

論文間の参照情報を考慮した 学術論文要約システムの開発

難波 英嗣† 奥村 学† 神門 典子‡

†北陸先端科学技術大学院大学
‡学術情報センター

発表内容

1. サーベイ論文作成における参照情報の利用
2. 参照情報を用いた関連論文の組織化
3. 書き換えによる抄録の読みやすさの向上

1. サーベイ論文作成における参照情報の利用

1.1 背景・目的

- サーベイ論文を用いた特定分野の研究動向調査は非常に効率的
- しかし特定の分野のサーベイ論文は必ずしも存在するとは限らない



特定分野の複数の論文から、サーベイ論文を自動的に作成するシステムの開発

1.2 サーベイ論文作成の手順

1. 特定分野の論文の収集
2. 論文間の関係の解析
 - 論文間の類似点、相違点の抽出
 - 収集した論文の組織化
3. 一貫性を考慮した文書の生成



論文間の参照情報の利用

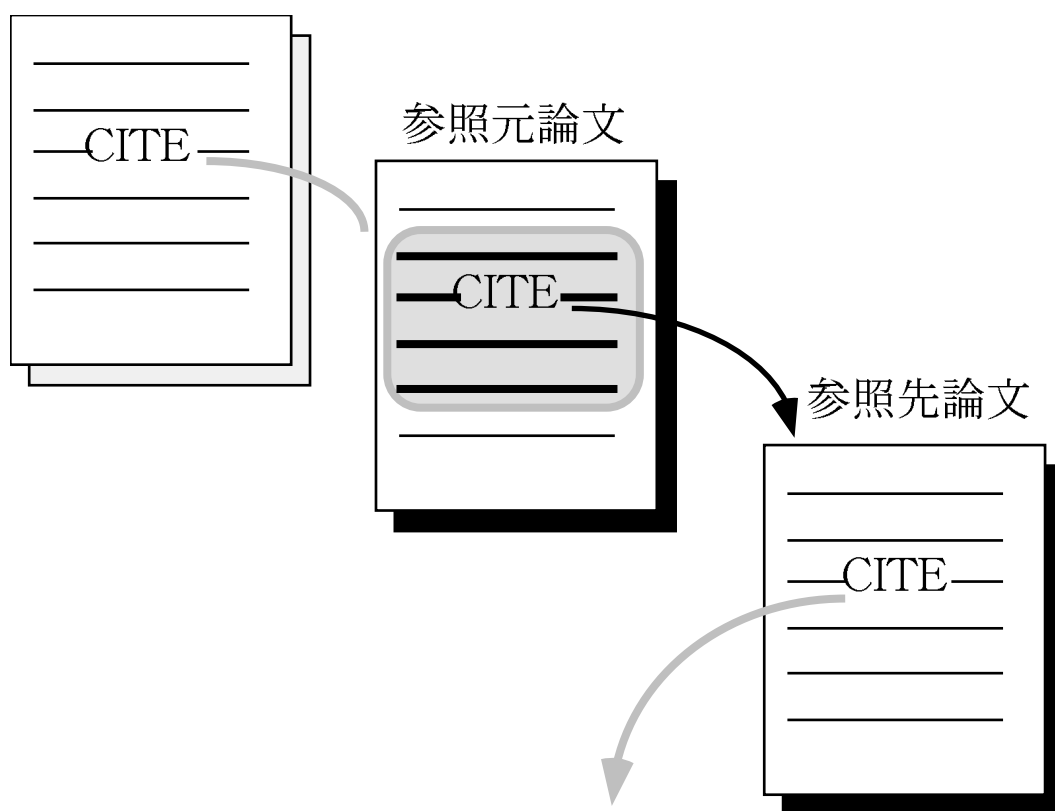
1.3 サーベイ論文作成における参照情報の利用

(Bond96) から一部抜粋

1. In addition, when Japanese is translated into English, the selection of appropriate determiners is problematic.
2. Various solutions to the problems of generating articles and possessive pronouns and determining countability and number have been proposed (**Murata:1993**).
3. The differences between the way numerical expressions are realized in Japanese and English has been **less studied**.
4. **In this paper, we propose** an analysis of classifiers based on properties of both Japanese and English.
5. Our category of classifier includes both Japanese *josūshi* ‘numeral classifiers’ and English partitive nouns.

[難波・奥村 '99]

(自然言語処理 v.6 n.5)(IJCAI '99)



参照情報

- 参照先論文の重要点
- 参照元/参照先論文間の差異を明示する有用な情報

参照タイプ

| | |
|-------------------------------|---|
| <u>問題点指摘型 (<i>type C</i>)</u> | 他の論文の理論や手法等の問題点を指摘する |
| 論説根拠型 (<i>type B</i>) | ある理論を提案する場合や仮定をする場合, その根拠となる論文 |
| その他型 (<i>type O</i>) | <i>type B</i> にも <i>type C</i> にも分類できない論文 |

参照箇所の抽出

1. In addition, when Japanese is translated into English, the selection of appropriate determiners is problematic.
2. Various solutions to the problems of generating articles and possessive pronouns and determining countability and number have been proposed (Murata:1993).
3. The differences between the way numerical expressions are realized in Japanese and English has been less studied.
4. In this paper, we propose an analysis of classifiers based on properties of both Japanese and English.
5. Our category of classifier includes both Japanese *josūshi* ‘numeral classifiers’ and English partitive nouns.

