りの 27 名については、被験者数の不足からテストを 5 つ受けてもらった。テスト一つにつき平均 7 人の被験者を割り振った。

調査の手続きとしては、被験者に文章課題を印刷した冊子を配布し、文章を読んでその文章において重要だと思う文に (指示された数だけ) ○印を付けてもらった。時間は 1 テストあたり 30 分程度であった。

2.2 調査の評価

Kappa 統計 今回の調査の評価として、被験者間での反応の一致の度合を見るため Kappa 統計と呼ばれる指標を導入した。(Jean Carletta, 1996; Sidney Siegel and N. John Castellan Jr., 1988) Kappa 値は、被験者間の一致度を反応が偶然に一致する場合を考慮して調整した値である。具体的には以下の式で定義される値である。

$$K = \frac{P(A) - P(E)}{1 - P(E)} \quad (1)$$

ここで、$P(A)$ は、被験者全体を二人組のペアの集合にしたとき、その集合のうち反応が一致してる組の割合である。 $P(E)$ は一致が偶然おこる期待確率である。

表 2: 分析表

	1	2	⋯	j	⋯	m	
1	n_{11}	n_{11}	⋯	n_{1j}	⋯	n_{1m}	S_1
2	n_{21}	n_{22}	⋯	n_{2j}	⋯	n_{2m}	S_2
⋮				⋮			⋮
i	n_{i1}	n_{i2}	⋯	n_{ij}	⋯	n_{im}	S_i
⋮				⋮			⋮
N	n_{N1}	n_{N2}	⋯	n_{Nj}	⋯	n_{Nm}	S_N
	C_1	C_2	⋯	C_j	⋯	C_m	

K は $P(A)$ を $P(E)$ で補正した値になっている。実際の計算手順は以下のようにする。まず、調査課題を多肢選択問題と見て、結果を調査 (質問) 項目と選択肢の種類のマトリックスで表現する。 例えば、表 2では行が項目 ($1 \leq i \leq N$) を、列が選択肢の種類 ($1 \leq j \leq m$) を表わす。n_{ij} は i 番目の項目に対する j 番目の答を選択した被験者 (投票) 数を表わす。つぎに i 番目の項目における被験者間一致度 S_i を以下の式 2で求める。

$$S_i = \frac{\sum_{j=1}^{m} \binom{n_{ij}}{2}}{\binom{k}{2}} \tag{2}$$

各 S が求まったら、その平均を $P(A)$ とする。(式 3)

$$P(A) = \frac{1}{N} \sum_{i=1}^{N} S_i \tag{3}$$

一方、各選択肢 C_j についてその選択肢が選択される期待値 P_j を以下で求める。

$$p_j = \frac{C_j}{N \cdot k}$$

ここで N は総項目数、k は被験者の総数である。要するに p_j は選択肢 C_j の得票率の期待値である。期待確率 $P(E)$ は以下で与える。

$$P(E) = \sum_{j=1}^{m} p_j^2 \tag{4}$$

なお、$\sum_{j=1}^{m} n_{ij} = k$ である。

次に今回の調査を上記のマトリックス形式に翻訳してみる。調査の課題は、文章のなかから、重要だと思われる文を指示された数だけ指摘することであった。列には文番号、行は第一番目、第二番目、第三番目等の重要文とする。ただし、第一、第二、第三は重要度の順序ではなく、文の出現順序に対応させた。 従って、質問項目としては、「第一番目に出現した重要文はどれですか。」、「第二番目に出現した重要文はどれですか。」、「第三番目に出現した重要文はどれですか。」と聞いていることになる。 表 3に今回の調査結果の K 値を載せた。(K 値は、文章単位で計算する。) また、K 値とデータの信頼性との関係は表 4(Carletta et al., 1997) のようになる。分野間で多少の違いはあるものの、全体的に非常に悪い値になった。例えば、春秋の K 値は 0.122 なのでマークアップの信頼性は「ほとんどなし」ということになる。ちなみに、ここでいう信頼性とはデータの再現性をさす。(Krippendorff, 1980)

