

F-015

知識ベースと未知語処理による算数問題の意味理解

Understanding Arithmetic Problems Using Knowledge-Base and Unknown Word Processing

杉本 晃紀† 稲葉 栄美子† 渡部 広一† 河岡 司†

Akinori SUGIMOTO† Emiko INABA† Hirokazu WATABE† Tsukasa KAWAOKA†

1. はじめに

人間は会話中にあいまいな表現や抽象的な表現を受け取った場合にも、連想することにより適切に判断し、会話を続けることができる。これは、言語の意味や概念同士の関係を知識として習得しているためであり、これらは知識であると同時に常識でもある。常識には量的なものや数、時間に関するもの、感覚や感情に関するものなどがある。人間の要求を理解できるコンピュータを実現するには、これらの常識をふまえて判断することができる常識判断メカニズムを組み込むことが必要である。

本研究では、常識判断のうち数量に関する知識を集めた知識ベースおよびシソーラス¹⁾、概念ベース²⁾・³⁾、同義・反語辞書などを用いた未知語処理により算数問題の意味理解を行う手法を提案し、「算数問題の意味理解システム」を構築する。

2. 算数問題の意味理解システム

(1) 算数問題の定義

本研究では、数量と質問文を含む問題を算数問題と定義する。また、一般問題（小学校1年生から4年生までの加算、減算、乗算、除算および四則演算混合の文章題）と特殊問題（仕事算、数列などのある程度定式化された文章題）を対象としている。

(2) システムの構成

図1に、意味理解システムの構成の概要を示す。本システムは大きく問題のパターン分類、個別演算システムによる意味理解に分けられる。前者は、前述した対象問題のどのパターンの問題であるかを正しく分類する処理であり、後者は分類された問題を個別の判断知識を用いて数式に変換し、答えを導く処理である。

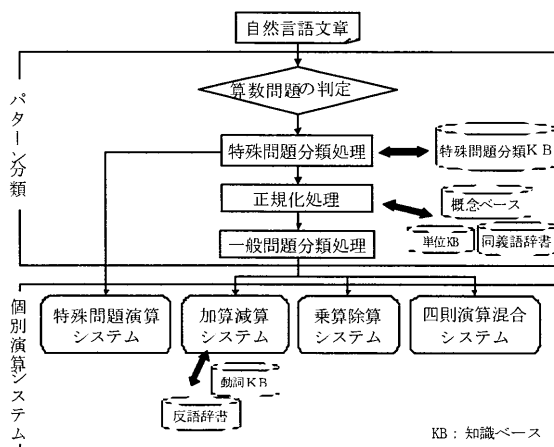


図1 算数問題の意味理解システムの基本構成

3. 未知語処理

各処理過程では様々な知識ベース（代表的な語のみ格納してあるデータベース）を用いる。しかし、会話や文章に出現する語の種類や表現は様々であり、これらの語に対しても対応するしくみが必要である。そこで、文章中に用いられる語で知識ベースにない語については未知語として扱い、概念ベースや関連度計算を用いて代表語の中で最も関連が強いものに置換する処理を行う。この処理を未知語処理と呼ぶ（図2）。この処理により少量の知識から様々な語に対応することを実現している。

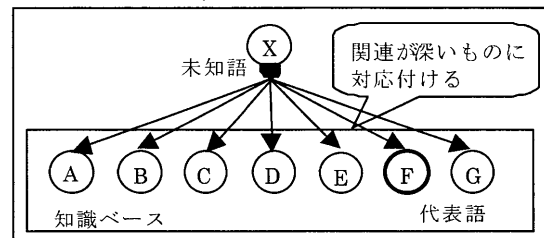


図2 未知語処理のイメージ図

3.1 概念ベース

ある語 A をその語と関連の強いと考えられる語 a_i と重み w_i の対の集合として定義する。

$$A = \{(a_1, w_1), (a_2, w_2), \dots, (a_m, w_m)\} \quad (1)$$

ここで、 a_i を 1 次属性と呼ぶ。また便宜上、 A を概念表記と呼ぶ。このような属性の定義された語（概念）を大量に集めたものを概念ベースと呼ぶ。ただし、任意の 1 次属性 a_i は、その概念ベース中の概念表記の集合に含まれているものとする。すなわち、属性を表す語もまた概念として定義されている。したがって、1 次属性は必ずある概念表記に一致するので、さらにその 1 次属性を抽出することができる。これを 2 次属性と呼ぶ。概念ベースにおいて、「概念」は n 次までの属性の連鎖集合により定義されている。