(Bond96) 中の (Murata:1993) に関する
参照箇所 (文 2~4)

cue word を用いた参照箇所 の抽出

- ・ 照応詞 “This”, “In this”, “For this”, “These”, “In these”, “On these”
- ・ 接続詞 “But”, “However”, “Although”
“In spite of”
- ・ 一人称代名詞 “we”, “our”, “us”, “We”
- ・ 三人称代名詞 “they”, “their”, “them”
- ・ その他 “Furthermore”, “Additionally”
“Nevertheless”, “Again”

計 86 個 の cue word

citation の前後 2 文以内に cue word が出現したら, citation の出現した文から cue word の出現する文までを参照箇所として抽出.

→ 抽出精度 約 80%

参照タイプの決定

1. Various solutions to the problems of generating articles and possessive pronouns and determining countability and number have been proposed (Murata:1993).
2. The differences between the way numerical expressions are realized in Japanese and English has been **less studied**.
3. In this paper, we propose an analysis of classifiers based on properties of both Japanese and English.

・ cue word(計 160 個) を選定し、それに基づいたルールを使ったタイプ決定

C type 決定ルール

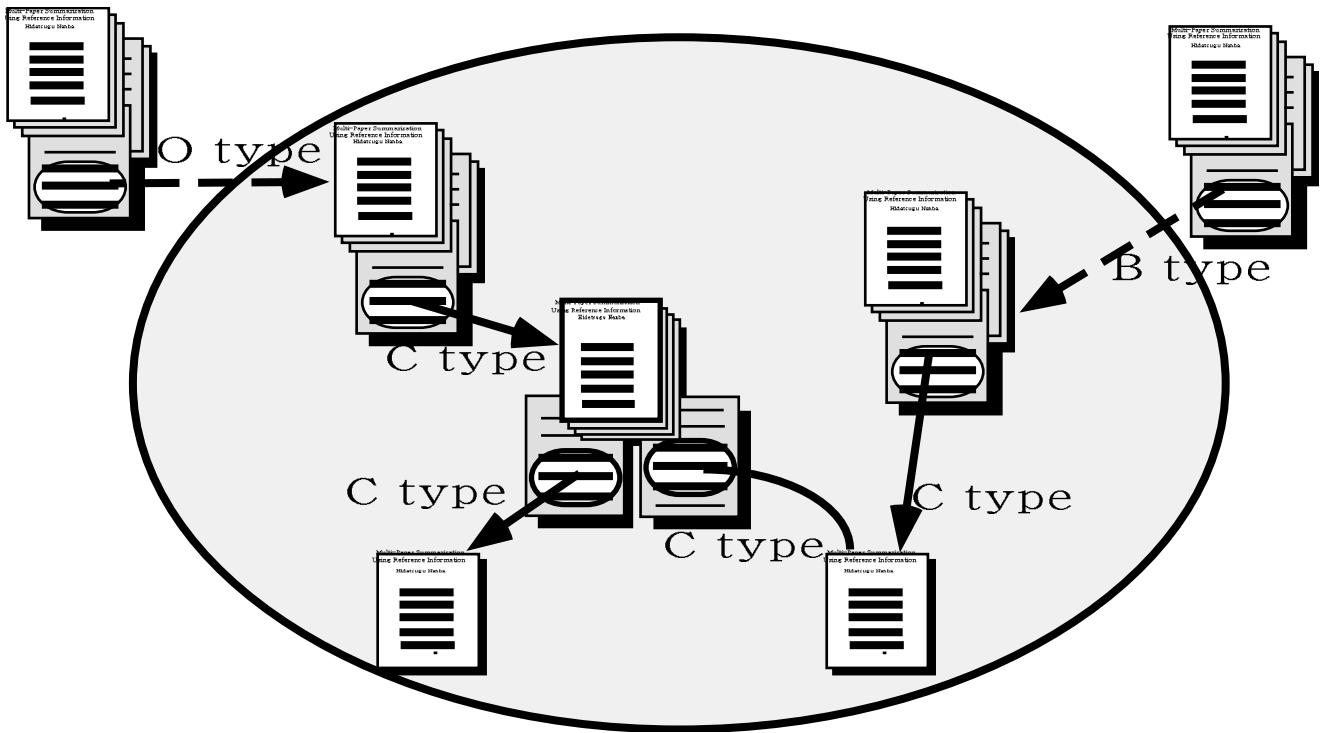
- cite の出現した文より後 3 文以内に
”However,” → C
- cite の出現した文から後 2 文以内に
”less studied” → C …

B type 決定ルール

- cite の出現した文から前 2 文以内に
”based mainly on” → B
- cite の出現した文から前 3 文以内に
”apply to” → B …

return(O)

参照タイプ決定精度 → 83%(評価用データ)



