自 然 言 語 処 理 98−10
（1993. 11. 19）

語の意味分類の出現傾向を考慮したキーワード抽出の試み

鈴木 斎 *　　増山 繁 *　　内藤 昭三 **

{suzuki@smlab., masuyama@}tutkie.tut.ac.jp,　naito@atom.ntt.jp

* 豊橋技術科学大学知識情報工学系
**NTT 基礎研究所

〒441 愛知県豊橋市天伯町字雲雀ヶ丘1 − 1
豊橋技術科学大学知識情報工学系 増山研究室
Tel. 0532 − 47 − 0111 (ext. 893)

概要

本稿では，シソーラスを使って，テキスト中の語の意味分類の出現傾向を考慮することによるキーワード抽出法を提案する．本方法では，シソーラスに基づき語の意味分類の出現傾向を調べることにより，まず各文，各段落中の話題を推定し，その結果を用いてテキスト全体のキーワードの抽出を行う．計算機実験を行ない，従来の方法と比較し，キーワード抽出における本手法の有効性を確認した．

Examination of Keyword Extraction using Thesaurus in Japanese text.

Hitoshi SUZUKI*, Shigeru MASUYAMA*, Shozo NAITO**

*Toyohashi Univ. of Tech., **NTT Basic Resarch Lab.

Abstract

This paper proposes a new method for extracting keywords in a text. The method uses a thesaurus for classifying words into semantic categories and takes count of the occurrence pattern of the semantic categories for words in a text. The method estimates a topic of each sentence or paragraph using the occurrence pattern of the semantic categories, and then extracts keywords for the whole text. We evaluated the performance of the method comparing with a conventional method through computer experiments.

1 はじめに

本稿では，シソーラスを使って，テキスト中の語の意味分類の出現傾向を考慮することによるキーワード抽出法を提案する．

一般にキーワード抽出においては，まず，各語に何らかの方法によりその重要度を表す重みを付し，重みの大きい順に与えられた個数のキーワードを選ぶ．

従来行なわれて来た重み付けの方法としては，しばしば語の累積頻度が使われた [1, 2, 3]．この場合には一般に，テキストの分野を限定し，キーワードとしては不適当な語を排除する操作が必要となる [2, 4]．またシソーラスの使用も提案されているが，それらは意味分類の出現の頻度だけを使用する [5] ものや，単語間の関係を調べる手段とする [4] ものに限られていた．

我々は，日本語文章の語彙的結束構造を解析するためのデータ構造として結束チャートを提唱した [7]．本報告では，この結束チャートをキーワード抽出に応用し，キーワードとして不適当な語を排除する問題の解決を計る．

具体的な方法は，語の出現頻度，及び，結束チャートのデータに基づき，文単位，段落単位と順にキーワードを抽出し，最後に，それらのキーワードを基に文章全体のキーワードを決定する．

実験用テキストには日経サイエンスの記事を10例使用した．各テキストは，段落が幾つか集まり大段落を形成している．

2 結束チャートとキーワード抽出

結束チャートの生成には，シソーラス上の意味分類を用いた．シソーラスには，角川の「類語新辞典」[6](収録語彙数 57415 語) を使用した．この辞典では，各語には 10 進 3 桁の意味分類番号が付されている．各意味分類は，大，中，小の三段階に分類されていて，必要に応じて小分類は，さらに細分されていることもある [6]．

結束チャートは，テキスト構造を視覚的にとらえることを目的として提唱された [7]．これまでは人間の理解を容易にするために，主に意味分類には中分類を使用し，処理するテキストの長さに応じて大段落，あるいは，段落を単位として結束チャートを作成してきた．

2.1 予備調査

キーワードの自動抽出を行なう前に，キーワード抽出というタスクに対しては，どのような詳細度の結束チャートが有効であるかを検証するために，以下の予備調査を行った．調査項目は，以下のとおりである．

- どのレベルの意味分類 (小，中，大分類) が，もしくは，分類の組み合わせがキーワード抽出に対して効果的か．
- 文，小段落，大段落の中の，どの単位に対して作成した結束チャートの利用が効果的か．

2.2 予備調査の結果

その結果として，キーワードの抽出には，中分類のみ，大分類のみ，中分類，大分類の同時使用，あるいは，小分類，中分類，大分類を同時に使用して作成した結束チャートのデータでは情報量不足となり，有用な手段となり得ないことがわかった．

