言語表現と統計グラフの相互変換に関する基礎検討

小泉 尚之 †　松下 光範 ††　松田 昌史 ††　馬野 元秀 †

† 大阪府立大学 大学院理学系研究科 情報数理科学専攻
〒 599–8531 大阪府堺市中区学園町 1 – 1
†† 日本電信電話 (株) NTT コミュニケーション科学基礎研究所
〒 619–0237 京都府相楽郡精華町光台 2 – 4

1　はじめに

近年、我々は多くの情報を簡単に手に入れることができるようになった。しかし、得られた情報は文章、図表、画像、音声など、様々なメディアで表現され、その構成や形式も様々である。そのため、これらの雑多な情報を利用する際に、利用者には様々なスキルが要求され、スキルの習熟の差による情報格差が生じている。そこで、様々な形式で表現された情報を利用者の要求に適した表現に変換して提供する技術が求められている。[1]

我々はそのような技術のひとつとして、言語情報と統計グラフの相互変換技術について研究を進めている。このような技術の実現には、言語情報と統計グラフとを対応付ける意味レベルでの処理が必要となる。そのため、実際に使われる語彙の分析が不可欠であるだけでなく、語彙の意味を正確に得るためには文章構造の解析も不可欠である。そこで、本研究では被験者実験を通じ、人がグラフを文章に符号化する際に使用する語彙や文章構造について調べた。さらに、その文章から元のグラフを推測させ、文章の適切さについての評価を行った。この結果を踏まえ、計算機上で言語情報を数値データに変換する方法の考察/議論を行う。

なお、本研究は NTT コミュニケーション科学基礎研究所におけるインターンシップで行われた研究を発展させたものである。

2　被験者実験

本実験はグラフを説明した文章の収集とその文章の適切さを評価することを目的としている。そこで、グラフを文章で説明する「説明課題」と文章からグラフを推測する「解読課題」の 2 種類の課題を考案した。

2.1　方法

2.1.1　参加者

実験は大阪府立大学の学生 41 名（男性 17 名、女性 24 名）を 3 つの集団に分けて行った。3 つの集団はそれぞれ別々の日に実験を行い、1 日目は 14 名（男性 6 名、女性 8 名）、2 日目は 14 名（男性 5 名、女性 9 名）、3 日目は 13 名（男性 6 名、女性 7 名）であった。参加者の平均年齢は 19.2（最小 18、最大 24）歳であり、在籍学部、文・理系について調べた。日本語に関する実験のため、参加者は日本人に限定した。また、参加者への実験報酬と課題の成績を連動させた。

2.1.2　実験刺激

実験には計算機で作成した 81 種類のグラフを実験刺激として用いた。一辺が 300 ピクセルの正方形を縦横に 3 等分、全体として 9 等分した後、各列から 1 マス選択し、選ばれた 3 マスの中心点をつなぎ、スプライン関数で平滑化する（図 1 上）。この 27 種類のグラフに対し、ランダムに上下に 20 ピクセル（6.7%）の振動を加えたもの（図 1 左下）と、ランダムに上下に 50 ピクセル（16.7%）の振動を加えたもの（図 1 右下）を合わせた 81 種類を実験刺激とした。

2.1.3　匿名性の保持と一貫性の保証

参加者には実験開始前に ID カードを配付した。これは匿名性の保持と問題の一貫性を保持するためのものである。各課題では、ID によって回答する問題番号をあらかじめ一意に決めておくことで問題の一貫性を保持した。問題を配付/回収する際には、ID のみが書かれた封筒に用紙を入れ、実験者に参加者の ID がわからないように配付/回収した。

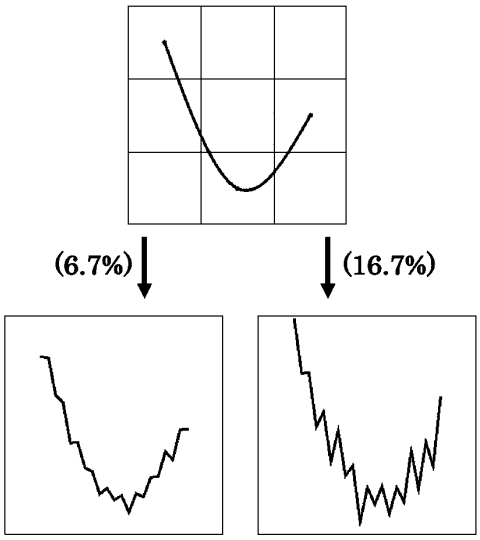

図 1: 参加者に提示するグラフの例
ただし、実際の実験で使用したグラフには罫線は引かれていない

2.1.4 手続き

参加者に各課題について教示をした後、説明課題を 40 分で行った。20 分の休憩を挟み、もう一度解読課題の教示をした後、解読課題を 25 分で行い、最後に実験に関する事後質問紙への回答を行った。部屋での席の配置は統制しなかったが、実験中は質問以外の発言を禁止した。