PRESRI

<http://galaga.jaist.ac.jp:8000/pub/tools/sum>

File Edit View Go Bookmarks Options Directory Window

Back Forward Home Reload Load Images Open... P

Location: <http://galaga.jaist.ac.jp:8000/pub/tools/sum/cgi-bin>

[PRESRI Web page]

show type C reference relationships

[Takeda94]
(9407008)

[Bond94]
(9511001)

[Bond96]
(9601008)

[Bond96]
(9608014)

ABST

REFERENCE AREA

[Murata93]
(9405019)

ABST

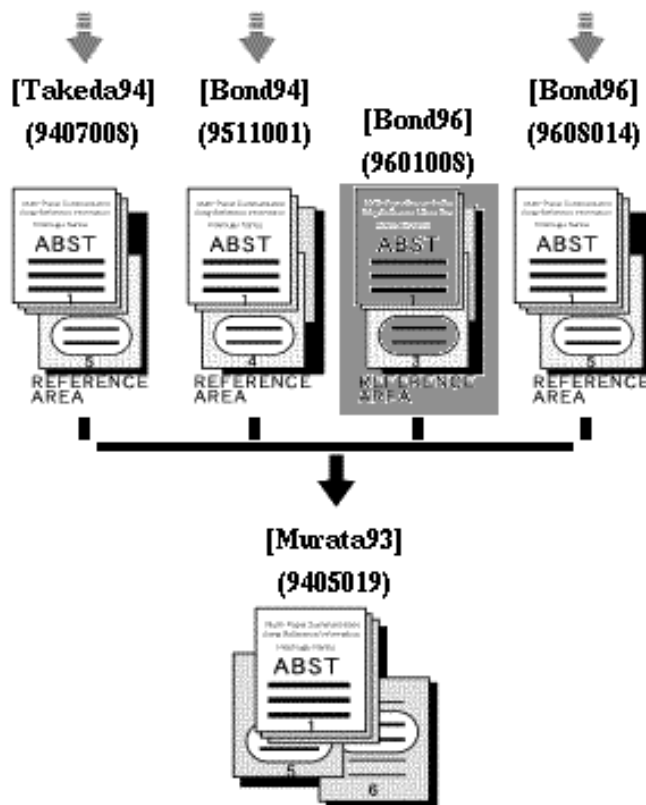
REFERENCE AREA

[PRESRI Web page]

show all reference relationships

Now displaying **type C** reference relationships.

Now, this system displays **only the papers in the same domain.**



File Edit View Go Bookmarks Options Directory Window Help

Back Forward Home Reload Load Images Open... Print... Find... Stop

Location: <http://galaga.jaist.ac.jp:8000/pub/tools/sum//cgi-bin/RA.cgi?9608014+94050>

9608014 --> 9405019 Murata:1993a

Various solutions to the problems of generating articles and possessive pronouns and determining countability and number have been proposed \cite{Murata:1993a ,Cornish:1994,Bond:1995b}.

The differences between the way numerical expressions are realized in Japanese and English has been less studied \cite{Asahioka:1990}.

9511001 --> 9405019 Murata:1993a

Recently, \namecite{Murata:1993a } have proposed a method of determining the referentiality property and number of nouns in Japanese sentences for machine translation into English, but the research has not yet been extended to include the actual English generation.

9407008 --> 9405019 murata93

This type of information is rarely available from a syntactic representation of a Japanese noun phrase, and a set of heuristic rules\cite{murata93 } is the only known basis for making a reasonable guess.

Even if such contextual processing could be integrated into a logical inference system, the obtained information should be defeasible, and hence should be represented by green nodes and arcs in the TDAGs.

これまでの研究の問題点

- 直接参照関係にはないが、関連のある論文間の関係が扱えない。



- 明示的に示されていない論文間の関係を明らかにする。

→ 関連論文の組織化 (分類・整理)

