2I1-2

テキストの一語への凝縮の試み
Towards Condensation of Text into a Word

鈴木 雅実[*1] 石先 広海[*1] 服部 元[*1] 小野 智弘[*1] 滝嶋 康弘[*1]
Masami SUZUKI Hiromi ISHIZAKI Gen HATTORI Chihiro ONO Yasuhiro TAKISHIMA

[*1] KDDI 研究所
KDDI R&D Laboratories, Inc.

Recently we proposed a new concept of human communication called “Concise Communication” that will arouse empathy between people via condensed expression (named as “concise expression”), like as short keywords or known phrases including proverbs or short verse (haiku etc.). From this viewpoint, we are making attempts of estimating suitable candidates of condensation (or concise abstraction) for given text. In this article, our first trial is described for one of benchmarks as estimating “kanji of the year” with related discussion.

1. はじめに

長い記事や文書などの内容を圧縮するテキスト要約技術の進歩により，短時間で情報を閲覧することが容易となってきている．これをさらに進めると，対象を一つの文や一つのキーワードに，さらには漢字一文字に象徴させることも射程に入る．しかし，テキスト要約技術は論理的な意味のまとまりを保持することに主眼を置くため，通常はパラグラフのようなテキスト生成に留まる．これに対して対象を一言で表すような象徴的な表現を抽出する方法については未だ知られていないが，実際「今年の漢字」の例では，毎年の世相を表わす漢字が投票で選ばれ，ニュース等にも取り上げられほど注目度は高い．この考えを拡張し，年間の世相以外の限定された対象（毎日／特定の地域の話題等）についても短く凝縮された言語表現（コンサイス表現）が適切に選択できれば，それに共感することが可能となる．筆者らはそれをコンサイス・コミュニケーションと名付けている[鈴木 2012]．本稿では，その一環として入力テキストを一語（今回は漢字一文字）に凝縮する方法についての試行と今後の展望を述べる．

2. コンサイス・コミュニケーションの一環として

2.1 対象の凝縮による直観的な理解

冒頭に述べたようなテキストの要約ではなく，一言で直観的に把握できるようなコンサイス表現は，その名の通り対象の特徴を凝縮することにより，論理的ではあっても冗長な説明文と比較して強いイメージ喚起力を持つものと考えられる．そのようなコンサイス表現を通じた高感度なコミュニケーションを歴史・文化的な営みの中で人間は実践しており，それを ICT 的にさらに促進する余地があり得るであろう．コンサイス表現のタイプとしては，語レベル（漢字や熟語など）のものと句または文レベルのものに分類できる．後者には，一般によく知られている諺や日本文化の中で特に発達を遂げた短詩（短歌・俳句・川柳など）を含むものとする．形式は様々であるが，コンサイス表現がどれほど共感を獲得できるかについては，文脈や状況等の外的条件の他に言語表現自体が持つ種々のレトリック的な特性が寄与するものと考えられる．この点については後述する．

2.2 コミュニケーションの活性化

前項に記したような対象を凝縮したコンサイス表現で直観的に把握することが ICT 的な手段で支援できるようになれば，それを通じた共感を生み出すことができる．例えば，オンライン・コミュニティで別々に書き込んだ内容を凝縮した結果に共通点があれば，それを各人に通知することにより，さりげない気づきを通じてコミュニケーションの活性化に寄与し得る．ここで重要なのは，単に共通のキーワードが用いられたということではなく，発言内容を凝縮したコンサイス表現を媒介として，結びつく可能性を示すことにある．

連絡先：鈴木 雅実（KDDI 研究所），msuzuki@kddilabs.jp

3. コンサイス表現への凝縮

3.1 凝縮する対象と凝縮のタイプ

テキストだけでなくマルチメディア的なコンテンツ全般をも対象とした凝縮を将来の目標と考えているが，その基本はテキスト的な手がかりから凝縮を行うことである．例としては，ある個人の一日分のツイートや，限定された地域／コミュニティ内での特定の話題に関する発信情報など，時空間的に範囲を絞った対象を凝縮することを目指す．凝縮結果としてのコンサイス表現には 2.1 に記したような種々のタイプがあり，理想的には対象やユーザの属性等に応じて適切なものを選択することが望ましい．

