テキスト自動要約に関する最近の話題

奥村 学[*,**]，難波英嗣[**]

[*] 東京工業大学 精密工学研究所

[**] 北陸先端科学技術大学院大学 情報科学研究科

2000 年 7 月

IS-TM-2000-001

北陸先端科学技術大学院大学

情報科学研究科

〒 923-1292 石川県能美郡辰口町旭台 1-1

oku@pi.titech.ac.jp, nanba@jaist.ac.jp

ISSN 0918-7561

要旨

本稿では，1999年の解説の後を受け，テキスト自動要約に関する，その後の研究動向を概観する．特に，最近注目を集めている3つの話題を中心に紹介する．

1 はじめに

電子化されたテキストが世の中に満ち溢れる現状から，テキスト自動要約研究が急速に活発になり，数年が早くも経過している．研究の活発さは依然変わらず，今年も ANLP/NAACL に併設する形で要約に関するワークショップが開催された他，10 月に開催予定の ACL では要約に関するセッションも企画されている．また，日本では，国立情報学研究所の主催する評価型ワークショップ NTCIR-2 の 1 つのサブタスクとしてテキスト自動要約が企画されており，日本語テキストの要約に関する初めての評価として，また，Tipster における SUMMAC に続く要約の評価として関心を集めている．

本稿では，1999 年の解説 [28] の後を受け，テキスト自動要約に関する，その後の研究動向を概観する．特に，最近注目を集めている，以下の 3 つの話題を中心に紹介する．

- より自然な要約作成に向けて
- 複数テキストを対象にした要約手法
- より良い要約の評価方法を目指して

2 より自然な要約作成に向けて

ここ 1, 2 年テキスト自動要約研究者が関心を持っている話題に，人間にとってより自然な要約を目指すというものがある．

これまでの要約手法である重要文抽出には，問題点として，テキスト中の色々な個所から抽出したものを単に集めているため，抽出した複数の文間のつながり (首尾一貫性) が悪いことが指摘されている．抽出した文中に指示詞が含まれていても，その先行詞が要約中に存在しない可能性があったり，また，不要な接続詞があったりするということだが，こういうことが起きていると，読みにくいということはもちろんだが，最悪の場合，要約テキストの内容を読み間違えてしまう可能性もある．また，文を重要として要約に含める際，他の文とは独立に抽出を行なっており，そのため，結果として要約中に抽出された文の内容に類似のものがいくつも含まれるということが生じる可能性がある．

このような，これまでの要約手法の問題点を受けて，「より読み易い要約」，「より冗長性の少ない要約」を目指す動きが近年活発になっており，また，人間の自由作

成要約 (human-written summary) を元に要約手法を検討する動きも盛んになってきている．

人間が自由に要約を作成する際，原文に基づかず一から要約を「書く」場合もあるが，多くの場合，原文を元に，原文の断片を適切に「切り貼り」し，その後それに編集を加えることで，要約を作成しているという観察を元に，そういった人間の要約作成過程を計算機上にモデル化しようという研究も，後述するように (2.2 節) 始まっている [16]．人間の要約作成モデルに基づく要約手法なら，人間の要約に (ある程度) 近い要約を作成できる可能性があり，注目すべき研究と言える．

もう一つ特筆すべき研究として，自然言語生成システムを利用した要約手法の提案も始まっている [23, 5]．詳細は 3 節で述べるが，複数テキスト中の重要個所を，FUF/SURGE という生成システムにより，つなぎ合わせることで要約として生成している．

要約の過程は，大きくテキストの解釈 (文の解析とテキストの解析結果の生成) と (テキスト解析結果中の重要部分の) 要約としての生成に分けられるとされてきたが，これまでの研究では，要約を生成するということは実際にはほとんど実現されていなかった．今後，より自然な要約作成を目指す過程で，自然言語生成技術の利用は不可欠となっていくであろう．

これまでも，要約の読みにくさ，首尾一貫性の悪さに対しては，対処法が提案されてきているが (Mathis ら [22] や [28] の 2.3 節を参照)，いずれも ad hoc な手法という印象が強い．これに対して，抽出した重要文集合を書き換える (revise) ことで，文間のつながりの悪さを改善し，より読み易い要約作成を目指す研究が最近試みられたりしている [33]．まだまだ技術的に難しい問題がいろいろあるが，興味深い．

また，重要文抽出ではなく，文中の重要個所抽出，不要個所削除による要約手法はすでに [28] で紹介しているが，この要約手法も，より自然な要約を作成するための第一歩と言える．2.3 節で紹介する「要約の言語モデル」は，この要約手法を統計的に定式化した枠組とも考えられる．

2.1 冗長性の少ない要約に向けて

複数テキスト要約では，複数のテキストから抽出した内容を要約とする際，内容が重複することを避ける手法がとられることが一般的である．単一テキストを対象にした要約作成でも，要約中に類似した文が含まれていれば冗長であり，冗長性を削減することで，他の有用な情報を要約に加え，要約中の情報の密度を増すことができる．近年単一テキストの場合にも，要約中の冗長度を下げ，同じ長さの要約に，

