TR-I-0218

言語データベースから抽出した知識データの分布

Distribution of Knowledge Data Extracted from a Language Data Base

江原暉将
Terumasa EHARA

1991.05.31

内容梗概

自然言語処理のための言語学的知識を構築する方法として、母国語話者による内省に基づく方法と、言語データベースから知識を抽出する方法がある。後者の方法は客観性に優れ、定量的な知識を得ることができる反面、内省では当然現われると考えられる知識が言語データベース中に全く存在しないか、存在してもその度数が小さいという問題がある。これは、データベースの過少性と呼ばれる。本報告では、知識をその知識に含まれる単語数で分類し、さらに、単語の分布にZipfの法則を仮定して、知識データの分布を求める。さらに、処理対象からの標本として言語データベースを構築した場合、処理対象中で未知知識データが出現する割合である未知データ率について考察し、データの過少性を具体的に示す。

Abstract

There are two methods to construct linguistic knowledge-base for natural language processing. One method is to construct the knowledge-base by the native speaker's introspective evaluation and the other method is to extract the knowledge from langage database. The latter is more objective and quantitative, but has less-data problem. This is the fact that there are many knowledge data which is supposed to appear in the database by the native evaluation but actually they do not appear at all or appear unfrequently. Linguistic knowledges can be categorized by the number of words included in the knowledge. Using this categorization, the distribution of knowledge data extracted from language database is caluculated assuming that the distribution of words is Zipf's distribution. The coveradge of knowledge-base constructed from language database which is considered a sample from a population is computed.

1．はじめに

いかなる技術にしても、その技術を達成するには、対象にしようとしている自然現象を詳細に知ることが必要であり、自然言語処理技術の場合も例外ではない。自然言語処理が対象とするものは人間の用いる言語であるから、言語現象に関する知識を得ることが必要となる。従来、これらの知識は、主として、母国語話者の内省に頼って構築されてきたが、内省には揺れがあり、また定量性に欠ける欠点があった。一方、言語データベースを利用して、知識を構築する方法は、客観性に優れ、定量的な知識を得ることができる利点がある。しかし、言語データベースを用いた言語研究には、以下の問題点があることが指摘されている［Atkinson］。それは、データの過大性と過小性である。前者は、同一のデータがデータベースの中で繰り返し現れることを指し、後者は、内省によれば当然現れると思われる言語現象が全く現れないか、現れても少数であることを指す。この問題点は、データベースを用いた自然言語処理システムにも当てはまる。前者は、データベースの効率が悪さであり、後者はデータベースから抽出された知識の被覆率が小さいことである。

本報告では、言語データベースから知識を抽出する場合に、知識データがどのように分布するかを考察し、上記問題点を検討し、解決法を探る。

2．言語処理における知識データの種類

本節では言語知識を単語数と言語数によって分類し、どの様な知識がありうるかを示す。

知識は、まず言語数によって、１言語の知識と２言語の知識に分けられる。（多言語の知識も考えられるが、ここでは、最大２言語に限る。）１言語の知識は、日本語なら日本語だけの知識であり、２言語の知識は、例えば日本語と英語の対応知識である。ＡＴＲ対話データベース（ＡＤＤ）［江原］は１言語と２言語の知識を含んでいる。次に、知識が幾つの単語によって構成されているかで分類する。単語数１の知識は、いわゆる単語辞書である。また、単語数２の知識は、例えば、単語間の係り受けの知識である［柿ケ原］［江原２］。［Sumita］で利用している「名詞＋の＋名詞」の知識は、３個の単語から成るが、「の」の部分は固定しているので、これも単語数２の知識と見なせる。一般に機能語は各種の言語でその数が限られているため、ここでは、単語として、内容語のみを考え、機能語は単語数の計算には含めない。さらに複合語も全体で１語と見なす。そうすると、単語数と文節数は同じものになる。そこで、文節間の係り受け知識は、単語数２の知識となる。次に文に関する知識について考える。文は長さに長短があるので、単語数（文節数）を幾つと決めることが出来ない。しかし、述語文節が１つである単文に限れば、文節数は、格要素文節数＋１となる。ここで言う格要素としては、自由格や副詞も含んでいる。格要素文節数は１ないし４ぐらいであるから、単文の知識は、単語数２ないし５くらいの知識と考えられる。平均的には単語数３程度ではないだろうか。次に文と文との係り受け関係の知識は、２文間の関係であるから、平均的には、単語数６程度の知識と考えられる。［工藤］で研究されている単文間の関係を用いた結束性の算出方式はこの種の知識の利用と考えられる。また、文節間の係り受けとその対訳を利用した翻訳方式［幸山］は言語数２単語数２の知識を利用している。また、いくつかの翻訳単位を定めた、パターン方式の翻訳手法［古瀬］［Witkam］は、最大、言語数２の文の対応知識を利用している。［佐藤］でも、単文を翻訳単位としているので、言語数２の文の対応知識である。以上をまとめると、表２．１が得られる。