本研究では、複数の国語辞書から、各見出し語を概念表記、その見出し語の説明文中の自立語を 1 次属性として抽出し、出現頻度に基づく重みを付加した約 4 万の概念からなる概念ベースを基に、新聞などから抽出した概念表記や属性を加え、質の向上を目的にした精練操作（属性の追加・修正など）を施し、更に、概念間に成り立つ一般的なルールに基づく適切な重みを付加した約 9 万の概念からなる概念ベース³⁾を構築し利用している。

3.2 関連度計算

関連度⁴⁾とは、関連の種類しか特定できない意味ネットワークのようなものとは違い、概念と概念の関連の強さを定量的に評価するものである。具体的には概念ベースを用いて概念連鎖により概念を 2 次属性まで展開したところで、最も対応の良い一次属性同士を対応付け、それらの一致する属性の重みを評価することにより算出するものである。

† 同志社大学大学院 工学研究科

Graduate School of Engineering, Doshisha University

関連度は、概念間の関連の強さを 0 と 1 の間の数値で表す。表 1 に例を示す。

表 1 関連度計算結果の例

概念 A	概念 B	概念 A と概念 B の関連度
自動車	車	0.72
自動車	馬	0.03

4. パターン分類

パターン分類は意味理解システムの上位段階の処理であり、自然言語の文章がどのパターンの算数問題であるのかを判断する。

4.1 判断知識

判断知識は入力文を分類するために必要な知識を集めたものである。その知識を以下に述べる。

(1) シソーラス

シソーラスとは、一般名詞の意味的用法を表す 2700 個の意味属性（ノード）の上位-下位関係、全体-部分関係が木構造で示されたものである。ノードに属する名詞として約 13 万語（リーフ）が登録されている。

(2) 同義・反語辞書

同義・反語辞書は、同義と考えられるもの約 20 万組、反語の関係にあるもの約 17000 組を格納したものである。パターン分類ではこの辞書のうち、同義語辞書を用いる。

(3) 特殊問題分類知識ベース

定義した算数問題の内、特殊問題を特定する際に用いる知識を格納したものを特殊問題分類知識ベースという。

知識として、問題の種類（仕事算、流水算など）を特定するキーワードを格納した（表 2）。

表 2 特殊問題分類知識ベース

問題の種類	判断基準
数列問題	数列、数の並び、項、...
仕事算	時間の単位、時間を表す語の数、...
...	...

(4) 単位知識ベース

物には単位がある。例えば、ジュースは「本」という単位で数え、体積は一般的に「ml, l」という単位で量る。このようにある名詞がもつ単位を格納したものを単位知識ベースという。

単位は数を数えるときの単位、体積を量るとき単位、重さを量るとき単位などその尺度により数種類に分類できる。この単位を与える尺度のことを「単位尺度」と呼び、計 9 種ある（表 3）。

表 3 単位尺度および単位の種類

単位尺度	単位
個数	個、つ、羽、匹、本、冊、...
体積	l, ml, cm ³ , cc, ...
重量	t, kg, g, mg, ...
...	...

4.2 パターン分類処理

パターン分類の処理は、算数問題の判定、特殊問題分類処理、正規化処理、一般問題分類処理の順に行う。

(1) 算数問題の判定

算数問題であるか、算数に関係ない文章であるかの判定を行う。算数問題である条件は、①文章の中に数量が出てくる、②質問文があり、ある概念の数量について質問して

いる、である。これら 2 つの条件を同時に満たすものを算数問題として扱う。

(2) 特殊問題分類処理

4.2 節 (1) で算数問題として取り扱うと判定された問題をさらに特殊問題とそれ以外の問題に分類する。入力された文章が特殊問題分類知識ベースのキーワードを含むならばその特殊問題と特定し、各特殊問題演算システムに渡す。

(3) 正規化処理

4.2 節 (2) の分類より、特殊問題でないと判定された問題を一般問題のいずれかの演算問題に分類するために、文章の中から問題パターンの特定に必要な情報を取得する必要がある。このために以下の手順で正規化を行う。必要な情報とは、文章中で述べられている演算対象の概念とその数量の組み合わせであるので、（概念 A, 数量 X a）となるように正規化する。

