2－3　自然言語処理による大気環境研究の歴史的分析

片谷　教孝、若月　玲（山梨大学工学部）

1.　はじめに

科学技術の進歩を振り返り、各分野ごとにその歴史的な流れを知ることは、その分野の研究者にとって今後の研究の方向づけの参考となるばかりでなく、歴史学的な関心の対象でもある。科学史はそれ自体が科学の一分野として重要な位置を占めるとともに、他の分野の科学に対して、その考え方や進むべき方向について重要な示唆を与えてきたといえる。特に環境問題のような社会問題の場合は、問題の本質が自然科学のみならず、人文・社会科学の分野に根ざしている場合も少なくないため、それに対する研究のアプローチを歴史的に分析することが問題の解明に役立つことも考えられる。

科学史は一般には人文科学の一分野に位置づけられる場合が多く、文系の世界と見られる場合が多いが、上記のような理由から、自然科学の研究者にとっても、科学史をたどることは有益な面が多いと考えられる。例えば土木学会では、土木史が一つの研究分野として古くから認知されてきており、過去の土木技術の遺構を分析して現代の技術との比較を行うなどの方法によって、技術の評価や新たな技術のヒントを得るなどの成果が上げられている。

本研究は、大気環境分野における研究の歴史的な流れを、論文タイトルから自然言語処理の技法を用いて分析しようとするものである。筆者らは科学史の専門家ではなく、また大気環境研究に携わった期間もそれほど長いわけではない。したがって従来からある科学史の研究技法ではなく、情報処理の分野で最近進歩が著しい自然言語処理の技法を採り入れて、新しい試みを行っている。本報ではその第一段階の結果を報告する。

2.　分析方法

一般に科学史において時系列的な流れを分析・把握しようとする場合は、100 年以上のスケールを対象にする場合が多く、その手法は多くの文献資料を参考に、分析者の判断によって前後のつながりを整理していく形がとられる。本研究の方法はそれらと全く異なり、論文タイトルに現れる単語の頻度をカウントすることによって、頻度の高い単語がその時代における注目度の高さを示すとみなすものである。この方法は、特定の学会誌など、同一分野である程度統一された編集形態で継続的に刊行されてきたものに対象が限られるため、長期的な分析には適していないが、分析者の主観がはいる余地が少ないため、より客観的な分析が可能となる利点がある。

ここでは大気環境学会誌の第１巻から第30 巻までに収録された論文のタイトルを対象とし、自然言語処理の形態素解析の手法によって単語を切り出した。自然言語処理の手法の概要と本研究で用いた方法の詳細については、次章に示す。

3.　自然言語処理の概要

近年になって自然言語処理の研究が進み、その手法についての解説書も多く出版されるようになってきた。本研究は自然言語処理の手法自体を研究対象としているわけではないが、本研究で用いた手法の理解の助けとするために、本章では自然言語処理の技術的な概要について述べる。

3.1 自然言語処理の歴史

初期のコンピュータにおいては、扱える言語はコンピュータが理解できる機械語と、翻訳によって機械語に置き換えることができる

コンピュータ言語（高級言語）に限られていた。このことは、コンピュータの利用範囲を限られたものにするとともに、それらの言語を習得した人間のみがコンピュータを利用できるという状況を生み出していた。しかし、コンピュータ技術の進歩とともに利用範囲が広がり、人間のコミュニケーションの道具としても用いられるようになってくると、人間が日常的に使用する言語をコンピュータ上で使用できることが、徐々に必要条件となってきた。その最も顕著な例がワードプロセッサである。また日常的に使用する言語の相互間の翻訳をコンピュータ上で行う、いわゆる機械翻訳も、ここ 20 年ほどの間に急速に実用の域に達してきている。また、近年のインターネットの普及にみられるように、誰もがコンピュータにふれるような状況のもとでは、むしろコンピュータ言語を使用する局面のほうが特殊であるといえるようになってきつつある。

