関連する複数新聞記事からの重要文抽出法

岡崎 直観* 松尾 豊† 石塚 満*
東京大学大学院情報理工学系研究科

〒 113-8656 東京都文京区本郷 7-3-1
Tel: 03-5841-6755
Fax: 03-5841-8570
e-mail: okazaki@miv.t.u-tokyo.ac.jp
URL: http://www.miv.t.u-tokyo.ac.jp/~okazaki/

概要

WWW の急速な発展により相当数の情報がオンラインで入手可能になり，Web 検索サービスに代表されるような情報検索技術により探したい情報を瞬時に得ることができる．このような状況は大変便利であるが，情報氾濫として問題視されている．このような現状を背景とし，大量のテキストを簡略にまとめてくれる複数テキスト要約が脚光を浴びている．

テキスト自動要約では文章中から重要な箇所を必要な量だけ抜き出してくる重要箇所抽出が基本であるが，複数文書の要約ではテキスト集合の中に同じ内容の文が含まれている可能性があり，重要な箇所を含みつつも内容の重複を避けることも必要である．そこで，テキスト中に含まれる語の共起関係を分析し，要約に含めるべき共起関係をできるだけ取り込むような文の組み合わせを求めることで，複数文書要約システムの構築を行った．このような最適化問題を解くことによって，従来の重要文抽出方法と比べて，原文に含まれる内容を網羅的に捕らえながら内容の重複を最小限に抑えることができる．本発表では，我々の考案した複数新聞記事に対する重要文抽出法を紹介する．

1 はじめに

唐突ではあるが，こんな課題が出たとしよう.「ハイジャックに関して調べてレポートにまとまよ」– この手の課題が出たときにやることといえば,

- Web で「ハイジャック」を検索する
- 「ハイジャック」に関する資料を図書館で探す
- 「ハイジャック」に関する新聞記事を検索する
- 辞書で「ハイジャック」を調べて定義を並べる

が月並みなところであると思われる．いずれにしても，ハイジャックに関する資料を「探して」→ それらの資料を「読んで」→ 頭の中に留められた内容を「書く」という手順を辿りそうである．

*東京大学大学院情報理工学系研究科電子情報学専攻
†産業技術総合研究所

そこで，98 年–99 年の毎日新聞のデータベースで「ハイジャック」を検索してみると，144 件の記事が見つかり[1]，検索エンジンである Google[2]では，実に 50,500 件のウェブページが一瞬にしてヒットする．このように，オンライン文書化や情報検索技術の進歩によって欲しい情報を手軽に「探す」ことができるようになった．

しかし,「読む」「書く」ということに関してはどうであろうか？　一瞬にして検索された 50,500 件のウェブページのすべてを，幾ばくの月日が流れようとも隈なく読み，各ページで記述されている情報を整理し，有用と思われる箇所を参照しながらレポートにまとめている自分の姿を想像したくはない．テキスト自動要約技術はこのような仕事を，レポートにまとめるところまで代わりにやってくれるコピーロボットを作ることも目標のひとつである．

しかしながら，推敲して「書く」という行為をコンピュータ上で実現することは難しい．また，ユーザーの望む形態で文章を纏め上げることも難しい．このようなことからテキスト自動要約は，それぞれのテキストで述べられていることをダイジェスト風にまとめ，ユーザーがその文章を読むべきかどうかの判断材料を提供したり，ユーザーにとって未知の情報との出会いを支援するというアプローチをとる場合がある．

本発表では，テキスト自動要約の基礎を紹介した後，我々が考案した重要文抽出法を紹介する．なお，我々は国立情報学研究所情報学資源研究センターの支援により開催されているワークショップ NTCIR のテキスト自動要約タスク (TSC) に参加している．

2　テキスト自動要約概観

2.1　文書自動要約の概観

要約とは，原文の大意を取りまとめる処理，またはその結果としての文章のことを指し，先に述べたような氾濫した情報の中で，短時間で原文の内容を把握することを支援するものである [1] .

