観点に基づいた新聞記事の重要文選択に関する心理実験と考察

難波 英嗣, 奥村 学
北陸先端科学技術大学院大学 情報科学研究科
e-mail:{nanba,oku}@jaist.ac.jp

1 はじめに

膨大な数のテキストから効率良く情報を得るために, 要約を利用するという方法がある. この要約を, テキスト中から重要と思われる文を抜き出す, いわゆる抄録をコンピュータで自動生成するという研究がある[1]. 要約研究では, テキスト中の個々の文の重要性をどのように評価するかが問題となり, また様々な提案がなされてきた. しかし一口に重要と言っても何が重要であるかは読み手によって異なってくる. 例えばあるテキストが「○×会社の女性雇用の現状」について書かれている場合, ○×会社に興味を持つ読み手は 「最近の○×会社の動向」の一部ととらえるであろう. またある読み手は 「一般的な女性雇用問題の一例」と考えるかもしれない. このように書き手がひとつの主題で書いたテキストを, 読み手側で様々な主題として受け取る可能性がある. これまで多くの要約研究では, 生成される要約の正解はひとつであるという仮定が前提にあった [3, 4]. しかしひとつのテキストも, 読み手の観点に応じて色々な抄録が存在しうるものと考えられる.

この問題に関連して, 要約研究者の間で近年動的な要約生成の必要性が認識されつつある. 「動的な要約生成」とは, 「要約の利用される状況で, ユーザの要求に応じた要約を生成する」ことである. たとえば, 情報検索において, ユーザがクエリを入力し, 検索されたテキストが適切かどうかを判断する際に要約を用いる場合を考えると, 要約はユーザが入力したクエリに即したものになっている必要がある. 本研究は, 要約研究において動的な要約手法に関する研究のための基礎的な調査として位置付けられる.

1.1 ひとつのテキストからの複数の抄録作成

Paice は自動要約生成研究で, 要約の正解に関して以下の指摘をしている. [2].

> **Edmundson の手法 [4] は正解の文書セットが 1 つのみで, もっと低いレベルの agreement で, 人間が抽出してもよいというものは正解として含まれていない.**

あるテキストに要約が複数存在しうる理由のひとつに, 本研究では読み手の主観の違いが挙げられる. 数人の被験者にテキストに対する抄録を作成してもらう時, 予め観点を与えなければ, 観点を与えた場合と比べて被験者が重要と考える文にばらつきがあると仮定できる. この仮定について, 本研究では日本語情報検索システム評価用テストコレクション BMIR-J1[5] 中の 25 テキストを用いて調査を行った.

2 BMIR-J1 を用いた心理実験

2.1 実験の目的

抄録生成において被験者に予め観点を与える/与えないで, 被験者が重要であると考える文にばらつきがあるか, またばらつく場合にはどのような傾向でばらついているのかを調べるために心理実験を行った.

2.2 実験方法

実験は 2 種類行った.

2.2.1 被験者に予め観点を与えない抄録作成 (実験 1)

まず, 実験 1 では被験者に観点を与えずにテキスト中の個々の文の重要度を 3 段階 (A, B, C) で評価してもらい, 次にどのような観点で重要度を判定したのかを示してもらった. 文の重要度の評価の基準として以下のものを与えた.

- A :評価が A のものだけ読めば, 記事のおおよその内容が把握できる.
- B :その文そのものは著者の主張ではないが, あればなお詳細に記事の内容が伝わる.
- C :その文がなくても、全体的文意が十分に伝わる.

[1] 文抽出による自動要約生成研究の近年の動向については [6] を参照されたい.

また評価の際, 以下の事項をインストラクションとして与えた.

- **指示代名詞の取り扱い**
  重要文の中の先行詞がほかの文全体にわたっている場合 (文脈照応の場合), 先行詞の文の評価を重要文のものと 同一の評価にして下さい.
- **接続詞について**
  重要文が接続詞で始まっていて, 前の文と意味的に強いつながりがあると思われる場合は、双方の文の評価を同じにして下さい.

2.2.2 被験者に観点を与えた抄録作成 (実験 2)

実験 2 では, 予め被験者に観点を与え，その観点に基づいてテキスト中の個々の文の重要度を評価してもらった. 観点を与える」以外は, 前節で述べた実験 1 のインストラクションを与えた. 観点については次節で述べる.

2.3 情報検索ベンチマーク :BMIR-J1

実験用テキストと実験 2 で与える観点として BMIR-J1 を利用した. ここで簡単に BMIR-J1 について簡単に説明する. BMIR-J1 は日本語用の情報検索ベンチマークデータで情報処理学会データベースシステム研究会の 情報検索システム評価用データベース構築ワーキンググループ」で作成された. ベンチマークは以下の 3 つの基本要素から構成されている.

