情 報 学 基 礎 21－4
（１９９１．５．２８）

英文科学技術抄録文からの機能語の抽出

松尾 文碩
九州大学工学部

本稿では，英文科学技術文献に対する情報検索システムの転置索引が不要語除去法でつくられる場合について，INSPECテープから統計的な方法でつくった不要語集合が人手でつくった既存の不要語集合より優れていることを示す．更に，このINSPECからの不要語集合に基づく機能語辞書の特徴を述べ，この辞書による科学技術抄録文理解システム実現の可能性を示す．

EXTRACTION OF FUNCTION WORDS FROM ABSTRACTS AND PRE-COORDINATED CLAUSES OF SCIENTIFIC AND TECHNICAL LITERATURE

Fumihiro Matsuo

Faculty of Engineering, Kyushu University
Hakozaki, Fukuoka 812, Japan

The keywords in an inverted index of practical large scale information retrieval systems are selected from single words occurred in documents by removing stop-words. First, this paper presents that a set of stop-words made from INSPEC-tapes by a statistical method is superior to the existing human-made sets in the inverted indexes constructed by them. Then, the properties of function words base on the stop-words from INSPEC-tapes is shown. Finally, the possibility of implementing an understanding system of abstracts of scientific and technical literature is discussed.

１．まえがき

大規模データの代表的なものに，科学技術文献についての２次文献ファイルがある．２次文献ファイルは，機械可読形の抄録誌というべきもので，そのレコードは書誌的事項と抄録とから成る．現在の商用情報検索サービスにおけるデータベースの大半は，２次文献ファイルから構築されている．現在の情報検索（ＩＲ）システムの主題検索能力は，キーワードの転置索引に依存している．文書からキーワードを自動的に抽出することを，自動インデキシング(automatic indexing)というが，大規模実用ＩＲシステムで採用されている方法は，最も単純な不要語除去法である．これは，文書の内容を同定する能力がないと考えられている語の集合をあらかじめ用意しておき，本文あるいは抄録等に現われる語のうち，この集合に含まれない語をキーワードとするのである．この集合を不要表(stop list)あるいは否定辞書(negative dictionary)と呼び[1]，その要素を不要語(stop word)または機能語(function word)と呼んでいる．ＩＲシステムにおける否定辞書は，人手によってつくられていて，数量的評価はほとんどなされなかった．

本稿では，代表的な英文の科学技術２次文献ファイルであるINSPECテープ[2]からつくった英文科学技術文書用の否定辞書と既存の否定辞書との数量的比較について報告する．更に，この否定辞書から得られた原形で約1,000の機能語辞書の特徴を述べ，この辞書に基づく科学技術抄録英文の理解システムについて，その実現の可能性を論じる．

２．不要語除去法による自動インデキシング

ＩＲシステムにおける主題検索のための索引は，キーワードを見出し語とする転置索引(inverted index)である．キーワードの選び方は，使用言語の影響を受け，英語と日本語の場合では違いがある．ここでは，英語の場合について述べる．英語の場合，転置索引では，見出し語，すなわちキーワードは単一語(single word)であり，各キーワードに対してそれが関連する文書の集合への対応づけがなされている．しかし，単一語だけを用いた検索では，検索の精度(precision)が悪いため，キーワードを組み合わせた用語による検索が必要である．そのため，各単一語が関連する文書集合に対する演算として，ブール演算と隣接演算(adjacency operation)がある．したがって，実質的には句の形での検索が可能であり，この句は事後結合句(post-coordinated clause)と呼ばれる．文書データベース作成時に，インデクサ(indexer)が文書に付与するキーフレーズを事前結合句(pre-coordinated clause)と呼ぶが，英語の場合，これは補助的な役割しか果たさない．

自動インデキシング(automatic indexing)とは，ここでは文書に転置索引のためのキーワードを付与することであるとする．英語については，1950年代のLuhnの先駆的研究以来，自動インデキシングの研究が盛んに行われてきた．しかし，大規模実用ＩＲシステムで採用されている方法は，前節で述べたように最も単純な不要語除去法である．２次文献ファイルのＩＲシステムでは，抄録や標題などに現われる語のうち，不要語でない語をキーワードとする．

英語の共通機能語としては，250単語からなるRijisbergenの否定辞書[3]が有名である．しかし，Dialogのような大規模商用システムでは非常に少数の単語のみを不要語としているため，文書に現われるほとんどすべての（異なり）単語がキーワードである．英語の専門語は，ギリシャ語あるいはラテン語に基づく造語がある反面，平易な単語の句であることが多く，高頻度の平易な単語というだけで，不要語とするのには危険がある．特に，科学技術文献の検索では，システムの利用者は検索漏れを恐れるのである．

