自然言語処理 64－1
（1987. 11. 20）

キーワード自動抽出と重要度評価

木本 晴夫
ＮＴＴ情報通信処理研究所

言語処理・知識処理・統計処理を用いる新しいキーワード自動抽出法として語特徴評価法を提案する。従来のキーワード自動抽出はフリーターム方式か統制キーワード方式を用いて行われていた。これらの方法では適切なキーワードとともに、適切なキーワードの３～５倍もの不適切なキーワードが抽出されていた。語特徴評価法は、これらの不適切なキーワードを大幅に削除して精度の高いキーワード自動抽出を実現することを目的としている。本方法は統制キーワード方式によって抽出したキーワード候補語について、個々の語の、文章中やシソーラスにおける特徴を抽出して、語の特徴によって、そのキーワード候補語が必要なキーワードか否かを評価する方法である。ここで、語の特徴として次ぎにかかげるものを採用している。それらは、並立に表現された語、連体修飾語、強調表現された語、シソーラスにおける上位語、シソーラスにおいて上位・下位の関係にある語、語の文章中における出現位置、語の文章中における出現頻度である。語特徴評価法を用いることにより、抽出される不適切なキーワードの数を従来の方法と比較して１／４にできることを実験によって確認した。また語の特徴に基づいてキーワードの相対的重要度を評価した結果、上位の１０語の中に必要なキーワードの９５％を入れることができた。

Automatic Indexing and Evaluation of Keywords

Haruo KIMOTO
NTT Communications and Information Processing Laboratories
405C, 1-2356 Take, Yokosuka-Shi
Kanagawa, 238-03, Japan

In this paper, a new method for automatic indexing is proposed. Almost all of the existing automatic indexing systems adopt the free term method. The free term method extracts as keywords all of the nouns in the texts, excluding the words in the user-defined dictionary. As a result, the free term method extracts many unnecessary keywords as well as the necessary ones. The new method provides two functions: one to delete the unnecessary keywords extracted by the existing method, and the other to rank all keywords extracted by the existing system. The new method adopts: linguistic processing, knowledge processing, and statistical processing. Indexing newspaper articles was evaluated. Of the keywords extracted using the free term method, 90% proved to be unnecessary. The new system was implemented and using deleting function that ratio was reduced to 50%. Experimentaly it was confirmed that manually indexed keywords contain 20 to 30% unnecessary keywords. On the other hand, by using the ranking function, 95% of the necessary keywords were included in the top ten keywords. The new method is thus the most accurate extracting method developed to date.

1．はじめに

キーワード自動抽出は、大量の文献情報データベースから必要な情報を早く引き出すための必須の技術である。現在、実用に供されているキーワード自動抽出システムはそのほとんどが不要語除去方式のものである［1］。この不要語除去方式はさらにキーワードを辞書によってコントロールするか否かによって２つの方式に分けられる。キーワードをコントロールする方式は統制キーワード方式［2］であり、コントロールするためにシソーラスを利用する。コントロールしない方式はフリーターム方式［3］、［4］である。２つの方式ではフリーターム方式のものがより多く利用される傾向にあるが、その理由の１つはコントロールのための辞書を準備しなくてもよいことであり、従って辞書のメンテナンスも不要だからである。

フリーターム方式を用いるにしても、統制キーワード方式を用いるにしても、これらの方式は本来、文章の意味を理解してキーワードを抽出しているものではないので、キーワードとしては不適切な語（以下、ノイズと呼ぶ）を数多くキーワードとして抽出してしまう。例えばフリーターム方式で一般の新聞記事からキーワード抽出をした場合では、１つの記事から１００個前後のキーワードが抽出されて、そのうち適切なキーワードは僅か１０個であるという事例も珍しいことではない。キーワードのノイズが多いと文献検索に時間がかかるとともに、不必要な文献を数多く抽出してしまい、その後で人が必要な文献を捜さなければならず、人に負担を強いることになる。

このような現用のシステムの問題点を解決するために、文献にあらかじめ、その文献の属する分野情報を付与しておいて、この分野情報とキーワードとによって文献を検索する方式が検討されつつある。この方式は、文献に新たな情報を人手によって付与して文献検索の精度を高めるものである。

本稿では、抽出されたキーワードのノイズを大幅に減らして、文献検索の精度を高める方法を提案する。キーワードのノイズを減らす方法として２通りの方式を考案した。まず、統制キーワード方式を用いて抽出した語をキーワード候補語と呼ぶことにする。ひとつの方式は、キーワード候補語が必要なキーワードであるかそうでないかを判定して必要なキーワードを抽出するものであり、もうひとつの方式は、キーワード候補語に得点を与えて、順位付けをしてキーワードを相対的に評価してキーワードの足切りを行うものである。

