

コホモロジーで意味解析 Semantic analysis using co-homology

□コホモロジーとは、もともと []¹と呼ばれる図形（のような）の性質を調べるためにアレクサンドル・グロタンディークが考案した計算手段です。特に有名なヴェイユ予想を証明するための道具としても使われています。この道具を言語の意味解析をするために利用しようと考えたわけです。それはテキスト言語の意味を抽出するには、形態素の意味概念だけではなく、係り受けを含めた構文解析や照応解析での結束構造や文としての知識をなす最小単位、そして文脈解析での文間関係や []²、その上文章や文書としての主題や話題に対する意味の概念が各層別に存在するためである。しかし、ただ単にそのまま利用可能な程、言語空間もコホモロジーも簡単なものでなく、いろいろ工夫が必要になります。

言語空間を考える時、まずテキスト文書を我々は考えます。文書を集合と考えると、その部分集合は文章（段落のような…）になる。その文章の部分集合が文であり、その部分集合が連文節である。そのまた部分集合が []³であり、そのまたまた部分集合が形態素になる。形態素の部分集合が文字になり、それが文書という全体集合の要素であることは自明なことである。ここでは意味解析を考えていくので、日本語や中国語などのように文字自体に意味を含んでいたとしても、他言語との共通手法ということも考え、文字を最小単位とはせず、「意味のある最小単位」である []⁴を集合の要素とする。

上記で述べた文書から形態素までの部分集合への分割化は、視点を変えると解析学の微分と似ています。文書を n 次元多様体と考えると、それを微分 ∂ という概念で次元を下げることによって文章になる…とする。そうすると以下も同じように形態素まで微分 ∂ という作用子（関数）で分割階層化ができる。微分積分（解析学）で習った微分概念を集合など別なものでも使う…というアイデアが大事です。（株価のグラフを微分すると…局所な上昇、下降率が出る…など。）また、この関係を写像と考えると、微分 ∂ は写像 ∂ : 文書 \rightarrow 文章となり、表記を変えると ∂ (文書) = 文章のことで、 ∂ は写像の関数や作用子になる。

ここでこの []⁵がどのような性質を持っているかという、文書 \supset 文章 \supset 文 \supset 連文節 \supset 文節 \supset 形態素という各集合間での制限写像と全射という性質を持ち、また始域で定義された演算（係り受け関係など）を終域へ写すことができれば準同型写像という性質を持つことになる。文書を n 次元多様体としたが、文書の深層レベルでの基底が n 次元であることを意味しており、また意味レベルで意味概念の要素が既約単語（素語）の族集合と考えると開集合同士の貼り付けされたものとみることができ、これは n 次元位相多様体である。すなわち、各階層の要素（開集合）どうしを貼り合せて作った n 次元位相多様体の次元を

¹ 多様体：ここでは有限位相多様体を指す。manifold

² 結束性：文間に於ける意味レベルでの概念関係。coherence

³ 文節：係り受けは文節単位である。dependency relation of clause

⁴ 形態素：意味ある最小単位。一文字や助詞、副詞も形態素なので単語とは区別する。

⁵ 写像：ここでは写像という狭い概念ではなく、もっと広い射という概念を考えている。

下げる写像を []⁶ と考えることによって、以下に示すようなコホモロジーの計算手法が使えるような階層化されたものになる。では、少しコホモロジーとは何か…をみていきましょう。

co-homology とは、 $H^i(X, R) = R^n$ という X が加群や多様体、層などの場合、実数空間 R で表現し、 X が「 i レベル」で n 次元である…という解が得られるのがコホモロジーの意味である。代表的な例でいうと、球面なら「 i が 1 で解が 0 次元」、トーラスであれば「 i が同じく 1 で解が 2 次元」となる。ここで「 i が 1」という意味は、球面もトーラスも []⁷ という 1 次元のもので「どう書けるか」を意味している。球面は、南極と北極のような特異点…といっても地球のように特定されない…がどこでも特異点が存在し、線で書いた輪は特異点で消滅してしまうので、0 次元となる。片一方、トーラスは、ドーナツの穴と同じ方向にひとつ線がひけて、輪切りの輪でもうひとつ線がひける。当然、その線は移動しても特異点がないので消滅しない。すなわち、2 次元となる。

ここで次元をもう少し考察をしよう。代数やベクトルなら「次元」とか「基底」というが、群なら「ランク (標数)」といい、トポロジーや位相多様体などでは「開基」で、コホモロジーなどでは「基」という。後者の方がより一般性があり、広い意味を持つ。すなわち、次元や基底、基と定義されているものは、集合や群そして多様体などを []⁸ 基準であると考えることができる。我々の住む世界では、長さ・角度・面積・体積・曲率などが測るもので、これを計量という。形式的な媒介変数 (パラメータ) によって定まる長さをリーマン計量といい、その測る方法を定めた多様体をリーマン多様体というものがあるが、測る対象が違う位相空間上では、もっと広くそして離散による基準がないと言語の意味は測れない。この次元という概念からより広く一般性のある基へ導いていくわけだが、次元を減らすという概念もここまでくると狭い概念であったことが判ると思います…が。

