E-2

複数文書からのテキスト断片抽出法
Extracting Textual Units from Multi-documents

岡崎 直観† 松尾 豊‡ 石塚 満†
Naoaki Okazaki Yutaka Matsuo Mitsuru Ishizuka

1. まえがき

近年多くの文章がオンラインで手に入るようになった．このような状況は便利である一方，どの情報が自分にとって必要なのか選別するだけでもコストがかかるという問題を生んでいる．いわゆる情報オーバーロード問題の解消を目的としている研究にテキスト自動要約がある．要約とは文章の中から重要と思われる情報を抜き出して縮約版の文章を作成することである [1]．要約を作成する際の「重要」という観点は要約を用いるユーザーやタスクによって異なる．たとえば，文章から何らかの意思決定に関する知見を得たい場合，要約システムが意思決定に関してクリティカルな内容を出力するのが望ましいが，要約システムの精度や人間とシステムとのインタラクションの問題もあるので，意思決定を下すための情報として要約を利用する．先に述べたのと同様に，文や文の集合としての文章の重要度は意思決定を行うユーザーや目的によって異なるため，やはり一般化が難しい．

そこで，要約タスクを新たな知識や情報をすばやく得ることとして一般的し，複数の文章中に含まれている情報を決められた制限文字数内でできるだけ詰め込むような文の集合を選ぶテキスト断片抽出法を考案した．従来の方法ではテキスト断片に対して割り当てた重要度を元に要約を作成するため，内容をできるだけ広くカバーするというアプローチではなかった．また，従来方法である高頻度の事象に着目した要約で得られる内容は，もともとユーザーにとって既知のものかもしれない．そこから一歩を踏み出して新たな知識を得たい場合，我々の手法は有効であると考える．

我々の手法では，要約というタスクを語の共起グラフ上における辺被服問題に置き換えたが，辺被服問題はNP完全問題であるため，コストに基づく高速仮説推論手法を用いて要約を直ちに得られるようにした．

2. 提案手法

今まで述べたように，複数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テキスト断片抽出法を考案した．以下ではテキストからの抽出単位を文としてタスクの定式化を行うが，他の抽出単位でも同様である．

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 [2]．したがって，多くの語と語の関係を明らかにするような特徴を持つ文が元の文章の内容を広くカバーする．このことを語の共起グラフ上で見ると，各文によって語と語のリンクがカバーされていき，その中でも多くのリンクを被覆するような文が元の文章の内容を広くカバーする．

†東京大学大学院情報理工学系研究科, The University of Tokyo
‡独立行政法人産業技術総合研究所, AIST

複数文書要約では，内容のカバーとともに冗長な内容を省くことも重要である．このことを共起グラフ上で考えると，ある文を要約文として抽出した場合，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であるということである．したがって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．

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

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．

$$\sum s_j l_j \quad L \qquad (2)$$

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．

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．

$$G \quad x_1, x_2, ..., x_k \qquad (3)$$

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，

$$x_{13} \quad h_{s_1}, x_{220} \quad h_{s_1}, x_{223} \quad h_{s_1} \qquad (4)$$

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，

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

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．

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_i} 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 [3] 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

$$39h_{s_1} + 77h_{s_2} + 54h_{s_3} + ... \quad 500 \qquad (6)$$

（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. 結果と考察

図1に我々の文抽出法により生成した要約の例を示す．紙面の都合で，要約前のオリジナルの記事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1: 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．

ハイブリットカーに関する記事を集めた要約（図1）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．

また，要約結果を割愛するが長野五輪での日本人選手の優勝に関する記事集合では，競技の結果を伝える内容の他に優勝秘話も要約に含まれていた．要約の読み手にとっては競技の結果は既知である場合があるので，このような逸話が選ばれたのは我々の抽出法の特徴を示している．

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……」という表現がよく用いられる．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，削除するか簡潔な表現に置き換えるべきである．

また要約対象の文章集合によっては，文章を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る．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．

このような課題に対しては，記事集合の中に含まれる小トピックをクラスタリングする，文を単位に抽出するのではなく節単位にするなど，さらに工夫が必要である．

4. 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テキスト断片抽出法を紹介した．要約システムとして応用するには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，要約課題データを用いました．また，形態素解析器として奈良先端科学技術大学の茶筅を利用させて頂きました．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] Halliday, M.A.K, Hansa, R, *Cohesion in English*, Langman, 1976.

[3] 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.

[§]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，論旨が簡潔にまとめられていることが多い．