これらの判定、評価を行うために個々のキーワード候補語について、次のような語の特徴の有無を認定し利用する。語の特徴は、語と語の上位下位関係

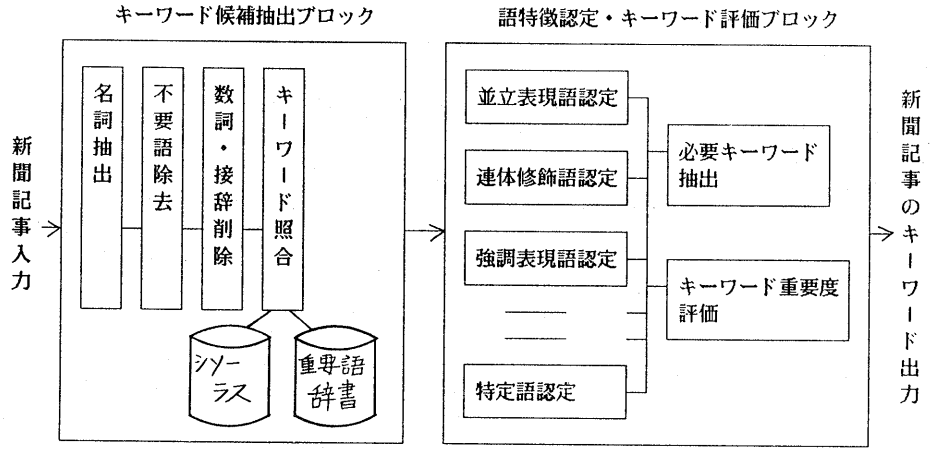


図1．　INDEXERシステムの全体構成

・語の文章中における表現・インデクサとよばれるキーワード付与の専門家が用いるルール・語の文章中における出現位置・語の文章中における出現頻度等である。これらの、語の特徴を組み合わせてキーワードの判定・評価を行う。語の特徴を利用してキーワードを判定する方式を語特徴判定方式と呼び、同じく、キーワードの重要度を相対的に評価する方式を語特徴評価方式と呼ぶ。

これらの方式を実現する実験プログラム（以下、ＩＮＤＥＸＥＲと呼ぶ）を作成してその評価を行い良好な結果を得たので報告する。キーワード抽出の対象は一般新聞紙の全ての分野の記事である。すなわち、政治、経済、国際、産業、商品、スポーツ、社会面の記事である

２．ＩＮＤＥＸＥＲシステムの構成

ＩＮＤＥＸＥＲシステムの全体構成を図１に示す。ＩＮＤＥＸＥＲシステムは２コの大きなブロックから構成されている。ひとつのブロックは統制キーワード方式によってキーワード候補語を抽出する部分であり、もうひとつのブロックはキーワード候補語について語の特徴を抽出して、語の特徴に基づいて、そのキーワード候補語が必要なキーワードかどうかを判定したり、キーワード候補語の相対的重要度を評価する部分である。