我々が文書の要約を作成する過程を考察してみると，おおよそ，

(1) 文書内容を理解する

(2) 重要だと思われる箇所を選別する

(3) 抜き出された断片を繋ぎ合わせ，文章としての整合性を持たせる

の 3 ステップを踏む．このうち，(2) のステップで行われる**重要箇所抽出**は比較的簡単なこともあって，自然言語処理の分野では 1950 年代から研究されており，自動要約技術の根幹をなすものである．

2.2　重要箇所抽出

重要箇所抽出とは，入力されたテキスト断片[3] をある基準を用いて評価し，重要度上位の箇所を抜き出して要約を出力する方法である．これは人間が長めの文章を読むときに，重要だと思われる箇所に下線やマーカーをつけるのと同じ原理である．重要箇所抽出の多くの場合はテキスト内容の理解は行わず，テキスト中の表層的な並びの裏に潜む現象を捕らえ，重要箇所を推定する．重要度を決定するファクターとしては，キーワードの出現頻度 [2, 3]，文書中あるいは段落中での位置情報 [4]，文書のタイトルやメタデータ [4]，文書中の手がかり表現 [4]，文間の関係を解析した文書構造 [5]，文あるいは単語間のつながり情報 [6]，文間の類似性の情報 [7] など，様々なものが提案されている．これらは単体で用いられる場合もあるが，複数のファクターの最適な組み合わせを見つける研究 [8] もある．

[1] 98 年は全日空機ハイジャックが起こった年であった.
[2] http://www.google.com/
[3] 段落，文，節，フレーズなど用いられる単位は研究によって様々である

2.3 複数テキストの要約

我々は，しばしば関連する複数の文書に出くわすことがある．先の「ハイジャック」のレポート例では,「ハイジャック」に関連する資料を収集し，たくさんのオンラインドキュメントを得ている．このように，あるキーワードに関して検索して得られた文章集合やウェブページ，インターネット掲示板のスレッドやメーリングリストなど，要約対象の文章が複数ある場合の要約を，**複数テキスト要約**と呼ぶ．

単一テキスト要約の場合 2.2 で述べたような重要箇所抽出を用いるだけで十分であることが多い．しかし，要約対象が関連する複数の文書になると，重要箇所として抽出した内容が重複してしまう恐れがある．そこで重要箇所の抽出と同時に，テキスト集合の共通点と差異を認識し，重要箇所中で冗長な部分を削除したり，他のテキストとの相違点を明確にすることが望まれる [10].

3 提案手法

3.1 目標と概要

以上のようなことを踏まえ，複数文書を対象とし，元の文書の内容をできるだけ広くカバーしつつ，冗長な内容をできるだけ除外する重要文抽出法を目標とした．文章においてそれぞれの文は，その文で用いられている語と語の関係を明らかにしていく [9]．そこで，テキスト中に含まれる語の共起関係を分析し，要約に含めるべき共起関係をできるだけ取り込むような文の組み合わせを求めることで，複数文書要約システムを構築することにした．このような組み合わせ最適化問題を解くことによって，重要度上位の文から抜き出してくる従来の重要文抽出方法と比べて，原文に含まれる内容をより網羅的に捕らえることができる．

3.2 重要文抽出問題の定式化

具体的な方法についてであるが，まず，複数文書に対して語の共起グラフを構築する．図 1 は要約の対象となる毎日新聞の記事 4 件[4] の記事中に含まれる語の共起関係を示したものである．各ノードは語を表し，リンクは語の共起頻度が 2 回以上であることを示している．共起回数が多い語の組はできるだけ近くに配置するようにし[5]，文書ごとのリンクを異なる濃さで表示している．

さて，このグラフ上でどのような特徴を持つ文が重要であるか考えてみる．先にも述べたように，文章においてそれぞれの文は，その文で用いられている語と語の関係を明らかにしていく．このことをこのグラフ上で見ると，各文がリンクをカバーしていくことに相当する．したがって，多くの語と語の関係を明らかにするような文，つまり，多くのリンクをカバーするような文を抽出してやれば良さそうである．さらに，ある文が選ばれた場合，その文と同じようなリンクをカバーする文を選んでも冗長な情報になってしまう．選ぶべき文は組み合わせ的に決まるものであり，次のような最適化問題に帰着する．

$$\min . f = \sum_{i \in K} \mathrm{cost}_i x_i \tag{1}$$

ただし，K はリンクの集合，cost_i はリンク i が要約に含まれないときのペナルティコスト，x_i はリンク i が要約に含まれなければ 1，そうでなければ 0 である 0–1 変数である．

