自然言語処理 110－11
（1995. 11. 18）

隣接文字の分散値を用いた定型表現の自動抽出

下畑さより 杉尾俊之 永田淳次
{sayori,sugio,nagata}@kansai.oki.co.jp
沖電気工業（株）研究開発本部 関西総合研究所
〒540 大阪市中央区城見 1-2-27

あらまし 本論文では、「文字列が意味のあるまとまりである時、その文字列の前後には様々な文字が出現する」という考えに基づき、隣接文字の分散の度合を基準として、表現の単位として有効な文字列の抽出を行なう方法を提案する。分散値は、文字列に隣接する文字の種類と各文字の生起確率によって計算する。計算にはエントロピー基準を導入した。また、本手法を約4,000万字の新聞記事データに適用した結果を紹介し、本手法の有効性を示す。

和文キーワード 自動抽出, 定型表現, コーパス, N-gram, エントロピー

Extraction of Frozen Patterns from Corpora through Entropy Thresholds

Sayori SHIMOHATA Toshiyuki SUGIO Junji NAGATA
{sayori,sugio,nagata}@kansai.oki.co.jp
Kansai Laboratory, Research & Development Group
Oki Electric Industry Co., Ltd.
1-2-27 Shiromi, Chuo-ku, Osaka 540, Japan

Abstract This paper describes a method to extract frozen patterns from corpora. This is done by calculating entropy values of characters before and after units and extracting units with high entropy values. The experimental results applied to newspaper articles containing 40 million characters are also discussed in this paper.

英文 **key words** automatic extraction, frozen pattern, corpus, N-gram, entropy

1 はじめに

近年、計算機能力が向上し、多くの言語資源の利用が可能となったことから、大量の言語データから使用頻度の高い語句や表現を自動的に抽出する研究が盛んになっている。

Churchらは、相互情報量を用いて“nurses”と“doctors”、“set”と“off”のように、関連のある単語対を抽出する方法を提案した[1]。また、Smadjaらは、相互情報量に単語間の相対距離の情報を加え、有効なcollocationを抽出した[2]。北らは、仕事量基準という概念を導入し、1まとめにすることでどれだけ処理効率が上がるかという観点から、使用頻度の高い連続単語を定型表現をとして抽出した[3]。

これらの方法はいずれも、対象となるテキストが単語単位に分割されていることを前提としている。そのため、日本語のように分かち書きする習慣のない言語では、事前に形態素解析処理を行なっておく必要がある。しかし、大量のテキストすべてに形態素解析を行なうには、多大な時間を要する。しかも、未知語が含まれていることなどにより形態素解析に失敗すると、有効なデータが得られないことになる。

これに対して長尾らは、テキストに対してn-gram統計処理を行ない、文字数の順、および出現回数の順に文字列を抽出する方法を提案した[4]。この方法は、形態素解析などの前処理を行なうことなく文字列を抽出できるという利点がある。しかし、文字数と出現回数によりテキスト中の文字列を網羅的に抽出するため、断片的な文字列がかなりの割合で混在するという問題があり、こうした文字列をいかに効率良く、正確に選別するかが課題となっている。これに対し、相互に重複する部分文字列を除去する方法[5]、ヒューリスティックを用いて抽出対象を限定する方法[6]、出現頻度を正規化する方法[7]などが提案されている。

2 分散値による文字列の抽出

2.1 従来技術の考察

従来研究において、テキストからの文字列抽出基準は、基本的に単語や文字数といった構成要素数と出現回数であった。しかし、構成要素数と出現回数では、文字列の有用度を計れない場合がある。

例えば、文字列“abc”の出現回数がnであるとき、“abc”の部分文字列“ab”の出現回数は常にnより大きくなる。この時、構成要素数を優先すると“abc”が有効であると判断され、出現回数を優先すると“ab”が有効であると判断されることになる。しかし、実際には構成要素数が多いものが有用な場合もあるし、出現回数の多いものが有用な場合もあり、どちらを優先すべきかは、一概に決められない。

また、構成要素数と出現回数が同じでも、一方はある文字列の部分文字列で、もう一方は独立した意味を持つ表現の単位であるような場合がある。これらの文字列の差を数値化し、有用なものとそうでないものを選別することはできない。

