手順テキストを対象とした比較対照要約

高木 優　　藤井 敦
東京工業大学大学院情報理工学研究科計算工学専攻

1 はじめに

ある事柄に関する体系的な解説には，その事柄に固有の観点に基づく多面的な記述や，他との対比による特徴付けが有効である．例えば，「第 47 回衆議院議員総選挙」に対するウィキペディア記事は，「争点」，「立候補者」，「選挙結果」などのセクションで構造化され，「選挙結果」に関する記述では前回選挙との比較が行われている．

複数文書要約（Multi Document Summarization）や比較対照要約（Comparative/Contrastive Summarization）は，ある事柄に関する多面的な記述や他と対比する記述を自動的に生成する処理であり，ある出来事に関する新聞記事を対象とした複数文書要約 [1] や，ある商品に関して評価が異なるレビュー文の対を特定する比較対照要約の手法 [4] が提案されている．さらに，ある目的に関する類似の手順集合から一つの代表的な手順を生成する複数文書要約の手法 [8, 9] も提案されている．しかし，手順を対象とした比較対照要約の手法は存在しない．

本研究は，ある目的に関する類似の手順集合から複数の代表的な手順を生成して，各手順の特徴や実行するための条件を可視化する比較対照要約の手法を提案する．図 1 は，本研究が目指す比較対照要約の例である．図 1 にはバジルの育て方に関して 4 つの異なる代表的な手順が記されている．例えば，図 1 の左上 till soil から一番下の harvest までの操作が一つの手順を記しており，till soil の上にあるラベルから，種からガーデンで育てる場合の手順であることがわかる．また，手順間で共通する操作はまとめてあるため，各手順の共通点と相違点が構造的にわかる．

2 関連研究

比較対照要約における先行研究 [4] の手法は，文書を基本単位である文に分割した後に，比較対照できる情報を持った文のペアを抽出し，それらのペアを内容の重要性を考慮しながら，列挙することにより要約を生成する．しかし，本研究の目的とする比較対照要約の生成には以下の二つの観点により，先行研究の手法を適応できない．

1. 比較する軸
   先行研究では意見の肯定・否定といった事前に決まった軸で比較を行う．本手法では，pot/garden などの自明ではない軸での解析が必要となる．
2. 比較する単位
   先行研究では，最小単位である文のペアを比較する対象の単位としている．本手法での比較する対象の単位は till soil → add compost to soil といった操作系列であり，比較単位とする操作系列がいくつの操作を含むかについて自明ではない．

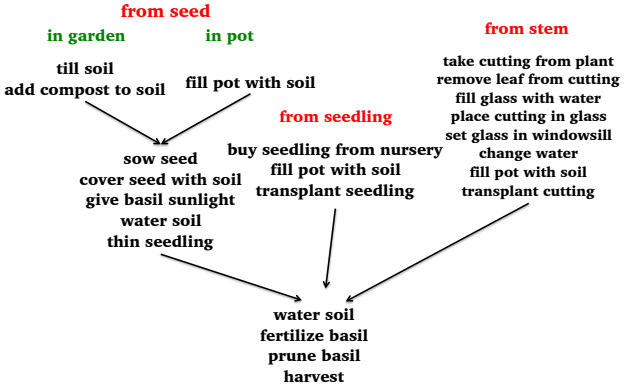

図 1: バジルの育て方に関するフローチャート

Jung ら [3] は，特定の状況で連鎖して起こる行動とそれに付随した情報を持った知識体系の構築手法を提案している．しかし，初めから最後まで一通りの手順を記載する要約目的ではなく，頻出する部分的な行動連鎖を抽出することが目的である．旭ら [7] は，ブログ記事からある目的の下で頻出する行動の連鎖を抽出する手法を提案している．しかし，事前に決められた抽出対象に限定した辞書を作成していることや手法の技術的詳細が記述されていないという問題がある．

山肩ら [8] と難波ら [9] は同一の料理における複数のレシピから，代表的な調理手順を生成する手法を提案している．どちらの研究も一つの代表的な料理レシピの生成を目的としているため，本研究の目的とは異なる．

3 提案手法

3.1 提案手法の概要

図 2 は，提案手法の流れである．同一目的に関する異なる手順を記述したテキスト集合を入力とし，まずはじめに各手順テキストから操作表現を抽出する．つぎに，抽出した操作表現に対して，同義の操作表現をまとめる．その後，図 1 での till soil → add compost to soil のような結束性の強い操作系列を生成する．以後，この操作系列をブロックと呼ぶ．最後に，生成したブロックをつなぎ，各手順に対する条件を付与することで，フローチャートの生成を行う．