表２．１　知識データの言語数と単語数による分類

言語数	単語数	知識の例
1	1	単言語の単語辞書
	2	単語（文節）間係り受け辞書［柿ケ原］［江原２］
	3	単文の辞書
	6	単文間の関係の辞書［工藤］
2	1	対訳単語辞書
	2	係り受け対の対訳辞書［Sumita］［幸山］
	3	単文の対訳辞書［古瀬］［Witkam］［佐藤］
	6	単文の対の対訳辞書

３．知識データの分布

ある言語データベースの中での単語データの分布はZipfの法則か、それを詳細化したMandelbrotの法則［Mandelbrot］に、近似的に従うことが経験的に知られている。これらの法則は次のように表される。ある言語データベースの中の単語の異なり数をＭとする。各異なりデータを度数の大きい順に並べたときの順位をｒとすると、第ｒ位の語のデータベース内での度数ｎ（ｒ）は次の式に従う。ここで、ａ、ｂ、ｃは定数である。

$$n(r) = a / r \qquad (3.\ 1)$$

$$n(r) = a / (r + c)^{b} \qquad (3.\ 2)$$

（３．１）がZipfの法則であり、（３．２）がMandelbrotの法則である。順位Ｍのデータの度数は１であるから、（３．１）では、ａ＝Ｍとなる。また、（３．２）では、$a = (M + c)^{b}$となる。単語数１の言語知識の分布の例として、図３．１にＡＴＲ対話データベースから抽出した国際会議電話対話の単語データに対する順位と度数のグラフを示す。この時の異なり語数Ｍと延べ語数Ｎは

Ｍ＝　　４，７９７
Ｎ＝１７１，３０３

である。

では、単語数２の知識データはどのような分布に従うであろうか。単語数１の知識データの分布がZipfの法則に従うと仮定して、単語数２の知識データの分布を求めよう。単語数２のデータの例として、係り受けデータを考察し、片方の単語を係り元、他方の単語を係り先と呼ぶ。しかし、これは１例であり、単語数２のデータであれば、以下の議論は一般的に成立する。例えば、２語組（バイグラム）の分布にも当てはまる。さて、単語数１の知識データの分布はZipfの法則に従うとしたから、係り元および係り先の分布は共にZipfの法則に従う。この時、係り受けデータの分布を求めよう。係り元、係り先の異なりデータ数を共にＭと仮定する。すると、順位１の係り元、係り先の度数ｎ（１）は（３．１）式から共に

$$n(1) = M$$

となる。ここで、順位ｒ１位の係り元が順位ｒ２位の係り先に係る係り受けの度数をｎ（ｒ１，ｒ２）と書き、

$$n(r1, r2) = \frac{a}{r1 * r2} \qquad (3.\ 6)$$

が成立すると仮定する。これは、係り元と係り先が独立に生起すると仮定したときの係り受けの度数である。

$$r = r1 * r2 \qquad (3.\ 7)$$

と置いて、この係り受けデータとしての順位をＲ（ｒ）とする。これを求めよう。まず、（３．６）の右辺分母の値がｒとなるｒ１、ｒ２の組合せの数ｋ（ｒ）を求める。Ｎを自然数の集合として、ある自然数ｒに対して、集合

$$D(r) = \{(r1, r2) \mid r1 * r2 = r,\ r1 \in N,\ r2 \in N\} \qquad (3.\ 8)$$

を定義すると、＃（Ａ）で集合Ａの要素数を表わすとして、

$$k(r) = \#(D(r)) \qquad (3.\ 9)$$

となる。ｋ（ｒ）はｒを素因数分解

$$r = p_1^{n1} * p_2^{n2} * \cdots * p_L^{nL} \qquad (3.\ 10)$$

したとき、

$$k(r) = (n1 + 1) * (n2 + 1) * \cdots * (nL + 1) \qquad (3.\ 11)$$

となる。このとき、

$$\Sigma[1 \leqq i \leqq r - 1]\ k(i) < R(r) \leqq \Sigma[1 \leqq i \leqq r]\ k(i) \qquad (3.\ 12)$$