テキストから文字列を抽出する場合の基準は、構成要素数や出現回数だけでなく、文字列のまとまりとしての強さや前後の文字列との関係をうまく表現できるものでなければならない。

2.2 基本的な考え方

本論文では、テキストから抽出された文字列が意味のある表現のまとまりであるか否かを、その文字列の前後に出現する文字の分散の度合を基準に判断する方法を提案する。これは、「文字列が意味のあるまとまりである時、その文字列の前後には様々な文字が出現する」という考えに基づくものである。なお以下では、語句や定型表現のように表現の単位となる文字列を表現ユニットと呼ぶ。

分散の度合は、ある文字列に隣接する文字の種類と各文字の生起する確率で表す。隣接文字の種類が多いほど、また、隣接文字の生起確率が均等であるほど、分散の度合は大きくなる。逆に、隣接文字の種類が少ないほど、また、隣接文字の生起確率が偏っているほど、分散の度合は小さくなる。つまり、分散の度合の大きい文字列はその位置で分割される表現ユニットである可能性が高く、小さい文字列はその文字列は隣接文字を含むより長い文字列の一部（つまり断片的文字列）である可能性が高くなる。

分散の度合の計算には、エントロピー基準を用いる。エントロピーは事象の不確定さ、乱雑さを表す量で、等確率で選択肢が多いほど増大し、偏った確率で選択肢が少ないほど減少する性質がある。

分散の度合は文字列の長さや出現回数に依存しないため、文字数や出現回数の違う文字列どうしを比較することができる。また、前後の文字列との関係で値が変化するので、文字数や出現回数が同じ文字列の比較も可能となる。

3 表現ユニット抽出アルゴリズム

前章の議論に基づき、テキストから表現ユニットを抽出する方法について説明する。表現ユニットの抽出は、大きく以下の4つのプロセスからなる。

1. テキストからの候補文字列の抽出

2. 各候補文字列の隣接文字情報の獲得

3. 候補文字列のエントロピー値の計算

4. エントロピー基準による表現ユニットの抽出

以下、各プロセスについて具体的に説明する。

3.1 n-gram 統計による候補文字列の抽出

候補文字列の抽出は、テキストから表現ユニットの候補となる文字列とその出現回数を求める処理である。ここでは、テキスト中に2回以上出現したすべての文字列を候補文字列として抽出する。

この処理には、長尾らの提唱したn-gram統計による文字列の抽出方法[4]を利用する。この方法によれば、任意の文字列長nについて、テキスト中に2回以上出現した文字列と出現回数を洩れなく抽出することができる。

例えば、

"a b c a b a b c b a"

のようなテキストを入力した場合、図1に示す文字列が候補文字列として抽出される。図中の( )内の数字は、入力テキストにおいて文字列の先頭文字が出現した位置を示す。

候補文字列	出現回数
a (1,4,6,10)	4
a b (1,4,6)	3
a b c (1,6)	2
b (2,5,7,9)	4
b a (5,9)	2
b c (2,7)	2
c (3,8)	2

図 1: 候補文字列の抽出結果

3.2 隣接文字情報の獲得

隣接文字情報の獲得は、候補文字列の前後に出現する文字の種類と生起確率を求める処理である。ここでは、後接文字の場合を例に説明する。

文字数lの文字列 $S = c_1, c_2, ...c_l$ は、文字数$l-1$の文字列 $S' = c_1, c_2, ...c_{l-1}$ の後ろにc_lが出現する状態を表す。すなわち、文字列S'の後接文字にc_lが出現する状態である。この時、S'の後ろにc_lが出現する回数$f(S', c_l)$は、Sの出現回数$f(S)$と等しい（式1）。このことから、3.1で得られたすべての候補文字列SをS'と最後の文字に分割し、S'ごとに最後の文字の種類と出現回数を集計することにより、後接文字の種類と出現回数を求めることができる。また、S'の出現回数を$f(S')$とすると、S'の後ろにc_lが生起する確率$P(S', c_l)$は式2で求められる。

$$f(S', c_l) = f(S) \tag{1}$$

$$P(S', c_l) = \frac{f(S', c_l)}{f(S')} \tag{2}$$

具体的な手順は、以下の通りである。まず、文字数$l(l \geq 2)$の候補文字列を、$1 \sim l-1$文字めの文字列とl文字めの文字に分割し、$1 \sim l-1$文字めの文字列ごとにl文字めの文字の種類と出現回数を集計する。これにより、各候補文字列について2回以上出現した後接文字の種類と出現回数が求められる。次に、候補文字列の出現回数から各後接文字の出現回数の合計を減じることにより、1回ずつ出現した後接文字の種類と出現回数を求める。そして、上述の処理によって得られた後接文字の種類と出現回数から、式2を使って各後接文字の生起確率を求める。