より多くの情報を含められるよう考慮した要約手法がいくつか提案されている．

Baldwin ら [3] は，照応解析に基づき，query-sensitive で indicative(指示的) な要約を作成する手法を提案している．テキスト中の文を選択するのだが，検索要求中の句がすべて要約の中にカバーされるように選択する．テキスト中の句がその句と相互参照していれば，検索要求中の句はカバーされているとする．

文を選択する基準は，その文により新たにカバーされる (すでに選択された文ではカバーされていない) 検索要求中の句が多い文を選択する．この文選択をすべての句がカバーされるまで繰り返す．これにより，要約の冗長性を最小にしている．

Baldwin らの手法は，なるべく冗長な参照句を含まないように文を選択していることに相当する．また，先行詞を要約中に含まない代名詞は，可能なら先行詞に置き換える，不要と考えられる，前置詞句，同格の名詞句，関係節は除去するなどの後処理も施している．

MMR(Maximal Marginal Relevance)[6, 7] は，テキスト検索，単一テキスト要約，複数テキスト要約において利用可能な尺度であり，検索要求との適合度と，情報の新規性 (すでに選択されたものとの異なり度) をともに考慮する尺度である．MMR は，テキスト検索を例にすれば，以下の式で定義される．

$$MMR(Q, R, S) = Argmax_{D_i \in R \setminus S}[\lambda Sim_1(D_i, Q) -$$

$$(1 - \lambda) max_{D_j \in S} Sim_2(D_i, D_j)]$$

ここで，

Q: 検索要求，

R: システムによって検索された (ランク付けられた) テキスト群，

S: すでに選択された R の部分集合

MMR を用いた要約では，query-relevant な要約を作成するが，単一テキスト要約では，検索要求に関連するパッセージ (文) を抽出した後，それらを MMR で再順序付け，要約の長さまで文を選択し，原文での順序，MMR のスコアの順序等を元に出力する．MMR を用いることで，要約は互いに (最大限) 異なる文により構成される．MMR を用いた複数テキスト要約は，3 節で紹介する．

加藤ら [29] は，放送ニュースを対象にした重要文抽出法として，まず 1 文目 (リード文) を抽出した後，それ以後の文のうち，リード文と内容が重複しない文を重要として抽出する手法を提案している．内容の重複は，文間の単語の対応の度合を元

に計算している．この手法は，重要文抽出に，テキスト中での位置情報と MMR の考え方を併用していると言うことができる．

石ざこら [31] は，同一の事象を表す表現が複数回テキスト中に出現した場合，2 回目以後の出現を重複部分として削除する手法を提案している．

2.2 人間の自由作成要約を目指して

人間は，単に重要文を抽出するだけでなく，それらを編集することで要約を作成していると考えられる．Jing ら [15, 16] は，人間の自由作成要約と原文の対応を分析し，抽出された文を編集する 6 つの操作を同定している．それらは，不要な句の削除 (文短縮)，(短縮した) 文を他の文と結合する (文の結合)，構文的変形，句を言い替える (語彙的言い替え)，句をより抽象的/具体的な記述に置き換える，抽出した文を並べ替える，の 6 つである．一方，人間が原文に基づかず，一から書いている文も自由作成要約には含まれており，その割合は，300 要約を調べたところ，19%であったと報告している．

Jing ら [16] は，人間の自由作成要約の分析から得られた 6 つの編集操作を用いた，「切り貼り」に基づく要約手法を提案している．システムは，抽出された重要文を編集し，不要な句を削除し，結果として残った句をまとめ上げることで一貫性のある文を作成する．Jing らの切り貼りに基づく要約システムは，まず重要文を抽出した後，抽出した文を，6 つの操作で (文短縮，文の結合のみが実装されている) 編集し，その結果を要約として出力する．文の結合に関しては，対応コーパスを分析し，人手で規則を作成して実現している．文の結合は，2 つの構文解析木に対する，結合，部分木の置換，ノードの追加という操作が TAG 上の操作として実装されている．

一方，文短縮は，抽出された重要文から，不要な句を自動的に削除するが，人間の自由作成要約と原文の対応コーパスから得られた統計情報，構文的知識，文脈情報を利用して，削除する句を決定している [13].