が成り立つ。ここで、Σ［ｐ（ｉ）］ｆ（ｉ）は述語ｐ（ｉ）が成立するｉの範囲に亘る関数ｆ（ｉ）の和を意味する。ここで、（３．１１）式のＬがｒの対数に比例すると仮定し、各（ｎｉ＋１）は２以上であるから、これを２で近似すると、ｃ、ｃ’を定数として、

$$k(r) \fallingdotseq 2^{c' \cdot \log(r)} = c * r \qquad (3.\ 13)$$

となる。（３．１３）が近似的に成立すると仮定する。すると、（３．１２）は

$$c((r-1)^2/2 + (r-1)/2) < R(r) \leqq c(r^2/2 + r/2) \qquad (3.\ 14)$$

となる。そこで、

$$R(r) \fallingdotseq c'' * r^2 \qquad (3.15)$$

が成立する。これを逆に解くと、係り受けの順位Rのデータに対する、rの値は、

$$r = c''' * R^{1/2} \qquad (3.16)$$

となる。そこで、係り受けの順位Rのデータの度数は、

$$n(R) = a / R^{1/2} \qquad (3.17)$$

となる。つまり、係り元、係り先の分布がZipfの法則に従う場合、係り受けの分布は、$b = 1/2$、$c = 0$とおいたMandelbrotの分布に近似的に従うことが分かる。実際、図3．2に示す、係り受けデータの分布は上記近似の妥当性を示している。この議論を拡張すると、以下の性質が予測される。

性質1　言語データベースから単語数nの知識データを抽出したとき、この知識データベースの度数分布は

$$b = 1/n, \quad c = 0$$

のMandelbrot分布で近似される。　（性質終）

4．知識データベースの被覆率

言語データベースから知識データを抽出して、知識データベースを作成した場合、このデータベースが処理対象に含まれる知識を十分覆っていることが必要である。例えば、言語データベースから単語を抽出して、単語辞書を作成した場合、処理対象に含まれる単語をどれだけ含むかが問題となる。もし、単語辞書に含まれない語が処理対象に存在すると、未知語となり、処理が不成功となることが予測される。本節では性質1を仮定して、この問題を考察する。

言語データベースを処理対象からの無作為標本とする。処理対象に含まれる知識の全体をDPとし、その（延べ）要素数をNPとする。また、言語データベースから抽出された知識の全体をDとし、その（延べ）要素数をNとする。DPの異なり要素の分布が、$c = 0$のMandelbrot分布に従うとすると、異なり要素数をMPとして、

$$NP = \Sigma\,[1 \leqq r \leqq MP]\ \{a / r^b\} \qquad (4.1)$$

が成立しなければ成らない。また、$r = MP$において、$n(r) = 1$となると仮定すると、

$$1 = a / MP^b$$

であり、これから

$$a = MP^b$$

となる。そこで、

$$n(r) = (MP/r)^{b} \qquad (4.\ 4)$$

となる。一方、異なり要素の度数は整数値を取らなければならないから、順位rが集合

$$Rm = \{r \mid m \leqq (MP/r)^{b} < m+1\} \qquad (4.\ 5)$$

の要素である場合は、

$$n(r) = m \qquad (4.\ 6)$$

が成立すると仮定する。これは、図3．1、図3．2から妥当な仮定であることが分かる。Rmの要素数＃（Rm）は、

$$\#(Rm) \fallingdotseq MP * \frac{(m+1)^{1/b} - m^{1/b}}{(m+1)^{1/b} * m^{1/b}} \qquad (4.\ 7)$$

