人間の認識特性に基づく時事情報からの情報抽出*

4Q－7

井上孝史, 稲垣博人, 中川透†
NTTヒューマンインタフェース研究所‡

1　はじめに

電子化されたテキストの増大にともなって、計算機によるテキストからの情報抽出の研究が盛んである [1]。これらの研究においては、抽出する情報構造の妥当性に関しては考慮されていない。本稿では多数の人間が個人差なく認識できる概念構造を認知実験によって求めるというアプローチを提案する。

また、時事情報からの情報抽出に用いる概念構造を求めるために行なった認知実験について報告する。

2　人間の認識のゆれを考慮した概念構造を求める方法の提案

計算機によるテキストからの情報抽出では、抽出する情報に関する概念的な構造 (以下、概念構造と呼ぶ) を設定し、その概念構造に当てはまる情報を抽出する。これまでさまざまな概念構造が提案され、計算機処理に用いられてきた (例えば [2])。その場合、概念構造は研究者の内省に基づいて決められ、システムの評価は、研究者によって与えられる正解と、計算機処理で出力された概念構造が一致する割合によって行なわれる。しかし、現実には人間による正解の割り当てにゆれが生じることも多く、正しい評価ができない。また、そのような構造に基づいて抽出された情報が多くの利用者にとって有効であるとは言えない。そこで我々は、概念構造が大多数の人間によって個人差なく認識できるかどうかを認知実験によって検証し、より適切な概念構造を求めるというアプローチを提案する。

本研究では時事情報として新聞記事を対象とする。まず、新聞記事の作成者が用いる記事作成法をもとに概念構造を仮定する。これは、新聞記事作成者が時事情報に関する効率の良い情報伝達の手法を知っており、それが人間の認識特性を反映していると考えるられるからである。新聞記事の作成者は、**本記**、**背景**、**解説**の3つの要素 (**機能的役割**と呼ぶ) によっておおまかな記事の構成を行ない、また個々の事実に対しては何時 (When)、どこで (Where)、誰が (Who)、何故 (Why)、何を (What)、どうした (How) のか、のいわゆる **5W1H** の概念を含むように作成する。そこで我々は機能的役割と 5W1H からなる構造を、時事情報からの情報抽出に用いる概念構造と仮定する。次に仮定した概念構造に対する人間の認識のゆれを認知実験によって調べ、より適切な概念構造を求めていく。

*Information Extraction from News Texts Considering Indivisual Variation.

†INOUE Takafumi, INAGAKI Hirohito and NAKAGAWA Toru.
‡NTT Human Interface Laboratories.

3　認知実験

仮定した概念構造が人間の認識する概念構造として客観的かつ適切なものであるかどうかを検証するために、機能的役割と 5W1H に関して認知実験を行なった。

3.1　機能的役割に関する実験

段落役割の割り当て

17人の被験者に新聞記事7件を読ませ、本記、背景、解説という3つの機能的役割のいずれかを各段落ごとに割り当てさせた。実験の結果を表1に示す。表中の**一致度**は、[**一致度** = 着目する役割の割り当てが多数意見であった段落においてその役割を割り当てた延べ人数 / (被験者の数 × 着目する役割が多数意見であった段落の数)] である。

役割	本記	背景	解説
多数意見である段落数	7	3	17
一致度	94%	67%	70%

表 1: 段落ごとの機能的役割の割り当て (全 27 段落)

文役割の割り当て

同じ記事に対して文ごとの役割の割り当てに関する実験を行なった。この実験ではより細かい役割を設定した。実験の結果は表2の通り。

以上の実験の結果から以下のことがわかった。

1. 段落役割の割り当て実験では本記の一致度は94%と非常に高かったが、背景と解説の一致度は70%前後でそれほど高くなかった。また本記の割り当てが多数意見だったのはすべて第一段落めであった。
2. 文役割の割り当て実験では、「主要な出来事」では一致度92%だが、その他の役割についてはあ

文の役割	多数意見である文数	一致度
主要な出来事	7	92%
主要な出来事に付随する出来事や事実	26	57%
出来事の起こった理由や状況	6	53%
出来事が起こったことによる結果	7	45%
未来の予測、推測	17	62%
過去の出来事	4	58%
特別な用語などの定義、説明	4	61%
書き手の意見	1	29%

表 2: 文ごとの機能的役割の割り当て (全 66 文)

まり一致していない。また、「主要な出来事」が過半数を占めたのはすべて第一文であった。

実験の結果から、本記という要素に対する人間の認識には個人差がないといえるが、背景、解説は明確に区別して認識されておらず、これらについては機能的役割を見直す必要がある。

3.2 5W1H に関する実験

機能的役割に関する実験で、第一文が主要な出来事であるという点で人間の認識に個人差がないことから、記事の第一文に関する 5W1H についてさらに実験を行なった。記事 10 件の第一文を抜き出して 18 人の被験者に提示し、そこに述べられている出来事の 5W1H に該当する部分を示させた。この時、文の構造の複雑さを変化させて提示し、人間の認識のゆれについても調べた。文の構造の複雑さのレベルとしては、重文 (対等な主語述語関係が複数ある文)、 複文 (従属節がある文)、単文 (一述語文) の 3 つのレベルを考慮した。

実験の結果、単文では高い確率で 5W1H の認識が一致する (表 3) が、複雑な文では人間の認識にゆれが観測された (重文における認識のゆれの例を表 4に示す)。この結果から、新聞記事の第一文について次のことが言える。

1. 単文では Why を除く 5W1H に対する人間の認識にゆれがない。
2. 複雑な文には複数の事実が述べられていて、個々の 5W1H については人間の認識が一致しているが、文全体としての 5W1H の認識にはゆれがある。これは、各人の持つ知識や関心によってどの事実が重要と感じるかに差が生じるからだと考えられる。

4 まとめ

テキストからの情報抽出において、多数の人間が一致して認識できる概念構造を認知実験によって求めるアプローチを提案した。また実際に新聞記事を対象にして行

	When	Where	Who	Why	What	How
文数	11	2	19	0	15	19
一致度	100%	89%	97%	–	89%	100%

表 3: 単文 (19 文) における 5W1H 認識の一致度

重文 (2 文)	
5W1H の取り方のパターン	パターンを取った人数
～は (Who) ～し、 (How) ～は ～した	3 人
～は ～し、 ～は (Who) ～した (How)	4 人
～は (Who1) ～し、 (How1) ～は (Who2) ～した (How2)	11 人

複文 (4 文)	
5W1H の取り方のパターン	パターンを取った割合
～は ～し、 (How) ～した	0 人
～は ～し、 ～した (How)	10 人
～は ～し、 (How1) ～した (How2)	8 人

表 4: 複雑な文における 5W1H 認識のゆれの例

なった実験について報告した。そこでは、新聞記事の第一段落が本記、第一文が主要な出来事であることについて認識の一致が見られた。また、単文に関する 5W1H に関して人間の認識が一致することがわかった。

今後は今回一致が見られなかった機能的役割について再検討、実験を行ない、よりよい構造を求めるとともに、計算機で 5W1H を抽出する方法を確立していく。

参考文献

[1] Paul Jacobs.(ed.), *Text-based Intellingent Systems*, Lawrence Erlbaum Associates,Hillsdale,NJ,1992.

[2] Paul Jacobs and Lisa Rau., "SCISOR: Extracting information from on-line news", *Comm. ACM*,33(11),1990.