2.2 説明課題

各参加者に対し、81 種類のグラフからランダムに 6 種類を選び、記入用紙にそれぞれ文章で説明させた。ここで、各グラフと記入用紙にはあらかじめ通し番号を入れておき、グラフと文章が一意に対応付け出来るようにした。各参加者に対し、6 種類のグラフと対応する記入用紙の合わせて 12 枚を同時に配付した。

回答時間は合計で 40 分とし、1 問当たりの回答時間は統制しなかった。また、回答順序に制限はなく、時間の許す限り文章の追加/変更を自由に行えるようにした。ただし、文章の収集を目的としているため、説明する際には図や記号の使用を禁止した。なお、回答は手書きで行い、筆記具は実験者が用意した。

2.3 解読課題

説明課題で得られた記入用紙を説明課題後の休憩時間を利用して 2 部コピーしておいた。各参加者に対し、説明課題で得られた記入用紙からランダムに 12 種類を選び、選択用紙から元のグラフを推測させた。ただし、選ばれた文章には推測を行う参加者自身が書いた文章が含まれないように操作した。説明課題と同様に、記入用紙と選択用紙にはあらかじめ通し番号を入れておき、文章と選択したグラフが一意に対応付け出来るようにした。各参加者に対し、12 種類の文章と対応する選択用紙の合わせて 24 枚を同時に配付した。

選択用紙には、正解のグラフと共に、8 種類のグラフがランダムに描かれており、合計 9 種類のグラフの中から文章に合致するグラフを選択させた。ここで、選択する際にはポイントの自由分配とし、1 問につき 100 ポイントとして 9 種類のグラフの中から文章に合致すると判断したグラフにポイントを自由に割り振らせた。そして、正しいグラフに割り振ったポイントの 5 倍を得点とし、ポイントを割り振った参加者とその文章を書いた参加者の両方に得点を加算した。

回答時間は合計 25 分とし、1 問当たりの回答時間は統制しなかったが、次の問題に着手した後に、前の問題には戻れないように統制した。

3 実験結果と分析

3.1 実験結果

説明課題の結果、得られた文章数は 246 個、総文字数は 30577 字となり、1 つの文章の平均文字数は 124.3 字であった。また、この 246 個の文章に対し、形態素解析エンジン Mecab[2] を用いて単語に分解した結果、単語は 948 種類、総単語数は 20931 個となり、1 つの文章の平均単語数は 85.1 個であった。

解読課題の結果、間違いの文章数は 19 個であり、その平均文字数は 95.5 字、単語は 261 種類、平均単語数は 67.0 個であった。

3.2 分析

説明課題で得た 246 個の文章からグラフを文章に符号化する際に使用する語彙や文章構造について調べた。

3.2.1 語彙

得られた文章に対し、Mecab を用いて形態素解析を行い、名詞、動詞、名詞と動詞を組み合わせた

表 1: 単語と出現回数（一部抜粋）

名詞	回数
グラフ	392
右	266
上	215
下	190
1	180
山	158
0	153
左	135
方	130
ギザギザ	127
2	99
点	96
谷	89
左端	86

動詞	回数
下がる	104
上がる	72
終わる	42
書く	40
描く	37
見る	35
始まる	33
はじまる	31
おわる	30
折れる	29
いう	28
位置する	28
おる	20
とがる	20

表 2: 分類別の単語と出現回数（一部抜粋）

左右	回数
グラフ	103
右	93
山	61
上	57
左	54
1	48
下	48
方	43
目	43
回	34
下がる	34
ギザギザ	30
位置	30
半分	30

全体	回数
グラフ	217
右	122
上	88
1	87
高い	76
山	75
点	72
ギザギザ	70
下	65
方	57
左	56
よう	54
谷	52
形	52

複合動詞を抽出した。その結果、得られた語彙とその出現回数を表 1 に示す。抽出された名詞は総単語数が 505 個、出現回数の累計が 7081 回であった。動詞は“ いる ”、“ する ”、“ ある ”、“ なる ”、“ いく ”、“ れる ”の 6 単語を stopword とし、集計から除外した。また、複合動詞は動詞として抽出した。抽出された動詞は総単語数が 257 個、出現回数の累計が 1369 回であった。

3.2.2 文章構造

得られた文章の構造を人手により、左から右に向かって説明をする「左右」、全体的な形で説明する「全体」、左から右に向かって説明した後に全体的な形を説明する「左右+全体」、全体的な形で説明した後に左から右に向かって説明する「全体+左右」、それ以外の「その他」の 5 種類に分類した。その結果、「左右」が 69 個、「全体」が 131 個、「左右+全体」が 11 個、「全体+左右」が 20 個、「その他」が 15 個となった。

各分類ごとに形態素解析を行い、名詞、動詞、複合動詞を抽出した。その結果を表 2 に示す。この結果より、両方の上位に現れる場所や特徴を表す名詞は、グラフを文章に符号化する際の着目点と考えられる。また、逆に片方の分類にしか現れない単語は、文章構造を分類するための分類語と考えられる。ここでは、着目点を表 3、分類語を表 4 に示す単語として分析を行った。ただし、「左右+全体」、「全体+左右」に対する分類語は「左右」、「全体」と同じであり、文章中の分類語の出現順から判断できると考え、分類語の選出は行わなかった。