この理由としては，通常の文章においては，主題が一つの大分類に集中するためと，数詞等を含む大分類の出現頻度が高くなり過ぎるため，その他の重要と考えられる意味分類の重要度が小さく見積もられてしまうためと考察される．

また，今回の実験対象としたテキストでは，文を結束チャートの作成単位とすると分類の出現傾向が不明確になることがわかった．この理由については，第 6 節で考察する．

そこで，本報告では意味分類のレベルは小分類を使用し，小段落，または，大段落を作成単位として結束チャートデータを生成し，その結果をキーワード抽出に利用する．

なお，抄録のように文単位の抽出を行なう場合には，単語を抽出するキーワード抽出の場合とは異なり，小分類を使用するよりも中分類を使用する方がよい結果が得られることが考察されている [8]．

2.3 キーワード抽出における問題点

キーワードとして不適切な語 (数詞，人称代名詞，接頭語，接尾語等) の排除を精度よく行なうためには，結束チャートのデータ作成を正確に行なうことが重要である．

ただし, 数詞等も, 一律にキーワード候補から排除されるべきではなく, あくまでも他の候補との相対的な重要度判定の結果のもとでキーワードの候補から削除されるべきである.

そこで, 本報告では, キーワード候補への重み付けパラメータを使用し, キーワードとはならない通常の数詞等の意味分類の排除可能性を評価する. パラメータ値の決定については, 第 4 節で述べる.

3 キーワード抽出規則

3.1 被験者による抽出規則例

アンケート調査の結果, 被験者達は, キーワードを選択する際に, 以下のような基準を用いて, キーワードを選択していることがわかった.

- 出現頻度の高い語を選択
- 話題の中心を指し示す語を選択
- 話題の中心と関連のある語を選択
- 文節中の語幹を選択
- 一文章中で 5 個から 10 個選択

上記結果の中で最後に挙げた基準は, 文章の長さに依存しない.

3.2 キーワード抽出規則

そこで, 本手法では, 以下の規則を設けることにより, 人間の選択基準を模倣することとした. しかし, 一文章中から, 限定した個数のみ, キーワードを選択することは, 被験者の行った結果と本手法による抽出結果とを適合させるための手段とする以外の利点が考えられないため, 今回は, 採用しないこととした.

- シソーラスを使用し出現頻度の高い意味分類中の語を選択
- シソーラスにより, 関連する分野の出現傾向を考慮し, 語を再選択
- 辞書中の名詞, 形容詞, 動詞等の語幹を選択

本手法の特徴は, 文章中の文, 段落の単位ごとの話題の推定に意味分類の出現傾向を使用することである. そのために, テキスト全体での出現頻度は高いが, 一つの段落だけに出現する意味分類, 幾つかの段落で出現するが, 出現段落間に空きがある意味分類などの意味分類の出現傾向を調べる. また, 後者の場合には, 特に, 段落の空きの長さ, 何分割されているかなどによる重み補正を行なう. 本報告では, 以下の方法を用いて重み付けを行ない, キーワードを抽出した.

4 キーワード抽出アルゴリズム

今回, 実験を行ったテキストに対しては, その文章の長さより大段落を意味のまとまりとすることが効果的であると判断された. よって, 以下のアルゴリズムで, '段落' とは, いくつかの形式段落 (以下 '小段落' と呼ぶ) からなり小見出しの付けられた '大段落' のことを指している.

また, 意味分類が同じ段落数出現した場合では, 意味分類の出現傾向を考慮に入れ, 飛び飛びの段落に分散して出現するよりも, 連続する段落に集中して出現する場合の方を重みを大きくするようにした.

1. シソーラスを辞書として使用し, テキストの単語分割を行ない, 文, 段落の区切りを抽出する.
2. 上記結果に基づき結束チャートを作成し, 出現意味分類の総数, 初出段落番号, 最終出現段落番号, 最大非出現段落数を算出し, 各意味分類ごとの出現段落総数を数え上げる.
3. 各意味分類の出現回数を, 最大出現回数で割ることにより, 相対出現率を計算し, その値を各意味分類の基本重みとする.
4. 意味分類の出現段落パターンによる重み補正を行なう (具体的な方法は 4.1節に述べる). 各段落毎に 4 個の意味分類を選択する. ただし, タイは以下の方法により解消する: 重みが同一の場合は, 同一段落内で一番最後に現れる意味分類を選択する.
5. 上で選択した 4 個の意味分類のいずれかに属し, かつその段落に 2 回以上出現した語を各段落ごとのキーワード候補とする.
6. 上記結果において 3 個以上の段落で, キーワード候補として選択された語をテキスト全体のキーワードとする.