表 3: 被験者判定の Kappa 値。ただし、判定者数には重複あり。

分野	K 値	記事数	判定者数
春秋	0.122	25	183
社説	0.156	25	184
報道	0.255	25	183

表 4: K 値の信頼性解釈 (Carletta et al., 1997)

K 値			信頼性
	<	0	POOR
.0	–	.20	SLIGHT
.21	–	.40	FAIR
.41	–	.60	MODERATE
.61	–	.80	SUBSTANTIAL
.81	–	1.0	NEAR PERFECT

多数決による候補の足切り したがって、得票があったすべての文を正しい判定 (要約文) と見なすには信頼性の観点から無理がある。そこで本稿では得票数に応じて要約文候補の足切りを行うことにした。これには以下のような手順を踏む。 基本的には、ある一定数の 得票のあった文を当選 (要約文) と見なし、それ以外の候補への投票を無効にする。具体的な進め方としては、ある分野について全体がある一定の K 値に達するまで当選に必要な得票数を 1 から順に増やしていく。それに満たない文への投票を無効にする (集計データから除く)。また、文章中の文への投票がすべて無効になったら、その文章も無効にする。

表 5: K 値による足切り効果

K 値	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
春秋	0.25(23)	0.37(21)	0.50(21)	0.55(20)	0.59(18)	0.73(10)	0.75(7)	1.00(1)
社説	0.20(24)	0.35(22)	0.49(20)	0.55(20)	0.62(18)	0.68(12)	0.87(5)	0.95(3)
報道	0.26(25)	0.38(25)	0.52(24)	0.62(23)	0.65(23)	0.76(13)	0.82(9)	1.00(5)

K 値を 0.1 から徐々に上げていくと、表 5に示すような結果が得られた。括弧内は有効文章数。注意すべきことは、 表 5の結果は実際に観測されたデータではなく、多数決票の K 値であることである。 したがって、われわれは多数決による項目分類の信頼性について評価していることになる。

2.3 要約手法

本節では、要約文抽出方法について述べる。基本的な方針は C4.5 (Quinlan, 1993) による決定木を使って、文章中の各文を要約文か否かに分類することである。決定木とは予め分類済みのデータから属性情報とある基準によって、そのデータを正しく分類するようなルール郡を抽出することが目的である。ただし、属性の選び方は基本的に経験的であり、理論的な指針というものはない。 今回は要約研究で比較的一般的なものを参考にして決めた。(Kupiec et al., 1995; Paice and Jones, 1993; Edmundson, 1969; Zechner, 1996)

2.4 属性

以下に今回実験で用いた属性とその説明をする。ここでは文を 1 ケースに対応させる。したがって、属性は文単位で与える。

- **分野** 意味：文の属する分野。属性値：「随想」「社説」「報道」のどれか。
- **テキスト内位置** 意味：文章内の文の位置。属性値：先行する文の数を文章中の文の総数で割った値。(Edmundson, 1969)
- **見出しとの類似度** 意味：文と文章の見出し (タイトル) との類似度。属性値： TF·IDF で定義。具体的には、見出し T と文 S について、類似度を以下の式で与える。

$$SIM(T,S) = \sum_{w \in W(T)} NF(w,S) \cdot IDF(w)$$

表 6: 文の属性表現

```
C,0.941,0.000,28,1,2.900,0.333,Y
E,0.000,0.717,31,1,6.366,0.000,Y
G,0.167,0.339,26,1,5.966,0.600,N
```

ここで w は見出し中の名詞を表わす。また NF、IDF は以下のように与える。

$$NF(w,S) = \frac{F(w,S)}{MAX\_F(S)}$$

$F(w,S)$ は w の S における頻度。$MAX\_F(S)$ は S において最高頻度の名詞の頻度。

$$IDF(w) = \frac{\log \frac{N}{DF(w)}}{\log N}$$

N は (S が属する) 文章の総文数。$DF(w)$ は w が出現した文の総数。

- **テキスト内 $TF \cdot IDF$** 意味：文の内容的独立度。(どれくらい他の文と内容的に違うこと言っているか。) 属性値：これも TF·IDF で定義。

$$D(S) = \sum_{w \in W(S)} NF(w,S) \cdot IDF(w)$$

S は文で、S に含まれる名詞 w について NF と IDF を計算する。いづれも定義は前と同じ。

- **態度表現の型** 意味：態度表現の型。属性値：タイプ 1 (= 態度表現なし)、タイプ 2 (= 話者の主観を表わす述語表現 (「必要だ」「希望する」等))、タイプ 3 (= 終助詞「か」「よ」「ね」)。
- **文の長さ** 意味：文の長さ。属性値：文の文字数。(Kupiec et al., 1995)
- **段落内位置** 意味：段落おける文の位置。属性値：(段落中における先行文の数)/(段落の文の総数)