さらに，要約では文字数を指定されることが多く，要約の長さに関する制約が加わる．

$$\sum s_j l_j \leq L \tag{2}$$

[4]“ハイブリット, カー, 発売, 開発” に対して 98 年–99 年の毎日新聞を検索して得られた記事集合の一部である．

[5]グラフ作成には Graphviz（http://www.graphviz.org/）を用いたが，ノードの配置問題は必ずしも最適解が得られるわけではない．

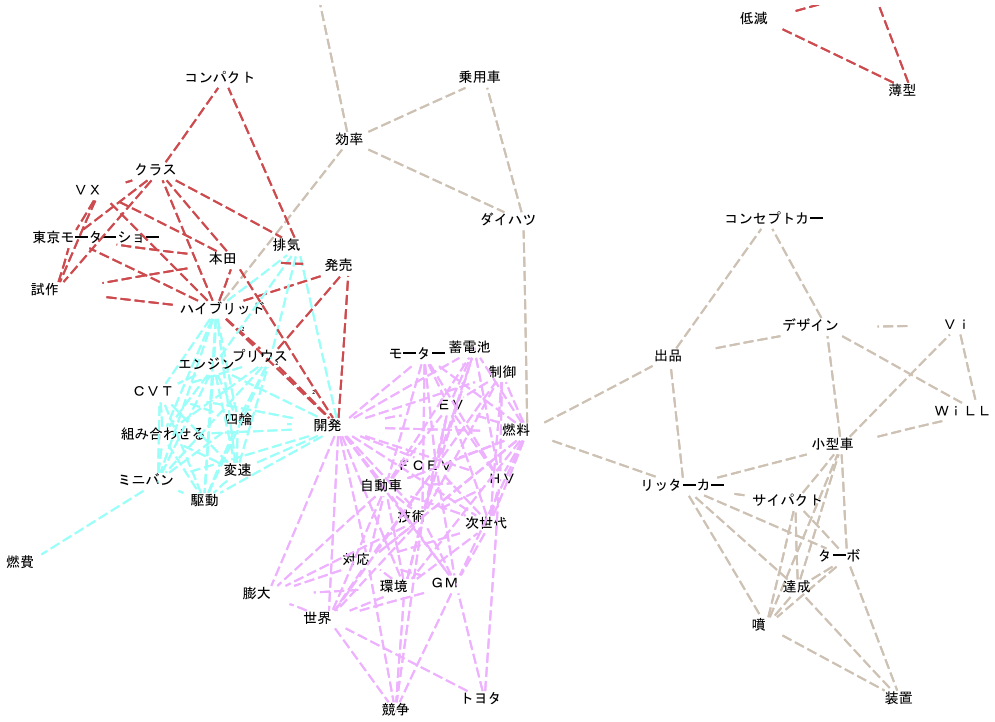

図 1: 「ハイブリットカーに関する記事集合」の共起グラフ．語をノードとし，２回以上の共起関係をリンクで示している．ノード間の距離は共起回数が多いほど近くなるようになっている．

ただし，s_j は文 j を選択するときは 1，選択しなければ 0 をとる 0–1 変数である．$s_j = 1$ のときは文 j に含まれるすべてのリンク i に関して $x_i = 0$ に，そうでなければ $x_i = 1$ になる．また，l_j は文 j の文字数で，L は要約文の文字数の上限を表す．

3.3　仮説推論問題への置き換え

このように定式化すると，複数文書要約問題は上述の仮定のもとで，文を要約に含めるか含めないかという組み合わせ最適化問題で表すことができる．これは，式の制約以外は次のような仮説推論問題で表すことができる．

リンクの総数を k，文の総数を m とする．まず，満たすべきゴールは「リンク 1 からリンク k まですべてのリンクが考慮されている」ことを表す G である．

$$G \leftarrow x_1, x_2, ..., x_k \tag{3}$$

文 j を選択するという仮説は h_{s_j} で表し，コストは 0 とする．例えば，文 1 でリンク 13 番，リンク 220 番，リンク 223 番がカバーされているとすると，

$$x_{13} \leftarrow h_{s_1}, x_{220} \leftarrow h_{s_1}, x_{223} \leftarrow h_{s_1} \tag{4}$$

と記述することができる．一方，選択されないリンク i に対しては，

$$x_i \leftarrow h_{emp_i} (i = 1, ..., k) \tag{5}$$

という仮説 h_{emp_i} を便宜的に用意しておき，ペナルティコストを与える．このペナルティコストの値が大きいほど，そのリンクは要約に含められる可能性が高くなり，逆にペナルティコストの小さ

いリンクは，要約に含めなくても全体のコストへの影響は少ない．そこで，このコストの値はリンクの両端のノード（語）に応じて決定することにし，今回は両端の語の tf*idf 値の平均を採用した．

以上で,「カバーされないリンクの数がもっとも小さくなるように文を選択する」という問題を記述できたことになる．しかし，このままではすべての文を選択することでコストが最小値 0 となってしまう．要約では文字数の制限が本質的であるので，この制限を入れなければならない．

文字数の制限を入れるには文を選択する仮説 h_{s_j} にコストを付与すればよい．しかし,「リンクがカバーされないこと」と「文字数」とは異なる性質のコストであり，指定された文字数に合わせたコスト値を探すのは大変である．むしろ，決められた文字数の中で，もっともリンクをカバーするような文を選ぶというように，文字数は制約として考えた方が適当である．

