複数ブログ記事からの関連事象を含む概要記事の自動生成
（一部抜粋）

2.2 手段の概要

流行のキーワード（これを「第一キーワード」と呼ぶことにする）を含む複数のブログから 1 文単位で文を引用し，最長 7 文を用いて 1 つの記事を作成する．以降の節では文の引用基準を説明する．

2.2.1 第一文

第一キーワードを含む段落（複数の文からなる）を抜き出し，その名詞の出現回数を数え，$n(w)$とする．ただし，wは名詞であり，形態素解析には JUMAN[5]を用いた．その後，その段落の文ごとに次式にしたがってスコアを定める．

$$score = -\alpha L + \sum_{w \in W} n(w) \quad (1)$$

ここで，Lは文の文字数であり，Wはその文に含まれる名詞を示し，αは係数でありここでは 0.4 とした．このスコアが最も高い文を第一文として引用する．

2.2.2 第二文

第一文とほぼ同様の方法で文にスコアをつける．ただし，$n(key)$を負とする．keyは第一キーワードの単語を示す．ここで，$\alpha = 0.3$とした．

また，任意の文 A と文 B の両方に存在する単語の個数を，文 A および文 B の短い方の単語の個数で割ったものを類似度として定義する．第一文との類似度が閾値(0.9)より低い文の中で最大のスコアを有する文を第二文として引用する．「それがなぜ流行のキーワードであるのか」を説明する文を引用することが，第一文と第二文の引用基準の狙いである．

2.2.3 第三文

まず，2.2.1 節で計算した$n(w)$の大きなwから$M(=7)$個抜き出し第二キーワード群とする．第二キーワード群を含む段落の文の中で，それ以前の引用文（この場合は第一文と第二文）と類似度が閾値(0.8)未満で最も高い文を第三文とする．

2.2.4 第四文

第三文で用いられた第二キーワードを含む段落から新たに$n(w)$を計算し，2.2.2 節と同様の方法で$score$を得る．ここで，$\alpha = 0.35$とした．それ以前の引用文との類似度が閾値(0.8)未満で，スコアの最も高い文を第四文とした．

2.2.5 第五文と第六文

第三文および第四文とほぼ同じ選択基準で引用する．ただし，第三文および第四文に含まれる第二キーワードを持つ文は候補から除外する．

第一キーワードの関連事象を補足するのが，第三文から第六文の狙いである．また，偶数文目ではキーワードとなる単語の頻度をマイナスにしているが，これは直接キーワードが出てこない文を採用しやすくするためである．

2.2.6 第七文目

2.2.1 節と同様の方法で$score$を計算する．第一キーワードを含む段落から「形容詞（形容動詞を含む）」と「判定詞（だ・である等）」の両方を含む文のみを抜き出す．さらにそれ以前の引用文と類似度が閾値(0.8)未満でスコアの最大のものを第七文とする．事象に関する簡潔な感想を得て記事に完結感を出すことが目的である．

2.2.7 記事の棄却

以上の基準により出来上がった記事が「五文未満」もしくは「引用元が三箇所未満」となっていた場合には，記事が生成しづらいキーワードだったものと見なし，記事自体を棄却する．

2.3 引用元の明示

著作者を明示するため，各文の末尾から引用元に対してリンクを張る．気になる文が出てきたときに元の記事を見に行くこともできる．

参考 URL

[1] http://d.hatena.ne.jp/hotkeyword
[2] http://d.hatena.ne.jp/saussure/
[5]http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html