1. 対象文書 :新聞記事 600 件
2. 検索要求文
3. 正解集合

BMIR-J1 の正解集合は, 検索クエリとそれに対する対象文書の正解レベル (A, B) を割り振ってある. 正解集合には, 入力クエリについて検索される文書として正解ではないが参考程度という意味で C レベルも設定されている. 正解レベルについて以下に示す.

- A : 正解, 主題も一致
  「〜を主題とする記事」に該当
- B : 正解, 主題は別
  「〜について書かれた記事」に該当
- C : 不正解, 参考として

本実験では, 正解レベル A のものを 18 テキスト, 正解レベル B を 7 テキスト選択し[2], 実験 2 で与える観点は BMIR-J1 で設定されている入力クエリを用いた.

[2] 正解レベル B のテキストは, BMIR-J1 のファイル番号, 09010128, 09290221, 10160113, 10270121, 11010211, 11270131, 12020005, 12080116 である.

2.4 実験条件

実験に使用したテキストと観点を表 1に示す.

表 1: 実験で用いたテキストと実験 2 で与えた観点

テキスト名 (文の数)	予め与えた観点
09010128(17)	農薬
09010220(21)	流通革命
09060260(20)	飲料品
09080130(18)	菓子メーカー
09110011(24)	減税
09110109(39)	携帯電話
09200105(16)	女性の雇用問題
09280214(29)	傘下企業の統廃合
09290221(22)	ビデオデッキ
10040238(26)	製販一体化
10070114(45)	菓子メーカー
10130127(30)	任天堂
10130184(18)	菓子メーカー
10160113(41)	飲料品
10270121(26)	国内航空大手 3 社
10300039(17)	国内航空大手 3 社
10300042(25)	国内航空大手 3 社
11010014(42)	流通革命
11010211(36)	農薬
11220214(62)	任天堂成長神話
11270131(18)	経営多角化の事例
12010042(16)	減税
12010222(17)	菓子メーカー
12020005(39)	任天堂
12080116(36)	携帯電話

20〜30 才の男女 12 人 (大学院生) を被験者とし, 6 人づつ 2 グループ (グループ A, グループ B) に分けて実験を行った.

実験用 25 テキストを 2 つのセット[3]に分けた. まず, グループ A の被験者にセット 1, グループ B にセット 2 を与え, 実験 1 を行った. 1 週間時間をあけて, 今度はグループ A にセット 2, グループ B にセット 1 を与えて実験 2 を行った.

なお, 実験前に被験者を集めて簡単な説明を行ったが, 基本的には Web を用いて実験をした[4]. 個々の実験は 2 週間の期間を設け, その期間内に被験者に実験を行ってもらった. 実験期間中は被験者間での実験に関する話をすることを禁じた.

[3] 13 テキスト (セット 1) と 12 テキスト (セット 2)
[4] http://www.jaist.ac.jp/~nanba/sum/BMIR/

3 実験結果

3.1 被験者間の評価の一致度

被験者間の重要度の評価がどの程度一致しているかを調べるために percent agreement[1] を用いた．Gale らは，percent agreement を以下のように定義している．

$$percent\,agreement = \frac{\text{観測された同意者の数}}{\text{可能な同意者の数}} \quad (1)$$

本実験では 6 人中 4 人以上が"A"と評価した文を重要文として選択する．選択される文において，"A"と評価した被験者の数/6 がその文の percent agreement となる．同様にある文が重要文として選択されない場合，"A"以外の評価を行った被験者の数/6 がその文の percent agreement となる．個々の文において percent agreement を算出しそれらの平均をとったものが，そのテキストにおける被験者の percent agreement として計算される．2 つの被験者グループ A，B の実験 1, 2 における percent agreement を表 2に示す．

表 2: 被験者間の評価の一致度

	実験 1	実験 2
set 1	0.847	0.757
set 2	0.748	0.827
total	0.800	0.790

グループ A … (set 1, 実験 1)(set 2, 実験 2)
グループ B … (set 1, 実験 2)(set 2, 実験 1)

表 2より，グループ A の方がグループ B と比較して percent agreement が高い，つまりデータの一致度が高いと言える．また，実験 1(観点を与えない) で用いたテキスト全体の percent agreement は 80.0%, 実験 2(観点を与えた場合) で用いたテキスト全体の percent agreement は 79.0%となった．

