5U-05　係り受け規則に基づく複合語キーワード生成

安藤一秋†，岡田真‡，獅々堀正幹‡，青江順一‡
†香川大学工学部，‡徳島大学工学部

1. はじめに

キーワード自動抽出は，情報検索における索引語の決定[1],[2],[3]やテキスト自動要約[4]での重要個所特定など幅広い分野で利用されている．現在までに提案されたキーワード抽出法では，文書の内容を適確に表現する語は必ずその文書中に出現する[1]という仮定の下に，文書内に存在する単語自身をキーワードとして抽出していた．しかし，この手法では，キーワードとなるべき語が原文中に存在せず，キーワードの構成単語が分離して存在する場合や，文書内容から推論・導出されるような抽象的な語（主題語）である場合に対処できない[2],[3]．本稿では，人間が離れた単語を合成して複合語キーワードを生成する点[2],[3]に着目し，係り受け規則に基づく複合語キーワード生成について報告する．

2. 係り受け規則に基づく複合名詞生成

横山ら[5]は，34種の意味素性を定義し，それらに対応する2字漢字を組み合わせることにより，複合語における意味素性間の接続性を調べる実験を行った．また，宮崎[6]は，単語間の係り受け関係を調べることにより，高精度の複合語自動分割を実現した．

本研究では，横山と宮崎らの規則を参考に，係り受け関係を利用した複合語生成規則を50種構築した．但し，今回の実験では，簡易な処理での複合語生成を実現するために，意味素性やシソーラスは用いない．以下に，規則の例を示す．

例1）AをBする　→　AB
例2）AするB　→　AB
例3）Aに対するB　→　AB
例4）AするためのB　→　AB
例5）Aに関するB　→　AB

但し，A, Bは，普通名詞の1個以上連続を意味する．

上記の規則と学術論文のアブストラクト50編を用いて，予備実験を行ったところ，生成された複合語の約50%がキーワードとして不適切であると判断された．この中には，“仮定場合”などの複合語として成立しないものが多数含まれていたので，これらの生成を抑制するために，何らかの制約を導入する必要がある．そこで，生成規則に対して，3つの制約を設けた．

1)　冗長語・不要語による制約

“提案”，“報告”など論文特有の表現，“場合”，“有無”などの複合語の構成要素になり難い語を人手で収集した．これらの語が複合語の構成要素として照合された場合は，複合語を生成しない．

2)　サ変名詞による制約

規則とのマッチングによる生成では，係り受け関係が十分に把握できない場合がある．従って，生成される複合語ができるだけ項関係をもつような制約を導入する．ここで，項関係とは，名詞が述語（動名詞，形容名詞）に対して，主語や目的語の文法関係を結ぶ場合を意味する．本稿では，特に，複合語中の述語をサ変名詞に限定し，サ変名詞を全く含まない場合は，規則を発火させない．

3)　係り受けに対する制約

「人間と計算機を比較する→計算機比較」のように規則が並列構造の一部と照合し，そこから複合語が生成されると文意と異なる語が生成されるかもしれない．そこで，規則がマッチする前後の品詞に着目し，係り先の曖昧性を生じる「と」，「や」，「の」などの助詞が存在する場合は，規則を発火させない．

3. 共通語を考慮した重要度計算

上記の規則によって生成された複合語の中からキーワードを抽出するために，重要度を付与する．文書中に「情報検索」，「検索質問」，「検索システム」など，共通する単語を含む複合語がいくつか存在する場合は，お互いに何らかの関連性をもち，文中での重要性を高めていると考えられる[7]．

本稿では，生成された複合語（キーワード候補）と文中の語との関連度を共通形態素の割合で表し，関連度に応じた重みをキーワード候補へ付加することで，キーワード性を高める方法をとる．

文書中に出現する語tとキーワード候補kの関連度を$R(t, k)$とするとき，kの重要度$W(k)$を以下の式で表す．

$$W(k) = \sum \{N(t) \times R(t,k)\}$$

$$R(t,k) = C(t,k) / L(t)$$

$N(t)$：tの正規化頻度
$C(t, k)$：kに対するtの共通形態素数
$L(t)$：tの形態素数

N(情報)=0.3，N(抽出)=0.2，N(情報検索)=0.15，N(キーワード抽出)=0.1と仮定する．このとき，キーワード候補「情報抽出」に対する共通形態素数は，

C(情報，情報抽出)=1；　C(抽出，情報抽出)=1；
C(情報検索，情報抽出)=1

Compound Keyword Generation Based on Dependency Relation
† Kazuaki Ando, Faculty of Engineering, Kagawa University, Takamatsu-shi, 761-0396 Japan
†† Makoto Okada, Masami Shishibori, Jun-ichi Aoe, Faculty of Engineering, Tokushima University, Tokushima-shi, 770-8506 Japan

となる．また，「情報抽出」に対する関連度Rは，

R(情報，情報抽出)=1/1=1；

R(抽出，情報抽出)=1/1=1；

R(情報検索，情報抽出)=1/2=0.5

となる．以上により，重要度Sは以下のようになる．

S(情報抽出)={N(情報)×R(情報，情報抽出)}＋
{N(抽出)×R(抽出，情報抽出)}＋
{N(情報検索)×R(情報検索，情報抽出)}
=0.3×1＋0.2×1＋0.15×0.5＋0.1×0.5=0.58