例えば図1の"ab"は出現回数が3であるが、これは、"a"の後接文字が"b"となる場合が3回であることを表している。同図より"a"の出現回数は4回であるから、後接文字は2種類で、"b"が3回、それ以外の文字が1回となる。また、各後接文字の生起確率は各々$\frac{3}{4}$、$\frac{1}{4}$となる。

前接文字についても同様である。文字数$l(l \geq 2)$の候補文字列を$2 \sim l$文字めの文字列と1文字めの文字に分割し、$2 \sim l$文字めの文字列ごとに上述の処理を行なえばよい。

図1の候補文字列から得られた前接文字の情報を図3に、後接文字の情報を図2に示す。"*"は1回しか出現していない隣接文字を表す。

3.3 エントロピーの計算

隣接文字情報から、各候補文字列の前方エントロピー、および後方エントロピーを計算する。文字列Sの隣接文字集合を$W(S) = \{w_i \mid w_1, ...w_n\}$、隣接文字$w_i$の生起確率を$P(S, w_i)$とする時、エントロピー$H(S)$は、以下の式で求められる。

$$H(S) = -\sum_{i=1}^{n} P(S, w_i) \cdot log P(S, w_i) \tag{3}$$

エントロピーH(S)の値は$0 \leq H(S) \leq \infty$で、すべての隣接文字が等確率で出現する時に最大となり、常に1つの隣接文字が出現する時に最小となる。

$$H(S) = -\sum_{i=1}^{n} \frac{1}{n} \cdot log(\frac{1}{n}) = \ln n \tag{4}$$

文字列	前接文字	出現回数	生起確率
a	b	2	2/4
	*	1	1/4
	*	1	1/4
a b	*	1	1/3
	*	1	1/3
	*	1	1/3
a b c	*	1	1/2
	*	1	1/2
b	a	3	3/4
	*	1	1/4
b a	*	1	1/2
	*	1	1/2
b c	a	2	2/2
c	b	2	2/2

図 2: 前接文字情報

文字列	後接文字	出現回数	生起確率
a	b	3	3/4
	*	1	1/4
a b	c	2	2/3
	*	1	1/3
a b c	*	1	1/2
	*	1	1/2
b	a	2	2/4
	c	2	2/4
b a	*	1	1/2
	*	1	1/2
b c	*	1	1/2
	*	1	1/2
c	*	1	1/2
	*	1	1/2

図 3: 後接文字情報

$$H(S) = -\sum_{i=1}^{n} \frac{n}{n} \cdot log(\frac{n}{n}) = 0(n = 1) \qquad (5)$$

図1の文字列についてエントロピーを計算した結果を、図4に示す。（ここでは、対数の底を１０とする。）

文字列	前方	後方	出現回数
a	1.04	0.56	4
a b	1.10	0.64	3
a b c	0.69	0.69	2
b	0.56	0.69	4
b a	0.69	0.69	2
b c	0	0.69	2
c	0	0.69	2

図 4: 候補文字列のエントロピー

3.4 表現ユニットの抽出

ここでは、候補文字列を表現ユニットとそうでない文字列に分類する処理を行なう。

本方式では、分散の度合の大きい文字列ほど表現ユニットとしての性質が強いと判断する。すなわち、3.3で求めたエントロピーの値によって候補文字列のソートを行ない、エントロピーの高い文字列を表現ユニットとして抽出する。この時、前方エントロピーと後方エントロピーのうち低い方の値を有効値とする。例えば、閾値を設定しエントロピーが閾値を越える文字列を表現ユニットとする、エントロピーによるソートを行ない上位x%(あるいはx個)を表現ユニットとするなどの方法が考えられる閾値として適当な値は、用いるテキストの質や量によって異なるが、一般的には、テキストの量が多くなるほど閾値も高くなると考えられる。