3.2 観点を [与える/与えない] で選択された重要文の一致度

次に，観点を与えて重要文を選択した場合と，観点を与えずに重要文を選択した場合で，選択された文がどの程度一致しているのかを調べた．結果を表 3に示す．

表 3: 観点を与えた場合と与えない場合での選択された要約文の一致度

テキスト名 (文の数)	選択された重要文の数 実験 1	実験 2	一致度
09010128(17)	3	2	1
09010220(21)	3	5	3
09060260(20)	5	5	1
09080130(18)	4	2	1
09110011(24)	5	5	4
09110109(39)	6	5	5
09200105(16)	2	3	1
09280214(29)	6	3	3
09290221(22)	3	2	1
10040238(26)	4	3	2
10070114(45)	11	5	5
10130127(30)	5	5	2
10130184(18)	6	5	4
10160113(41)	9	7	3
10270121(26)	3	4	3
10300039(17)	5	6	4
10300042(25)	2	6	2
11010014(42)	11	8	7
11010211(36)	7	2	1
11220214(62)	5	6	3
11270131(18)	4	4	4
12010042(16)	7	4	4
12010222(17)	4	4	2
12020005(39)	5	5	3
12080116(36)	9	5	5
計	134	111	74

実験 1 における一致度

$$\frac{74}{134} \simeq 0.552$$

実験 2 における一致度

$$\frac{74}{111} \simeq 0.667$$

表 3より，実験 1 において重要文として選択された全 91 文のうち，観点を与えた時にも同様に選択された文の割合は 55% (74/134)．実験 2 において重要文として選択された全 78 文のうち，観点を与えない時にも同様に選択された文の割合は 67% (74/111) となった．これらの結果から，観点を与えるのと与えないのとでは，選択される重要文に明らかに差が出てきていると言える．また, 表 3から非常に大まかな分析をすれば, 実験 1 で被験者に選ばれた文の数が実験 2 で選ばれた数より多いことは, 以下の可能性が考えられる.

表 4: 実験 1 の被験者の観点と実験 2 で与えた観点

テキスト名 (文の数)	実験 1 での被験者の観点	予め与えた観点 (実験 2)
09010128(17)	農業事業で生き残りを図る住友化学工業	農薬
09010220(21)	ソフトバンクの新流通システム	流通革命
09060260(20)	低カロリービールの動向	飲料品
09080130(18)	コメ不足に苦しむ食品メーカー	菓子メーカー
09110011(24)	減税による政府のリストラ支援	減税
09110109(39)	DDI の経営	携帯電話
09200105(16)	山善の LEP プロジェクト& 女性社員の社会進出	女性の雇用問題
09280214(29)	大手商社の不況対策	傘下企業の統廃合
09290221(22)	松下の経営悪化	ビデオデッキ
10040238(26)	東京電気とテック電子の合併	製販一体化
10070114(45)	コメ不足による和菓子メーカーの経営悪化	菓子メーカー
10130127(30)	ナムコによるニッカツの再建	任天堂
10130184(18)	越後製菓のミニショップ方式による売上向上	菓子メーカー
10160113(41)	米国流デフレ型経営	飲料品
10270121(26)	日航のリストラ	国内航空大手 3 社
10300039(17)	国内航空会社の経営悪化 & 航空会社の経営再建策	国内航空大手 3 社
10300042(25)	航空会社の不況対策	国内航空大手 3 社
11010014(42)	価格破壊 & 流通革命	流通革命
11010211(36)	西友の生き残り作戦	農薬
11220214(62)	任天堂の危機	任天堂成長神話
11270131(18)	製紙会社の多角経営による経営維持	経営多角化の事例
12010042(16)	減税による政府のリストラ支援 & 土地譲渡益課税の検討	減税
12010222(17)	和菓子メーカーの小豆不足対策	菓子メーカー
12020005(39)	企業の将来のビジョンと経営目標	任天堂
12080116(36)	カシオ, ポケベルで進出	携帯電話

あるテキストでは実験 1 の方が被験者間の観点のばらつきが多く, 結果として選択される重要文の数も増える

表 4に, 実験 1 で被験者が抄録作成の際に想定した観点と, 実験 2 で被験者に与えた観点を示す. 実験 1 で被験者に示してもらった観点を著者の主観的な判断でまとめた. ひとつにまとめきらないものは"&" を挟んで並べた.

観点が複数存在すると判断されたものは, 25 テキスト中 4 つあった. 実験 1 で得られた観点と 2 で与えた観点の比較は次節で行う.