原文は，構文解析され，構文解析木中の必須要素と考えられる部分は印が付けられ，後の処理で削除され，文法的でない文が作成されることを防止する．次に，文中の句で話題ともっとも関連するものを決定する．また，対応コーパスを構文解析した結果を用いて，どの句がどういう条件でどの程度削除され易いか (たとえば，主動詞が 'give' のとき，'when' 節が削除される確率) を計算する．また，句が短縮される (部分が削除される) 確率，句が変化しない確率も合わせて計算される．そして，必須でなく，話題とあまり関係がなく，人間が削除している確率がある程度

ある句を削除の対象とする．

人間の削除個所との一致度に基づく評価では，平均で 81.3%の精度を得ており，すべての前置詞句，節，to 不定詞，動名詞を削除する場合を baseline と考えるなら，その精度は 43.2%だった．また，システムは平均で文の長さを 32.7%短くしていたが，人間の場合は 41.8%だった．システムの出力における誤りの原因は，50 文を分析した結果では，8%が構文解析誤りによるものだった．

この Jing らの研究と同様，(重要文抽出ではなく，) 人間が自由に作成した要約のコーパスに基づいた要約研究が近年数多く見られる．これらの研究では，人間の自由作成要約と原文を対応付けた (aligned) コーパスが必要であるため，要約と原文の間の対応づけ (alignment) を行なう手法に関する提案もいくつか見られる．

Jing ら [15] の対応づけプログラムは，人間の自由作成要約中の句を原文中の句に自動的に対応付ける．要約中で隣接する 2 単語は，原文中でも隣接して現れ易い，遠く離れた文中に現れないというようなヒューリスティックスを元にした HMM に基づいており，要約中の各単語が原文中のどこに位置するかを Viterbi アルゴリズムにより決定する．50 要約中の 305 文に対する対応関係を人手で調査したところ，93.8%の文で正しい対応関係を得ていると報告している．

Marcu[21] は，原文と自由作成要約をともに，出現する単語のベクトルで表現し，その間の類似度をコサイン距離で計算する．そして，自由作成要約と類似度がもっとも大きくなるように，原文から節を削除していくことで，対応する抜粋を決定している．

Banko ら [4] は，文を単位とし，文を文中の単語の出現頻度のベクトルで表し，ベクトル間の距離で文間の類似度を計ることで，自由作成要約中の文と原文中の文をもっとも類似度が大きくなるように対応付けている．Banko らと Marcu の手法はともに，abstract から抜粋 (extract) を生成することを目的としているため，対応させる単位が文，節と大きい．

望主ら [34] も，自由作成要約を原文と対応付けるツールを作成し，対応結果から，自由作成要約，重要文抽出による要約の相違点の分析を行なっている．また，[28] で紹介した加藤らは，要約知識の自動獲得を目的に，単語の部分一致を考慮した DP マッチングによる対応づけ手法を示している．

このようにして，自由作成要約と原文を対応付ける (あるいは，対応する抜粋を生成する) と，自由作成要約と抜粋の間の比較 ·分析が可能になる．

Marcu[21] は，人間の要約に含まれる内容をすべて含むように，テキストの抜粋を作成する場合，どの程度の長さの抜粋が必要であるかを調査している．新聞記事を対象にした場合，対応する要約と比べ，抜粋は 2.76 倍の長さが必要であるとい

う結果を示している．この結果は，抜粋中の冗長性を除去したり，さらに文をより短くするなど，抜粋をさらに加工する必要があることを示しているとも言える．

また，Jingらは，自由作成要約は，対応する抜粋と比較すると，52%の長さであるという報告をしている．Goldsteinら [12] の報告では，平均して抜粋の長さは，自由作成要約に比べ，20%長くなるという．

2.3 要約の言語モデル

原文と自由作成要約の組がコーパスとして大量に存在するなら，人間の要約過程を模倣するようにモデルを訓練することが可能である．KnightとMarcu[17]は，このような考え方に基づき，文要約 (文短縮) において，文法的で，しかも，内容としては原文の情報の重要な部分を維持するような手法を2つ示している．2つの手法は，確率的noisy-channelモデルと決定木をそれぞれ用いている．入力として，単語列 (1文) を与えると，単語列中の単語の部分集合を削除し，残った単語が要約を構成する[1]．

確率的noisy-channelモデルは，統計的機械翻訳の場合と同様，次の2つのモデルで構成される．

- Source Model:
  要約を構成する文 s の確率 $P(s)$．文sが生成される確率を示す．この確率は，文法的でない文の場合低くなり，要約が文法的であるかどうかの指標となる．単純にはbigramでモデル化される．

- Channel Model(Translation model):
  単語列の組 $\langle s,t \rangle$ の確率 $P(t|s)$．要約sがより長い単語列tとなる確率．原文中の各単語が要約に出てくる確からしさを示しており，各単語の確からしさの積をその単語列が要約となる確からしさとする．重要な内容を保持しているかどうかの指標となる．

Knightと Marcuは，上の2つの確率を単語列に対してではなく，それを構文解析した結果得られる木に対して計算している．$P_{tree}(s)$ は，木sを得る際に利用される文法規則に対して計算される標準的な確率文脈自由文法のスコアと，木の葉に現れる単語に対して計算される標準的な単語のbigramのスコアの組合せである．