上記の属性を用いて、文を表現した例を表 6に示す。1 行が 1 文に対応しており、属性値はコンマ「,」で

区切ってある。最初から順に「分野」(C: 春秋；E: 社説；G: 報道)「テキスト内位置」「見出しとの類似度」「文の長さ」「態度表現の型」「テキスト内 TF·IDF」「段落内位置」「分類」(Y: 要約文；N: 非要約文) を表わす。

2.5 評価実験

本節では、上記要約手法の実験と評価について説明する。実験の目的は、各 K の足切りレベルと要約モデルの性能の関係を見ることである。 K 値が 0.5 を越えるデータは有効文章数が少いので実験対象から外した。例えば、表 5を見ると K が 0.7 の時すべての分野について有効文章数が一桁代に落ているのがわかる。

さて、評価実験は以下のようにして進めた。

Step 1 まず、実験用データを分野別に用意し、それぞれの分野ごとに要約文を 40 文、非要約文を 200 文、無作為に実験用データから抽出し分野別の評価用データを作成する。

Step 2 次に、評価用データを 10 等分し 9 ブロックを学習用、残り 1 ブロックをテスト用データにする。(*10-fold cross validation*) 学習用データを使って決定木の学習をおこない、テスト用データで決定木の評価をおこなう。

Step 3 どのブロックをテスト用にまわすかで 10 通りの場合ができるので、そのそれぞれ場合について学習と評価をおこなう。

Step 4 最後に各場合の評価の平均を取って全体の評価とする。

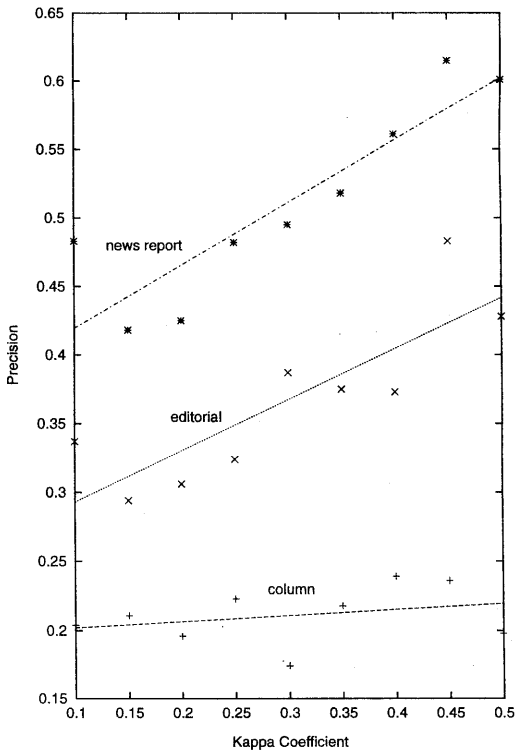
さらに評価用データの選び方に偏りをなくすため、Step 1-4 の実験サイクルを 50 回繰り返して、その平均を取った。ただし、評価値は、要約文が正しく分類された割合、つまり、適合率 (precision) で定義した。

評価実験の結果を図 1に、またその詳細を表 7に示す。多少の振動はあるものの全体的に明確な傾向が見られる。 K 値が上るとそれに応じて要約モデルの精度が上昇しており、 K 値と要約モデルとの間に相関傾向が認められる。ただし、相関傾向は分野間でかなり差があり、特に column(春秋) 記事ではまったく相関が認められない。このような差の原因として、用いた属性セットが分野によっては最適ではないということが考えられる。 あるいは、分野 (ジャンル) の性質として、そもそも要約文を持たないということもありうる。