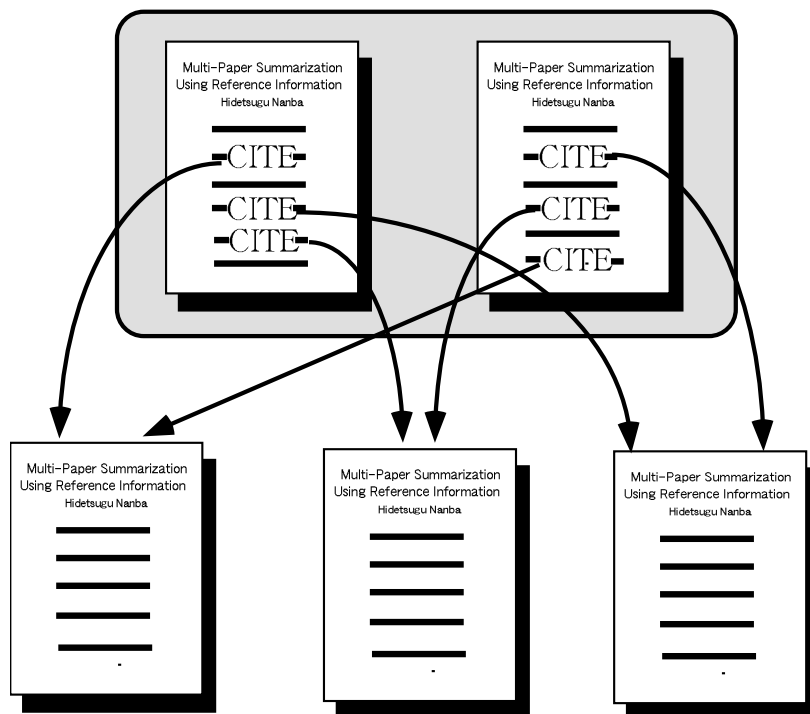
2. 参照情報を用いた関連論文の組織化

サーベイ論文作成実現の第1歩として…
既存のサーベイ論文のバージョンアップ

1. 論文データベース中からサーベイ論文を検索する。
2. サーベイ中で引用されていない新しい論文を付け加える。
→ 関連論文の組織化の技術
3. 付け加えた文が前後の文とつながりが良くなるよう書き換える。
→ [難波・奥村 '99](情処技報 NL)

2.1 組織化手法の提案

書誌結合 (Kessler 1963) の利用

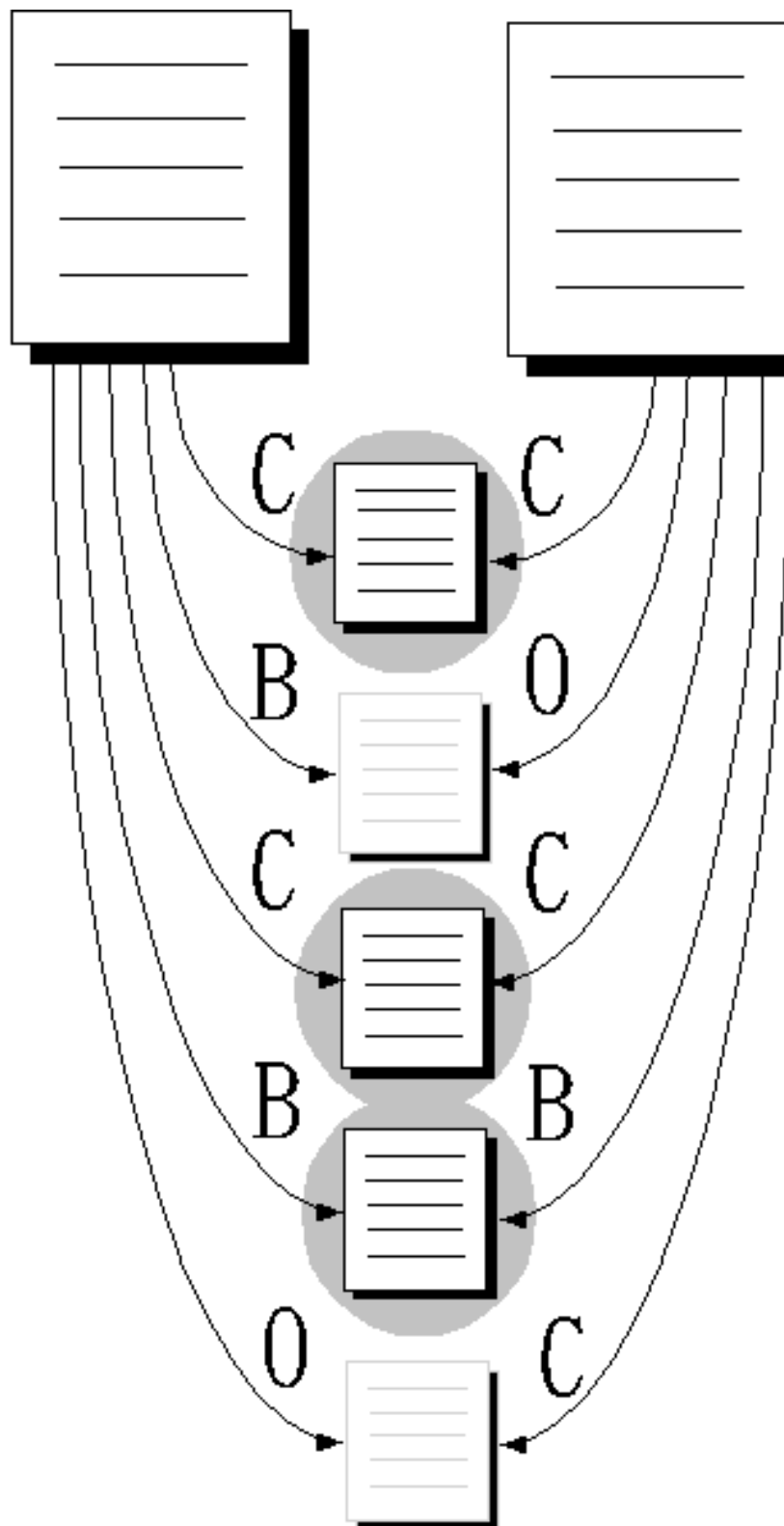


しかし、2つの論文が同じ目的で参照しているわけではない。



参照タイプを考慮した書誌結合

参照タイプを考慮した書誌結合



関連論文の組織化に書誌結合を用いる理由

語の共出現による関連論文の組織化

2 論文間で共出現する語の数が多いほど
類似



2 論文に出現する語が似ていても、
何処が似ているかを示すのが困難

参照タイプを考慮した書誌結合による組織化



- 共に他の論文を **type C** で引用
2 論文が共通の問題意識を持っている。
- 共に他の論文を **type B** で引用
2 論文が同じ理論やツールを用いている。