2 種の置き換え法の強調による高速仮説推論法 [11] では，変数間の制約をある程度自由に記述することができる．したがって，式 2 に相当する制約，例えば

$$39h_{s_1} + 77h_{s_2} + 54h_{s_3} + ... \leq 500 \tag{6}$$

（文 1 が 39 文字，文 2 が 77 文字，文 3 が 54 文字，...，全体の文字数が 500 文字以内の場合）を別に記述しておく．

このように，各複数文章要約問題に対して知識ベースを生成し，G を証明するような仮説の組を求めることで，要約となる文の集合を求めることができる．

3.4 システムの実装

要約対象となる文章は茶筅[6] を用いて形態素解析を行い，形態素と品詞の同定を行った．語の共起グラフを作成する際には，名詞，動詞，未知語だけを用いている．システムの実装には C 言語を用いた．また，実験にあたっては TSC の dryrun で出題された新聞記事集合を用いた．

4 結果と考察

図 2，図 3 に我々のシステムが実際に生成した要約の例を示す．紙面の都合で，要約する前のオリジナルの文は割愛させていただく．

> ハイブリッド車の開発はトヨタ自動車が先行し、昨年１２月に「プリウス」を発売。…昨年１０月の東京モーターショーで、本田は１０００ＣＣクラスのハイブリッド車の試作車「Ｊ－ＶＸ」＝写真＝を展示。
>
> トヨタ自動車と米ゼネラル・モーターズ（ＧＭ）は１９日、次世代低公害車の本命として期待されている燃料電池電気自動車（ＦＣＥＶ）など、環境対応型の先進技術車を共同開発することで合意したと日米で同時発表した。…共同開発するのは、燃料の水素と空気中の酸素を化学反応させて発電し、モーターで走るＦＣＥＶのほか、ガソリンエンジンと電気モーターを併用するハイブリッド自動車（ＨＶ）、蓄電池でモーターを動かす電気自動車（ＥＶ）などをめぐる幅広い技術。
>
> ガソリンと電気を組み合わせたハイブリッドカーはこれまでプリウスの１５００ＣＣだけだったが、より大きなパワーが必要なミニバン向けに、２４００ＣＣのエンジンとモーター、無段変速機（ＣＶＴ）を組み合わせたハイブリッド車初の四輪駆動方式を新開発した。
>
> 日産自動車は直噴（直接噴射式）ディーゼルターボエンジンの小型車「サイパクト」で３リッターカーを達成した。

図 2: システムが作成した要約の例．要約のソースは「ハイブリッドカー」に関する毎日新聞の 4 記事.

[6] http://www.chasen.org/

第１８回冬季オリンピック長野大会第４日の１０日、長野市のエムウェーブで行われたスピードスケート男子五百メートルで清水が優勝、今大会の日本勢金メダル第１号となった。今大会からインコース、アウトコースをスタートとする２回のレースの合計タイムで争われることになった同種目。
１１日、長野冬季五輪フリースタイルスキー・モーグル女子。海外遠征先からも真っ先に電話していた昌昭さんの病に、里谷選手は揺れた。一息ついて、力強く「自分のためにも滑りました」と言い切った。
長野冬季五輪第９日の１５日、白馬村ジャンプ競技場で行われたスキー・ジャンプのラージヒル（K点１２０メートル）で、船木和喜（デサント）が優勝、原田雅彦（雪印）が３位に入った。
○…１５日のジャンプ・ラージヒル個人戦で、２回目の２５番スタートで１３６メートルを飛んだ原田雅彦の飛距離や１回目との合計得点・順位などがすぐに電光掲示板に公表されず、その後の順位変動や終了後のメダル速報が混乱、終了約１０分後にようやく速報された。
日本のジャンプ団体戦優勝は初めてで、日本選手団が夏・冬季五輪を通じて獲得した金メダルは通算１００個となった。

図 3: システムが作成した別の要約の例.「長野，オリンピック，日本，金」をクエリに与えて毎日新聞を検索した結果得られた 7 記事が要約のソースとなっている.