もうひとつ、 X である加群や多様体や層などの対象物とその関係だが、これはなんでも良いというわけではない。何故かと云うと、 i というレベルが必要であるので、ある対象物を準同型写像という []⁹ で対象物の系列 \rightarrow を作り、その系列の順番を意味するものが i で、代数的には整数 $i \in \mathbb{Z}$ で表すが、ここでは文書から形態素までの各層を示すので $H^i(D, B)$ と表す。 R が B になったのは、**Basis** のことで、基がこのホモロジーの解であることを示している。解析的概念で作る ∂ (制限写像) や δ (包含写像) の準同型写像などがよく教科書に出てくるものである。すなわち、準同型写像とは、同型写像である始域から終域へ写像する恒等的な同相写像を含み、制限写像や包含写像という階層化された写像で始域の構造を保持しながら終域へ写すものが必要になる。

$H^i(D, B)$ を取ると解はどうなるか…を考えてみよう。ここで D はある文書で、 B は基

⁶ 準同型写像 : **homomorphism** といい、始域の構造を保ち終域へ写像するもの。

⁷ 線 : 0 次元が点で、1 次元が線、そして 2 次元が面。

⁸ 計る : 測るや量るとい言葉の意味すべてを含む。

⁹ 射 : 関数より写像、写像より射の方が概念が広い。

であり、 i は文書～形態素のことである。言語空間を位相空間として考えるので []¹⁰とは云わず、基と称する。この基とは、各対象物やそれらの射並びに関手の意味的概念で、意味タグで表されている。この意味的概念を求めることがコホモロジーの目的である。

ここでは i には、文書、文章、文、連文節、文節、形態素（、文字）という言葉が入ることになる。 $H_{\text{文書}}(D, B) = B_{\text{文書}}$ となり、文書の基を求めることになる。また、文書から文章への準同型写像は、 $\partial_{\text{文書}}$ と書き、 $\partial(\text{文書}) = \text{文章}$ となる。前述で、線で「どう書けるか」が、球面やトーラスを解析した結果であると述べたが、文書を文章でどう表されるか、文章を文でどう表されるか、文を連文節でどう表されるか、連文節を文節でどう表されるか、文節を形態素でどう表されるか…が、意味を解析することになる。そして、その解が各々の階層で基として表現されることになる。それを表すのが各階層で分かれた110種類以上もある []¹¹である。

最後に、球面とトーラスの解析を線で行ったが、それでは面ではできないのか…という疑問が湧く。球面とトーラスは共に面である2次元なので同じ答えが返ってきて解析をした意味がないので、ここでは正三角錐を考えてみよう。正三角錐を面で解析すれば、解は4面と出る。では、その面を線で解析すれば、3と出る。では、その線を点で解析すれば、2と出る。同じ正三角錐でも面や線、点という違った次元のもので解析すれば、違った答え4、3、2と出ることが判った。言語解析も同じで、文書を文章から文字までの各系列で解析すれば、違った答えが返ってくる。それが意味解析である。形態素ごとに意味があり、その順序という []¹²で意味が違ってくるのは誰しもが知っているが、その「並べ方」すなわち「係り受け」で意味が変わることが意味を解析する重要なキーワードになる…ということだけでなく、文書から形態素までの各層での意味の概念も各々相違する。従来の言語解析や検索エンジンでは形態素だけで意味を解析しようとして限界に陥っている。係り受けも含めて意味解析をしようという傾向が出てきたことは嬉しいが、形態素解析の []¹³で係り受け解析をしても、上記の通り、意味がないことが判るでしょう。係り受けは文節という単位で成されるので、文節層の基で意味解析が表現されなければならない。文節層での基とは何か…とは、深層格であるが、従来の深層格は意味を表現するには不十分である。当社オリジナルの「意味タグ」は、意味概念の定義されたセットを階層的に持っており、大規模辞書を使った解析手法とは違い、エコなリソースを目指している。エコとは、軽く、速く、精度が良いことである。【第5版】

¹⁰ 次元：標数や基など空間が違えば次元の言葉も概念も変わる。

¹¹ 意味タグ：Semantics(株)が発案したテキストの意味を表現する記述言語関係子。

¹² 構造：文書の構造には各層の相違する構造がある。

¹³ 延長線上：形態素解析と構文解析、そして文脈解析や意味解析は意味レベルが違う。