TSC4: 意見要約コーパスとそれを用いたワークショップ

奥村学
(東京工業大学 精密工学研究所)
平尾努
(NTT コミュニケーション科学基礎研究所)
難波英嗣
(広島市立大学 情報科学部)

概 要

テキスト自動要約の評価型ワークショップ TSC では, 過去 2 回にわたり, テキスト集合から要約を作成する複数テキスト要約課題を行なってきている. 4 回目の今回は, その 1 つの課題として, 「質問応答 (QA) の要素技術」としての自動要約の評価を念頭に, 解答となるテキストを, 解答 (の断片) を含むであろうと考えられるテキスト集合を与えて作成するものを検討している. 特に, 人の意見を尋ねる質問 (「自衛隊のイラク派遣についてどういう意見があるの?」) を対象にするが, 解答となるテキスト (意見要約) コーパスおよび, 要約中の文とテキスト集合中の文の対応づけデータ等, 付随するタグ付コーパスの作成方針等について述べる.

1 はじめに

テキスト自動要約は，1950 年代から研究されている研究分野であるが，1990 年代後半から急速に研究が活発になり，今日に至っている．しかし，システムの出力である要約をどのように評価するかに関しては明確な基準がなく，従来評価が難しいとされてきた．しかし，研究が活発化するに伴い，評価方法を議論し，基準を明確にしようという動きも活発になり，アメリカでは現在，DARPA TIDES プロジェクトの一貫で，DUC という要約の評価を行なう会議が毎年開催されるようになっている．日本でも，日本語テキストの要約の評価を目指す動きが本格化し，テキスト自動要約の評価型ワークショップ TSC が開催されてきている.

本稿では，引続き行なわれる予定である TSC4 について課題等を説明する.

後述するように, テキスト自動要約に関する評価型ワークショップは日米で並行して行なわれている. アメリカにおける DUC では今年から大きく方向を転換している. 日本における TSC では, 過去 2 回同様, 複数の新聞報道記事を対象にした複数テキスト要約コーパスを引続き作成する予定だが, それとは別に, アメリカでの動きと同様, より具体的な応用に即した課題として, 意見要約課題を今回予定している. 本稿では, その背景, 狙う点, 課題の内容等について主に説明する.

以下ではまず，アメリカで開催されている DUC の動向について触れた後，TSC4 意見要約課題の概要を説明する.

2 日米におけるテキスト自動要約の評価型ワークショップ

2.1 DUC

TIDES プロジェクト (`http://www.darpa.mil/ito/research/tides/index.html`) の一貫として 2001 年に始まった DUC (Document Understanding Conference) (`http://www-nlpir.nist.gov/projects/duc/`) は, NIST が主催しており, 毎年開催されるテキスト自動要約システムの評価型ワークショップである. これまでに 4 回の評価が行なわれたことになる.

今年の DUC(DUC 2005) は, 過去 4 年間の経験, 成果を元に, 大きな方向転換を行なっている. 過去 4 回の DUC により, 新聞記事の generic な要約に関しては, すでに一定の成果が上がっており, その継続は必要ないという判断である. そして, 新たな方向性として, 要約を必要とする実際のニーズに基づき, より実際的なシナリオから課題を設定しようとしている.

広い分野におけるさまざまなジャンルのテキストを対象に, 実際的な応用として, 多くの情報源からの情報を統合したレポート作成をあげ, 情報要求に対する解答となるテキストをテキスト集合から合成する課題を設定している. 具体的には, 自然災害 (たとえば, 台湾での地震) に関する状況レポート作成を目的として, 何が起き, どういう地域, どの程度の人々が影響を受けたか, (救援として) 何が必要とされているか, 社会的, 政治的, 地理的な制約はどのようか, などの情報を盛り込んだテキストを自動作成する. この課題は, パイロットスタディであり, 2006 年も同じ課題が予定されている.