[1]原文と抜粋の組のコーパスから重要文抽出のためのモデルを学習する手法については [28] の2.2節ですでに紹介している．

確率的な channnel モデルでは，拡張テンプレートを確率的に選択する．たとえば，NP と VP を子ノードとして持つノード S に対して，確率 $P(S \rightarrow NPVPPP|S \rightarrow NPVP)$ を元に，子ノード PP を追加する．

そして，単語列 t からそれに対応する要約 s を選択する際，$P(s|t)$ を最大にするものを選択する．これは，$P(s) \times P(t|s)$ を最大にする s を選択することと同じである．原文中の単語列の部分集合で，上の 2 つの確率の積を最大にするものを Viterbi ビームサーチを用いて選択する．

Ziff-Davis コーパス中の 1067 組の文を対象にし訓練を行なっている．拡張テンプレートは，原文と要約文をともに構文解析し，その木の対応関係から抽出している．

一方，決定木に基づく手法としては，原文に対応する木 t を与えると，それを要約文に対応する，より小さな木 s に書き換えるモデルを示している．拡張した決定的 shift-reduce 構文解析の枠組に基づき，空のスタックと，入力の木 t を入れた入力リストを用いて処理を開始し，より小さな木へ書き換えるべく，shift (入力リストの先頭をスタックへ移動), reduce(スタック上の k 個の木を組み合わせて新たな木を構成し，スタックにプッシュ), drop(入力リスト中の構成素を削除) の操作を繰り返し実行する．

決定木に基づく手法は，noisy-channel モデルに基づく手法よりも，より柔軟であり，原文の構造と要約文の構造が著しく異なる場合にも対処可能である．どの操作を選択するかは，訓練データ (原文-要約文の組の集合から構成される操作の系列の集合) から，決定木学習を行なうことで学習される．

このように，文要約のモデルを，訓練コーパスから自動的に訓練することで得る手法は，Witbrock と Mittal[27] が，原文と abstract の組で直接訓練した確率モデルを適用したのが，最初の研究とされる．これ以外は，前節で紹介した Jing らの研究や，[28] で紹介した，文中の重要個所抽出，不要個所削除による要約手法を含め，いずれも，人手で作成した，あるいは半自動で得た規則を元に，冗長な情報を削除したり，長い文をより短い文に縮めたり，複数の文をまとめたりしている．

堀之内ら [35] は，「日本語らしく，かつ意味的に重要個所を含む」ように，文を短縮する統計的手法を示している．日本語らしさの評価のために n-gram モデル，意味的に重要個所を含むかどうかの評価のために idf をそれぞれ利用している．この 2 つを重み付けした重要度を文中の断片に与え，重要度の小さい断片を繰り返し削除することで文を短縮していく．

小堀ら [30] は，あらかじめ原文から抽出された重要文節データを元に学習した決定木を用いて重要文節を抽出する手法を示している．

2.4 要約における言い替え，書き換えの役割

2.2節で述べたように，人間の要約過程は，単に重要文を抽出するだけでなく，それらを編集する操作が含まれていると考えられる．この編集の操作には，書き換え (revision) や言い替え (paraphrase) が含まれている．本節では，書き換えや言い替えが用いられた要約研究を概観する[2]．

抽出した重要文集合である抜粋を書換える目的には，少なくとも次の2つがあると考えられる．

1. 文の長さを短くする

2. 抜粋を読み易くする

片岡ら [36] は，連体修飾節を含む名詞句を「A の B」の形に言い替えることで要約を行なう手法を示しているが，これは前者に該当すると言える．また，[28] で紹介した，概念辞書等を用いて語句を抽象化する言い替えを行ない要約する手法である「抽象化，言い換えによる要約手法」(3節) や，加藤，若尾らのような手法 (6節) は，言い替えを行なうことで，文字列を削減する要約手法と言うことができる．

また，Mani ら [19] は，抜粋を書き換えることで，質の向上を目指している．3つの操作，elimination, aggregation, smoothing を示している．それらを抜粋に繰り返し適用することで，抜粋の読み易さを低下させずに informativeness を向上できたと主張している．このことから，Mani らの主眼は，書き換えにより，要約内の情報の量を向上させること (抜粋中の不要な個所を削除することで，他の個所の情報を要約に加える) であると言える．

elimination が Jing らの文短縮，aggregation と smoothing が文の結合にそれぞれ対応している．elimination では，文頭の前置詞句，副詞句を削除する．smoothing には，読み易さ (首尾一貫性) を改善するための操作が一部含まれる．

一方，後者の研究としては，難波ら [33] の研究がある．難波らは，人間に抜粋を書き換えてもらう心理実験を行ない，抜粋の読みにくさの要因を分析した後，要因ごとに読みにくさを解消するための書き換えを定式化している．接続詞を追加したり，削除したり，また，冗長な単語の繰り返しを代名詞化したり，省略したり，逆に，省略されている単語を補完したり，などである．そして，そのうちいくつかを実装している．

[2]川原 [32] は，人間の要約作成過程において，どのように書き換えが役割を果たしているかを調査している．

3 複数テキストを対象にした要約手法

これまでの複数テキスト要約研究では, あらかじめ人間が用意した比較的小規模なテキスト集合をシステムの入力として, 要約を作成するのが中心的であったと言える. しかし, 近年情報検索システムの検索結果を直接要約システムの入力に用いるなど, 大規模なテキスト集合を入力とした, より実用性の高いシステムがいくつか提案されている.