となる。m＝1つまり度数1の異なり要素数は、

$$\#(R1) \fallingdotseq MP * \left(1 - \frac{1}{2^{1/b}}\right) \qquad (4.\ 8)$$

となる。b＝1の時は、総異なり数の1／2が度数1であり、b＝1／2の時は、3／4が度数1である。このように、言語データベースから抽出した知識データには度数の少ないデータが大量にあることが示された。これは、データの過少性を意味しており、この点に対する対策が必要である。

言語データベースDは母集団DPの無作為標本であるから、DPのある要素がDに存在する確率は、

$$p = N/NP \qquad (4.\ 9)$$

である。そこで、DPでの度数がmの異なり要素がDにサンプルされない確率は

$$p(m) = (1-p)^{m} \qquad (4.\ 10)$$

となる。Dにサンプルされない要素は未知データとなる。DPでの度数がmの要素のうち、未知データとなる要素の延べ度数の平均をf（m）と書くと、

$$f(m) = m * \#(Rm) * p(m) \qquad (4.\ 11)$$

となる。そこで、未知データの総延べ度数の平均fは

$$f = \Sigma\ [1 \leqq m \leqq MP]\ f(m) \qquad (4.\ 12)$$

となる。DPの総要素数に占める未知データの割合を未知データ率と呼び、uと

おくと、

$$u = f / NP \qquad (4.13)$$

となる。p（m）はmの巾のオーダーであり、#（Rm）はmの多項式のオーダーであるから、f（m）はmの巾のオーダーで小さくなる。そこで、MP'をMPより小さい適当な数として、

$$f \fallingdotseq \Sigma\,[1 \leqq m \leqq MP']\ f(m) \qquad (4.14)$$

と近似できる。

N／NP＝1／2とした時の、b＝1と1／2に対するf（m）の値を表4．1に示す。ここで、MP'＝5と仮定した。ただし、#（Rm）とf（m）は表中の数値にMPを乗じた値である。

表4．1　未知データ数の値

b	m	#（Rm）	p（m）	f（m）
1	1	1／2	1／2	0．250
	2	1／6	1／4	0．083
	3	1／12	1／8	0．031
	4	1／20	1／16	0．013
	5	1／30	1／32	0．005
1／2	1	3／4	1／2	0．375
	2	5／36	1／4	0．081
	3	7／144	1／8	0．018
	4	9／400	1／16	0．006
	5	11／900	1／32	0．002

表4．1より、fの値は、

$$f = 0.382 * MP \qquad (b = 1) \qquad (4.15)$$
$$f = 0.482 * MP \qquad (b = 1/2) \qquad (4.16)$$

となる。一方NPは次の様にして求まる。（4．1）（4．4）式より

$$NP = \Sigma\,[1 \leqq r \leqq MP]\ (MP / r)^{b} \qquad (4.17)$$

である。この式の右辺を度数mがMP'より小さい部分と大きい部分に分ける。

$$R' = \Sigma\,[1 \leqq m \leqq MP']\ \#(Rm) \qquad (4.18)$$

とおくと、

$$NP = \Sigma\,[1 \leqq r < MP - R']\ (MP / r)^{b} + \Sigma\,[1 \leqq m \leqq MP']\ m * \#(Rm) \qquad (4.19)$$

となる。（４．１９）式の右辺第１項をＮＰ’と第２項をＮＰ’’とおく。ＮＰ’を積分で近似すると、

$$NP' \fallingdotseq \int [1 \leqq r \leqq MP - R'] \ (MP/r)^{b} \ dr$$

$$= MP^{b} \frac{1}{1-b} \{ (MP - R')^{1-b} - 1 \} \qquad (b \neq 1)$$

$$MP * \log (MP - R') \qquad (b = 1)$$

（４．２０）

となる。ｂ＝１と１／２について、計算する。

$$MP' = 5 \qquad (4.21)$$

と置く。表４．１より

$$R' = \begin{cases} MP * 0.833 & (b = 1) \\ MP * 0.972 & (b = 1/2) \end{cases} \qquad (4.22)$$

