2F4-OS-01a-7

# テキストに対応するコンサイス表現の選択について

## Towards Selecting Appropriate Concise Expressions for Given Text

鈴木 雅実[*1] 鍋島 弘治朗[*2] 石先 広海[*1] 服部 元[*1] 滝嶋 康弘[*1]
Masami SUZUKI Kojiro NABESHIMA Hiromi ISHIZAKI Gen HATTORI Yasuhiro TAKISHIMA

[*1] KDDI 研究所 KDDI R&D Laboratories, Inc.
[*2] 関西大学 Kansai University

This manuscript describes our current approach to text condensation: from input news article into certain concise expressions like as popular proverbs. Recent trials showed that there will need some drastic improvements on estimating really important associative words, shared between input text and concise expression candidates.

## 1. はじめに

筆者らは「コンサイス・コミュニケーション」と名付けた共感のスタイルを提案，その特質と近未来的な支援について研究開発を行っている[鈴木 2012]. 今回は 2013 年度に発表した年間の代表的なニュース記事群を漢字一文字に凝縮する試行[鈴木 2013b]に続いて，新聞記事等のテキストを入力として，それに対応するコンサイス表現（ことわざ等）を選択するための手法考案と実験経過および今後の課題等について報告する.

## 2. 背景と目的

### 2.1 これまでの経緯

2012 年度の本大会において初めて提唱したコンサイス・コミュニケーションとは，対象を簡潔に言い表すような言語表現の持つイメージ喚起力を媒介として共感を得ることである. すなわち，テキスト内容を人間が読んだ際に，その示唆内容を直観できるコンサイス表現（例「渡に船」）を提示すれば，それを通じて理解が促進されるようになるものと考えられる.

人間の場合テキスト理解から直観的に連想するようなコンサイス表現を，元テキストから自動推定することはチャレンジングであるが，コンサイス表現として想定する各種の名言やことわざのリストを一定数用意して，その中から該当するものを順位付きで選択するような仕組みを考案・試行することを当面の課題とした.

### 2.2 目標

そこで，初期目標として，人間が幾つかのコンサイス表現を連想可能な入力テキスト（ニュース記事：下記の例を参照）に対して，100 種程度のことわざ表現の中から適切と思われるような候補（複数）を順位づけた場合に，上位 10 位以内に該当するものが入る割合を 6 割程度にすることをねらいとしている. 直近の試行ではその半数程度に留まっており，改善の余地が大であるが，その現状のアプローチと課題について以下に述べる.

入力例 【アンジェリーナ・ジョリーのおば，乳がんで亡くなる

米女優アンジェリーナ・ジョリーさん(37)の母方の叔母デビー・マーティンさんが 2013 年 5 月 26 日、米カリフォルニア州の病院で乳がんのため 61 歳で亡くなった. ジョリーさんは 14 日に乳がんのリスクを高める遺伝子異変が見つかり、両乳房の切除・再建手術を受けたことを公表. マーティンさんは 04 年に乳がんと診断された後に，ジョリーと同様の遺伝子異変が見つかっていたという. 】

対応するコンサイス表現の例 【転ばぬ先の杖】

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

## 3. 試みたアプローチと実験経過

筆者らの考察[鈴木 2013a]によれば，任意のテキストは，その中に含まれる重要語との連想関連性の強い語（コンサイス・タグ）によって特徴づけられる. ここでコンサイス・タグとは，テキストの意味特徴を表す語として，階層的な意味体系（例えば国立国語研究所の「分類語彙表」）において，抽象的（上位）過ぎず，具体的（下位）過ぎない中間層の語として予めリストに登録した語彙である. タグの種別としては，感情（50 個）およびテーマ（250 個）を初期設定した. テキスト中の重要語がリスト中の各語と一定規模の参照用コーパス内で共起する頻度（相対共起頻度）を積算することにより，コンサイス・タグを導出可能である.