4.1 出現段落パターンによる重みの補正

本報告では, 以下の重み補正を行なった.

$$w_n = \frac{a}{e-b+1} \times w_o \times \delta$$

w_n: 段落に対する補正後の意味分類の重み

a: 各意味分類の出現段落総数

e: 最後に現れた段落番号

b: 最初に現れた段落番号

w_o: 補正前の意味分類の重み

δ: 出現段落の空きによる補正係数 (具体的には, 以下 2 を参照).

1. 意味分類の出現パターンを考慮に入れ, ある意味分類が全テキスト中において集中して出現する場合に重みを大きくするように重み補正を行なう。

2. 段落の空きに応じて重み (δ) を掛ける. 段落の空きは, 1, 2, 3 以上, の三タイプとした. つまり, n を段落の空きの長さとすると

$$\delta = 1 - k\min(n, 3)$$

但し, k は, 0.1, 0.05, 0.01 の 3 つの値に対して実験した.

5 実験

日経サイエンス 10 編に対し, 本手法でキーワードを抽出したものと, 被験者 (本学学生 5 名) による抽出結果とを再現率 R_r 及び, 適合率 R_p で評価した結果を示す.

実験テキストは, テキスト長が EUC コードで平均 28569 byte, 段落数が平均 43 であった.

キーワードの個数は, 本手法では自動的に決まる. 一方, 人手による場合は, 適当と思われるだけの個数キーワードとして抽出してもらった結果を元にして, 辞書中に存在しない語を削除し, 全員が選択したものを「被験者が抽出したキーワード」と呼び, 正解キーワードとみなした.

また, 比較のために, 同一の辞書を使用し, 意味分類情報を無視し, キーワードとなり得ない意味分類 (通常の数詞, 動詞等) を削除した後に, 一段落中で, 3 回以上出現した語をキーワードとする手法 (手法 2) の結果も合わせて報告する (これは, [2] の手法を簡略化したものである).

$$R_r = \frac{\text{被験者と本手法の抽出結果で共通なキーワード数}}{\text{被験者が抽出したキーワード数}}$$

$$R_p = \frac{\text{被験者と本手法の抽出結果で共通なキーワード数}}{\text{本手法で抽出したキーワード数}}$$

5.1 入力テキスト例

今回の実験に使用したテキストの中から一つ例を挙げる. 実際のテキストは全部で 15 の大段落からなる. その内の第 1 大段落のみを以下に示す.

タイトル「視覚の脳内機構」[9]

生物の長い進化の歴史の中でも, 大脳皮質の発達の過程は, 最も成功したものの1つといえる。哺乳類以下の脊椎動物では, 大脳皮質は存在するといってもほんのわずかにすぎない。それが下等哺乳類になると突然目立ちはじめ, 食肉類では脳内でも優位を占めるようになる。さらに霊長類ではそれが爆発的に増大し, ヒトになると, 脳全体をほぼ完全に包んでしまい, 大脳皮質以外の脳の部位を, すっかり影の薄い存在にしてしまっている。

動物のある器官の重要性は, 器官の大きさよりも, 動物がその器官にどの程度依存しているかによって表わされるが, 大脳皮質への依存度は, 哺乳類の進化に伴って急激に増加してきた。大脳皮質を切除したハツカネズミは, ちょっと見ただけでは正常なハツカネズミとほとんど区別がつかないが, 大脳皮質のないヒトはまさに植物人間で, 話すことも, 見ることも, 感じることもできない。

この大きな, そしてまた高等動物にとっては必要欠くべからざる器官の実体は, まだほとんど解明されていない。それは, 大脳皮質が構造ばかりでなく, 機能の面でも非常に複雑なためであるが, それだけではなく, 神経生物学者たちの機能についての洞察が, しばしば誤っていたことも原因となっている。大脳皮質は, 基本的な構成要素であるニューロンが非常に多数集まり, 相互に複雑に連絡しあったこみいった構造になっているが, 研究の技術が進歩し, そうした複雑な構造の中でのインパルスの流れや, ニューロン相互の連絡部位であるシナプ