膨大なテキスト集合を対象に複数テキストの要約を行う際, テキストの内容に応じて, あらかじめテキスト集合をいくつかのグループに分けておき, グループごとに要約を作成するというのが一般的と考えられる. 最近の複数テキスト要約研究は, このような対象とするテキスト集合の性質により，以下の 2 種類に分類される.

1. 要約システムに与えるテキストはどれもユーザの目的に合致しているが, 似たような内容のテキストが複数含まれる可能性がある. そこで, テキスト (あるいはテキスト中のパッセージ) 間の類似度を考慮し, お互いに内容がなるべく重複しないような要約を作成する.

2. 情報検索の結果得られたテキスト集合を要約システムの入力に用いるような場合, そのテキスト集合には, ユーザの目的と合致しないものが数多く含まれている可能性がある. 従って, テキスト集合を自動的に分類し, グループごとにラベルとして，グループ中のテキスト集合の要約を付与する. ユーザは, 自分の必要なテキストがグループに含まれているかどうかを, 付与されたラベルを見て判断する.

[28] では複数テキスト要約のポイントとして, 以下に示す 3 点を挙げたが, 本節でもこの 3 点に沿って最近の研究動向をいくつか紹介する.

- a 関連するテキストの自動収集
- b 関連する複数テキストからの情報の抽出
  - $b-1$ 重要箇所の抽出
  - $b-2$ テキスト間の共通点の検出
  - $b-3$ テキスト間の相違点の検出
- c テキスト間の文体の違い等を考慮した要約文書の生成

分類 1 の要約手法には Goldstein ら [11], Radev ら [25], Barzilay ら [5], McKeown ら [23], Stein ら [26] のものがある.

Goldstein ら [11] は新聞記事を対象とし, 記事集合中からある検索クエリに関するパッセージを抽出, 収集し (a), それらを並べて要約を作成する MMR-MD(Maximal Marginal Relevance Multi-Document) という手法を提案している. 検索されたパッセージを単純にクエリとの適合度の高い順に並べただけでは, パッセージ間で重複する個所が存在する可能性があり, 要約として望ましくない. そこで, MMR-MD では, クエリに対するパッセージの適合度を考慮しつつ, すでに上位にランクされているパッセージと類似度の低いもの (重複個所が少ないと思われるパッセージ)(b-3) を選択して順に出力することで, 冗長性の少ない複数テキスト要約の作成を行っている. また, パッセージの出力順序を決める際, 記事が書かれた日時なども考慮している.

Radev ら [25] は新聞記事集合をあらかじめクラスタリングし, 各クラスタごとに要約を作成する手法を提案している (a). まず, クラスタ中の記事の各文の重要度を計算し, 次に要約率に応じて記事集合から重要度の高い数文を抜き出し, 抜き出された文を記事の書かれた日付順に並べて, 要約として出力する. 文の重要度は, クラスタの特徴を表す語を文が含む割合 (b-2), 文の位置 (lead)(b-1) により決定する. また, Goldstein らと同様, 自分よりスコアの高い文と内容が重複するような文はスコアを下げることで, 冗長性の少ない要約の作成を目指している (b-3).

Barzilay ら [5], McKeown ら [23] は, 複数の新聞記事間で言い回しは異なるが同じ内容の一連の文を, 7 種類の言い換え規則を用いて同定している (b-2). 抽出された一連の文は, 構文解析器を用いて述語項構造に変換され, 文間で共通な句が抽出される. その後, 文生成器を用いて，抽出された共通語句を統合し, 最終的な要約文として出力する (c).

Stein ら [26] は, あらかじめテキストごとの要約を作成し (b-1), 作成された要約をクラスタリングし, 似たような内容の要約をグルーピングしている. さらに, 各クラスタ中で最も代表的な要約をクラスタの要約として抽出する (b-2). また, クラスタの要約同士の類似度を計算し, 隣接する 2 つの要約の類似度が高くなるよう並べ替えて出力している.

分類 2 の要約手法には Eguchi ら [9], Fukuhara ら [10], Ando ら [2], 上田ら [37] のものがある.

Eguchi ら [9] は, WWW 上のテキストを対象にした適合性フィードバックに基づく検索システムを構築している. このシステムでは, 検索結果 (a) をテキスト間の類似度に基づいてクラスタリングし, 各クラスタごとにクラスタに多く含まれる語と, そのクラスタを代表するテキストのタイトルを, そのクラスタの要約として出力する (b-2). ユーザに出力されたクラスタを選択してもらい, そのクラスタに含ま

れるテキストを用いて適合性フィードバックを行っている.