そこで，同様なタグが付与されたテキスト同士は共通な特徴を持つとの仮定の下に. 次のような手順の適用を試みた. なお，入力テキスト（ニュース記事等）に対して，該当するコンサイス表現の側では，ことわざ表現のように短い語の組合せが比喩的に用いられる場合も多いため（例：「猿も木から落ちる」），直接の共起では適切なタグ語が導出されない可能性が高い. そこでコーパス中でことわざ表現が引用されるような形で現れるテキスト等も参考に，現段階では人手でコンサイス・タグを付与している.

### 3.1 処理の流れ

全体の処理過程の概略は図1の通りである. 以下その流れを追って順次説明する.

1) 入力テキストの形態素解析に
重要語の候補としてテキストから自立語列を抽出
2) 重要語の選定
tf・idf 基準で重要語を絞り込む
この際に，テキストの構造情報（パラグラフなど）を利用して，ニュース等の段落の最初の文に重みを付けることも考慮する.
3) 連想語としてのコンサイス・タグの導出
重要語とタグ語が外部コーパス（10 万件のブログ）において共起する度合い（相対共起頻度）を相互情報量基準で求め，それを各タグ語について積算（語数で正規化）した値で順位付けすることにより，最大 5 個程度のコンサイス・タグとする.
4) コンサイス表現に対するコンサイス・タグの付与
現在は人手で付与（上述参照）. 事前に DB 登録可能.
5)入力テキストおよびコンサイス表現間でのタグの突合せ
各 3～5 個のコンサイス・タグの組合せの類似度を算出し，入力テキストに対する類似度順にソート・ランキング出力する. 類似度計算については様々な案が想定されるが，後述するようにコーパスを用いてタグ語同士の間の距離を算出，対となる語間の距離の和の最小値により求めている.

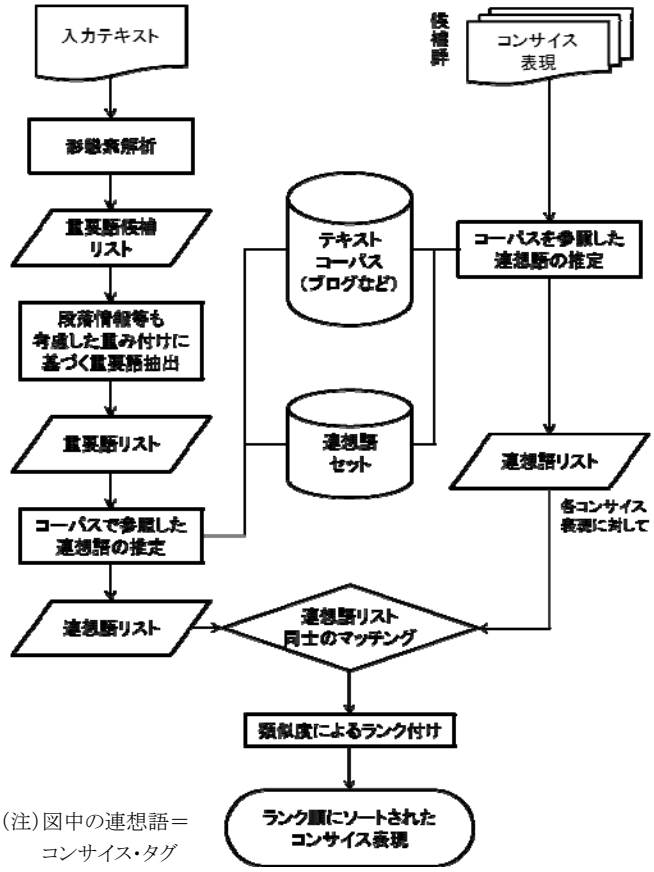


図1 テキストに対応するコンサイス表現の選択フロー概略

### 3.2 実験結果

ニュース記事 50 件を対象に，入力テキストからの重要語抽出を経て，コーパスを用いたコンサイス・タグの推定を行った．一方，ことわざ表現についても人手でコンサイス・タグを付与することにより，入力テキスト側のコンサイス・タグ（連想語）リストと類似度の高い順にコンサイス表現をランキング出力し，その結果を検討した．2.2 節に示した例についての処理の途中経過は次の通りである．なお，コンサイス・タグ同士の類似度については，タグ語と共起する語の重なり分布により，タグ全体を階層的にクラスタリングした結果の目視により，概念的に近いと思われるタグ語が局所的に近傍に位置することが確認できたことから，それに基づきタグの組合せ間の類似度を計算した．