3.2 漢字一文字への凝縮の試行

前項に記したように，対象に応じてそれを凝縮した種々のコンサイス表現を示すことにより納得性の高い共通認識を得る試みの検証を行う上で，まず実際に世間的に認知されており，ベンチマークとなり得る事例に倣うこととした．それは，一年を振り返って，その年を漢字一文字で表すとしたら？というイベントである [*1]．数万～数十万人規模の投票により「今年の漢字」が決定されるが，上位 20 位までは漢検の Web サイトに掲載されている．多数により選択された漢字の傾向について定性的に分析した結果，次のような傾向が見られた[Suzuki 2012]．

a) 上位の漢字の多くはニュース記事等に高頻度で出現する．
　新聞記事 DB に収録された主要記事の一覧より
b) 複数の事象と結び付くような漢字は選ばれ易い．
　例）金：金メダル／金字塔／税金・・・
　＝直接または間接に複数の概念と結びつくような漢字
c) 人間の感情（情緒）との結び付きがある漢字は選ばれ易い．
　例）愛・悠・笑・・・ （有名人の名前の一部の場合も）

[*1] 今年の漢字，日本漢字能力検定協会（漢検）
http://www.kanken.or.jp/project/edification/years_kanji.html

そこで，まず年間のニュース記事ダイジェスト 7 年分（2006 年～2012 年，各年平均で 2000 件弱の記事数）*[2] から抽出した話題語を構成する漢字の出現頻度によるランキングを導いた．その手順の概略は次の通りである．

(1) 年間のニュース記事を形態素解析した結果より，tf･idf 基準で話題語候補を抽出．
(2) 各年の話題語候補を比較して，特にその年に突出して（平均よりも 1.7 倍以上）頻度の高い語を話題語とする
(3) 抽出した話題語を構成する漢字の頻度（df）により年別の漢字ランキングを作成．

以上のような順位推定結果と投票による順位を次のように比較した．まず比較対象は，2006 年～2012 年の「今年の漢字」への投票による上位 20 漢字と，上記のようなニュース記事に基づく単純推定結果（ベースライン）のうち投票結果と重複したもの（年毎の平均で約 3 漢字）を除く上位 20 漢字であり，両者ともものべ数は 140 である．

ここで，上述したような観察に基づき，漢字そのものの特性が何らかの形で人の判断に影響を与え，実際に目に触れるような話題語（構成漢字）の頻度とは様相が異なる投票結果に結びついたものと考えてよいと思われる．そこで，個々の漢字の持つ内在的な特性を次の 3 つの観点から分析し，上記の比較対象とした漢字群の間で対照させた．その結果を表 1 に示す．また表 2 は，2011 年の場合を例に取り両者を対照させたものである．

表 1 「今年の漢字」に見るコンサイス定量化指標値の比較

評価指標	値	観察結果
評価性	[＋]／[－]／[±]	[＋]または [－] の評価極性を持つ漢字の割合が，投票結果では 78%（+／－の割合はほぼ 6:4）に達したのに対し，推定結果では 32%に留っており，対照的．
情緒関連性	0～3 の 4 段階	1 漢字当たりの平均値は，投票結果の 1.10 に対して，それに含まれない推定結果の方は 0.37 で，顕著な差が見られた．
多義性	1～3 の 3 段階	1 漢字当たりの平均値は，投票結果の 1.51 に対して，それに含まれない推定結果の方は 1.71 で，両者の間に差は見られなかった．

表 2 具体例による比較対照（2011 年の例）

ランキング	10 位までにランクされた漢字
投票による集計結果	絆･災･震･波･助･復･協･支･命･力
上記の推定結果*	大･電･日･発･島･福･子･者･被･故

* 投票結果の 20 位までと一致した 6 漢字を除く

これらの比較結果は，上記の定性的な観察に基づく仮説をある程度裏付けるものと言える．そこで，この分析結果に基づいて各漢字に対応する重み付けを反映させることにより，ランキング順位の調整を図ることが可能となる．ただし，情緒関連性が高い漢字のうち幾つか（例えば 2011 年の第 1 位「絆」など）はニュース記事中にはほとんど出現せず，そのような候補漢字を推定するためには別の方法を導入する必要がある．また，今回漢字そのものの多義性については明確な差が出ていないが，実際の事象との結び付きの観点等から再検討が必要と思われる．