Fukuharaら [10] も, Eguchiらと同様に検索結果をクラスタリングし (a), クラスタごとに要約出力を行っている. Fukuharaらは, テキスト中の単語の出現頻度分布を考慮し, クラスタごとの話題を表す語とそれらを含んだ文を抽出する. さらに, 抽出された文を, 焦点-主題連鎖を考慮して並べ替え, クラスタごとの要約として出力している (b-2).

Andoら [2] は, ベクトル空間モデルを用いて新聞記事集合の記事間の類似度を計算し, それらを semantic space と呼ばれる 2 次元空間上に配置し表示するシステムを構築している. semantic space 上では各記事はドットで表現され, また, 話題の似た記事は semantic space 上で隣接して配置される. マウスで semantic space 上のドットを指せば, そのドット (記事) と関連のあるドット (記事) が強調され (a), さらに関連記事中の頻出単語 (topic term) や頻出単語を多く含む文 (topic sentence)(b-2) が表示される.

上田ら [37] は, クラスタリングによりある程度同じ話題でまとめられたテキスト集合を対象に, 各クラスタの特徴を表す文を自動的に作成する手法を提案している (a). 上田らも Barzilay ら, McKeown らと同様に, テキスト中の各文を構文解析し, テキスト間で構文木同士を比較することで, テキスト間の共通個所を同定するという手法を提案している (b-2). 構文木の比較には 2 種類の方法を提案している. 1 つは, 例えば 「フーバー社が携帯電話を発売」という文を, 意味的に等価な 「携帯電話がフーバー社から発売」などに構文レベルで変換し, 同一内容の異なる 2 文を同定し, クラスタのラベルとして出力するという方法である. もう 1 つは, シソーラスを用いて 「ホウレンソウからダイオキシンが検出された」と 「白菜からダイオキシンが検出された」の 2 文から 「野菜からダイオキシンが検出された」といったように, より抽象度の高いレベルで融合し, ラベルとして出力するという方法である.

4 より良い要約の評価方法を目指して

人間の作成した抜粋との比較による評価には批判が大きいことは [28] ですでに述べたが, 近年この評価方法を改良する試みもいくつか現れている.

[28] ですでに紹介したように, Jing ら [14] は, 人間の被験者の作成した抜粋との比較による評価と, 要約を利用して人間がタスクを行なう場合の, タスクの達成率による評価の, 2 つの評価方法を分析し, 評価結果に影響を与える要因を同定することを試みているが, その結果少なくとも次の 2 つの点において, これまでの人間の抜粋を用いた評価方法は問題であるとの知見を得ている.

1. 人間の抜粋との比較による評価では，要約率を変化させると，システムの評価がかなり変化する．このため，特定の要約率でシステム間の性能の比較をする意味がどの程度あるのかは疑問が残る．

2. テキスト中に類似の内容を含む文が複数存在する場合，どちらの文が正解として選択されるかにより，システムの評価は大きく変化する．

そして，Jing らは，2 つ目の問題点に対する解決策として，人間が選択した重要文を用いて評価を行なう際，正解と一致した場合正解数 1，一致しない場合 0 として再現率，精度を計算するのではなく，正解数を被験者間の一致の度合として計算する手法を提案している．たとえば，5 人の被験者中 3 人，2 人がそれぞれ一致して選択した文が存在する場合，これまでの評価方法では，前者をシステムが選択した場合正解数 1(過半数以上の被験者が選択しているので)，後者では 0 となるが，提案する手法では，システムの正解数は，前者では 3/5，後者では 2/5 となる．

Mittal ら [24] は，要約率の違いによるシステムの評価の違いに関して，さまざまな要約率における精度を求めた上で，情報検索の評価で用いられている 11 点平均精度 (11 point average precision) のように，複数の要約率での精度の平均として結果を示すべきであるとしている．

さらに，コーパスとするテキスト集合の違いも精度に影響を与えることから，コーパスの要約のしやすさを計る指標として，ランダムに文を選択して要約を作成した場合の精度をベースラインとして示すべきであると主張している．そして，システムの性能を評価する場合，

$$p' = \frac{p - b}{1 - b}$$

(ここで，p, b, p' はそれぞれシステム，ベースライン，補正後のシステムの精度) のように，ベースラインを用いて補正した精度を用いるべきであるとしている．

Radev ら [25] は，文の utility という概念を用いた評価方法を示している．文の utility は，文がそのテキストの話題に対してどの程度適合した内容であるかを示す尺度であり，[0-10] の値をとる[3]．