なお操作表現を抽出し，同義の操作表現をまとめるまでは，Jung ら [3] の手法を改変した手法を用いる．

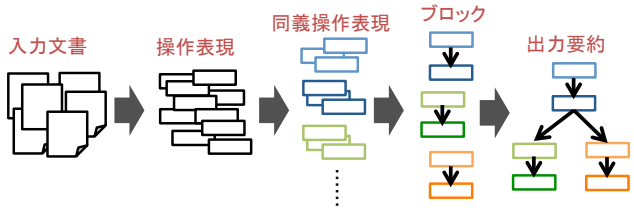

図 2: 提案手法の流れ

3.2 操作表現の抽出

各テキストから手順を構成する一連の操作を述語項構造（述語，目的語句，前置詞句の組）として抽出する．テキスト内の文には操作の補足説明のような操作表現でない抽出すべきでない文も存在するため，抽出すべき文を選択する必要がある．手順テキストには，操作を記述する文は命令文で記述されるという特徴があるため，本手法は命令文を抽出する対象とする．

本手法では，既存の構文解析器（Stanford Parser[1]）を用いるが，構文解析器は命令文に対して，出現頻度の低さと通常の文とは大きく異なる文法から解析を誤ってしまう [2]，そのため，文頭の品詞情報等などから，事前に命令文の判定を行い，適切な位置に主語の補完をし，構文解析を行う．その後，構文解析器の結果を基に，述語，目的語句，前置詞句の組を抽出する．

3.3 同義操作表現の特定

抽出した操作表現に対して，操作の異表記問題と曖昧性の問題を解消することで，同義の操作表現を特定する．操作の異表記問題とは，fill pot with soil と fill container with soil のように同義である操作が異なる表記をされている問題であり，操作の曖昧性問題とは，例えば，「種を植えるための穴を掘る操作」と「検査に用いる土を収集するために穴を掘る操作」のように質の異なる操作が，dig hole のように同じ表記をされてしまう問題である．

異表記問題については WordNet を基にした意味的な類似度を計算することにより解消し，曖昧性問題については，手順の前後文脈に対する類似度を計算することにより解消する．意味的な類似度と手順文脈類似度の線形和が閾値よりも高いものを同義であると判定する．

意味的な類似度は，述語，目的語句，前置詞句に分割して独立に計算し，それぞれの重みに従った線形和を値とする．各述語項の類似度は，NLTK[2] で実装されている WordNet を基にした単語の類似度手法の中から，最も結果の良かった path distance 類似度を利用して計算した．

手順文脈類似度は，計算したい操作表現ペアの前後の操作系列の類似度を計算する．当該の操作表現ペアの前後それぞれ k 個の操作表現を抽出し並べて操作表現列を作成する．作成した操作表現列のペアに対して，編集距離を計算し，距離を類似度に変換した値を手順文脈類似度とする．編集距離の各編集コストは，置換: $a\,(1 - sim(op_1, op_2))$，挿入: 1，削除: 1 として計算する．置換コストの類似度は意味的類似度を利用し，二つの操作表現が類似するほど置換するコストが小さくなるようにする．パラメタの値は k=3, a=2.0 とした．

3.4 要約生成

同義操作表現が特定された手順集合から，フローチャートの生成と各代表的な手順に対する条件ラベルの付与を行う．提案手法は，操作の共起性を利用することにより，結束性の強い操作系列（ブロック）を生成した後，ブロック間のつながりを特定することでフローチャートを生成し，各手順条件を付与することにより最終的な要約を生成する．

ここで，別の要約生成手法として，以下のような手法が考えられる．まず，各記事を記事のタイトルなどの情報を用いることで，図 1 の from seedling のような条件による記事分類を行い，条件毎で代表的な手順を生成した後に，代表的な手順の共通部分を統合し一つのフローチャートの生成を行う手法である．しかし，求める条件が記事のタイトルに記載されているとは限らず，分類に重要でないタイトルがつけられることも多いことや，手順に対する条件は，from seed, in pot のように複数の条件の組み合わせにもなりうることから，正確に記事を条件で分類するのは非常に困難である．

ブロックを構成する操作群は，以下の三つの要件があることがわかるため，このような要件を満たす操作を選択し集めることでブロックの生成を行う．

1. ブロックを構成する操作群は，同一の条件を共有する
2. ブロックを構成する操作群は，1. の条件では必須な操作である
3. ブロックを構成する操作群は，連続する操作系列である．