4 考察

4.1 BMIR-J1 の正解レベルが B のテキストの重要文選択の傾向

BMIR-J1 の正解レベルが B のものは, 「検索クエリについて書かれた記事」であるが, 主題はクエリとは別の可能性がある. 本実験において使用する 25 テキストのうち 7 テキストは正解レベルが B のものである. この 7 テキストについて実験 1 と 2 の重要文選択の傾向を調べた.

実験用全 25 テキストで, 実験 2 で選択された総文数は 111 で, 実験 1 で選択された総文数 134 の約 83%(111/134) と減少している. この傾向は正解レベル B のテキストに限定すればさらに強くなる. 7 テキストの実験 1 で選択された総文数が 40 であるのに対して実験 2 では総文数 27, 約 67.5%となる. 特に, 選択された文の数がわずか 2 文のテキストが 7 テキスト中 3 テキストもある. この 3 テキストは 2 文のうちの 1 文は,

いずれもテキストの第 1 文目が選択されている. これは, 新聞記事の一般的な傾向として, 第 1 文目に重要文が出現する可能性が高いことと一致する. 2 文のうちのもう一方は, 観点 (検索クエリ) の語が出現する文が選択されていた. 評価 B のいずれのテキストにおいても, percent agreement 自体は他のテキストの物とそれほど差はない. 従って, 被験者に適切でないクエリを与えて重要文を選択させようとすると, ほとんどの文が選択されないことになる.

これらの事項から, BMIR-J1 の正解レベル B のクエリでは必ずしも適切な要約が生成できるとは限らない可能性がある.

4.2 実験 1 と実験 2 の一致度に注目した重要文選択の傾向

次に全 25 テキストにおける観点を [与える/与えない] での選択される文の違いについて傾向を述べる. 分析の結果に基づき, その違いを大きく 5 つに分類した.

4.2.1 実験 1 で選択された文が実験 2 で選択された文を含んでいる (5)

- 実験 1 での被験者の観点がまちまちで (観点の拡散), 実験 2 と比較して抽出される文の方が多い (4)

  例えば 「流通革命による商品の価格破壊」に関する記事. 実験 1 では 6 人の被験者が要約の際に想定した観点が 「流通革命」と 「価格破壊」に分かれている. 実験 2 では 「流通革命」をあらかじめ与える.

- 実験 1 では具体的な例まで選択されている (1)

  「大手企業の傘下企業の統廃合」の記事. 実験 1 では不況対策と, その対策方法の記述箇所まで選択. 実験 2 で与えた観点は 「傘下企業の統廃合」.

4.2.2 実験 2 で選択された文が実験 1 の物を含んでいる (1)

- 実験 2 では実験 1 と比べて, 具体的な例の記述まで選択されている (1).

  「航空業界の経営危機とその対策」に関する記事. 実験 2 では, 「国内航空大手三社」を観点として与えている. 大手三社の具体的な記述の箇所も選択されている

4.2.3 実験 1,2 で選択された文がほとんど一致しない (7)

- 実験 1 と実験 2 の主題が異なる (5)

  「機械商社の山善が女性社員の職域を拡大するためのプロジェクトを推進」するという記事. 実験 1 では山善のプロジェクトの具体的な記述が選択. 実験 2 では 「女性の雇用問題」という観点を与えた. より一般的な文が選択されている.

- 選択の傾向がよくわからない (2).

4.2.4 実験 1,2 で選択された文が半分くらい一致する (5)

- ポイントとなる文は実験 1 と 2 で共通だが, 挙げている具体例が異なる. (2)

  「企業の経営計画」の記事. 実験 1 では花王やイトーヨーカドーといった企業を例に出している. 実験 2 では任天堂. 実験 2 で与えた主題が 「任天堂」.

- ポイントとなる文は実験 1 と 2 で共通だが, その他の文の選択傾向が異なる (1).

  「米国のデフレ型経営」の記事. 実験 1 ではデフレ型経営の説明文が選択されている. 実験 2 では主題として 「飲料水」を与えている. デフレ型経営でビールの例の記述箇所が選択.

- 選択の傾向がよくわからない (2).

4.2.5 選択された文がほぼ一致する (7)

- 実験 1 で被験者が想定していた主題と実験 2 で与えた主題がほぼ一致している (7).