3．否定辞書の評価

ＩＲシステムにおいて，否定辞書は転置索引を創成・更新するプログラムの入力あるいは基礎データである．これらのプログラムが大規模実用ＩＲシステムで採用されているような非統制語方式の不要語除去に基づいたものであるとすると，文書に現われる語のうち，あたえられた否定辞書にない語を見出しとする転置索引がつくられる．いま，この場合の否定辞書の評価について考える．当然ながら，ＩＲシステムにおける否定辞書は，それによってつくられる転置索引によって評価される．転置索引の検索能力を検索語が転置索引の見出しにある確率，すなわち検索者のキーワードが転置索引にヒットする確率であると定義すると，この確率を大きくすることだけが目的ならば，否定辞書は空集合の場合が最も良い．しかし，不要語除去法の目的は，転置索引の２次記憶量を不必要に大きくしないことである．

そこで，否定辞書は二つの尺度で測ることができる．一つは，転置索引の検索能力であり，もう一つは転置索引の２次記憶量である．もちろん，前者は大きい方が良く，後者は小さい方が良い．一般に，否定辞書を大きくすると，前者は悪く，後者は良くなるので，否定辞書の決め方は，一種のトレードオフであるとも考えられる．

ここでは，二つの尺度を次のように定義する．まず，検索能力は基本的には上述のように考えるが，ここでは否定辞書の影響だけをみるために，否定辞書がなければヒットしたのだが，否定辞書があるためにヒットしなかった確率を能力の劣化とみなし，それを１から引いたものを検索能力とする．ただし，検索者の出すキーフレーズは，一般性のある標本を得るのが困難なため，ここでは，

事後結合句＝事前結合句

とし，２次文献ファイルの文書の付与された事前結合句の集まりを，検索者が質問に用いた事後結合句の標本であるとする．２次記憶量については，否定辞書がある場合とない場合の２次記憶量の比を相対領域量と定義する．転置索引は，隣接演算のために，各見出し語に対して，その語が生起する文書の参照番号のほかに生起位置情報をもつものが多い．この型の転置索引の２次記憶量は，文書集合における見出し語の延べ単語数に比例する．そこで，この相対領域量は，文書における見出し語の延べ単語数と全延べ単語数の比で計算できる．

4．ＩＮＳＰＥＣテープによる評価

INSPECテープを用いて，次の３個の否定辞書について評価を行った．

S_D：Dialogの不要語集合，S_D＝ {an, and, by, for, from, of, the, to, with} ；

S_R：250単語のRijisbergenの英語共通機能語辞書から {eg, ie} を除いたもの；

S_F：INSPECテープから抽出した科学技術文書用否定辞書，$|S_F|=1,667$.

S_Dは，S_RとS_Fの部分集合であるが，S_RはS_Fの部分集合ではない．

評価のために用いたINSPECテープは，1973年から1982年の１０年間に配布されたもので，分野による差異をみるために，分野によって次の三つの文書の集合をつくった．

集合Ａ：物理学；

集合Ｂ：電気，電子工学；

集合Ｃ：制御工学，計算機科学．

Ａ，Ｂ，Ｃの文書数は，それぞれ97万，50万，32万であり，約20％の文書が複数の集合に属している．ここで，文書というのは標題と抄録を一緒にしたものである．INSPECテープは，自由索引句(free-indexing term)という事前結合句をもっている．検索能力の評価には，各文書集合に対応する全事前結合句の集まりを検索の事後結合句として使用する．こうして評価した結果を，表１に示す．

表1　検索能力／相対領域量

	S_D	S_R	S_F
集合A	0.987／0.786	0.976／0.619	0.983／0.554
集合B	0.989／0.788	0.980／0.615	0.983／0.541
集合C	0.988／0.785	0.981／0.599	0.984／0.540

／の前の数値が検索能力，後の数値が相対領域量を示す．

表1から，Dialogの9語の不要語の集合S_Dによって，転置索引の大きさが21～22%減少することがわかる．検索能力の低下は，1.1～1.3%である．248語のRijisbergenの機能語の集合S_Rによる転置索引は，S_Dによるものより検索能力は0.7～1.1%低下するだけで，大きさは更に21～24%減少する．

ところで，S_DやS_Rのように，語の品詞や意味に基づき人間の判断で否定辞書をつくろうとすると，S_Rより良いものを求めることが困難になる．例えば，冠詞，前置詞，接続詞，代名詞，助動詞の品詞をもつ単語を市販辞書から拾うと，223語の否定辞書ができる．この223語のうち，165語はS_Rに属している．この否定辞書による転置索引は，S_Rのものより検索能力が良く，相対領域量は悪い．しかし，その差はわずかで，比にして2%以下である．