ハイブリットカーに関する記事を集めた要約（図 2）では，特殊なヒューリスティックを導入していないにもかかわらず，新聞記事の LEAD 文[7] が多く抜き出されている．内容の重複なども見受けられず，ハイブリットカーに関する様々なメーカーの対応が簡潔にまとまった要約となっている．

図 3 は，長野五輪で日本人選手が金メダルを取ったことに関する記事集合を要約した例である．試合結果の速報を伝える内容に加え，日本人選手の勝利にまつわる逸話も要約に含められている点にも注目していただきたい．選手の優勝を伝える文が一部抜け落ちてしまっているが，このソースとなる文章は，クエリとして「長野五輪」「日本」「優勝」「金」を与えたときの検索結果であるため，日本人選手が優勝したという事実は，要約を読む側にとってすでに既知となっている可能性が高い．この記事集合にはもともと,「長野五輪」「日本」「優勝」「金」に密接な共起関係があったわけで，我々の要約手法では一度要約に含めた共起関係を再度取り入れることに反発するため，これらの LEAD 文は選択されにくい．代わりに，設定された要約文字数が埋まるまで，別の共起関係を要約に取り込もうとしている結果がうかがえる．

5　今後の課題

しかしながら要約という観点から眺めた場合，いくつかの問題点も見受けられた．ある事件に関する続報記事を集めた記事集合を要約した際,「15 日夜○○が××された事件で……」という表現がよく見受けられた．これから述べる事件がどの事件のものなのかを明示する意図の表現であるが，事件を詳細に記述する内容の文を要約の中に含めた場合，後続の文としてはこのような表現は冗長であり，削除するか簡潔な表現に置き換えるべきである．

また要約対象のテキスト集合によっては，テキスト収集したときに使ったクエリ以外にも，トピック的なまとまりを含むことがある．例えば「台湾大震災」に関する記事集合では，速報記事，震源を伝える記事，被害状況を伝える記事，諸外国の声明を伝える記事，被災地の状況のレポート記事など，様々な小トピックを含んでいた．このような場合，これらの小トピックに沿って要約文を並べ替えてやらないと，つながりの悪い要約文となってしまう．

このような問題に対処するため，記事集合の中に含まれる小トピックをクラスタリングしたり，文を単位に抽出するのではなく，節やフレーズを単位にするなど，要約としての応用には工夫が必要である．

[7]LEAD 文とは，新聞記事の本文中で一番最初もしくはその近辺に現れる文のことを指す．新聞記事における LEAD 文では，これから述べることが簡潔にまとめられていることが多く，LEAD 文を抜き出していくだけでそこそこ良い要約になるので，重要箇所抽出のベースラインシステムとしてよく用いられる．

6 結論

本発表では，語の共起グラフ上でのリンク被服問題を用いる重要文抽出法を紹介した．要約システムとして応用するには文の並べ換えや冗長表現への対応など，さらなる工夫が必要であったが，この抽出法では，元の文章に含まれる内容を広くカバーするとともに，元の文章に含まれる冗長な内容を削減することをいくつかの例を交えて示した．

謝辞

我々は国立情報学研究所情報学資源研究センターの支援により開催されているワークショップ NTCIR のテキスト自動要約タスク (TSC) に参加し，本研究にあたっては毎日新聞記事データ，要約課題データを利用させていただきました．ここに感謝の意を表明いたします．

参考文献

[1] Mani, I. *Automatic Summarization*. John Benjamins Publishing Company, 2001.

[2] Luhn, H. P. The automatic creation of literature abstracts. *IBM journal of Research and Development*, Vol. 2, No. 2, pp. 159–165, 1958.

[3] Salton, G. *Automatic Text Processing*. Addison-Wesley, 1989.

[4] Edmundson, H. P. New methods in automatic extracting. In *Journal of the Association for Computing Machinery*, 16(2), pp. 264–285, 1969.

[5] Marucu, D. From Discourse Structures to Text Summaries. In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pp.82–88, 1997.

[6] Barzilay, R. and Elhadad, M. Using lexical chains for text summarization. In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pp.10–17, 1997.

[7] Salton. G., Singlhal, A., Buckley, C., and Mitra, M. Automatic Text Decomposition Using Text Segments and Text Themes. In *Proc. of the 7th ACM Conference on Hypertext*, pp.53–65, 1996.

[8] Mani, I. and Bloedorn, E. Machine Learning of Generic and User-Focused Summarization. In *Proc. of the 16th National Conference on Artificial Intelligence*, pp.662–628, 1998.

[9] Halliday, M.A.K, Hansa, R, *Cohesion in English*, Langman, 1976.

[10] 奥村 学，難波 英嗣．文書自動要約に関する研究動向．自然言語処理,「テキスト要約のための言語処理」特集号，Vol.6, No.6, pp.1-26, 1999

[11] 松尾 豊，石塚 満．コストに基づく仮説推論の２種の連続値最適化問題への置換法とその強調による推論法．人工知能学会論文誌，Vol.16, No.5, pp.400–407, 2001.