人間が選択した重要文を用いた，これまでの評価方法は，正解と一致した場合正解数 1，一致しない場合 0 として再現率，精度を計算していたが，utility に基づく評価値は，システムが選択した文に対して人間が割り当てた utility の総和を，正解の文の utility の総和で割った値として計算する．これにより，Jing らの研究で指摘されている 2 つ目の問題点に対する解決策を与えている．

[3]generic な要約を考えた場合，テキスト中での文の重要度を示していると考えて良い．

これまでの評価方法では，システムが選択した不正解の文は，全く評価が得られなかったのに対し，Jing らが提案する手法と同様に，utility に基づく場合，たとえ不正解でも，その文がある程度の重要度を持つ場合，その重要度に対する部分的な評価が得られる点が異なる．ただ一つ正解が存在し，それとまさに一致することを要求されていたこれまでの評価に比べ，正解の文の utility にどのくらい近い utility の文を選択できるかで評価を行なう，utility に基づく方法は，より柔軟で，自然な評価方法と言える．しかし，このような 10 段階での重要性 (適合性) 評価を複数の被験者がゆれなく一致して行なえるか，その作業負荷は大きくないかという問題は存在する．

また，被験者間の一致の度合を J とすると，J は要約システムの精度の上限と考えられ，また，ランダムに選択した時の精度 R は下限と言える．そのため，Radev らも，Mittal らと同様に，システムの性能を計る値を示す際，普通に計算された値 S を単に用いるのではなく，これらの値で正規化した値

$$S' = \frac{S - R}{J - R}$$

を示すべきであるとしている．

Donaway ら [8] は，generic で indicative な要約の評価を行なう 3 つの尺度を示し，それらを比較，検討している．3 つの尺度とは，これまでの評価方法である再現率に基づいた評価尺度，「人間にも，システムにも，テキスト中の文にすべて順位をつけさせるようにして，その文の序列を比較して評価を行なう手法」，「人間の作成した正解要約の単語頻度ベクトルとシステムの要約の単語頻度ベクトルの間のコサイン距離で評価する方法」である．

Donaway らが示している 2 つ目の手法は，これまでの方法がテキスト中の文を重要/非重要の 2 つに分類して評価に利用していたのに対し，テキスト中の文数に分類して利用することに相当する．また，Radev らの手法は，10 に分類していることに対応する．

Donaway らは，実験の結果，3 つ目の評価方法が人間の要約との比較による評価方法としては，Jing らの指摘の 2 つ目の問題点に対する解決策ともなっており，もっとも優れていると結論づけている．

5 おわりに

1999 年の解説 [28] の後を受け，テキスト自動要約の研究分野において，ここ数年関心が高まっている話題を 3 つ紹介した．

テキスト自動要約は，必要性が高まっていることもあり，今後も活発に研究が進められていくことと思われる．今後は，複数テキスト要約だけでなく，さらに対象範囲を広げ，複数の言語で書かれたテキスト (translingual summarization)，複数のメディアの情報を対象にした (テキストだけでなく，画像や音声も対象にする) 要約 (multi-media summarization)，話し言葉の要約なども注目を集めそうである．今後も，テキスト自動要約の研究分野の動向には目が離せない．

最後に，新しい参考文献をいくつか紹介しておく．昨年出版された [20] は，この分野の論文を，古典から最新のものまで集めた論文集であり，テキスト自動要約の最初の研究とも言われる [18] も入っている．この分野で研究を始める人には必読と言える．

Tipster の Text Program Phase III の論文集 [1] も昨年出版されている．SUMMAC 参加システムの概要がいくつか収録されており，また，SUMMAC の dryrun の報告も含まれている．

参考文献

[1] *Proceedings of The Tipster Text Program Phase III*. Morgan Kaufmann, 1999.

[2] R.K. Ando, B.K. Boguraev, R.J. Byrd, and M.S. Neff. Multi-Document Summarization by Visualizing Topical Content. In *Proc. of the ANLP/NAACL2000 Workshop on Automatic Summarization*, pp. 79–88, 2000.

[3] B. Baldwin and T.S. Morton. Dynamic coreference-based summarization. In *Proc. of the 3rd Conference on Empirical Methods in Natural Language Processing*, pp. 1–6, 1998.

[4] M. Banko, V. Mittal, M. Kantrowitz, and J. Goldstein. Generating extraction-based summaries from hand-written summaries by aligning text spans. In *Proc. of the PACLING'99*, pp. 276–281, 1999.

[5] R. Barzilay, K. McKeown, and M. Elhadad. Information fusion in the context of multi-document summarization. In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 550–557, 1999.

[6] J. Carbonell, Y. Geng, and J. Goldstein. Automated query-relevant summarization and diversity-based reranking. In *Proc. of the IJCAI-97 Workshop on AI in Digital Libraries*, pp. 9–14, 1997.

[7] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proc. of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998.

[8] R.L. Donaway, K.W. Drummey, and L.A. Mather. A comparision of rankings produced by summarization evaluation measures. In *Proc. of the ANLP/NAACL2000 Workshop on Automatic Summarization*, pp. 69–78, 2000.