*[2] 参照元：YOMIURI ONLINE ニュース月録
http://www.yomiuri.co.jp/getsuroku/

4. 多様なコンサイス表現への凝縮に向けて

4.1 関連研究

これまでテキストの自動要約技術について多くの研究事例が報告され，一部は実用化の段階に到達している．その主眼となる方法は，テキスト中の重要語句を抽出し，それらの語句を含む文を繋ぎ合わせて要約テキストを生成することである．最近では，複数のテキストを要約するような高度な目標を掲げる事例もあるが，いずれにせよ論理的なテキストとして再構成するものが大半である．具体的には，内容網羅性や可読性の高い要約の生成を目的としている場合が多い（[西川 2010]など）．一方，テキストに対して各種のアノテーションを施し，それにより検索を容易とする仕組みの提案も行われている．テキストに含まれる語の分布や様々な形式的な特徴から，種々の属性を付与することがその目的で，アノテーションされた情報自体を人間が参照することは少ない．最近の研究事例としては，[富浦 2011]などが挙げられる．本研究では，従来のテキスト要約やアノテーションとは異なり，対象全体を特徴付けるような凝縮表現（漢字一文字など）に対応させることにより，それを見たユーザが直観的に理解･納得できるような共通認識を得ることを目的としている．

4.2 課題と展望

話題語の抽出を経て対象（年間の世相など）を漢字一文字に凝縮するためには，対象テキスト中の出現頻度の他に，凝縮表現（漢字）に内在する特性を反映させることが有力な手段となることが確認された．このことは漢字や熟語等ではなく，句レベルのコンサイス表現に凝縮する場合にも適用可能であり，現在そのような表現を一定数収集して，認知言語学的な観点からレトリック特性等を含む定量化指標を十数種設定し，値を同定する作業を進めている．また，これまで句レベル以上のテキストに対して，その中には直接含まれない感情属性等（100 程度のオーダー）をコンサイス･タグとして付与する方法を考案し，主観評価との比較も実施している [鈴木 2013].

コンサイス表現らしさ，すなわち対象テキストを凝縮した結果として共感／共通認識を得やすい効果を持つ程度については，上記のコンサイス表現の定量化指標とコンサイス･タグ（合せてコンサイス･ラベルと命名）と関係付けられる．その（半）自動獲得を目指しつつ，2 章で紹介したコンサイス･コミュニケーションの支援に結びつく種々の対象の凝縮とその活用に向けて研究を進めることが直近の課題である．最後に，本研究の遂行にご協力頂いている関西大学の鍋島弘治朗教授に謝意を表したい．

参考文献

[西川 2010] 西川仁，長谷川隆明，松尾義博，菊井玄一郎: 文の内容性と連接性を目的関数とする複数の評価文書の要約，言語処理学会第 16 回年次大会，B1-1, pp.39-4, 2010.
[冨浦 2011] 冨浦洋一･石田栄美: 学術論文検索の高度化のための論文アブストラクトのアノテーション，テキストアノテーションワークショップ･コンテスト（国立情報学研究所），2011.
[鈴木 2012] 鈴木 雅実，服部 元，小野 智弘: コンサイス･コミュニケーションとその支援に向けて，人工知能学会第 26 回全国大会，1N-2-OS-1b-4, 2012.
[Suzuki 2012] M. Suzuki, G. Hattori, C. Ono: Towards Concise Review of Online Communication with Reference to “Kanji of the Year”, *Culture and Computing 2012*, Hanzhou, 2012.
[鈴木 2013] 鈴木 雅実，石先 広海，服部 元，小野 智弘，鍋島弘治朗: テキストへのコンサイス･タグ付与とその主観評価，第 42 回ことば工学研究会，pp.9-15, 2013.