３．キーワード候補抽出

キーワード候補抽出には統制キーワード方式を用いた。シソーラスは新聞記事索引用シソーラス［５］を利用した。このシソーラスの構成語数は約８０００語である。

４．語の特徴抽出

抽出された個々のキーワード候補語について、次ぎに示すような、語の特徴の有無を文章中およびシソーラスから抽出する。

語の特徴

（１）並立表現語

並立表現語とは文中で、「ＡやＢ」、「ＡとＢ」、「Ａ、Ｂ」（Ａ、Ｂは漢字カタカナ列）のように表現されている語である。

（２）連体修飾語

連体修飾語とは文中で、「ＡのＢ」（Ａ、Ｂは漢字カタカナ列）のように表現されているＡの語である。ただしＢが用言性名詞でない場合である。

（３）強調表現語

強調表現語とは文中で例えば、「この机、大学まで…」の「机」のように強調表現されている語である。

（４）シソーラスにおける上中位語

シソーラスにおける上中位語とはシソーラスにおいて下位語がある語のことである。

（５）シソーラスにおける上位語・下位語ペア

ひとつの文章の中にシソーラスで上位・下位関係にある語が両方とも出現している場合に上位語・下位語ペアとする。

（６）文章中での出現位置

文字数を単位とした文章中での出現位置のこと。

（７）文章中での出現頻度

文章中での語の出現頻度のこと。

また、次ぎに示す特定語にはキーワード付与の専門家が用いるルールを適用するので、これらの特定語が文章中に出現すれば無条件にそれらの語を抽出しておく。

（８）特定語

学会名、企業名である。

５．必要キーワードの抽出

５．１　抽出方法

キーワード候補語の語の特徴に対して、次ぎに示すキーワード抽出表を適用して必要キーワードを抽出をする。必要キーワードを抽出した例を表１に示す。

キーワード抽出表

○：キーワードとする

×：キーワードから削除する

（1）×：並立表現語

（2）×：連体修飾語

（3）○：強調表現語

（4）×：シソーラスにおける上中位語

（5）○：シソーラスにおける上位語・下位語ペア

（6）×：文章中での出現位置が最初から数えて８６文字目以降の語

（7）○：文章中での出現頻度が４回以上の語

（8）○：特定語にキーワード付与の専門家が用いるルールを適用して新たに生成した語

［例］学会名（薬学会）から、新たに「学会」という語と学問名（薬学）を生成する

企業名から、新たにその企業の属する業界名を生成する

キーワードの抽出でひとつの語に対して○、×の両方がでたときは○を優先する。

５．２必要キーワード抽出の実験結果と抽出精度

必要キーワード抽出の実験結果を図２に示す。キーワード抽出の精度は一般に再現率と適合率によって評価される［脚注］。

ＩＮＤＥＸＥＲシステムによる必要キーワード抽出の精度を他のシステムと比較したものを表２に示す。表２からわかるとおりＩＮＤＥＸＥＲシステムは人間に近い精度を実現している。

［脚注］

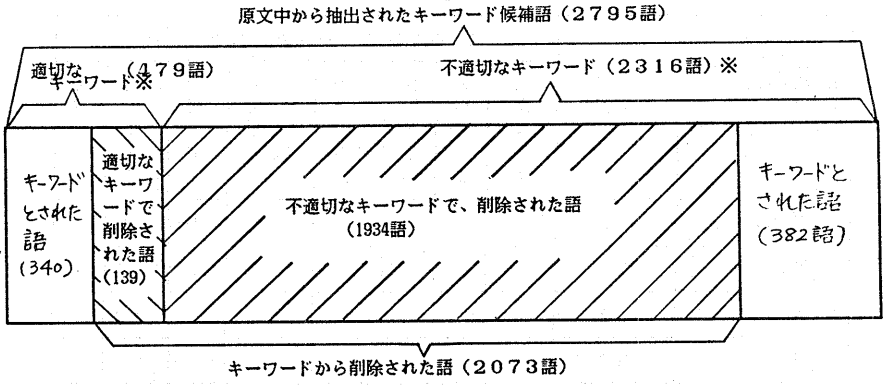
$$\text{再現率} = \frac{\text{抽出したキーワードの中で適切なキーワードの数}}{\text{適切なキーワードの数}}$$

$$\text{適合率} = \frac{\text{抽出したキーワードの中で適切なキーワードの数}}{\text{抽出したキーワードの数}}$$

表１．　必要キーワード抽出の例

統制キーワード方式によって抽出された語（注） 人手付けキーワード	評価結果※ K：キーワード D：キーワードから削除された語	語の特徴によるキーワード候補語の評価							
		並立表現語	連体修飾語	強調表現語	シソーラス上中位語	シソーラス下位語対上位語	出現位置	出現頻度	特定語
電気	K								
日本電力会社	K						K		K
エンジン	D	D			D				
ガス	D	D			D				
研修	D	D			D				
電力	K						D		K
研究	D						D		

※ひとつの語に対して”K”と”D”の両方の評価が有る場合は”K”を優先する
また、なんの評価もない場合は　”K”とする



※適切なキーワードとはあらかじめ人が付けたキーワードと一致した語であり、
不適切なキーワードとは一致しなかった語である
（参考）原文中に出現しない語も含めた人手付けキーワードの総数は７２３語である

図2．　必要キーワード抽出の実験結果

表2．　自動索引システムの精度比較

精度 / システム	再現率（%）	適合率（%）
フリーターム方式によるシステム	７０	１０
統制キーワード方式によるシステム	６５	２０
ＩＮＤＥＸＥＲシステム	５０	５０
マニュアルインデクシング（人）	７０～８０	７０～８０