図4の文字列をエントロピーによりソートすると図5のようになる。下線で示した値が有効エントロピー値である。

エントロピーが０の文字列　エントロピーが0となるのは、文字列の前あるいは後ろに常に同じ文字が出現する場合である。例えば図5の“bc”、“c”は、入力テキスト中で常に“abc”の形で出現している。このような文字列は、特に断片的文字列である可能性が高いと考えられる。

文字列	前方	後方	出現回数
a b c	0.69	0.69	2
b a	0.69	0.69	2
a b	1.10	0.64	3
a	1.04	0.56	4
b	0.56	0.69	4
b c	0	0.69	2
c	0	0.69	2

図 5: エントロピーによるソート結果

4 実験

4.1 実験の方法

本方式の有効性を検証するため、日本経済新聞の朝刊記事1年分 [8] (記事数 97,148、文字数 39,545,824字) を対象に、文字列抽出実験を行なった。

本実験では、stringの単位を句読点および改行とし、テキスト中に2回以上出現した文字列を候補文字列として抽出した。候補文字列に対してエントロピーを計算し、以下の条件に当てはまる文字列を除去して抽出文字列とした。

- 記号で始まる文字列
- 語頭にならない文字 (「ん、ゃ、っ」など) から始まる文字列
- 1文字からなる文字列
- エントロピーがどちらか一方でも0になる文字列

抽出された文字列の数は、表1の通りである。

	異なり
候補文字列	9,875,606
抽出文字列	2,632,917

表 1: 抽出文字列の種類

4.2 エントロピー値別抽出文字列の評価

本方式の妥当性を検証するため、エントロピー値別に抽出文字列の評価を行なった。具体的には、エントロピー値によって5段階に分類した文字列から各々1,000件を評価対象として取り出し、以下に示す評価基準に従って人手で評価した。

a 語・句・節のように、文法的な単位となる文字列

b a以外の意味的なまとまりを持つ文字列 (慣用表現など)

c 断片的で意味を持たない文字列

この基準において、a,b が我々の定義する表現ユニットになる。評価結果を表2に示す。

	文字列 (種類)	a	b	c
4以上	8,301	866	128	6
3 ~ 3.99	63,511	836	76	88
2 ~ 2.99	262,602	738	61	201
1 ~ 1.99	814,248	502	57	441
0 ~ 0.99	1,484,255	479	25	496

表 2: エントロピー値別評価結果

表2では、エントロピー値が高いほどa,bの割合が多くなり、低いほどcの割合が多くなっている。この結果は、「隣接文字の分散の度合が大きい文字列ほど表現ユニットである可能性が高い」とする我々の考えと一致している。

4.3 頻度順抽出文字列との比較評価

次に、同じ文字数の抽出文字列について、出現回数順にソートした結果と比較した。ここでは、文字数10の文字列 (61,753件) を対象とし、エントロピー値および出現回数の順にソートしたものから各々上位1000件ずつを取り出して評価を行なった。評価基準は、4.2と同じである。結果を表3に示す。

	a	b	c
エントロピー値順	808	87	105
出現回数順	657	100	243
(共通の文字列)	254	23	37

表 3: エントロピー順と出現頻度順の評価結果

エントロピー値によるソートでは、出現回数によるソートと比べてa,bの割合が多くなっており、断片的文字列の抽出が抑制されていることが分かる。

5 考察

前章の実験結果に基づき、本方式における抽出文字列の特徴を説明する。以降の表の項目欄において、"e値"はエントロピー値を、"n値"は出現回数を示す。特に断りがない限り、e値は有効エントロピー値を表す。

エントロピー値が特に高い文字列　表4は、エントロピー値によるソートで上位10件の文字列である。この図からも分かるように、エントロピー値の高い文字列は助詞や助詞相当語、接続詞のように機能的な働きを持つものが多い。また、単語より“名詞と格助詞”や“動詞と助動詞”のように比較的長い単位で高値となる特徴がある。これは、「様々な語句と接続する」という本方式における表現ユニットの定義が、これらの文字列の出現する条件に合致したためと考えられる。

また、句読点、文頭、文末を含む文字列のエントロピーも高くなっている。これは、候補文字列抽出の単位を句読点と改行とし、この位置で切れる文字列のエントロピーを最大値としているためである。句読点や改行があった場合、文字列がそこで分割されることは明らかであるが、他の文字列と比較するということを考えると、この処理は検討の必要がある。