スについての知識が蓄積されてくると，大脳皮質の理解について新たな展望が開けてくる。

この小論では，大脳皮質の一部であり，視覚に関係した皮質領野の中では最もレベルの低い一次視覚野（有線野，１７野ともいう）について，現在得られている知識を描いてみたいと思っているが，それは必然的に，視知覚の問題にも立ち入ることになる。なぜなら，ある器官の活動は，その器官の生物学的な目的とは切り離せないからである。

5.2 被験者によるキーワード抽出結果

テキスト [9] に対する被験者のキーワードの抽出結果は以下のとおりである．

「大脳, 網膜, 視神経, 脳, 神経, 眼, 細胞, 視覚, 視野, 組織」

5.3 本実験によるキーワード抽出結果

テキスト [9] に対する本手法のキーワードの抽出結果は以下のとおりである．

パラメータ ($k = 0.01$)

「皮質, 眼, 両眼, 眼球, 機能, 線, 方向, 垂直, 一方, 機構, 構造, 構成, 水平, 細胞, 視覚, 同様, 場合, 電極, 通常, 位置, 部位, 大脳, 網膜, 視神経, 脳, 神経, 非常, 対象, コラム, 刺激, 最初, 様相, 表面, 方法, 明暗, 優位, 外界, 外側, 連絡, 予測, 図式, グラフ, 視野, 領域, 反映, 投影, 放射, 投射, 主要」

パラメータ ($k = 0.05$)

「大脳, 網膜, 視神経, 脳, 神経, 非常, 皮質, 方向, 対象, 眼, 両眼, 眼球, 刺激, 垂直, 様相, 場合, 表面, 模様, 機構, 構造, 構成, 方法, 細胞, 水平, 明暗, 優位, 現在, 外界, 外側, 視覚, 電極, 平行, 連絡, 受容, 考え, 予期, 予測, 図式, グラフ, 視野, 領域, 重要, 必要, 有効, 反映, 投影, 放射, 投射, 位置, 部位, 研究, 解剖, 主要」

パラメータ ($k = 0.1$)

「周囲, 非常, 大脳, 神経, 網膜, 視神経, 脳, 皮質, 方向, 対象, 間隔, 垂直, 様相, 場合, 表面, 模様, 組織, 機構, 構造, 構成, 細胞, 実体, 水平, 明暗, 現在, 全体, 優位, タイプ, パターン, 外界, 外側, 体内, 視覚, 程度, 高等, 同様, 複雑, 単純, 左側, 関する, 連絡, 受容, 図式, グラフ, 特定, 通常, 周期, 視野, 領域, 中心, 周辺, 基礎, 運動, 使用, 反映, 投影, 光, 放射, 投射, 部位, 位置, 次元, 事実」

表 1 処理結果 再現率 (%)

	$k = 0.1$	$k = 0.05$	$k = 0.01$	従来の手法
テキスト例 1	83	42	83	58
テキスト例 2	33	75	58	42
テキスト例 3	78	78	67	67
テキスト例 4	44	78	56	78
テキスト例 5	90	90	90	67
テキスト例 6	80	60	40	40
テキスト例 7	43	86	43	14
テキスト例 8	44	78	78	67
テキスト例 9	50	88	50	50
テキスト例 10	50	75	75	75
平均	60	75	64	56

表 2 処理結果 適合率 (%)

	$k = 0.1$	$k = 0.05$	$k = 0.01$	従来の手法
テキスト例 1	36	19	31	10
テキスト例 2	15	26	24	7
テキスト例 3	19	26	23	7
テキスト例 4	8	21	16	12
テキスト例 5	18	17	15	7
テキスト例 6	31	30	20	4
テキスト例 7	20	43	20	3
テキスト例 8	13	33	41	8
テキスト例 9	15	25	18	6
テキスト例 10	12	30	33	6
平均	19	27	24	7

テキスト例 1 から 10 は, すべて日経サイエンスの記事

6 考察

- キーワードとして不適切な語のほとんどは，シソーラス中の幾つかの限定された，大分類と中分類に属していることが分かった．例を上げれば，大分類では性状，中分類では人称，人物，記号，態度，性格等である．

- 意味分類の重み付けをする為の結束チャートの作成単位は，テキストの長さに依存して変える必要がある．

  その理由は，テキストの長さにより，話題のまとまりの単位が異なるからである．たとえば，今回実験対象としたテキストでは，大段落が使われており，大段落単位で話題が一貫しているので，比較的明確な意味分類の出現パターンが現れる．一方，小段落単位では話題の一貫性が弱く明確な意味分類の出現パターンが現れてこない．