この重要度計算により，キーワード候補の共通語を含む語の重みが反映される．

次に，この重要度に対して補正を行うことで，よりキーワード性の高い語を抽出する．

「を用いた」，「に基づく」などの手がかり表現の前後には重要な語が出現することが多い[1]．また，表層情報のみを用いた方法では，「A の B」から AB を生成するより，「A を B する」から AB を生成する方が適切なキーワードを生成しやすいことを予備実験により確認した．これらを考慮し，キーワードkの補正重要度として$W'(k)=\alpha_i \times W(k)$を用いる．ここで，補正定数$\alpha_i$（1? α_i ?2）予備実験により決定した．

4. 評価

4. 1 実験データと評価方法

実験には，情報検索システム評価用テストコレクション[8]から自然・音声言語処理に関するデータファイルを抜粋し，タイトルと抄録部分（65 ファイル，総容量 37.5KB）を用いた．このデータは予備実験のものとは異なる．そして，5 人の被験者が上記データファイル中のタイトルと抄録を読み，生成されたキーワードの妥当性を次の 4 段階で評価した．

A：キーワードとして適切；　B：違和感がない；
C：少し違和感がある；　　D：不適切

本実験では，3 人以上が A 及び B 評価した複合語をキーワードとして妥当であると判断する．また，生成された複合語の数に対するキーワード数の割合を精度と定義する．

$$\text{精度 (\%)} = \frac{\text{キーワード数}}{\text{生成された総複合語数}} \times 100$$

4.2 係り受けに基づく規則の評価

本実験では，生成規則 50 個を用いた．また，2.で行った予備実験より，76 個の冗長語・不要語を人手で決定した．

実験の結果，3 つの制約を適用しない場合の精度は 38.2%であったが，全てを適用した場合は 61.2%に向上した．その内訳としては，係り受けに対する制約により 12.1%，冗長語・不要語による制約により 7.1%，サ変名詞による制約により 3.8%精度が向上した．

次に，キーワード候補を上位 15 個まで出力した際の精度を図 1 に示す．図 1 より，共通語を考慮した重要度付けによって最大で 4.9%（上位 3 個），さらに，重

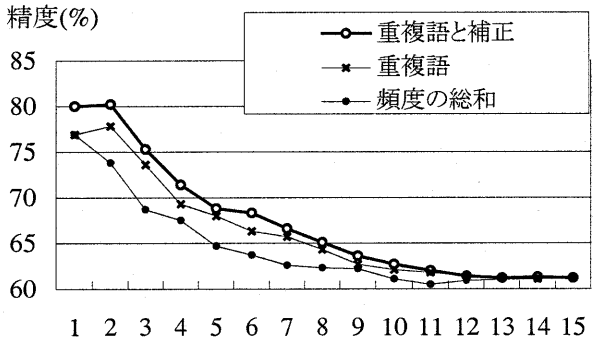

図 1 生成キーワードに対する精度

要度を補正することで最大 6.6%（上位 2 個）精度が向上することが確認できた．

以上から，3 つの制約および共通語，重要度補正が有効であることが分かる．

また，C や D の評価を下されたキーワード候補を分析した結果，複合語として不適切なもの，複合語が長すぎて違和感があるものがほとんどであった．複合語として不適切なものに関しては，意味素性の利用が必要であると考えられる．また，短縮に関しては，複合語の分割処理を利用し，冗長な構成語を削除する必要がある．いずれの場合も，精度向上を実現するには，今回用いた情報以外に，意味素性やシソーラスの利用が必要である．

5. まとめ

本稿では，人間が離れた文字列を合成してキーワードを生成する点に着目し，係り受け関係に基づく複合語の生成実験を行った．実験により，3 つの制約や共通語を考慮した重要度付けの有効性を確認した．しかし，これらの情報だけでは，これ以上精度を向上させることは難しいと考えられる．今後は，意味素性の利用や複合語の短縮方法を考案と検索精度に与える影響を調べる予定である．

参考文献

[1] 諸橋正幸，“自動索引付け研究の動向”，情報処理，Vol.25, No.9，pp.918-925，Sep.1984.
[2] 原正巳，中島浩之，木谷強，” テキストのフォーマットと単語の範囲内重要度を利用したキーワード抽出”，情処学論，Vol.38，No.2，pp.299-pp.309，1997.
[3] 永田昌明，木本晴夫，“重要概念抽出に基づく新聞記事からのキーワード生成”，第 37 回情処学全大，pp.1030-1031，1988.
[4] 奥村学，難波英嗣，“テキスト自動要約に関する研究動向”，自然言語処理，Vol.6，No.6，pp.1-26，1999.
[5] 横山昌一，佐久間一弘，“意味素性を用いた複合名詞の生成による分析”，計量国語学，Vol.20，No.7，pp.304-314，1996.
[6] 宮崎正弘，“係り受け解析を用いた複合語の自動分割”，情処学論，Vol.25，No.6，pp.970-979，1984.
[7] 亀田雅之，“擬似キーワード相関法による重要キーワードと重要文の抽出”，言語処理学会第 2 回年大，pp.97-100，1996.
[8] NACSIS Test Collection for IR Systems，学術情報センター，1999.