2.2 TSC

TSC(Text Summarization Challenge)(`http://lr-www.pi.titech.ac.jp/tsc/`) は, NTCIR のタスクの 1 つ

として，2000 年から 2001 年にかけて, 第 1 回 (TSC1) が開催された．要約システムの評価および要約データの蓄積を行なうことを目的としている．TSC は, 約 1.5 年で 1 回のペースで開催されてきており, これまでに 3 回開催されている．

TSC では, 過去 2 回にわたり, テキスト集合から要約を作成する複数テキスト要約課題を行なってきている．その 1 回目, TSC2 では, さまざまな種類のテキスト集合を対象として扱った．報道記事とその続報記事の集合 (formal run 30 トピック中 15 トピック)，ある話題の数値の動向, 推移を記述した一連の記事集合 (3 トピック), さらに, 報道記事だけでなく関連する社説記事を集合に含むものもあった．

2 回目の前回 TSC3 では, その中の「一連の続報記事集合」に対象を特化した．同様に, TSC2 のさまざまなテキスト集合のうち 1 つに着目し, コーパス作成, 評価型ワークショップを行なう試みとして, 加藤ら [2] のものがある．加藤らの試みは, TSC2 のテキスト集合中の「ある話題の数値の動向, 推移を記述した一連の記事集合」に特化し, その出力として, 単なるテキストとしての要約だけでなく, 可視化した形の出力も認めたものと考えることができる．後述するように, 本稿で述べる TSC4 の課題も, TSC2 で扱ったさまざまなテキスト集合のうち,「報道記事と関連する社説記事集合」を対象とした場合から派生していると考えることもできる．

3 TSC4: 意見要約課題

3.1 質問応答の解答としてのテキスト

2.1 節で, DUC では今年, 実際的な応用を指向した課題設定を行なっていること, また, レポート作成という状況で, 情報要求に対する解答となるテキストをテキスト集合から合成する課題を設定していることを述べた．

質問応答は, ある意味で「究極の」情報アクセス技術であり, 情報検索, 情報抽出, テキスト自動要約など, 現在研究されている他の情報アクセス技術は, 質問応答技術の要素技術と位置づけられないわけではない．現状の質問応答では, 対象とする質問の解答は, 固有名や名詞が中心であり, 質問応答技術は, テキスト (パッセージ) 検索技術と, 情報抽出技術の統合技術として実現されている感がある．しかし, 質問は現状で扱われている以外に多様なものがあり, ある単語の意味, 説明を尋ねる質問 (「DNAって何ですか?」), 何かの仕方を尋ねる質問 (「シュクリームの作り方を教えて」), 人の意見を尋ねる質問 (「自衛隊のイラク派遣についてどういう意見があるの?」), 何かの原因, 理由を尋ねる質問 (「バブルはどうしてはじけたの?」) 等では, 解答を含むテキスト (集合) から, 固有名や名詞よりも長いテキストの断片 (パッセージ) を抽出し, (必要なら) それらをまとめて 1 つのテキストを作り出す処理が必要になる．これらの処理は, テキストをより短くすることを要約というなら, 要約とは言えないだろうが, クエリに関連したテキストの断片を抽出する処理まで要約に含めるなら, 要約技術の範疇に入る処理と言える．

このような背景から,「質問応答 (QA) の要素技術」としての要約の評価を目指す課題設定を TSC4 でも行なう．テキストが解答となるような QA を, 解答を含むであろうと考えられるテキスト集合を与えて行なう形式を採る．テキスト集合と質問を与え, 解答のテキストを一定の長さで作ってもらう．

3.2 なぜ意見要約か?