ブロック生成の問題を最適化問題として一度に解くことは不可能である．複数文書要約の手法 [6] では，要約生成のための最適な文の選択を最適化問題として解くことがあるが，複数文書要約では一つの集合（要約）を生成するのに対し，本研究では，複数の集合（ブロック）を生成しなければならないため，解の探索空間が爆発的に増加してしまう．

そのため，本手法では，ブロックを構成する操作群の要件を用いて，ヒューリスティックな方法でフローチャートの生成を行う．まず，条件を共有する操作をまとめることにより，ブロックの候補となる要件 1 を満たす操作群を生成する．次に，要件 2 を基に各ブロック候補からひとつのブロックを生成する．最後に，要件 3 によりヒューリスティックな手法による誤りを修正しながら，ブロックのつながりを特定し，フローチャートを生成する．

要件 1 からブロックを構成する操作群は同じ記事に共起することが分かる．そのため，共起性の高い操作を集めることで，要件 1 を満たすブロックの候補を作成する．操作 op_i の要素を各記事に含まれるか否かによって，式 1 のように定義しベクトル化をした後，クラスタリングをすることにより，ブロック候補となる操作群の集合を得る．

$$op_{i,j} = \begin{cases} 1 & \text{操作 } i \text{ は記事 } j \text{ に含まれる} \\ 0 & \text{その他} \end{cases} \quad (1)$$

[1] http://nlp.stanford.edu/software/lex-parser.shtml
[2] http://www.nltk.org

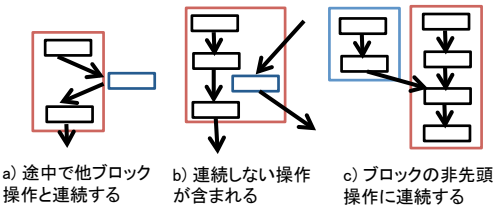

図 3: 操作の連続性に関する誤りパタン

階層クラスタリングを使用し，クラスタ間距離の計算方法は群平均法を，距離の基準はコサイン類似度を使用した．

次に，各ブロック候補から要件 2 を満たす最適なブロックをひとつずつ生成する．式 2 でブロック候補に対する部分集合を評価し，最もスコアの高い部分集合をブロックとする．

$$score(B) = \frac{\min_{op_i \in B} \|op_i\|_0}{\|\sum_{op_i \in B} op_i\|_0} \cdot |B| \qquad (2)$$

$\| \|_0$ は非ゼロの要素数を返す 0-ノルムであり，$\|op_i\|_0$ は，操作の出現する記事の総数になる．そのため，式の前半部分は，ブロック内の操作のいずれかが出現する記事数に対する 1 操作の最小記事出現数の割合になり，出現頻度の低い操作を含むと値が小さくなる．後半部分 $|B|$ は，ブロックに含まれる操作の合計数であり，含まれる操作数が多いほどスコアが高くなる．全体としては，特定の手順条件の記事で必須な操作をできるだけ多く含んだブロックが選出される．

ブロックは各ブロック候補から一つずつ生成されるため，ブロック候補の数だけブロックが生成される．その中から，スコアが閾値 1.0 より大きく，かつ全体の記事の 10 %以上に出現するブロックのみを要約に必要なブロックとして残す．

最後に，操作の連続性に対する要件（要件 3）を基に生成したブロックの誤り修正をするとともに，ブロック間のつながりを特定する．まず，元の記事での操作の連続性から操作間のつながりを計算する．正しくブロックが生成できていれば，ブロック内の操作は一つのパスでつながり，ブロックの最終操作が他ブロックの先頭操作につながっているはずである．しかし，現実的には図 3 のような誤りが含まれる可能性があり，それぞれ (a) ブロックへ操作の追加，(b) ブロックから操作の削除，(c) ブロックの分割を行うことで誤りを修正する．誤りの修正を行った後に，ブロック間のつながりを決定する．

生成されたフローチャートに対して，各手順の条件となるラベルを付与する．手順の条件は，from seed，in pot のように，各ブロックに対する条件の組み合わせで構成される．そのため，各ブロックに対して条件ラベルとなるフレーズを付与する．

まず，ブロックに対する重要な単語の特定を行い，その後フレーズ化を行う．ブロックでの重要な単語は当該ブロックで良く出現し．当該ブロックと排他的なブロックでは出現しにくいと考えられ，文単位での TD·IDF である TF·ISF を用いて評価を行う．図 1 の in pot ブロックの重要単語は，in pot ブロックでの TF · ISF 値から，in garden ブロックでの TF · ISF 値を引いた値の大きい単語とする．