となる。そこで、（４．２０）式より

$$NP' = MP * \log (0.167 * MP) \qquad (b = 1)$$

$$MP^{1/2} * \frac{1}{2} \{ (0.028 * MP)^{1/2} - 1 \} \qquad (b = 1/2)$$

（４．２３）

一方、ＮＰ’’は（４．２１）式の下で

$$NP'' = \begin{cases} 1.450 * MP & (b = 1) \\ 1.325 * MP & (b = 1/2) \end{cases} \qquad (4.24)$$

となる。そこで、

$$NP = MP * \log (0.167 * MP) + 1.450 * MP \qquad (b = 1)$$

$$MP^{1/2} * \frac{1}{2} \{ (0.028 * MP)^{1/2} - 1 \} + 1.325 * MP$$

（ｂ＝１／２）

（４．２５）

となる。ＭＰとＮＰの関係をｂ＝１の場合について、表４．３に示す。ｂ＝１／２の場合は、表４．３の範囲で、未知データ率は０．７程度であまり変化しない。

表４．３　異なりデータ数と延べデータ数、未知データ率の関係
（ｂ＝１）

ＭＰ	ＮＰ	u
$1*10^3$	$6.6*10^3$	0．１２６
$1*10^4$	$8.8*10^4$	0．０９４
$1*10^5$	$1.1*10^6$	0．０７５
$1*10^6$	$1.3*10^7$	0．０６２
$1*10^7$	$1.6*10^8$	0．０５３

実際には、ＭＰが増えるに従って、ｂの値も大きくなるので、未知データ率はこれらの値より小さくなる。表４．４に実際のデータによるＮ／ＮＰ＝１／２の時の未知データ率を示す。ただし、上記議論では、処理対象を母集団とし、言語データベースを標本としたが、表４．４では、言語データベース自体を母集団とし、そこから、抽出率１／２の標本を抽出する場合を仮定した。ＡＤＤとＡＰのデータは単語データであるから、ｂ＝１でモデル化でき、係り受けのデータはｂ＝１／２でモデル化できる。しかし、いずれも、表４．３やそのすぐ上の記述の値よりも実際のデータの方が小さくなっている。ここで、ＡＰ電のデータは［浦谷］によった。また、係受けのデータは新聞データベース［江原２］によるものである。

表４．４　実際の言語データに対する未知データ率

データ名	ＭＰ	ＮＰ	ｆ（１）	ｆ（２）	ｆ（３）	u
ＡＤＤ	2,875	67,000	1,146	420	222	0.027
ＡＰ電	75,000	3,500,000	25,000	?	?	0.007
係り受け「を」格	101,666	167,281	76,166	16,100	3,600	0.576

５．　おわりに

言語データベースから言語知識を抽出して、知識データベースを作成するときのデータの分布について考察し、度数の少ないデータが大量に存在する、いわゆるデータの過少性を示した。

この過少性に対処するためには、なんらかの単語のカテゴリー化が必要であろう。その１方法は［江原２］で提案されている。

参考文献

[Atkinson] Martin Atkinson et al.: Foundations of General Linguistics, P.33, George Allen & Unwin, London,1982.
[Mandelbrot] Benoit Mandelbrot: On the Theory of Word Frequencies and on Related Markovian Models of Discourse, Proc. of the Symposium on Applied Mathematics, Vol.12, pp.190-219, 1961.
[Sumita] Eiichiro Sumita et al: Example-Based Approach in Machine Translation, Proc. of Info Japan'90, Oct., 1990.
[Witkam] Toon Witkam: Analogy-Based MT, Presentation for ATR, July, 1990.
[浦谷] 浦谷則好：ＡＰ電の単語分布、私信、１９９０年。
[江原] 江原暉将ほか：ＡＴＲ対話データベースの内容、ＴＲ－Ｉ－０１８６、１９９０年１０月。
[江原２] 江原暉将：単語のカテゴリーを用いた係り受け整合度の平滑化、ＴＲ－Ｉ－０２１５、１９９１年５月。
[柿ケ原] 柿ケ原康二：係り受け関係を用いた文節候補選択処理、ＴＲ－Ｉ－０１０５、１９８９年９月。
[工藤] 工藤育男：文と文の結束性を捕らえるための知識、ＴＲ－Ｉ－０１３４、１９９０年２月。
[幸山] 幸山秀雄ほか：用例を用いた依存関係単位の翻訳、４１回情全大、４Ｓ－５、１９９０年９月。
[佐藤] 佐藤理史ほか：実例に基づいた翻訳、情研資、ＮＬ－７０－９、１９８９年１月。
[古瀬] 古瀬蔵ほか：変換主導型機械翻訳の実現手法、情研資、ＮＬ－８０－８、１９９０年１１月。