上述したように, テキストが解答となる質問応答にもさまざまなものが考えられるが, 今回は人の意見を尋ねる質問に限定する．今回意見要約にタスクを限定した理由は, DUC における方針と同様, 要約を必要とする実際のニーズに基づき, より実際的なシナリオから課題を設定しようとしたからである．

インターネットの普及に伴い，一般の多くの人々からの情報発信が盛んになり，その発信されている大量の情報を有効に活用したいという要求も高まっている．一般の多くの人々が発信する情報の中で, 特に注目を集めているのが, ものや会社等の対象に対する評判を含む, 人々の意見である．こうした状況を背景に，現在掲示板 (BBS) や blog (Weblog) が情報源として注目され, これらを定期的に監視し，そこから情報を抽出，発掘することで，一般大衆の「生の声」を製品開発，企業活動に反映しようという試みも見られる．

現状こういった目的で開発されたシステムの多くは, 意見を抽出し, それをそのまま列挙する形で提示するか, さらに, 抽出した意見に対し, マイニング処理などを行なう等して, その結果を提示しているものと考えられる．たとえば, 我々の開発しているシステム blogWatcher (`http://www.lr.pi.titech.ac.jp/blogwatcher/`) においても, ユーザが入力したキーワードに対する評判を検索できる機能を提供しているが, 図 1 に示すように, 現在のヴァージョンでは, 検索結果は, 意見が列挙されて提示されるだけである．数少ない, 意見を要約する試みとして, 立石ら [3] の研究があるが, これも領域が狭く限定されている上に, 提示する形態はテキストではなく, レーダーチャート形式になっている．

図 1: キーワード「レストラン」に対する評判検索結果

言うまでもなく, 大量の抽出された意見は, そのまま提示されるのではなく, 類似するものはまとめられ, また, いくつかの観点で分類され, ユーザにとって負担のない情報量で, 提示されることが望ましい. これこそ, 人の意見を尋ねる質問に対し, その解答となるテキストを作成する意見要約に他ならない.

また, PI(Public Involvement) の分野などにおけるアンケート分析では, 回答者の意見を分類, 整理した結果をレポートとしてまとめる作業が現在人手で行なわれている. 図 2 に, 想定される出力結果の例を示す. したがって, 同様に, この分類, 整理の自動化あるいは半自動化が望まれている.

TSC4 意見要約課題では, このような背景から, テキスト集合を与え, その中から意見を抽出し, 抽出した意見を分類, グルーピングし, 要約した上で, テキストとしてまとめ上げるシステムを構築することを目的に, そのための, コーパス作成を行なう. 関連する試みとして, アメリカでは, Wiebe ら [1] が精力的にプロジェクトを進めている.

3.3 意見要約の要素技術

意見要約技術は, 以下のステップから (少なくとも) 構成されていると考えられる.

1. 意見抽出,
2. 分類, 同一性判定,
3. 要約,
4. 出力

意見抽出では, テキスト中から, 意見が記述された部分を抽出する. このステップに関しては, 現在多くの研究が活発になされている (たとえば, [4]). 同一性判定では, 同一内容と考えられる意見をグルーピングする. 分類ステップでは, 何らかの観点を元に, 抽出した意見を分類する. 要約ステップでは, 分類された意見の中から, 出力すべきものを選択する (たとえば, 特徴的な意見を選択) といった処理を行なう. 最後に, 出力ステップでは, 分類され, 選択された意見を, どういう順番に整列して提示するかを決定したり等して, テキストとして意見要約を出力する.

次節で述べるように, 今回のタスクで作成するコーパスには,

- 意見要約として望ましいテキストはどのようなものか (正解要約),
- 意見要約に含まれるテキスト中の意見はどれか (正解要約と元テキストの対応付け),
- 同一内容と判断される意見はどれとどれか,
- テキスト中のどの部分が意見であるか,

などの情報を含んでいる. そのため, 意見要約の個々の要素技術の開発, 評価にも, このコーパスは利用できると考える.

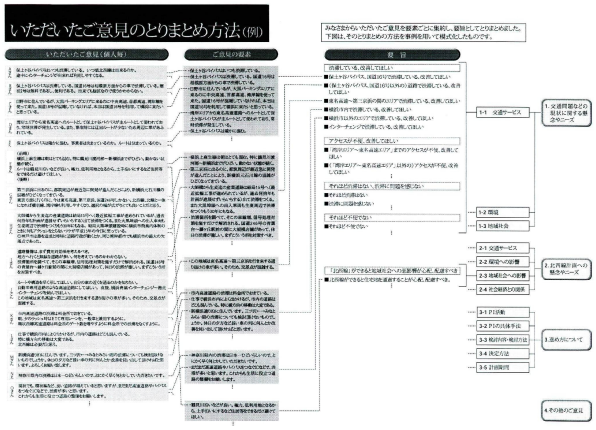
3.4 課題設定とそのための意見要約コーパス

課題の詳細および, そのために作成予定のコーパスは以下の通りである.
課題の詳細:

入力: 質問および, (noisy な) テキスト集合. そして, 出力テキストの長さ.
テキスト集合は, 平均 10 テキスト程度, 事実だけの報道記事も含まれる可能性がある. テキストは, 原則同一トピックに関する 4 紙 (毎日新聞, 読売新聞, 日本経済新聞, 朝日新聞) の 2 年分新聞記事データ (2002 年, 2003 年) からのものとし, 主に社説および読者の投書記事等とする[1].

出力: 指定した長さの意見要約 (テキスト)

[1]なお, これとは別途, 現在我々が収集中の blog(web 日記を含む) からのテキスト集合についても意見要約コーパスの作成を検討している. blog は, 現在約 700 万エントリ以上を収集済みである. 詳細は, `http://www.lr.pi.titech.ac.jp/blogwatcher/` を参照されたい.



(http://www.yokohama-nwline.jp/ref_pi/ より引用)

図 2: 想定される意見の取りまとめ結果

作成予定のコーパス:

> 意見要約コーパス, 要約中の文とテキスト集合中の文の対応づけデータ, 同一内容文の同定も含む.

また, これとは別に, テキスト集合中の各テキストにおいて, 意見部分にタグ付けを行なう.

ただし, 意見要約と言っても, 事実を全く含んではいけないわけではなく, たとえば, 意見の理由や背景となるような客観的事実は, 含める必要があるかもしれない. その辺りに関して, 作成者でゆれがなるべく小さくなるよう, 作業の統制が必要であると考える.

4 おわりに

NTCIR とは独立に運営している TSC4 について，その背景, 狙う点, 課題の内容等について述べた.

多くの研究者の方々の参加を期待したい. ただし, 評価の費用は参加者の自己負担とする. また, この課題の参加者は, 単に課題に参加し, 結果を提出し, 評価結果のフィードバックを受けるという形式ではなく, 我々との共同研究に参加してもらう形式とし, 何らかの貢献を期待される. そのため, 参加を希望されても, お断りする可能性があることをご了解頂きたい. また, 原則として, この課題のデータは, 無条件で公開する予定はない.

TSC の web ページは http://lr-www.pi.titech.ac.jp/tsc/ にあり，また，メイリングリストのアドレスは，tsc-ml@lr.pi.titech.ac.jp である．メイリングリストへの参加希望者は，web ページに記載されている情報を御参照頂きたい.

謝辞

本稿で述べた TSC4 の課題に関しては，これまでのタスク検討会に参加して下さった皆様との議論が大変参考になっています．議論に加わって下さった皆様に感謝いたします.

参考文献

[1] Claire Cardie, Janyce Wiebe, Theresa Wilson, and Diane J. Litman. Low-level annotations and summary representations of opinions for multi-perspective QA. In Mark T. Maybury, editor, *New Directions in Question Answering*, pages 87–98. AAAI Press/MIT Press, 2004.

[2] 加藤恒昭, 松下光範, 平尾努. 動向情報の要約と可視化に関するワークショップの提案. **情報処理学会自然言語処理研究会** *(NL-164-15)*, pages 89–94, 2004.

[3] 立石健二, 福島俊一, 小林のぞみ, 高橋哲朗, 藤田篤, 乾健太郎, 松本裕治. Web 文書集合からの意見情報抽出と着眼点に基づく要約生成. **情報処理学会自然言語処理研究会（***NL-163-1***）**, pages 1–9, 2004.

[4] 鈴木泰裕, 高村大也, 奥村学. Weblog を対象とした評価表現抽出. **人工知能学会セマンティックウェブとオントロジー研究会** *(SW-ONT-A401-02)*, 2004.