[9] K. Eguchi, H. Ito, A. Kumamoto, and Yakichi Kanata. Adaptive Query Expansion Based on Clustering Search Results. 情報処理学会論文誌, Vol.40, No.5, pp.2439–2449, 1999.

[10] T. Fukuhara, H. Takeda, and T. Nishida. Multiple-text Summarization for Collective Knowledge Formation. Workshop on Social Aspects of Knoledge and Memory, IEEE Systems, Man and Cybernetics Conference, 1999.

[11] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. Multi-Document Summarization by Sentence Extraction. In *Proc. of the ANLP/NAACL2000 Workshop on Automatic Summarization*, pp. 40–48, 2000.

[12] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In *Proc. of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 121–128, 1999.

[13] H. Jing. Sentence reduction for automatic text summarization. In *Proc. of the 6th Conference on Applied Natural Language Processing*, pp. 310–315, 2000.

[14] H. Jing, R. Barzilay, K. McKeown, and M. Elhadad. Summarization evaluation methods: Experiments and analysis. In *Intelligent Text Summarization*, pp. 51–59. AAAI Press, 1998. Technical Report SS-98-06.

[15] H. Jing and K. McKeown. The decomposition of human-written summary sentences. In *Proc. of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1999.

[16] H. Jing and K. McKeown. Cut and paste based text summarization. In *Proc. of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 178–185, 2000.

[17] K. Knight and D. Marcu. Statistics-based summarization – step one: Sentence compression. In *Proc. of the 17th National Conference on Artificial Intelligence*, 2000.

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

[19] I. Mani, B. Gates, and E. Bloedorn. Improving summaries by revising them. In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 558–565, 1999.

[20] I. Mani and M. Maybury, editors. *Advances in automatic text summarization.* MIT Press, 1999.

[21] D. Marcu. The automatic construction of large-scale corpora for summarization research. In *Proc. of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 137–144, 1999.

[22] B. Mathis, J. Rush, and C. Young. Improvement of automatic abstracts by the use of structural analysis. *Journal of the American Society for Information Science*, Vol. 24, No. 2, pp. 101–109, 1973.

[23] K. McKeown, J.L. Klavans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin. Towards multidocument summarization by reformulation: Progress and prospects. In *Proc. of the 14th National Conference on Artificial Intelligence*, pp. 453–460, 1999.

[24] V. Mittal, M. Kantrowitz, J. Goldstein, and J. Carbonell. Selecting text spans for document summaries: Heuristics and metrics. In *Proc. of the 16th National Conference on Artificial Intelligence*, pp. 467–473, 1999.

[25] D.R. Radev, H. Jing, and M. Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proc. of the ANLP/NAACL2000 Workshop on Automatic Summarization*, pp. 21–30, 2000.

[26] G.C. Stein, T. Strazalkowski, and G.B. Wise. Summarizing Multiple Documents using Text Extraction and Interactive Clustering. In *Proc. of the PACLING'99*, pp. 200–208, 1999.

[27] M. Witbrock and V.O. Mittal. Ultra-summarization: A statistical approach to generating highly condensed non-extractive summaries. In *Proc. of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 315–316, 1999.

[28] 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, Vol. 6, No. 6, pp. 1–26, 1999.

[29] 加藤直人，浦谷則好. 放送ニュースを対象にした重要文抽出. 言語処理学会第6回年次大会発表論文集, pp. 237–240, 2000.

[30] 小堀誠，田村直良. 段落中の接続関係と段落間の重要度配分による文章要約. 情報処理学会自然言語処理研究会報告, pp. 79–86, 2000. 136-11.

[31] 石ざこ友子，片岡 明，増山 繁，中川聖一. テレビニュース番組の字幕作成のための重複部削除による要約. 情報処理学会自然言語処理研究会報告, pp. 45–52, 1999. 133-7.

[32] 川原裕美. 要約文のパラフレーズの諸相. 佐久間まゆみ　編），文章構造と要約文の諸相, pp. 141–167. くろしお出版, 1989.

[33] 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報告, pp. 53–60, 1999. 133-8.

[34] 望主雅子，荻野紫穂，太田公子，井佐原均. 重要文と要約の差異に基づく要約手法の調査. 情報処理学会自然言語処理研究会報告, pp. 95–102, 2000. 135-13.

[35] 堀之内寛，山本幹雄. n-gram モデルと idf を利用した統計的日本語文短縮. 言語処理学会第 6 回年次大会発表論文集, pp. 364–367, 2000.

[36] 片岡明, 増山繁, 山本和英. 要約のための連体修飾節の “a の b” への言い換え. 情報処理学会自然言語処理研究会報告, pp. 37–44, 1999. 133-6.

[37] 上田良寛, 小山剛弘. 共通意味断片の抽出による複数文書要約. 言語処理学会第 6 回年次大会発表論文集, pp. 360–363, 2000.