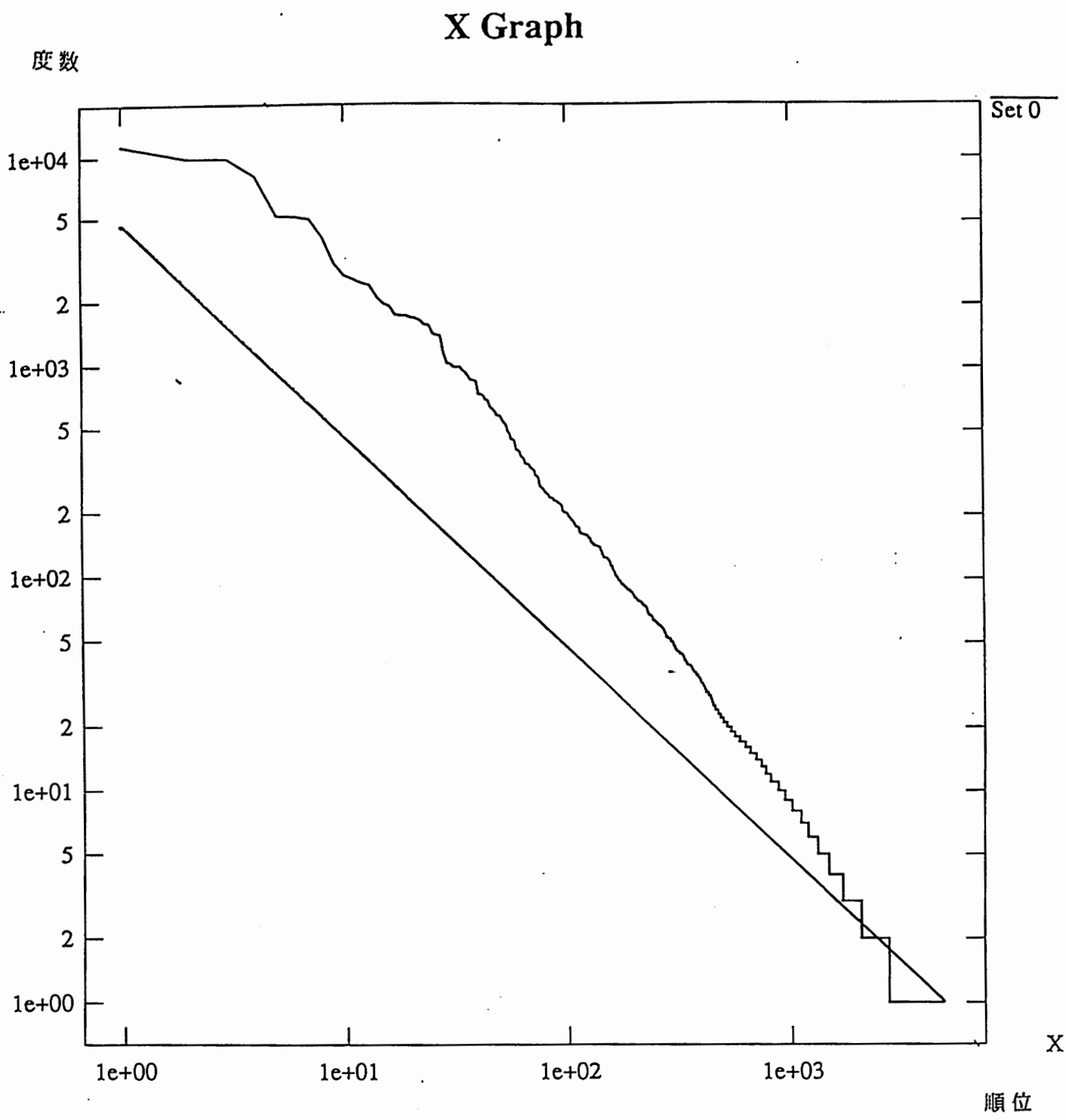

図３．１　単語数１の知識データ（単語データベース）の分布
ＡＤＤ　国際会議　電話対話
直線はｂ＝１のMandelbrot分布