フレーズとして記事タイトル内のフレーズを使用する．例えば，How to Grow Cinnamon Basil in Pot という記事のタイトルを構文解析することにより，grow, cinnamon basil, in pot という三つのフレーズを抽出する．各記事タイトルからフレーズを抽出し，フレーズに含まれる単語の重要度を基に評価を行い，最適なフレーズを選択する．

表 1: テストセット

番号	手順内容	文書ソース	記事数
1	バジルを育てる方法	eHow.com	50 件
2	パンケーキの作り方	allrecipes.com	30 件
3	twitter のフォロワーを増やす方法	eHow.com	14 件

表 2: 操作表現抽出の正解率

	完全一致	部分一致
述語	0.816	0.864
目的語句	0.580	0.762
前置詞句	0.620	0.756

4 評価実験

4.1 実験方法

Web 上から三種類の記事集合を収集し入力文書集合とした（表 1）．提案手法の各ステップは有効性をそれぞれ独立に評価するために，各ステップはそれまでの処理が正確に解けたと仮定して評価した．

操作表現の抽出と同義操作表現の同定に関しては，正解データを基に性能を評価し，簡単な誤りの分析を行う．

生成したフローチャートの評価には，Lupu ら [5] の基準を用いる．参照フローチャート内において，到達可能な任意の二つノード op_i, op_j が出力フローチャートでも到達可能であれば正解として，精度，再現率を計算する．

提案手法は，難波ら [9] と Jung ら [3] の手法と有効性を評価して比較する．難波らの手法は閾値よりも出現確率の高い操作を出力し，Jung らの手法は閾値以上の確率で連続する操作をつなげて出力する．最も f 値が高くなる閾値を選択した．山肩ら [8] の手法は，調理ドメイン固有の手法を使用しているため，対象外とする．

4.2 実験結果と考察

表 2 は述語項別の抽出結果の正解率を示している．部分一致も含めると，いずれの述語項も 7 割以上の正解している．誤り原因としては，文内に複数の述語項が存在する際に，第二述語項以降で正解率が大きく下がること，目的語の省略や前置詞句の係り先の特定に対する誤りなどがあった．

表 3 は，同義操作表現の特定を行った結果である．上段は意味的な類似度のみでの結果であり，下段は手順文脈類似度を混合した全体の類似度での結果である．手順文脈類似度の性能が低く，意味的な類似度と混合しても結果はほとんど向上しなかった．その原因としては，スコアの算出方法が厳しく，多くの操作表現ペアの類似度が 0 に近かったこと，そもそも誤りを含む類似度から算出していることが挙げられる．

意味的類似度の誤り原因は，単語の持つ多義性や，バジルの入ったポットを「バジル」とも「ポット」とも呼

表 3: 同義操作表現特定の結果

	精度	再現率	f-値
意味的類似度	0.564	0.342	0.426
＋手順文脈類似度	0.490	0.378	0.427

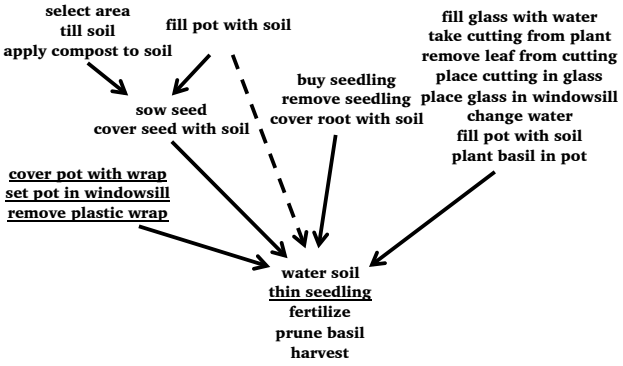

図 4: 提案手法で生成されたフローチャート

べるといった参照表現のゆらぎなどの原因があった．

表 4 は，生成したフローチャートの評価結果である．提案手法は，f 値が最も高く既存手法よりも有効であることがわかる．また，出力結果例（図 4）を見ると，いくつかの誤り（破線矢印と下線の引かれた操作）が存在するものの，複数の手順が提示できていることが分かる．