入力テキスト：アンジェリーナ・ジョリーのおば，乳がんで亡くなる
・・・
重要語抽出例：女優／叔母／病院／乳がん／リスク／遺伝子／異変／乳房／切除／手術／診断
コンサイス・タグ（連想語）の推定例： C0 検査／予防／病気
コンサイス表現候補のコンサイス・タグ付与例
「転ばぬ先の杖」 C1 後悔／怪我／予防
「渡りに船」 C2 対策／運／感謝
「玉に瑕」 C3 価値／失格／評価
マッチング順位（C0 と C1/C2/C3/・・・の間の類似度計算による）
C0 < C1 < C2 < C3 ・・・ （C0 に最も近い候補は C1）

この処理をサンプル・ニュース記事 50 件について 100 種のことわざ表現の中から，コンサイス・タグ（連想語）セットとのマッチングにより選択した．各テキストに正解コンサイス表現として与えた 1～複数個のコンサイス表現（ことわざ）が，10 位以内にランクされたケースは，約 27%であった．

## 4. 課題と今後の展望

### 4.1 問題点

3 章に述べた試行の結果はまだ満足できるものではなく．様々な観点から改善すべき段階にあることが明らかとなった．そこで解決すべき課題について検討・考察する．最重要と考えている問題は，入力テキストに対して適切な（質の高い）コンサイス・タグ（連想語）の組合せを推定する方法の探求である．関係する要因は次のように捉えることができる．

1）重要語の抽出方法

コンサイス・タグ推定の元となるテキスト中の重要語は，頻度情報以外は同等な重みで扱っている．一方，大量のテキストを分類する手法（LDA など）を用いて入力テキスト群を分類した場合，各クラスタへの帰属に影響する度合の強い／弱い語が存在する．その影響度を考慮して連想語としてのコンサイス・タグの推定を行う方が，テキストの特徴をより反映されると考えられる．

2) コンサイス・タグとしての登録語の取捨選択

現在用いているコンサイス・タグ一覧中の語は，3 章に述べた考え方で初期設定したものである．ところが，入力テキストから導かれたコンサイス・タグの導出結果を見ると，ノイズ的なものを除けば，次のような分布が存在している．

A 群: テキスト内容全体の意味理解に直結するような連想語（3 章の例では「予防」）

B 群:テキストの一部と類推関係を持つに過ぎない連想語 タグ語そのものが意味的に曖昧（例「説明」）または多義的（例「自然」）なものも含まれる．

そこで，極力 A 群の占める割合を高めるような取捨選択を経たコンサイス・タグのセットを用いるとともに，テキスト中の重要語の持つ固有の分類寄与度を考慮したコンサイス・タグの推定精度向上を図りつつある．

### 4.2 今後の展望

現在は，テキストの意味的な特徴をテキスト中の語の分布に基づく連想語の組合せに還元しているが，そのアプローチによる限界以前の問題点が明らかとなってきた．これに対して当然ながら，テキストの構造的な側面，すなわちストーリー展開やレトリック等の認知言語学的な観点からの分析も不可欠である．コンサイス表現についても，例えばことわざがその本意として理解されるための要件などを注意深く記述して，入力テキスト側との類似性を突き合わせる試みも必要である（実践中）．さらには，テキスト中では全く言及されない前提知識をも動員して能動的な理解を行う人間の思考活動への接近も視野に入れつつ，テキストをコンサイス表現に凝縮するというテーマ探求の奥深さをあらためて再確認したい．

### 参考文献

[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.

[鈴木 2013a] 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.

[鈴木 2013b] 鈴木 雅実，石先 広海，服部 元，小野 智弘，滝嶋 康弘: テキストの一語への凝縮の試み，人工知能学会第 27 回全国大会，2I1-2，2013.