次に、着目点と分類語の使用回数を調べた。着目点の平均出現回数は 1 人につき 35.3 回、1 つの文章につき平均 5.9 回であった。分類語の「左右」については、平均出現回数は 1 人につき 9.5 回、1 つの文章につき平均 1.6 回であった。分類語の「全体」については、平均出現回数は 1 人につき 11.5 回、1 つの文章につき平均 1.9 回であった。

さらに、分類語を用いて 246 個の文章の分類を行った。以下にその手順を示す。

- 「左右」と「全体」の分類語数が同じである場合は先に現れた方で分類する。
- 「全体」の分類語数が「左右」よりも多い場合は「全体」とする。
- 上記を満たし、かつ分類語数の差が閾値（ここでは 2）以下であり、「左右」が先に現れている場合は「左右+全体」とする。
- 「左右」の分類語数が「全体」よりも多い場合は「左右」とする。
- 上記を満たし、かつ分類語数の差が閾値（ここでは 2）以下であり、「全体」が先に現れている場合は「全体+左右」とする。
- 「左右」と「全体」のどちらの分類語も現れなかった場合は「その他」とする。

表 3: 着目点

着目点			
上	下	左	右
左端	右端	半分	真ん中
高い	ギザギザ	折れ線	

表 4: 分類語

分類語			
左右	目	回	値
	減少	増加	最後
	上昇	はじまる	始まる
全体	よう	形	書く
	凸	上がり	直線
	線分	型	横
	頂点	曲線	

分類の結果、正解率は約 66%であった。

4 考察

分析結果を踏まえ、正しく分類出来なかった文章と、解読課題で間違いがあった 19 個の文章についての考察を行う。まず、どの分類語も現れずに「その他」と分類されたものは 14 個であった。この理由として、使用する単語の個人差が挙げられる。「その他」が 1 人から複数回検出されるケースがあり、3 人の文章から 14 個中の 9 個が検出された。また、同じ意味の単語であっても、ひらがな、カタカナ、漢字による表記の違いで別の単語として扱うため、分類語を検出出来ないケースもあった。これらは分類語の追加によって解決できると考えられるが、出現回数の少ない単語を追加すると分類語としての一般性を損なう可能性がある。そこで、表記の違いや同じ意味の単語は意味ごとにグループ化するなどして追加する必要があると考えられる。

次に、「左右+全体」と「全体+左右」に関する間違いが 36 個あった。そのうち、「左右+全体」または「全体+左右」が正解であるものは 10 個であり、これら 1 個あたりの平均は、文字数が 163.4 字、単語数が 108.8 個、分類語の出現回数は「左右」が 2.5 回、「全体」が 3.5 回であった。全文章の平均と比べ、すべてについて多く出現している。また、誤って「左右+全体」または「全体+左右」と分類されたものは残りの 26 個であり、1 個あたりの平均は文字数が 135.4 字、単語数が 91.7 個、分類語の出現回数は「左右」が 2.2 回、「全体」が 2.3 回であった。全文章の平均と比べ、大差はないものの、分類語は両方ともやや多く出現している。これらのことから、分類語が多く出現する文章は分類が難しく、分類手順を改良する必要があると考えられる。

そして、解読課題で間違いがあった 19 個の文章に対し、着目点と分類語の平均出現回数を調べたところ、1 つの文章につき着目点は 4.9 回、分類語は「左右」が 1.9 回、「全体」が 2.0 回であった。全文章に対するそれぞれの平均（着目点が 5.9 回、「左右」が 1.6 回、「全体」が 1.9 回）と比べ、大差はない。しかし、平均文字数が約 30 字、平均単語数が約 20 個少ないことから、文章の情報量は少ないと言える。また、参加者の事後質問紙で各課題の難しさを 5 段階リッカート法で測定（1：とても簡単、3：どちらともいえない、5：とても難しい）した結果、説明課題の平均 2.9 に対し、解読課題の平均は 1.8 であった。このことから、解読課題は問題が比較的簡単であったと言える。特に、選択する際のランダムに選ばれる正解ではない 8 種類のグラフに難しさが依存しているため、文章のみからでは間違いの原因を特定することはできなかった。

5 今後の課題

本論文では、言語情報と統計グラフとを意味のレベルで処理するために、被験者実験を通じて実際に使われる語彙の分析と文章構造の解析を行った。その結果、文章の構造を語彙の出現頻度によっておおよそ分類可能であることを示した。

今後はこの手法の改良を行うと共に、この結果を基にして、文章構造から語彙の意味を決定し、その意味とグラフとを対応付けることによって、言語情報から数値データに変換するためのモデルの構築について検討していきたい。

参考文献

[1] 加藤, 松下: 情報編纂（Information Compilation）の基盤技術, 第 20 回人工知能学会全国大会, 1D3-2 (2006).

[2] http://mecab.sourceforge.net/index.html