図 4 における誤り原因について述べる．細かな参照フローチャートとの違いを除いた大きな誤りは三つである．

一つ目は，fill pot with soil から water soil への存在しないエッジを生成していることである．二つ目は，thin seedling の位置に対する間違いである．本来，thin seedling は種から育てるブロックに所属しており，cover seed with soil の直後に来るべき操作である．この二つの誤りは共通して water soil という操作に原因がある．この操作は，繰り返しの周期性がある操作であり，一つの記事中の多くのタイミングで記述される．そのため，操作の順序を決定する際に大きな問題となる．この問題は，一つの記事で複数回出現するような操作は特別な扱いをするといった対処が考えられる．

三つ目の誤りは，不必要なブロックが抽出された誤りである．種を蒔いた後にラップをし，発芽後にラップを外す操作系列は，一連の操作系列ではあるが，必須な操作系列ではなく，なくても代わりの操作系列が必要ではないため，要約に入れるべきではないと判断した．この誤りは，操作の必須性を単純な出現頻度を基にしたことによる誤りであり，その操作系列がなくても成立するのかどうかを考慮することで解決できる可能性がある．

また，今回は誤り原因とはならなかったが，「ポットで育てる場合」と「ガーデンで育てる場合」といった複数の手順が，一つ記事の中で記載される場合があり．本来は排他的な操作が同じ記事で存在するため，操作の共起性によるブロック生成や順序の決定をする際に大きな誤りの原因となりうる．この問題は，操作表現を抽出する段階で，各操作表現間の関係を特定し，構造化した操作表現の系列を抽出することで改善できる．その場合，if you will ··· のような表現が手がかりになる．

ブロックに対するラベル生成の結果（表 5）を見ると，5 ラベル中 3 ラベルが人手による正解ラベルと一致して

表 4: フローチャートの評価

	精度	再現率	f-値
Jung ら	0.243	0.631	0.351
難波ら	0.562	0.414	0.477
提案手法	0.639	0.722	0.678

表 5: 条件ラベル生成の結果

正解ラベル	重要単語（トップ 3）	出力ラベル	結果
in garden	garden, compost, spade	garden basil	不一致
in pot	fill, pot, potting	in pot	一致
from seed	seed, inch, cover	from seed	一致
from stem	fill, stem, water	from stem	一致
from seedling	from, seedling, nursery	sweet basil	不一致

いる．不一致であった 2 ラベルも重要単語の特定まではできており，最適なフレーズが記事のタイトル中に存在しなかったことが誤り原因であった．そのため，統語情報などを用いた本文中からのフレーズ生成手法が必要であると考えられる．

5 おわりに

本研究では，同一の目的に関する手順テキスト集合から，複数の代表的な手順を比較対照するための新たな要約手法を提案した．評価実験の結果，目標とする要約生成において提案手法は既存手法よりも有効であることが分かった．

参考文献

[1] Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. Information fusion in the context of multi-document summarization. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 550–557, 1999.

[2] Tadayoshi Hara, Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. Exploring difficulties in parsing imperatives and questions. *The 5th International Joint Conference on Natural Language Processing*, pp. 749–757, 2011.

[3] Yuchul Jung, Jihee Ryu, Kyung min Kim, and Sung-Hyon Myaeng. Automatic construction of a large-scale situation ontology by mining how-to instructions from the web. *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 8, No. 2, pp. 110–124, 2010.

[4] Hyun Duk Kim and ChengXiang Zhai. Generating comparative summaries of contradictory opinions in text. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pp. 385–394, 2009.

[5] Mihai Lupu, Florina Piroi, and Allan Hanbury. Evaluating flowchart recognition for patent retrieval. *The Fifth International Workshop on Evaluating Information Access*, pp. 37–44, 2013.

[6] Ryan McDonald. A study of global inference algorithms in multi-document summarization. *Proceedings of the 29th European Conference on IR Research*, pp. 557–564, 2007.

[7] 旭直人, 山本岳洋, 中村聡史, 田中克己. 行動連鎖を用いた情報検索支援と web からの行動連鎖の抽出. 電子情報通信学会・日本データベース学会・情報処理学会第 1 回データ工学と情報マネジメントに関するフォーラム, A7-2, 2009.

[8] 山肩洋子, 今堀慎治, 杉山祐一, 田中克己. レシピフローグラフを介したレシピ集合の要約と特徴抽出. 電子情報通信学会技術研究報告 DE 研第 1 種研究会 データ工学と食メディア, Vol. 113, No. 214, pp. 43–48, 2013.

[9] 難波英嗣, 土居洋子, 辻田美穂. 複数料理レシピの自動要約. 電子情報通信学会 言語理解とコミュニケーション研究会, Vol. 113, No. 338, pp. 39–44., 2013.