表1のS_Fは，INSPECテープから統計的な方法によって求めたもので，語数はS_Rより大きいにもかかわらず，S_Fによる転置索引はS_Rのものより，検索能力，相対領域量ともに優れている．検索能力は，S_DとS_Rのほぼ中間にあり，2次記憶の大きさは，否定語がない場合に比べ，半分近くに減少している．

5．INPECテープからのS_Fの作成

この節では，INSPECテープからどのようにしてS_Fを求めたかについて，簡単に述べる．詳細は，文献4を参照していただきたい．

ここでも，基本的な考えは，

事後結合句＝事前結合句

とし，INSPECテープの文書の付与された事前索引句の集まりを，検索者が質問に用いた事後結合句の標本であるとしたことである．そこで，抄録と事前結合句に出現する各語について，事前索引句における生起頻度と抄録における生起頻度の比を求め，この比がある閾値θより小さい語を不要語とした．すると，文献集合が決まると，θによって否定辞書$S(\theta)$が求まる．$S(\theta)$によってつくられる転置索引は，θが大きくなるにつれ，検索能力，相対領域量ともに小さくなる．しかし，集合A，B，Cについて実測してみると，検索能力の方はθが0.4くらいまではゆっくり低下し，検索能力が0.9になるのはθ＝0.2～0.25で，θが0.4より大きいときの能力の低下はなだらである．それに対し，相対領域量の方はθ＝0.03～0.07までに0.5に低下し，θがそれより大きくなるにつれ，ほぼ線形に低下することがわかった．すなわち，θによって転置索引の大きさを自在に制御でき，またその大きさを小さくしても，検索能力は破局的に悪化しないことがわかった．また，θが小さい場合，検索能力の低下はわずかなのに，転置索引はかなり小さくなることがわかった．

詳細に調べてみると，相対領域量が0.5付近でもっとも変化するのは，否定辞書に‘of’を含むか含ま

ないかであることがわかった．そこで，集合A，B，Cに対し，'of'を含むようにする最小のθを選び，否定辞書をつくった．これらを，それぞれS^A，S^B，S^Cとする．S^A，S^B，S^Cのそれぞれの文書集合に対する検索能力／相対領域量は，それぞれ0.986／0.544，0.985／0.544，0.982／0.502であり，表1と比較してみると優れた値であることがわかる．しかし，この欠点は否定辞書が分野によって異なり，否定辞書が大きいことである．実際，$|S^A|=132,610$；$|S^B|=75,959$；$|S^C|=58,686$である．

いま，これらの共通部分

$$S_c = S^A \cap S^B \cap S^C$$

を考える．このとき，$|S_c|=11,667$．そこで，S_cから，それぞれ対応する文書集合における最も生起頻度の高いr個の語によって，否定辞書$S_r{}^A$，$S_r{}^B$，$S_r{}^C$をつくる．すると，それらがつくる転置索引は，いずれもrが2,000以上になると検索能力，相対領域量ともに低下しなくなることがわかった．そこで，

$$S_r = (S_r{}^A \cap S_r{}^B \cap S_r{}^C) \cup \{to\}$$

による転置索引を集合A，B，Cについて評価してみると，この場合も，rが2,000以上になると検索能力，相対領域量ともに低下しない．更に，$|S_r|/r$もr＝2,000付近では，変化は比較的平坦ではあるが，r＝2,000でちょうど極大になる．そこで，

$$S_F = S_{2000}$$

とした．前にも述べたように，$|S_F|=1,667$．つくり方から，S_Fは科学技術文書用の共通否定辞書として使うことができると考えられる．

6．S_Fからつくられる機能語

S_Fは，語の変化に関して閉じていない．そこで，語の変化に関して閉じるように，"完備化"を行った．こうして得られた語の集合W_fを英文科学文献用機能語辞書，あるいは単に機能語辞書という．ここで，

$$|W_f|=2,841.$$

このうち，原形は，1,143語である．W_fがS_Rなどと大きく違うことは，2,841語のうち，動詞が第一義的な品詞である語が，1,851語であり，全体の65%を占める．このことがW_fの大きな特徴である．

科学技術抄録文は，機能語と専門語から成り立っていると考えることができる．そこで，W_fがその意味で機能語の集合であるかどうかをみるために，1985年配布の1年分のINSPECテープを用いて，調査を行った．この場合は，分野によって分けていない．

まず，事前索引句は専門語から構成されると仮定し，1985年1年分の事前索引句の集合から単一語の辞書W_tをつくった．すると，

$$|W_t|=84,690,$$
$$|W_f \cap W_t|=1,700.$$

そこで，1年分の抄録文に生起するW_fとW_tの延べ単語数を調べた．抄録文の数は，314,248であり，全延べ単語数Tは，

$$T=7,088,690.$$

この抄録文集合における，辞書Xに属する単語の延べ単語数を$T[X]$で表わせば，

$$T[W_f]/T=0.5364;$$
$$T[W_t]/T=0.9409;$$

$T[W_f \cap W_t] / T = 0.5151$;