3.2 自然言語処理の構成要素と本研究で利用する要素

自然言語処理は、数多くの要素技術の集合である。主な要素としては、辞書、シソーラス、文法、形態素解析、かな漢字変換、構文解析、意味解析、文脈解析、文生成などがある。本研究で用いている方法は、自然言語処理全体からみれば、ほんの一部分にすぎない。具体的には辞書と形態素解析である。以下ではその内容を詳しく述べる。

3.3 形態素解析

形態素解析とは、通常の言語を構成する外見的な要素（語）に分解するための手法であり、自然言語処理の第一段階として用いられる。ほとんどの語を網羅する辞書が用意されている場合には、形態素解析は文を辞書に登録された語に分解することに限定される。それでも日本語の場合には、英語などの多くの言語と異なり、語と語の間に空白のような形態的な切れ目がないので、何らかの方法で語を切り出さなければならない。一つの文をそれを構成する文字の集合と考えれば、語はその部分集合（部分文字列）であり、文から切りだされた語の集合は、もとの文と一致することが要求される。

日本語の場合は、接頭語・接尾語の存在や、活用による語尾の変化があるため、それらによる語のバリエーションがすべて辞書に登録されていない限り、語処理によって辞書との対応を図らなければならない。これも形態素解析の重要な部分の一つである。

語の切り出しにおいては、文法的なルールのほかに、ヒューリスティクスも用いられる。最も典型的なヒューリスティクスは、連続する漢字は一つの語である場合が多いというものである。本研究の方法も基本的にこれにしたがっている。このようなヒューリスティクスは他にも数多く存在するが、ここでは省略する。

4. 本研究における形態素解析の方法

本研究において採用した方法は、前述のヒューリスティクスにしたがった語切り出しと、辞書とのマッチングによる切り出しの併用である。ただし本報で報告するのは、ヒューリスティクスのみを用いた結果である。具体的な方法としては、図１に示すように、漢字またはかなの２文字以上で構成される名詞（固有名詞を含む）を切り出して、出現頻度をカウントした。もちろん１文字のみから成る語もあるので、２文字以上とすることによる検索もれも生じる。しかし試行的分析の結果から、１文字の語を切りだすと切りだされる語数が多くなりすぎて、分析に支障が生じることがわかったため、今回は２文字以上とした。そして時系列的な変化を見るために、上記期間を 10 年ずつ３期に分け、期間ごとの頻度を求めた。また、同義語は合算して集計したが、この同義語の判定は、現在のところ筆者らの経験的な判断によっている。

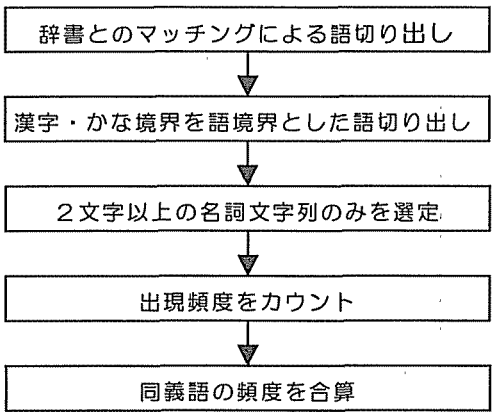

図１　本研究における語切り出しの方法

なお、論文誌に掲載された論文には、多くの場合キーワードが付与されている。このキーワードは論文の内容を要約的に示す単語または熟語であるので、本研究のような目的に利用できる可能性もある。しかし今回は以下のような理由により、キーワードは利用しなかった。

・初期の大気汚染学会誌にはキーワードがない
・キーワードの語統制が行われていない。
・キーワードは１論文あたり５語程度であり、統計的な分析に必要な語数が得られない。

5.　分析結果

ここでは大気環境学会誌（旧・大気汚染研究）の第１巻から第 10 巻（1966～1975 年度）を第１期、第 11 巻から第 20 巻（1976～1985 年度）を第２期、第 21 巻から第 30 巻（1986～1995 年度）を第３期とする。表１は期別の集計結果を示したものである。また図２は、主要な単語について、時系列的な出現頻度の変化を論文 100 件あたりの出現頻度により示している。