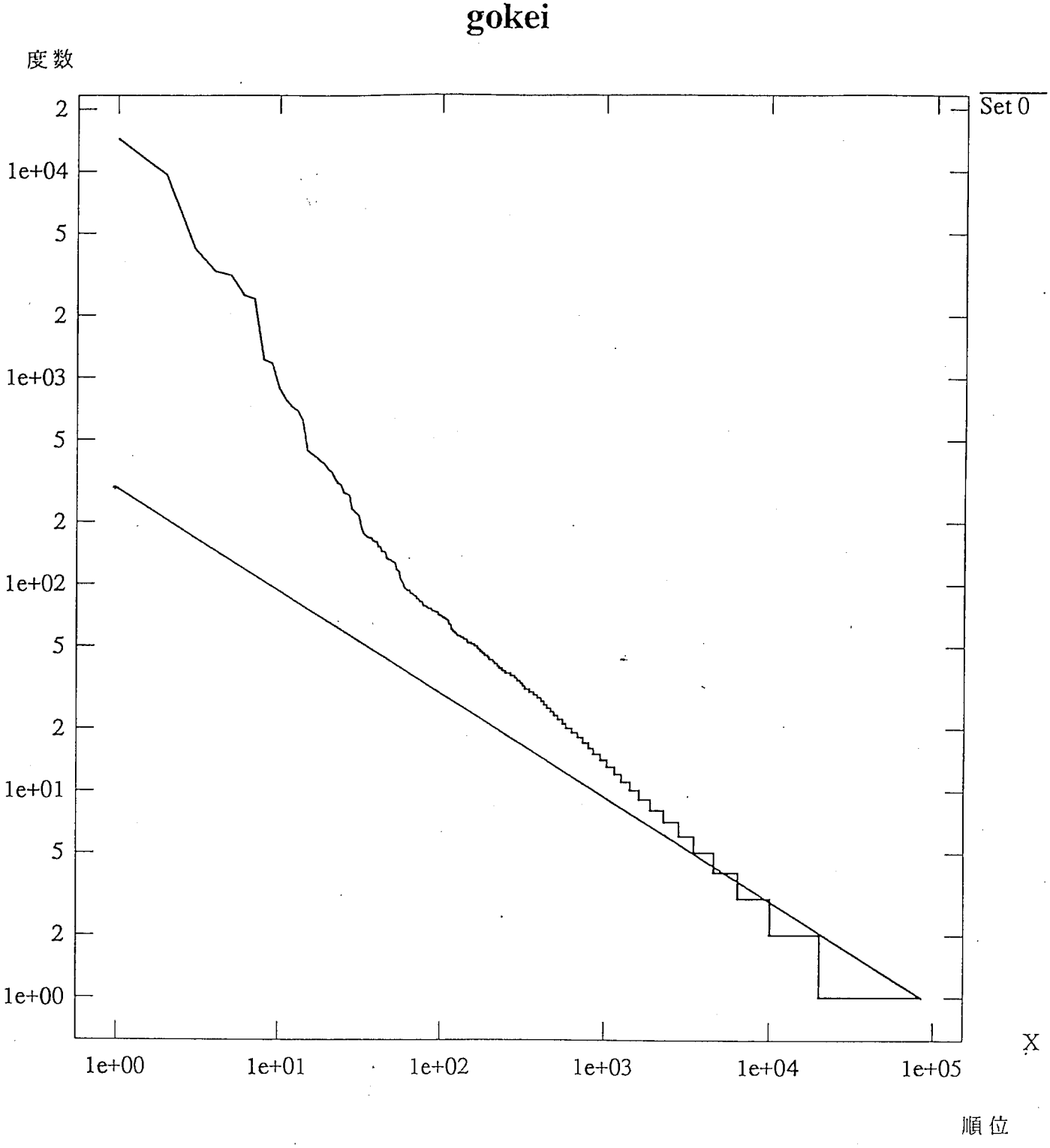

図３．２　単語数２の知識データ（係り受けデータベース）の分布
「が」格の係り受けデータ
直線はｂ＝１／２のMandelbrot分布