2.2 ハンドシミュレーションによる 提案手法の検討

研究対象

CiteSeer/ResearchIndex

(<http://www.researchindex.com>):

- WWW 上の PostScript, PDF ファイルを集めて作成されたデータベース
 - フルテキスト：約 20 万件
 - References から抽出された書誌情報データ：約 340 万件
- 論文間の参照・被参照関係を解析
- 論文中の引用箇所も解析

Discourse Structure に関する論文への適用

1. (Moser and Moore 1995)
2. (Jurafsky et al. 1998)
3. (Passonneau and Litman)
4. (Grosz)

↓ 参照

(Grosz et al. 1992)

目的：1～4の (Grosz et al. 1992) に関する引用を著者の観点毎に分類したい！

書誌結合による分類

| 2 論文 | 類似度 |
|------|-------------|
| 1-2 | 3(2) |
| 1-3 | 4(3) |
| 1-4 | 5(4) |
| 2-3 | <u>1(0)</u> |
| 2-4 | 4(3) |
| 3-4 | 5(4) |

() は (Grosz et al. 1992) を除いた数



論文 2 と 3 のつながりがあまりなさそうである、ということ以外はわからない

参照タイプを考慮した書誌結合による分類

| 2 論文 | 類似度 | type 毎の一致数 | | |
|------|-------------|------------|--------|--------|
| | | type C | type B | type O |
| 1-2 | 3(2) | 2 | 0 | 1 |
| 1-3 | <u>2(1)</u> | 1 | 0 | 1 |
| 1-4 | 5(4) | 1 | 1 | 3 |
| 2-3 | <u>1(0)</u> | 1 | 0 | 0 |
| 2-4 | 4(3) | 1 | 0 | 3 |
| 3-4 | <u>2(1)</u> | 1 | 0 | 1 |



論文 3 と他論文との類似度は小さい。

→ 論文 3 は論文 1,2,4 とは異なる観点？

実際、論文 3 はテキスト・セグメンテーションに関するもので、他の論文とは分けた方がよい。

考察

- 参照タイプ毎の一致の重みについて
 - type O の一致は類似度計算に含めてよいか？
 - type B と type C の一致は等価に扱ってよいか？
- 各論文の引用数の違いについて
- 1 つの論文が論文中で複数回引用されている場合

3. 書き換えによる抄録の読みやすさの向上

既存のサーベイ論文のバージョンアップ

1. 論文データベース中からサーベイ論文を検索する。
2. サーベイ中で引用されていない新しい論文を付け加える。
→ 関連論文の組織化の技術
3. 付け加えた文が前後の文とつながりが良くなるよう書き換える。
→ [難波・奥村 '99](情処技報 NL)

複数論文から重要箇所を抽出して単純に並べただけでは、文書として読みづらい文書を読みにくくする要因

- 複数テキスト要約

- 論文の著者毎の文体が異なる。
- 言い回しの違い。etc.

- 単一テキスト要約 [Paice 1990]

- 抽出した文中に代名詞などが含まれている場合，その先行詞が抄録中に存在する保証がない。

本研究では、最初に単一テキスト要約の読みにくさの解消を試みる。

関連研究 (1)

- ・ テキストの読み易さ [Klare 1963]

入力:人が正しく書いたテキスト
テキストの読み易さ (難易度) を,
語彙レベル (語彙の難易),
構文レベル (文構造の複雑さ)
という側面から捉え, 数値化.

- ・ 抄録の読み易さ

入力:計算機により自動的に作られたテキスト→他の観点も考慮

↓

文脈レベル (文間のつながりの良さ)
(Halliday & Hasan 1976, 森岡 1963)

関連研究 (2)

抄録の文間のつながり

- [Paice 1990]
 - 照応詞を含む文の前の数文を要約に追加
 - 接続詞の削除
 - 動詞の時制や態の修正
- [山本他 1995]
 - 重要文として抽出される文の先頭に指示語や接続詞が出現する場合、あるいは、抽出される文の主語が省略されている場合、前文も要約に追加



非重要文を抄録に含めてしまう点で問題
読みにくさの要因を網羅的に調べていない



- 書き換えによる抄録の読み易さの向上
- 心理実験により読みにくさの要因を網羅的に調査

関連研究 (3)

抄録の書き換え 1

テキストと人手で作られた要約との比較

- ・ [川原 1989]

「言語サイズ (※)」という尺度を用いて分析。

※ 語，句，節などの統語的な単位

- ・ [Jing 1999]

「人間は、基本的には原文中の語句を cut and paste して要約を作成する」という仮定を、HMM を用いて計算機的に実証。

→ 80%程度は対応づけ可能。

関連研究 (3)

抄録の書き換え 2

書き換えの目的

1. 文の長さを縮める
2. 文のスタイルを変える
3. 要約文書を読みやすくする

- [Jing 1999]