文字列	前方e値	後方e値	n値
しかし、	6.26	6.36	576
ただ、	6.04	6.04	419
だが、	5.55	6.86	958
また、	5.43	5.56	260
などの	5.28	5.63	427
一方、	5.20	6.14	462
への	5.17	5.24	421
日本の	5.24	5.15	342
から	5.11	5.11	771
には	5.10	5.26	662

表4: エントロピーによる上位10件

文法的単位ではない表現ユニットと判断された文字列　前章の実験でbと判断された文字列は、文法的な単位とは認識されないが、表現上あるいは文の構成上意味があると考えられる文字列である。今回の実験では、この項目で興味深い結果が得られた。表5は、bに分類された文字列の例である。引用符を使った表現や新聞記事特有の定型表現のように、用法的な特徴を持つ文字列が多く抽出されている。これらの文字列は、呼応表現やパターン翻訳といった言語処理への応用が有効であると考えられる。

断片的文字列と判断された文字列　表6は、エントロピー値4以上で断片的文字列であると判断された文字列である。多くの文字列と組合せ可能な漢字や助詞となるひらがなが、文字列の先頭および最後にくる場合がほとんどである。今回対象としなかった1文字の候補文字列にもこれに該当するものが多い。頻出文字を含む文字列のエントロピーをどう扱うかは、今後さらに検討が必要である。

文字列	e値	n値
氏は「	5.20	1377
」と呼ばれる	5.40	371
」だった。	4.95	148
」から「	4.91	219
はこれまで	4.93	1853
【ニューヨーク２８日＝	3.19	27
日（上）――――・	3.65	40

表5: 文法的単位ではない表現ユニットの例

文字列	前方e値	後方e値	n値
で日本の	5.20	5.53	332
に日本の	5.22	5.48	313
山正	5.09	5.18	542
田博	4.99	4.99	223
の旧	5.22	4.96	375
市）に	4.96	5.02	236

表6: 断片的文字列の例

出現回数順抽出文字列との比較　表7、表8は、4.3での各ソート結果による上位10件の文字列である。表中のn順は出現頻度によるソートでの順位を、e順はエントロピーによるソートでの順位を示す。

文字列	e値	n値	n順
（シンガポール支局）	5.43	229	43
しなければならない。	5.09	361	19
することで合意した。	4.85	280	24
クリントン米大統領が	4.46	103	221
に拍車をかけている。	4.36	98	241
するよう求めている。	4.33	120	174
になる可能性もある。	4.33	101	228
したにもかかわらず、	4.25	74	359
する可能性が大きい。	4.20	71	381
されることになった。	4.19	69	403

表7: エントロピーによる上位10件

エントロピー値によるソートでは、出現回数順でも

文字列	e値	n値	e順
ことを明らかにした。	1.34	923	3437
ウルグアイ・ラウンド	3.08	777	238
日本電信電話（ＮＴＴ	0.02	639	6769
リストラクチャリング	0.40	608	5750
ラクチャリング（事業	0.03	561	6764
チャリング（事業の再	0.01	552	6770
（６月２９日）取締役	1.18	549	3805
国内総生産（ＧＤＰ）	2.67	484	637
ることを明らかにした	0.17	462	6525
なければならない」と	2.68	435	628

表 8: 出現回数による上位 10 件

文字列	種類	e値	n値
リストラクチャリング	30	0.40	608
～（	9	0.16	574
～（再構築）	7	1.91	8
～（事業	10	0.13	559
～（事業の再構築）	84	2.67	548
～の	5	1.61	5

表 9: 「リストラクチャリング」を含む文字列の情報

上位の文字列が出現しているが、出現回数によるソートでは、「日本電信電話（ＮＴＴ」「ラクチャリング（事業」「チャリング（事業の再」のように、エントロピー値での順位が極端に低い文字列が出現している。これらは、ある特定の文字と隣接することが多い文字列であると考えられる。

「リストラクチャリング」を例に説明する。「リストラクチャリング」を含む文字列の情報は表 9のようになっている。ここで、第 2 欄は後接文字の種類を示している。

「リストラクチャリング」はテキスト中に 608 回出現しているが、そのうち 574 回は「（」が後続している。それ以外の後接文字は 29 種類あるが、出現回数は合わせて 34 回である。このような特定の文字と隣接する文字列は、表現ユニットではなく、後接文字を含むより長い表現ユニットの部分文字列であると考えるのが妥当である。