表 7: K 値と精度の関係 (1)。括弧内は再現率 (recall)。

K	春秋	社説	報道
0.10	0.204 (0.113)	0.337 (0.195)	0.483 (0.307)
0.15	0.211 (0.119)	0.294 (0.167)	0.418 (0.262)
0.20	0.196 (0.118)	0.306 (0.189)	0.425 (0.267)
0.25	0.223 (0.127)	0.324 (0.198)	0.482 (0.307)
0.30	0.174 (0.092)	0.387 (0.249)	0.495 (0.322)
0.35	0.218 (0.117)	0.375 (0.271)	0.518 (0.366)
0.40	0.239 (0.138)	0.373 (0.253)	0.561 (0.395)
0.45	0.236 (0.134)	0.483 (0.349)	0.615 (0.466)
0.50	0.198 (0.114)	0.428 (0.316)	0.601 (0.462)

図 1: K 値と精度の関係 (2)。縦軸は精度 (適合率)、横軸は K のレベル。直線は回帰直線。 $Y = 0.197800 + 0.0440 * X$ (column = 春秋); $Y = 0.255844 + 0.3720 * X$ (editorial= 社説); $Y = 0.373789 + 0.4570 * X$ (news report= 報道).



2.6 まとめと課題

本稿では大学生を中心とした被験者112名について要約文指摘能力に関する調査を行い、そのデータをもとにした自動要約手法について述べた。要約問題は日本経済新聞95年の記事から随想(春秋)、社説、一面報道の各分野別に粒度の揃った記事を選び作成した。調査結果はKappa統計と呼ばれる尺度で評価し、さらに調査データから要約文判定の被験者間一致度を調べ、その高低に応じていくつかのデータセットを作成した。一方、自動要約の手法として、C4.5学習アルゴリズムを使い、上記データセットに対する要約モデルの生成とテスト実験をおこなった。実験の結果、K値と自動要約モデルの性能との間に相関傾向があることが認められた。

今後の問題としては、抽出要約文の利用方法が内容理解のしやすさの点からかなり限定されることであろう。ユーザに対して文脈なしの要約文の提示はあまりに唐突であり、それだけでは理解が困難である。したがって、要約文を内容的に完結させる文脈をどのように構成するかが今後の問題となる。

謝辞

要約文調査にあたっては以下の諸氏にご協力いただきました。この場を借りて、感謝します。高橋 秀明氏(文部省マルチメディア教育センター)、吉田 佐知子氏(文教大学)、芳賀 純氏(関西福祉大学)、宇津呂 武仁氏(奈良先端大学)、宮田 高志氏(奈良先端大学)。

References

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. 1997. The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, 23(1):13–31.

H. P. Edmundson. 1969. New Method in Automatic Abstracting. *Journal of the ACM*, 16(2):264–285, April.

E. Hovy. 1993. Automated Discourse Generation using Discourse Structure Relations. *Artificial Intelligence*, 63:341–385.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*, volume 5 of *The Sage* COMMTEXT *series*. The Sage Publications, Inc.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A Trainable Document Summarizer. In *Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Developmnet in Information Retrieval*, pages 68–73. Seattle, USA.

Kathleen McKeown and Dragomir R. Radev. 1995. Generating Summaries of Multiple News Articles. In *Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Developmnet in Information Retrieval*, pages 68–73. Seattle, USA.

Seiiji Miike, Etsuo Itoh, Kenji Ono, and Kazuo Sumita. 1994. A Full-text Retrieval System with a Dynamic Abstract Generation Function. *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 152–159. Dublin, Ireland.

Chris D. Paice and Paul A. Jones. 1993. The Identification of Important Concepts in Highly Structured Technical Papers. In *The Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 69–78. Pittsburgh, USA.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Sidney Siegel and N. John Castellan Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Second edition.

Hideo Watanabe. 1996. A Method for Abstracting Newspaper Articles by Using Surface Clues. In *Proceedings of the 16th International Conference on Computational Linguistics*, volume 2, pages 974–979, August. Copenhagen, Denmark.

Klaus Zechner. 1996. Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 986–989. Copenhagen, Denmark.

日本経済新聞社. 1995. 日本経済新聞95年CD-ROM版. 東京.