$T[W_f \cup W_t] / T = 0.9621$

である．つまり，W_fの語で，抄録の全延べ単語の53.64%を覆っていること；両方の辞書に出現しない延べ単語は，3.79%であること；W_fの延べ単語数の96.03%がW_tの単語でもあることがわかった．

しかし，$W_f \cap W_t$の単語が文中で機能語として使われているのか，専門語として使われているのかを判断するためには，文を理解する必要がある．約5,000文を読んだ限りでは，$W_f \cap W_t$の単語のほとんどが機能語として生起していることがわかった．また，事後結合句の99%には，W_fの語が生起しないことがわかった．つまり，大雑把にいえば，抄録文に現れるW_fの語は，意味的にも機能語であり，文中の残りの語は専門語である．

7．機能語辞書W_fによる抄録文理解

自動インデキシングの目的は，ＩＲのためのキーワードを自動抽出することである．ＩＲシステムでは，転置索引に各キーワードに関連する文書集合を置き，これらの集合に対するブール演算と隣接演算(adjacency operation)によって検索を行う．したがって，自動インデキシング法の評価は，それを使ったＩＲシステムの性能評価によって行われる．しかし，ＩＲシステムの評価は難しく，特に大規模ＩＲシステムの場合は，非常に困難である．このことがこれまで研究されたより複雑な自動インデキシング法が採用されない理由になっている．つまり，大規模ＩＲシステムにおいて，単純な方法より明らかに優れていることが確認された方法がないのである．また，このことが自動インデキシングの研究を困難なものにしている．いずれにしても，転置索引とブール／隣接演算に基づくＩＲシステムでは，自動インデキシング法の改良によって，システムの性能向上を図るのには一定の限界があるように思われる．ＩＲシステムの性能向上が目的であるならば，人工知能技法による文書理解に努力を傾けるべきであろう．これは，既存のＩＲシステムから文書に基づく知識ベースシステムへの道を開くことになろう．
一方，ＡＩの分野では，古くから自然言語理解の研究が行われてきた．しかし，現在の技術水準では，人間と同等の自然言語理解能力をもつ計算機プログラムをつくることは不可能のようにみえる．したがって，いま実現可能な言語理解プログラムの“理解能力”の程度は，あまり高いものは期待できない．だが，機能が上述の問題を解決するためのものであれば，比較的“浅い理解”で，実用的理解システムを実現しうる可能性がある．理解システムとしては簡単であるとはいえ，これを実現するためには，技術的突破口が必要である．機能語辞書W_fがこの突破口になりうる可能性がある．

ここで，抄録文の理解とは，抄録文からキーフレーズをノードとする意味ネットワークをつくることをいう．意味ネットワークでは，アークのラベルの種類が知識の記述力を決める．この種類は，記述すべき対象があたえられたとき，あらかじめ適当に決めることが普通であるが，これを機能語辞書W_fから抽出しようとするものである．前節の結果はW_fの構文・意味辞書をつくることによって，意味ネットワークのアークのラベル付けが可能であることを示唆している．W_fは，原形が1,143語であり，この程度の機能語数であれば，機能語についてかなり詳細な構文・意味辞書をつくることができる．この方法では，専門用語それ自体の意味は取り扱う必要はないが，文中でのキーフレーズの同定のために，事前結合句における高頻度単語を利用する必要があることがわかっている．

8．むすび

本稿では，INSPECテープから統計的な方法によって得られた否定辞書が人手でつくった否定辞書より性能が良いことを述べ，次いでこの否定辞書に基づく機能語辞書の性質を述べ，最後にこの機能語によ

る抄録文理解の可能性を論じた．この抄録文理解システムは，“浅い理解”しかできないが，現在のＩＲシステムにおける問題を解決できると考えている．

参考文献

1）Salton, G. and McGill, M. J.: Introduction to Modern Information Retrieval, p. 448, McGraw-Hill, New York (1983).

2）Aithison, T. M., Martin, M. D. and Smith, J. R.: Developments towards a Computer Based Information Services in Physics, Electrotechnology and Control, Inform. Storage and Retrieval, Vol. 4, No. 2, pp. 177-186 (1968).

3）Van Rijsbergen, C. J.: Information Retrieval (2nd ed.), p. 208, Butterworths, London (1979).

4）二村祥一，松尾文碩：英文科学技術文献情報に対する不要語除去法による自動索引，情報処理学会論文誌，第28巻，第7号，pp. 737-747 (1987).