7C-5

# Google API を用いた関連文生成の一手法

望月 英樹　阿部 裕行　伊藤 一成　Martin J. DÜRST
青山学院大学理工学部

## 1　はじめに

情報化社会の現在，一般の利用者が書き手となっているコンテンツが多様化してきている．多大な情報から効率よく目的に即した情報の入手する方法が提案されているが，入手した情報の理解に関する考案も必要となる．人間は書き言葉より，話し言葉の方が親しみやすく容易に理解できる．そこで，書き言葉を話し言葉に変換し，情報理解の負荷を低減させる方法が提案されている [1].

また，対話文形式による，類似コンテンツに対する差異情報の提供 [2] や Web コンテンツに対して，漫才の要素を取り入れて理解を容易にする方法が提案されている [3]．これらの手法は，公開されているコンテンツを対象として情報を収集し，文章を生成する方法やユーザが興味を持つ情報だけを対象とした方法である．本稿の目的は，一般ユーザが日常文章を書くことで，偏りのない様々な分野の関連情報を入手し，発想の拡大や知識の向上が期待できる手法の提案である．電子メール，ブログやソーシャルネットに代表される一般ユーザが日常的に作成した文章を対象とし，関連文生成を行う．

## 2　提案する関連文生成手法

自然言語処理や Web サービスが提供する情報を用いた処理を加え，関連文を生成する．また，単に機械処理結果を見せるのではなく，通常の会話のように親しみやすく，自然な形態を維持した文章を生成する．

### 2.1　基となる文章からの重要語抽出

はじめに，タイトルと本文からなるフォームを用意する．フォームの入力内容は，RDF 形式で保存される．RDF の生成及び処理には，我々が開発したライブラリ [4] を用いる．タイトルは `dc:title` に，本文は `dc:description` に保存される．我々のライブラリの特徴として，`dc:description` の内部は単に通常のリテラルで記述しておくのではなく，言語情報を GDA [5] により付与した形式の XML 構造をそのまま内包して保存する．GDA は，文法機能（主語，目的語，間接目的語），主題役割（動作主，非動作主，受益者など），修辞関係（理由，結果など）や照応関係を表すことができるタグセットである．GDA タグ付けされているテキストの文法情報から名詞を割り出し，tf–idf 法を用いて重要語を抽出する．

### 2.2　Google API を用いた情報獲得

文章生成する際の関連情報収集は Google API [6] を利用する．Google API は，Java や Perl など各々の開発環境の中で Web ページの検索やスペルミスチェックを行う仕組み，キャッシュページの取り出しなどを提供するサービスである．抽出した重要語を用いて Google 検索を行い，情報を収集する．この際，一語だけでの検索は，関連度の低い情報も比較的多く出てくるので，発想の拡大に繋げられる．一方，重要語群から上位数個を選び出して and 検索を行えば，関連度の高い情報のみが獲得できる．検索結果から得られた各々の情報も，入力内容と同じ RDF 形式で保存される．検索結果のタイトルは `dc:title`，要約文は `dc:description`，URL は `dc:url` に保存される．さらに，テキストと同様に `dc:description` と `dc:title` に GDA タグ付けを行い，文法情報を付加する．

### 2.3　文章生成

重要語が `dc:description` 内のある一文に含まれる場合は，一文を取り出してそのまま関連文として利用する．また，`dc:title` に含まれる場合や `dc:description` に単独で存在する場合は，タイトルを利用した文や，`description` や `URL` を用いた関連文の生成を行う．文章生成には，雛形となるフレームを複数用意しておく．フレームの一例を図 1 の上部に示す．変数に相当する部分は“ `$` ”で囲まれる．変数の名前には，フォームのタイトルに入力された文章を示す `$title$` やフォームの入力文から抽出される複数の重要語を並立助詞“ と ”で連結させた文に変換する `$keyword$` がある．重要語のある特定の単語一つのみを抽出したい場合は，`$keyword_n$`（`n` は番号）という形式で表現する．重要語の列は順序付けされており，重要度の大きい重要

---

**A technique for the generation of related sentences using the Google API**

Hideki MOCHIZUKI, Hiroyuki ABE, Kazunari ITO, and Martin J. DÜRST
College of Science and Engineering, Aoyama Gakuin University, 5–10–1 Fuchinobe, Sagamihara, Kanagawa 229–8558, Japan
{mochi, hiroyuki}@sw.it.aoyama.ac.jp,
{kaz, duerst}@it.aoyama.ac.jp

```
[ フレーム ]
$title$ には $keyword$ がよくでてきましたね.
$keyword_1$ のことで何か知ってますか？
$keyword_1$ といえば,
$keyword_1_googlerank_1_sentence_1$
という話があります.
この話をもっと詳しく知るために,
$keyword_1_googlerank_1_url$
をみていろいろ調べてみましょう.

[ 生成文 ]
未納問題には税金と所得と国民がよくでてきましたね.
税金のことで何か知ってますか？
税金といえば,
アフィリエイトで稼ぐと税金がかかります
という話があります.
この話をもっと詳しく知るために,
ここ
をみていろいろ調べてみましょう.
```