Reduction(1) と Combination(3) という 2 種類の書き換え手法の提案。

- [Mani 1999]

3 種類のオペレータによる抄録の書き換え。元の抄録の読み易さを低下させずに要約の informativeness を向上。1 と 3 を対象。



本研究では 3 を対象。抄録の読み易さの向上

研究の手順

1. 網羅的に抄録の読みにくさの要因を調査

心理実験で数人の被験者に抄録を読みやすく書き換えてもらい、抄録の読みにくさの要因を分析 (Halliday & Hasan や森岡の指標を考慮)

2. 要因毎に読みにくさを解消する方法を検討

3. 分析結果に基づき抄録を読みやすく書き換えるシステムを計算機上に実装

4. システムの評価

抄録の読みにくさの要因の調査

- 実験題目 ：
『新聞記事中の文の、人手による
重要性の評価と要約作成』
- 調査期間 ：1997年10月～12月
- 被験者 ：大学院生 12名
- 対象テキスト ：日本経済新聞 25記事
- 実験方法 ：
 1. 記事中の各文の重要度を3段階
(A,B,C) 評価.
 2. 評価 A の文のみを用いて要約を
作成 (修正箇所を記してもらう).

↓

343箇所の書き換え。このうち読みにく
さの修正に着目。

- 読みにくさの要因を5種類に分類
- 各要因毎の書き換え手法の検討

ア. 不要な表現/接続詞の不足による文間のつながりの悪さ

2. 業績が落ち込んでいる中で生産財から消費財へのシフトを進める会社にとって、全体の四割を占める四百人の女子社員の活性化が不可欠。
3. 専任社員を置くことで女性の職域を積極的に拡大していく方針だ。
15. 会社の期待は大きいですが、このような女性参加のプロジェクトの前例がなく、現在は悪戦苦闘中。
16. ~~しかし~~ 目標を二〇〇一年に置いて女性社員にとって満足できる仕事環境、会社にとって有効な人材の活用と両者に利益のある改革を進めている。

→接続詞的表現の [削除/付加]
(談話構造解析)

イ. 構文的な複雑さ

2. (修正前)

樫尾和雄社長自ら「将来の経営の核」に挙げる通信事業の第一弾であり、今後、携帯電話などに手を広げていくための布石でもある。



2. (修正後)

樫尾和雄社長自ら「将来の経営の核」に挙げる通信事業の第一弾である。

今後、携帯電話などに手を広げていくための布石でもある。

→2つの文をつなげる / 単文の分割

ウ. 冗長・不自然な繰り返し

2. 越後製菓が四月に始めた新商品シリーズ「越後ベスト一〇〇」が主婦に人気だ。
4. <para 同社>越後製菓</para>は米菓分野で八七年からN T Tのキャプテンなどを活用、スーパーなど小売店の売り場から工場に直接注文を受け即日製造し出荷、翌日には店頭に並べるという、いわば「ジャストインタイム」方式で売り上げを伸ばしてきた。

→代名詞化 / 主語の省略 / 指示詞の付加

エ. 情報の欠落 (照応・省略・代用等)

1. 消費者のブランド離れ、価格への敏感さで日本に先行している米国で、「デフレ型経営」によって成功している企業がある。
2. 自動車のクライスラーとパソコンのコンパック・コンピューターだ。
8. 企業の人員削減が消費意欲を沈滞させ、それがまた企業業績の足を引っ張るというデフレ型の悪循環に陥りつつある。
9. ~~その中で、~~クライスラーが大健闘しているのは、新製品投入の時期と価格・商品戦略がまさしく低成長時代に適応しているからだ。

→主語の補完 / 照応補完 / 照応詞の削除
(照応・省略補完)

→ 代用の補完 / 補足情報の付加
(情報抽出)