６．キーワードの重要度評価

６．１重要度評価方法

キーワード候補語の語の特徴に対して、次ぎに示すキーワード重要度評価表を適用してキーワードの重要度を評価する。キーワード重要度を評価した例を表3に示す。

キーワード重要度評価表

（１）並立表現語　　：評価値－５０点
（２）連体修飾語　　：評価値－３０点
（３）強調表現語　　：評価値＋８０点
（４）シソーラスにおける上中位語
　　：評価値－３０点
（５）シソーラスにおける上位語・下位語ペア
　　：評価値＋８０点
（６）文章中での出現位置が最初から数えて８６文字目以降の語
　　：評価値－５０点
（７）語の文章中での出現頻度
　　：評価値＋２０点／回
（８）特定語にキーワード付与の専門家が用いるルールを適用して新たに生成した語
　　：評価値＋８０点

表3．　キーワード重要度の評価例

順位	統制キーワード方式によって抽出された語（注）　人手付けキーワード	評価結果	語の特徴にるキーワード候補語の重要度評価							
			並立表現語	連体修飾語	強調表現語	シのソ上ー中ラ位ス語	シの下ソ上位ー位語ラ語対ス・	出現位置	出現頻度	特定語
1	日本電力会社	160						60	20	80
2	電力	70						−50	40	80
3	電気	40							40	
4	研究	−30						−50	20	
5	エンジン	−40	−50			−30			40	
6	ガス	−40	−50			−30			40	
7	研修	−60	−50			−30			20	

重要度評価表の評価値は次のようにして決めた。無作為に選んだ２００件の記事とその記事にたいしてあらかじめ専門家が付与したキーワードを評価値を決めるための基準とする。重要度評価表の各項目の評価値をパラメータとして変化させて、キーワードの重要度を評価し、専門家が付与したキーワードが上位１０位のキーワードの中にに最も多く含まれる場合のパラメータ値を最適パラメータ値として採用した。このパラメータ値のときに、専門家が付与したキーワードが上位１０位に含まれる割合は約９５％であり、新聞記事標本を別の無作為に選んだ１０００記事にかえても結果は同様であった。

表4．　キーワード重要度評価の実験結果

評価に使用された語	評価に使用された語の中に必要なキーワードが含まれている比率
上位の5語	80%
上位の10語	95%

6．2キーワード重要度評価の実験結果

キーワード重要度評価の実験結果を表4に示す。

6．3キーワード重要度評価の精度の分析

従来は抽出したキーワードの重要度を評価する技術は無かった。本稿で提案した重要度評価方法（以下、語特徴評価法と呼ぶ）の精度を明確にするために、別に簡単な重要度評価方法を想定して両者の比較を試みた。別の簡単な重要度評価方法とは、抽出されたキーワード候補を文章中で最初からでてくる順番に第１位、第２位、…としてゆく方法である。

この方法を出現位置評価法と呼ぶ。両者の精度を次ぎに示す２つの指標によって比較した。

指標１：必要キーワードを上位１０位のキーワードの中に含む割合

指標２：キーワード候補全体の中での必要キーワードの分布状況

指標１、指標２のおののついての精度比較を図３、と図４に示す。ここで精度の比較は、個々の新聞記事に対して、原文中に出現する語の中から付与されたキーワードの数の別におこなった。この理由は次の通りである。

①原文中に出現する語に限ったのは、キーワード付与の専門家は原文中に出現しない語でも文意から判断して必要な語はキーワードとして付与するが、現状の機械処理ではこれを実現する技術はない。

②キーワード数の別にしたのはキーワードの数が１個とか２個とか少ない場合は、それらの語はほとんどの場合、文章の最初の部分に出現しているとか、文章中で頻繁に出現するとかで語特徴評価法、出現位置評価法のいずれの方法で評価しても重要な語となり、有意差がでないためである。また、通常、専門家によって付与されるキーワードの数は５～６個であることと、機械によって抽出されるキーワード候補の数は２０個以上であることから、キーワードの数がある程度多い場合についての精度比較をすることが標準的である。

指標１による比較の結果から次のことが明らかになった。語特徴評価法はキーワードの数にかかわりなく、安定して、高い精度で重要度を評価している。一方、出現位置評価法はキーワードの数が多くなるにつれて重要度評価の精度がかなり低くなる。