- キーワードの自動抽出に用いる分類としては，小分類のみ，もしくは，小分類と中分類の同時使用が効果的であると考えられる．

  なぜなら，各テキストが扱っている語の分野が限定されているため，大分類をキーワード抽出の際に考慮すると，数詞等の特殊な語の含まれる意味分類の重みが他の意味分類より必要以上に高くなるためである．

- 重要な意味分類は基本的に大段落をまたがらないという性質を持つ．これは，重要な意味分類が局所的な話題を示すのに十分な性質を持つためといえる．

- 大段落中の小段落数や，文数に，ばらつきがあると抽出精度が劣化する．

  これは大段落内の小段落数や，文数が少ない時には，その前後の大段落と，まとまるようにテキストが構成されることがあるためである．

  このことは文単位の結束チャートを作成した際の精度についても抽出精度が劣化することも意味している．どれだけの文を挟んでの再出現であるかでは，大域的な文章構造をうまく反映しきれないといえる．

- 単語の出現頻度だけを考慮した方が精度が良い場合もある．

  なぜなら，未知語が，その下位範疇で既知の語となる場合には，シソーラスを用いた重み算出の精度を劣化させるので最長一致法だけでは単語の切り出しには，不十分となる．

  例として，ブラックホールは2語に分けられその，どちらも異なった意味分類に属する．また，単語の一部分のみが辞書に存在する例としては，ニュートリノ中のニューのみが取り出される例などが挙げられる．

7 おわりに

本稿では，語の意味分類の出現傾向を考慮することによるキーワード抽出法を提案した．段落単位の意味分類の出現傾向を考慮することにより，段落ごとの話題を抽出することができるので，本手法が，キーワード抽出の精度向上にも有効であることを確認した．

現段階では，シソーラス情報に依るキーワード抽出の向上の評価の為に一種類の辞書を使用しているだけである．また，辞書に未登録な語はキーワード候補から除外している．

今後は，キーワード抽出の精度向上をはかるために，意味分類を合成された単語にも適用できる方法を考え出す必要がある．

そのために，使用する辞書数を拡大し，シソーラスに対する未登録語の処理を組み込むことによりシソーラス情報の信頼性を向上させる予定である．

また，さらに適合率を向上させるためには，シソーラス情報だけでは不十分であり，主題に関する談話構造の解析などの談話処理が必要であると考えられる．

謝辞

「類語新辞典」の使用許可をいただいた (株) 角川書店，および，アンケート調査に協力していただいた方々に深謝する．

参考文献

[1] 荒木 雅弘，河原 達也，西田 豊明，堂下 修司，キーワード抽出に基づく意味解析による音声対話システム，電子情報通信学会，言語理解とコミュニケーション研究会技術研究報告，NLC 91-51，(1992)．

[2] 水野　聡, 島田 静雄, 中牟田 純, 近藤 邦雄, 佐藤　尚, 日本語キーワードの自動抽出手法, 情報処理学会, 自然言語処理研究会研究報告, NL 91–6, (1992).

[3] 鈴木 康広, 栃内 香次, キーワード密接方式自動抄録法の改良, 情報処理学会, 論文誌 Vol.29 No.3, (1988).

[4] 木本 晴夫, 日本語新聞記事からのキーワード自動抽出と重要度評価, 電子情報通信学会論文誌 D - I Vol. J74-D-I No.8 pp. 556–566, (1991).

[5] 構文解析より特許文からキーワードを 96% 自動抽出, 日経エレクトロニクス 10 月号, (1981).

[6] 大野　晋, 浜西 正人, 類語新辞典, 角川書店, (1981).

[7] 佐々木 一朗, 増山　繁, 内藤 昭三, 結束チャートの自動生成と日本語文章の語彙的結束構造解析への応用, 電子情報通信学会, 言語理解とコミュニケーション研究会技術研究報告, NLC93–8, (1993).

[8] 佐々木 一朗, 増山　繁, 内藤 昭三, 結束チャートとそれを利用した抄録への応用の試み, 情報処理学会, 自然言語処理研究会, 研究報告, (1993).

[9] D.H. ヒューベル, T.N. ウィーゼル, 視覚の脳内機構, 日経サイエンス 11 月号, (1979).