広域知能へ向けての複数テキスト文書の要約手法

石塚 満　　(協力者：岡崎直観, 松尾 豊, 松村真宏, 友部博教)
情報理工学系研究科　電子情報学専攻

概 要

今や情報流通，情報共有の基幹的インフラストラクチャになってきた WWW (World Wide Web) に知的能力を付与し，広域知能基盤に成長させるための一つのアプローチとして，本研究では複数テキスト文書の要約に関する研究を行った．ここでの手法の特色は，テキスト中に含まれる語の共起関係を分析してグラフ構造にし，コストに基づく仮説推論を援用した最適化処理を適用して，内容の重複する文を極力抑えながら出来るだけ重要な話題を網羅的に含みような要約を生成することである．

1　まえがき

WWW(World Wide Web) は今や情報流通，情報共有の最も重要な新しい情報インフラストラクチャになっきており，今後も発展，進化を続けていくことになる．流通，蓄積，共有される情報量は膨大であるので，今後は特に膨大な情報を有効に利用するための知的能力，知的メカニズム付与し，グローバルかつパーソナルな広域知能の基盤に成長させる必要がある．

情報の形態には画像，映像，音声といったマルチメディアもあるが，テキスト情報は意味を伝達する上で特に重要な役割を担っている．情報洪水と言われるような膨大な情報の中から，関心をもつトピックに関する重要で必要な情報を見出して提示するには，キーワード抽出，重要文抽出，要約作成などの技術の高度化が要請される．ここでは，これまでにあまり研究がなされていなかった複数テキスト文書の要約についての研究を行った．

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，本研究では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．ここでは，我々の考案した複数新聞記事に対する内容の網羅性を重視した重要文抽出法について概要を記す．

なお，本研究は国立情報学研究所 (NII) 主催により 2001–2002 年に開催された，ワークショップ NTCIR のテキスト自動要約タスク (TSC) のコンペティションに参加して行ったものである．

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，1) キーワードの出現頻度 [2, 3]，2) 文書中あるいは段落中での位置情報 [4]，3) 文書のタイトルやメタデータ [4]，4) 文書中の手がかり表現 [4]，5) 文間の関係を解析した文書構造 [5]，6) 文あるいは単語間のつながり情報 [6]，7) 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．

2.3 複数テキストの要約

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 提案の網羅性重視の要約手法

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 (”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) 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，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 が要約に含まれると 0，そうでなければ 1 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．

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

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

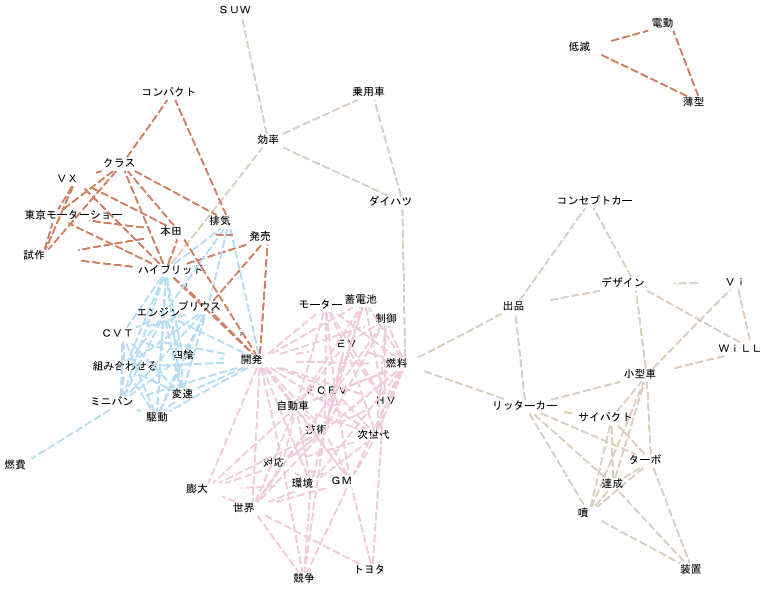

図 1: 「ハイブリットカーに関する記事集合」の共起グラフ.
(語をノードとし，2 回以上の共起関係をリンクで示している．ノード間の距離は共起回数が多いほど近くなるようになっている．)

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 コストに基づく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，式の制約以外は我々がこれまでに行ってきたコストに基づく仮説推論法 [11] で，効率的に解くことができる．スペースの関係で，ここではその詳細は省略する.

3.4 システムの実装

要約対象となる文章は茶筅 (http://www.chasen.org/) 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 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する.

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.

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

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)

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．

6 むすび

本研究では，語の共起グラフ上でのリンク被服問題を用いる内容の網羅性を重視した重要文抽出法を研究開発した．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．

この他に，内容のまとまりを重視した要約手法の研究開発も行った．

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.