図 1: フレームと生成文の例

語ほど番号が小さくなる．また，$googlerank_n$（n は番号）は Google 検索の n 番目の結果を示す．これらの変数は，“ _ ”による連結で拡張が可能である．例えば，$keyword_1_googlerank_2_title$ （図 2 参照）は，入力された文章の重要語の 1 番目の単語で Google 検索を行った第 2 位の結果のタイトルを意味する．雛形

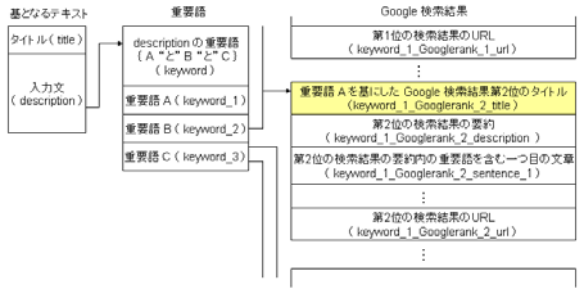


図 2: 変数の例

となるフォームにある“ $ ”で囲まれた変数に，単語や文章を当てはめて文章を生成する．同じフレームばかり用いると表現が単調となってしまう．複数のフレームを用意し，用途に合わせた関連文の生成を行う．例えば，ブログでは，専門家風，女子高生風，友達風といった様々な口調のフレームを取り入れ，多くの人が返信したように見せることができる．図 1 の下部に生成された関連文の例を示す．検索結果の情報は，テキストと同様に RDF 形式で保存される．Google 検索結果 1 件分の情報をフォームの入力とみなし，同様の処理を繰り返すことで，興味を持つ情報に対して，更なる関連情報の収集ができる．生成された関連文の表示例を図 3 に示す．

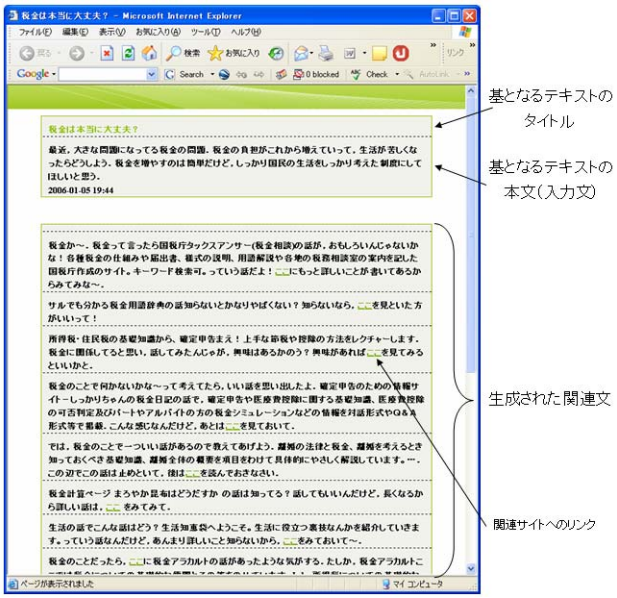


図 3: 生成された関連文の表示例

## 3 まとめと今後の展望

本稿では，Google API を用いた関連文生成について提案を行った．今後は，他の Web サービスの利用も応用として考える．例えば，はてな API による語彙情報 や Amazon API による商品情報などを利用することで，より有用な情報を取得できる．構造が単純であるため，新たな Web サービスの追加や様々な処理の組み合わせが可能となり，多様な文生成ができるだろう．また，今回は人が書いた文章に対して，自然な形式での文章生成で情報提供を行ったが，文章での提供以外に画像や映像による情報提供を加え，より有用な情報となることも考えられる．

## 参考文献

[1] 久保田 秀和, 山下 耕二, 福原 知宏, 西田 豊明: POC caster: インターネットコミュニティのための会話表現を用いた情報提供エージェント，人工知能学会論文誌，Vol. 17，No. 3，pp. 313–321，2002

[2] 灘本明代，田中克己：異メディアコンテンツの差異情報に基づく対話文自動生成，日本データベース学会 Letters Vol. 4, No. 2, pp. 57–60，2005

[3] 蓬莱博哉，灘本明代，田中克己：理解しやすさとユーモアを考慮した Web コンテンツの対話番組変換，日本データベース学会 Letters，Vol. 2, No. 2, pp. 29–32，2003

[4] 阿部 裕行，伊藤 一成，Martin J. Dürst: 自然言語処理と構造マイニングを併用したアノテーションデータの解析方法，第 68 回情報処理学会全国大会，2006

[5] Global Document Annotation (GDA) : http://www.i-content.org/gda/

[6] Google Web APIs : http://www.google.co.jp/apis/