第１期は昭和 40 年代の公害問題最盛期にあたる。この時期にはＣＯやＳＯ2 などが特に研究上の注目を集めていたことがわかる。第２期は昭和 50 年代で、窒素酸化物やオゾン、粒子状物質などに重点が移行していることがわかる。また手法面では、モデルが増えてきている点が注目される。

第３期は昭和 60 年代から平成にはいり、公害から環境創造への転換や、地球環境問題への注目などが特徴とされる時代であるが、大気環境研究にみる限りでは必ずしもそれらの傾向は現れておらず、わずかに酸性雨が新たに登場している程度である。

表１　論文タイトルから切り出された主要な単語

期別	第1期(1966～1975)		第2期(1976～1985)		第3期(1986～1995)	
総論文数	160		494		515	
主要な単語と出現回数	影響	39	影響	42	影響	48
	大気汚染	25	光化学	42	ガス	36
	ガス	21	大気中	42	大気汚染	36
	CO	21	研究	41	NO2	34
	SO2	13	ガス	36	モデル	28
	研究	13	モデル	33	オゾン	24
	大気中	13	大気汚染	31	エアロゾル	24
	環境基準	11	NO2	24	研究	24
	測定	11	オゾン	22	検討	23
	オキシダント	9	粉じん	22	評価	22
	NOx	8	NOx	21	粉じん	21
	光化学	8	エアロゾル	19	大気中	21
	NO2	7	SO2	16	オキシダント	19
	ばいじん	7	評価	16	濃度	19
	エアロゾル	7	オキシダント	15	拡散	16
	オゾン	6	濃度	14	酸性雨	15
	AIR	5	関係	13	SO2	14
	酸化物	5	検討	13	測定	14
	大気汚染物質	5	CO	13	汚染	13
	HEALTH	4	測定	12	現状	12
	いおう	4	解析	11	ディーゼル	11
	がん	4	昭和	11	推定	11
	アサガオ	4	大気汚染物質	11	スギ	10
	モルモット	4	拡散	10	解析	10
	関係	4	炭化水素	10	イオン	9
			分析	10	挙動	9
					分析	9

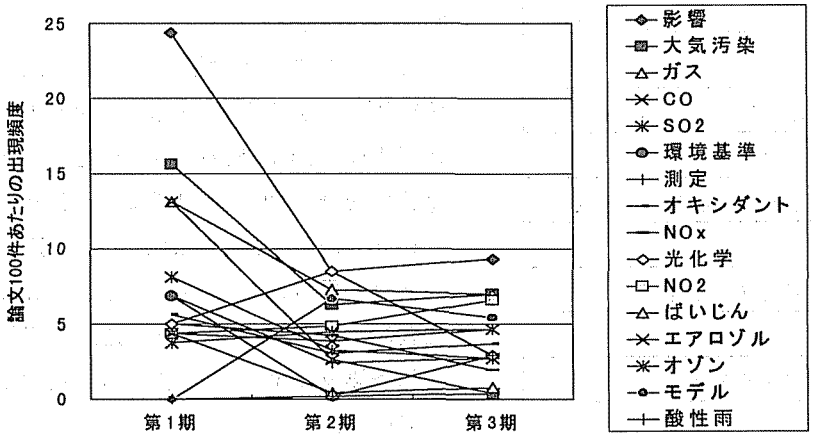

図２　主要な単語の出現頻度の時系列的変化

6.　おわりに

本研究はまだ初歩的な段階にあるが、それでも大気環境分野における研究の歴史的な流れを断片的には解明することができた。今後は語切り出しの方法の改良、辞書を用いた切り出し、単語分類方法の検討、他の歴史的事実との時間的整合関係の分析などによって、さらに詳細な分析を行う予定である。

謝辞:

本研究の実施にあたっては、山梨大学工学部平成９年度修士課程学生・石　真和君（現松下システムエンジニアリング（株）勤務）の協力を得ました。同君に深く感謝します。

参考文献:

野村浩郷；自然言語処理の基礎技術,(社)電子情報通信学会，1988.