オ. 不要な副助詞/副助詞の不足による 文間のつながりの悪さ

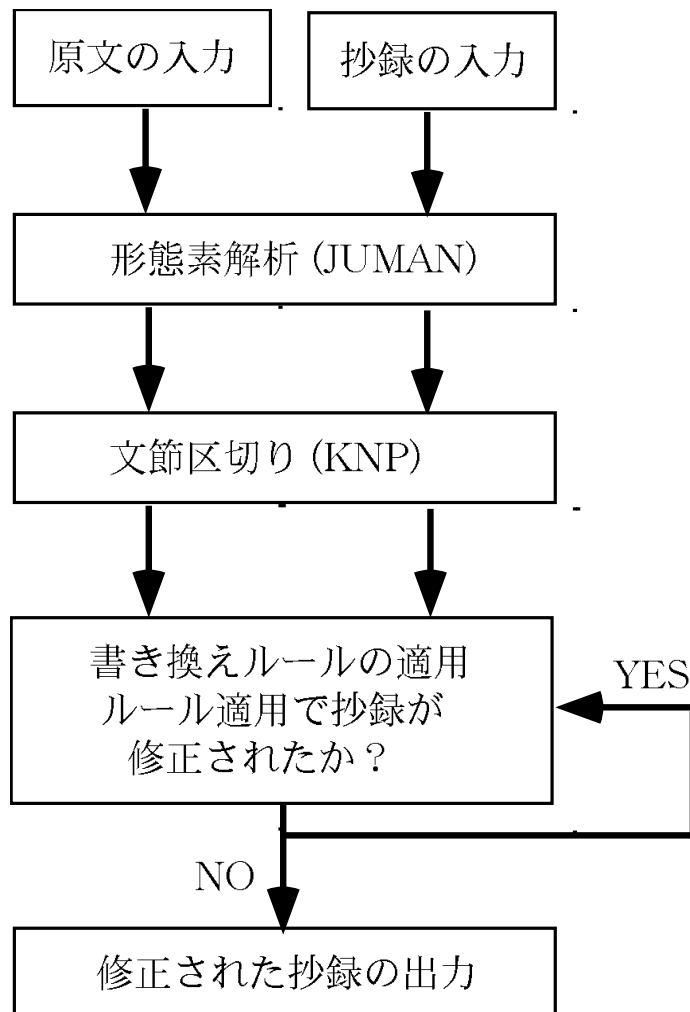
26. 今回のド・ムオイ書記長の訪日は両国が相互理解を深めるうえで、よい機会である。
29. ベトナムはアジアの安全保障の観点からも今後、重要性を増していく。
30. 経済協力の面でも長い目で発展を支援していく姿勢が必要だろう。

→副助詞の [削除/追加]

抄録書き換えシステムの構築

書き換えに基づく要約生成システム

… [Robin 1996][Mani 1999]



ア. 不要な表現による文間のつながりの悪さ

→接続詞的表現の削除

3. 専任社員を /置く /ことで /女性の /職域を /積極的に /拡大していく /方針だ。
15. 会社の /期待は /大きい /が、 /このような /女性参加の /プロジェクトの /前例が /なく、 /現在は /悪戦苦闘中。
16. しかし /目標を /二〇〇一年に /置いて /女性社員に /とって /満足できる /仕事環境、 /会社に /とって /有効な /人材の /活用と /両者に /利益の /ある /改革を /進めている。

接続詞リスト：(しかし、だが、けれども、ところが…)

ICOT の談話構造解析ツールを部分的に利用 [福本他 1990]

ウ. 冗長・不自然な繰り返し

→主語の省略

1. ゲーム / 機最大手の / 任天堂は / 十 / 月初め、
/ 九四年 / 三月 / 期の / 業績見通しを / 大幅に
/ 下方修正した。
4. 任天堂は 九五年 / 九月に / 新
しい / ゲーム機 / を発売する / 計画だが、 /
国内を / 代表する / 高収益企業と / なった
/ 同社の / 成長神話は / これからも / 続くの
か。

エ. 情報の欠落 (照応・省略・代用等)

→照応詞の削除

2. 自動車の /クライスラーと /パソコンの /コンパクト・コンピューターだ。
8. 企業の /人員削減が /消費意欲を /沈滞させ、 /それがまた /企業業績の /足を /引っ張ると /いう /デフレ型の /悪循環に /陥りつつある。
9. その中で、 /クライスラーが /大健闘しているのは、 /新製品投入の /時期と /価格・商品戦略が /まさしく /低成長時代に /適応しているからだ。

評価

評価方法：被験者の書き換えとシステムの書き換えとの比較による評価

抄録：日本経済新聞 1993年…15記事，
1995年…16記事 [野本他 1997] 過半数 (4名以上) の被験者が A と評価した文を抄録として利用

データの作成：1抄録あたり5人の被験者に読みやすく書き換えてもらう

表 処理毎の被験者の書き換え回数

| | 書き換え処理 | 回数 |
|---|--------------------------------|-----|
| ア | 接続詞的表現の削除 (11)/付加 (61) | 72 |
| イ | 2つの文をつなげる (2)/単文の分割 (6) | 8 |
| ウ | 代名詞化 (5)/主語の省略 (3)/指示詞の付加 (8) | 16 |
| エ | 主語の補完 (11)/照応補完 (8)/照応詞削除 (15) | 34 |
| | 代用の補完 (2)/補足情報の付加 (26) | 28 |
| オ | 副助詞の削除 (4)/追加 (5) | 9 |
| 計 | | 167 |

結果

$$Recall = \frac{\left(\begin{array}{l} \text{ルールで書き換えられた箇所のうち} \\ \text{被験者と一致した箇所数} \end{array} \right)}{\text{被験者が書き換えた箇所数}}$$

$$Precision = \frac{\left(\begin{array}{l} \text{ルールで書き換えられた箇所のうち} \\ \text{被験者と一致した箇所数} \end{array} \right)}{\text{ルールで書き換えられた箇所数}}$$

ルールと被験者の書き換え箇所の一一致度

| 書き換え | 評価 1 | | 評価 2 | |
|-----------|--------|-----------|--------|-----------|
| | Recall | Precision | Recall | Precision |
| ア (のべ 11) | 2/2 | 2/5 | 4/4 | 4/5 |
| ウ (のべ 3) | 0/0 | 0/0 | 0/1 | 0/1 |
| エ (のべ 15) | 4/5 | 4/7 | 4/5 | 4/7 |