出現回数順ソートでは「リストラクチャリング（」「リストラクチャリング（事業」のように、高頻度文字列の部分文字列はすべて高頻度となる。これに対して、エントロピー順ソートではこれらの文字列の値は抑制されている。また、「リストラクチャリング（再構築）」「リストラクチャリングの」のように表現のまとまりと考えられる文字列は、出現回数が低いにもかかわらず、上記の部分文字列より高い値となっている。この観点から、本手法では出現回数による文字列抽出と比べて、断片的文字列の排除に有効であることが分かる。

エントロピー値のギャップ　前述の内容とも関連するが、前方・後方分散値の差が大きい文字列は、断片的文字列である可能性が高い。

表 10は、２つのエントロピー値の差が大きい文字列の上位 10 件である。

有効エントロピー値が同じである時には、前後のエントロピー値のギャップが小さいものより大きいものの方が断片的文字列である可能性が高い。

今回はこの観点からの評価は行なわなかったが、これを基準として、断片的文字列の除去を効果的に行なうことができると考える。

文字列	前方e値	後方e値	n値
アジ	4.34	0.01	1099
を求	4.79	0.01	708
を進	3.90	0.01	581
質	3.64	0.02	625
サービ	4.26	0.01	502
ついては	0.01	3.45	583
は「きょう	4.63	0.02	351
このた	5.52	0.03	252
らだ。	0.03	5.47	238
ループ	0.02	3.12	374

表 10: エントロピーの差が大きいもの上位 10 件

6 まとめ

本論文では、「意味のある表現の部分は固定的であり、様々な語句と接続する」という考えに基づき、隣接文字の分散の度合を基準として表現の単位となる文字列を抽出する方法を提案した。分散の度合の計算にはエントロピー基準を導入した。この値は、隣接文字の種類と生起確率から求められるため、文字数や出現回数に依存することなく、文字列どうしを比較できるという特徴を持つ。また、特定の隣接文字に偏った生起確率を持つ文字列は値が低くなるという性質があり、従来技術で課題となっていた断片的文字列の除去にも適している。本手法を約4,000万字の新聞記事データに適用し、本手法の有効性を確認した。

今後は、本手法で得られた結果を言語処理に応用する方法を検討していく予定である。具体的には、辞書作成の自動化や形態素解析の前処理への利用を考えている。また、現在の手法では、連続した文字列のみを対象としているが、呼応や共起といった離れて存在する定型表現を扱えるような枠組を考えたい。

謝辞 本論文で使用したテキストデータは、日本経済新聞記事データCD-ROM94版の記事を使用しました。使用を許可して下さった日経総合販売（株）およびデータの研究利用に尽力して下さった皆様に深く感謝致します。

参考文献

[1] Church,K.W. and Hanks,P.: Word Association Norms, Mutual Information, and Lexicography, Computational Linguistics, Vol.16, No.1, pp22-29(1992)

[2] Smadja,F.A.: Retrieving Collocations from Text: Xtract, Computational Linguistics, Vol.19, No.9, pp143-177(1994)

[3] 北, 小倉, 森元, 矢野: 仕事量基準を用いたコーパスからの定型表現の自動抽出, 情報処理学会論文誌, Vol.34, No.9, pp1937-1943(1993)

[4] 長尾, 森: 大規模日本語テキストのnグラム統計の作り方と語句の自動抽出, 情報処理学会自然言語処理研究会報告 96-1, pp1-8(1993)

[5] 池原, 白井, 河岡: 大規模日本語コーパスからの連鎖型および離散型共起表現の自動抽出, 電子情報通信学会技術研究報告(NLC95-3), Vol.95, No.29, pp17-24(1995)

[6] 新納, 井佐原, 疑似Nグラムを用いた助詞的定型表現の自動抽出, 情報処理学会論文誌, Vol.36, No.1, pp32-40(1995)

[7] 中渡瀬, 木本: 統計的手法によるテキストからの重要語抽出メカニズム, 情報処理学会情報学基礎研究会報告 39-6, pp41-48(1995)

[8] 日本経済新聞記事データ CD-ROM94年度版, 日経総合販売（株）