指標２による比較の結果から次のことが明らかになった。語特徴評価法は出現位置評価法に比較して、必要なキーワードの重要度を精度よく評価している。

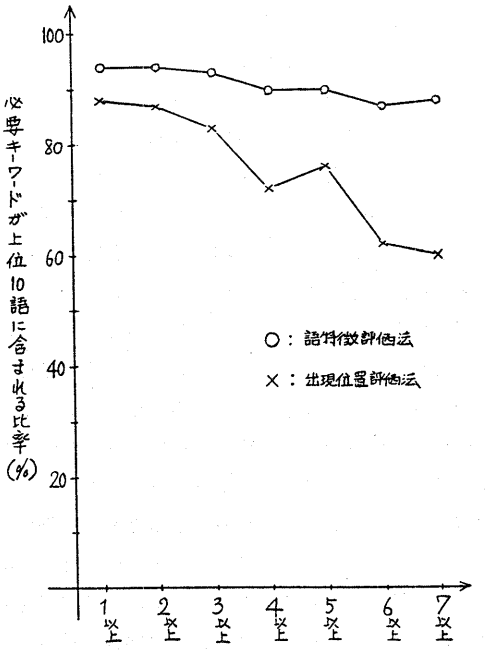


図3. 語特徴評価法と出現位置評価法の比較
－必要キーワードを上位10位のキーワードに含む割合

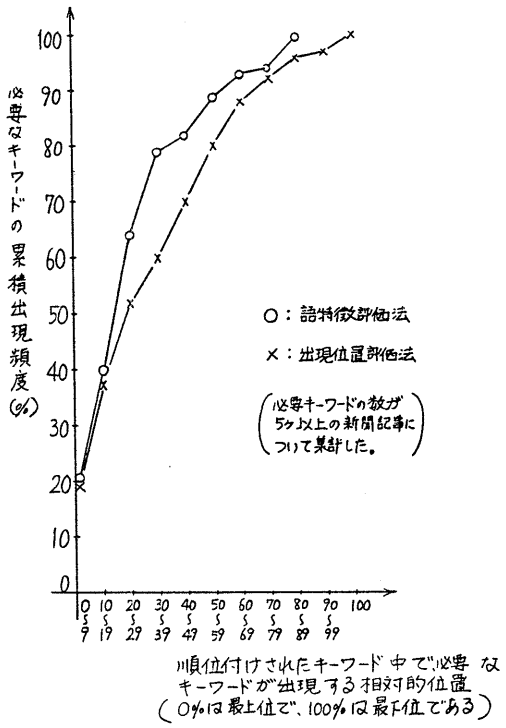


図4. 語特徴評価法と出現位置評価法の比較
－キーワード候補全体の中での必要キーワードの分布状況

７．むすび

従来のキーワード抽出技術は、必要なキーワードを抽出すると同時に多くの不必要なキーワードを抽出するという問題点があった。本稿ではキーワード抽出技術の精度を向上させるために、キーワードについて、文章中、並びに、シソーラス中での語の特徴を抽出して、その特徴を利用して、必要なキーワードを抽出する方式とキーワードの相対的な重要度を評価する方式とを提案した。これらの方式を実現する実験プログラムＩＮＤＥＸＥＲを作成し精度の評価を行った。必要キーワードの抽出においては人間に近いキーワード抽出精度を得た。キーワード重要度評価においては、従来は重要度評価をする技術は無かったが、必要キーワードを精度良く評価することができた。本来は、必要なキーワードは文章を理解して抽出するものであるが、本稿では文章を完全に理解しないでも、文章中の語の特徴を利用してキーワードを精度良く抽出できることを示した。今後は本技術を基盤技術として、自動要約・抄録、文章理解へと技術を発展させたい。

謝辞

本研究を進めるにあたって有益な御指導、御助言をいただいたＮＴＴ情報通信処理研究所自然言語処理研究部、寺島信義部長、坂間保雄主幹研究員、池原悟主幹研究員に深謝いたします。また、プログラミングに協力していただいたＮＴＴソフトウエア（株）の熊倉利昌氏に感謝いたします。

参考文献

［１］諸橋：”自動索引付け研究の動向”、情報処理、Ｖｏｌ．２５，Ｎｏ．９，ｐｐ９１８－９２４（１９８４）

［２］中園ほか：”日本語索引自動生成システム”情報処理学会「自然言語処理技術」シンポジウム、（１９８４．１１）

［３］絹川ほか：”日本語情報検索システムにおけるキーワードの自動抽出”、日立評論、６４、Ｎｏ．５，（１９８２）

［４］パンフレット：ＨＡＰＰＩＮＥＳＳ、（株）平和情報センター

［５］広木：”ニュース・シソーラス”、（１９８４）