4.3 実験 1 での被験者の観点と実験 2 で与えた観点との関係に基づいた重要文選択の傾向

例えば, テキスト 11010014 において実験 1 では被験者の観点は 「価格破壊」と 「流通革命」の 2 つに分かれていた. 一方実験 2 において, 観点として被験者に 「流通革命」を与えた. その結果, 実験 2 で選択された重要文は実験 1 での重要文にすべて含まれていた. このように, 実験 1 と実験 2 の観点の関係が選択される重

要文と関連があるのではないかと考え調査した. 今述べた例の他にも例えば, 観点が実験 1 と 2 でほぼ一致する場合は, 選択される文もほぼ一致するものと考えられる. そこで, 実験 2 で選択された文が実験 1 で選択された文にすべて含まれる場合について, 実験 1 と実験 2 の観点の関係を調べた. テキスト 12010042 では実験 1 の被験者の観点として 「減税による政府のリストラ支援」がある. 一方で実験 2 で与えた観点は 「減税」であった. これらは観点としてはほぼ等しいと考えられるが, 包含関係であるとは言いがたい.

実験 1, 2 で選択された文がほぼ一致しているものについて, 観点の関係を調べた. しかし, それらの関係は様々でまた事例数が十分でないことから傾向らしきものがつかめなかった.

5 結論

本研究では, 人手で抄録を作成する場合, テキストの読み手毎に主題が異なると考え, それに伴い作成される抄録文も主題毎に異なるという仮定を立てた. この仮定を実証するために, 12 人の被験者により心理実験を行った. 心理実験では BMIR-J1 という情報検索ベンチマーク内の 25 テキストを用いた. 実験では, 被験者に, まず観点を与えずに抄録を作成してもらった. 次に BMIR-J1 の検索クエリを観点とみなし, それを被験者にあらかじめ示すことで, ある特定の観点に基づいた抄録を作成してもらった. その結果, 被験者に観点を [与える/与えない] でテキストの抄録作成をしてもらった結果, 選択される重要文に違いが生じることがわかった.

自動要約生成研究において, 多くの場合システムが生成すべき要約の正解はひとつであるということを前提にしている. 今回の実験結果より, 要約生成にはテキストの観点を考慮する必要があり, またその観点に応じて正解の要約が必ずしもひとつであるとは限らないことを示した.

6 今後の課題

本研究では, 12 人の被験者を 2 つのグループに分けて実験した. しかし, 被験者間の一致度 (precent agreement) として十分高いとは言えない. また, 被験者に抄録を生成してもらう際, 特に選択する文の数を制限しなかった. そのため, テキスト毎に要約率が異なっている. 観点を [与える/与えない] による重要文の一致度を調べるために選択する文の数を一定にし, その上で percent agreement を計算するという手法も考えられる. また, 3 段階の評価のうち, A の物のみを利用して 6 人の被験者のうち 4 人以上が A 評価の物を重要文として選択したが, B 評価のものも含めた選択, あるいは A が 3 つ以上の文を選択する等, 色々なケースで選択される文の傾向を調べる必要があると考えられる.

謝辞

本研究では, 株式会社 日本経済新聞の協力によって, 社団法人 情報処理学会 ·データベースシステム研究会 · 情報検索システム評価用データベース構築ワーキンググループが, 1993 年 9 月 1 日から 12 月 31 日の日本経済新聞記事を基に構築した情報検索評価用データベース (テスト版) を使用した. この使用を許可して下さった同グループに感謝致します.

心理実験を行うにあたり, 貴重な御意見, 御助言を下さいました北陸先端大, 奥村研究室博士後期課程の望月源氏, 近藤恵子氏に感謝致します. また, 心理実験に御協力していただいた, 北陸先端大, 島津 ·奥村研究室の皆様方に感謝致します.

参考文献

[1] William Gale, Kenneth Ward Church, David Yarowsky. “Estimating Upper and Lower Bounds on the Performance of Word-Sense Disambiguation Programs”. ACL92. pp.249–256.

[2] Chris D. Paice “Constructing Literature Abstracts by Computer: Techniques And Prospe cts”. Information Processing & Management. Vol.26 No.1, pp. 171–186. 1990.

[3] Julian Kupiec , Jan Pedersen , Francine Chen. “A Trainable Document Summarizer”. SIGIR’95. pp. 68–73. 1995.

[4] H.P. Edmundson. “New methods in automatic abstracting”. Journal of ACM. Vol.16 No.2,pp. 264–285. 1969.

[5] 福島俊一, 他. 日本語情報検索システム評価用テストコレクション BMIR-J1. 自然言語処理シンポジウム 「次規模資源と自然言語処理」論文集, 1998.

[6] 奥村 学, 難波 英嗣. “テキスト自動要約技術の現状と課題”. 言語処理学会第 4 回年次大会併設ワークショップ 「テキスト要約の現状と将来」論文集. 1998.