まとめ

- 抄録の読みにくさについて、心理実験により調査
- 読みにくさ要因を 5 種類に分類
- 読みにくさを解消するルールを部分的に作成
- 被験者の書き換えとの一致度によりルールの有効性を確認

ただ、現在の NLP 技術では可能な書き換えが限られており、作成したルールの Coverage が低い

今後必要とされる技術

- 談話構造解析
- 照応・省略解析
- 固有名詞抽出

付録：ResearchIndexに関する基礎的調査

- 1論文につき、平均 34 の論文を参照
- そのうち 15%は ResearchIndex 上にフルテキストで存在

論文の種類

| 種類 | 件数 | 割合 (%) |
|------------------|-----|--------|
| Proceedings | 121 | 71 |
| Technical Report | 14 | 9 |
| Journal | 11 | 7 |
| Lecture Note | 4 | 2 |
| Ph.D. | 2 | 1 |
| その他 (書籍など) | 16 | 10 |
| 計 | 168 | 100 |

References

- Halliday, M.A.K., Hasan, R.
“Cohesion in English”, Longman, 1976.
- Jing, H.
“Summary Generation through Intelligent Cutting and Pasting of the Input Document”, Ph.D. Thesis Proposal, Columbia University, 1999.
- Kessler, M.M.
“Bibliographic Coupling between Scientific Papers”, American Documentation. Vol.14, No.1, pp.10–25, 1963.
- Klare, G.R.
“The Measurement of Readability”. Iowa State University Press, 1963.
- Mani, I., Gates, B., Bloedorn, E.
“Improving Summaries by Revising Them”.
the 37th Annual Meeting of the Association for Computational Linguistics(ACL-99), pp.558–565, 1999.
- Mathis, B., Rush, J., Young, C.
“Improvement of Automatic Abstracts by the Use of Structural Analysis”. 24
(2), pp.101–109, 1973.
- Minel, J., Nugier, S., Gerald, P.
”How to Appreciate the Quality of Automatic Text Summarization?
Examples of FAN and MLUCE Protocols and their Results on SERAPHIN”,
Intelligent Scalable Text Summarization, Proceeding of a Workshop , ACL,
pp.25–30, 1997.
- Nanba, H., Okumura, M.
“Towards Multi-paper Summarization Using Reference Information” The
International Joint Conferences on Artificial Intelligence(IJCAI-99) ,
pp.926–931, 1999.
- Paice, C.D.
”Constructing Literature Abstracts by Computer: Techniques And Prospects”.
Information Processing & Management. 26 (1), pp. 171–186, 1990.
- Robin, J., McKeown, K.
“Empirically Designing and Evaluating a New Revision-based Model for
Summary Generation”. Artificial Intelligence 85, pp.135–179, 1996.

- 乾裕子.
日本語文書可読性評価に関する研究. 平成7年度富士通株式会社委託研究報告, 1996.
- 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, 6, (6), pp.1-26, 1999.
- 川原祐美.
“9. 要約文のパラフレーズの様相「文章構造と要約文の諸相」”. くろしお出版, pp.141-167, 1989.
- 難波 英嗣, 奥村 学.
“論文間の参照情報を考慮したサーベイ論文作成支援システムの開発” 「自然言語処理」 Vol. 6, No. 5, pp.43-62, 1999.
- 難波 英嗣, 奥村 学.
“書き換えによる抄録の読みやすさの向上”. 情報処理学会 自然言語処理研究会 99-NL-133-9, pp.53-60 1999.
- 野本忠司, 松本裕治.
“人間の重要度判定に基づいた自動要約の試み” 情報処理学会自然言語処理研究会報告, 120-11, pp.71-76, 1997.
- 福本淳一.
“著者の主張に基づく日本語文章の構造化” 情報処理学会自然言語処理研究会報告, 78-15, pp.113-120, 1990.
- 森岡健二.
“文章構成法 -文章の診断と治療-” pp.124-132, 至文堂, 1963.
- 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, 2 (1), pp.39-55, 1995.

参照情報の説明に用いた論文

- Bond, F., Ogura, K., Ikehara, S.
“Classifiers in Japanese-to-English Machine Translation” *COLING'96*
125-130, 1996.
<http://xxx.lanl.gov/ps/cmp-lg/9608014>

- Murata, M., Nagao, M.
“Determination of referential property and number of nouns in Japanese sentences for machine translation into English” *TMI-93*, 1993.
<http://xxx.lanl.gov/ps/cmp-lg/9405019>

参照タイプを用いた書誌結合の説明に用いた論文

- Grosz, Barbara, and Julia Hirshberg.
“Some Intonational Characteristics of Discourse Structure”. *ICSLP92*, 429–432, Banff, Canada. 1992.
- Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K.
Johns Hopkins LVCSR Workshop-97 Switchboard Discourse Language Modeling Project Final Report, 1997.
- Passonneau, R.J., Litman, D.J.
“Intention-Based Segmentation: Human Reliability And Correlation With Linguistic Cues”
- Barbara Grosz. “Chapter 6 Discourse and Dialogue 6.1 Overview”, *Survey of the State of the Art in Human Language Technology*.