以下に演算対象の特定方法を述べる。

初めに、文章の中から演算対象である概念を特定する。このとき、単位知識ベースを用いる。まず、入力文を形態素解析して文章中の名詞すべてを抽出し、各名詞の持つ単位情報を単位知識ベースより展開する。例えば、「球根」ならば {個数: 個, つ, 重量: g, kg, ...} である。そして、文章中に出てくる単位とこの単位情報を参照し、一致する単位がある名詞が、演算対象である概念として特定される（図 3）。

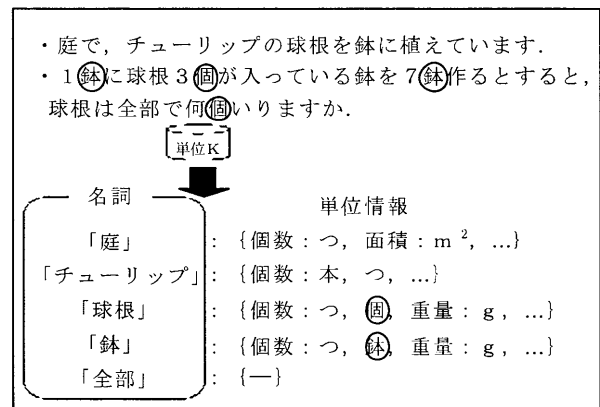


図 3 演算対象概念の特定の例

(4) 一般問題分類処理

4.2 節 (3) で得た正規化表現を元に、一般問題をさらに加減算、乗除算、四則演算混合のいずれの問題であるか判断し分類する。

基本的な考え方は、質問部の概念が条件部に 2 個以上あれば加減算、1 つならば乗除算となる。この考えを基に以下に各演算問題の条件を示す。ただし、 $n_{If=Q}$ を質問部の概念と一致する条件部の概念の個数、 $n_{If \neq Q}$ を質問部の概念と異なる条件部の概念の個数とする。

(a) 加算減算問題の条件

$$n_{If=Q} \geq 2, n_{If \neq Q} = n_{If}$$

(b) 乗算除算問題の条件

$$n_{If=Q} = 1, n_{If \neq Q} = 2$$

(c) 四則演算混合問題の条件

$$n_{If=Q} \geq 2, n_{If=Q} \neq n_{If}, n_{If=Q} = 1, n_{If \neq Q} \geq 3$$

5. 特殊問題演算システム

現在の特殊問題演算システムは数列、N進数、仕事算の問題に対して処理を行う事ができる。以下に問題が仕事算である場合の処理を述べる。

(1) 仕事算演算システム

仕事算演算システムは、正規化処理、仕事表に格納の順に行う事により、仕事算問題の意味理解を行う。

(2) 判断知識

判断知識は仕事算問題を理解するために必要な知識を集めたものである。具体的には、仕事をする上での状態（休む、働く等）を理解できる知識が必要となる。その知識として、以下の知識を作成した。

(a) 休暇知識ベース

「休憩」、「休む」といった動作や状態を表す語を休暇知識ベースに格納した。尚、休暇知識ベースの代表語は休暇である。

(b) 実働知識ベース

「働く」、「労働」といった動作や状態を表す語を実働知識ベースに格納した。尚、実働知識ベースの代表語は実働である。

(c) 総時間知識ベース

人間がある仕事を終らせた事を意味する動詞を総時間知識ベースに格納した。尚、総時間知識ベースの代表語は総時間である。

(3) 正規化処理

仕事算問題を意味理解するために必要な情報を文章から取得する。必要な情報とは文章中述べられている仕事をする人（以下主体と表記）、仕事をした期間、どのような動作をしたか（以下カテゴリと表記）なので、文章を（主体、期間、カテゴリ）となるように正規化する。

以下に（主体、期間、カテゴリ）の取得方法を述べる。

入力文を形態素解析し、解析結果から最初に期間を特定する。「1 日」、「3 時間」といったように数字と時間の単位のセットを抽出し、それを期間とする。次に主体の特定であるが、先ほど特定した期間を基準にして、前にあり、人の属性を持つ語を主体とする。最後にカテゴリの特定であるが、先ほど特定した主体より後ろにあり、期間に距離的に一番近い知識ベースの語があれば、その語をカテゴリとする。しかし、ここでは、どの知識ベースに属する語なのか重要であるので、カテゴリには知識ベースの代表語を格納する。また、知識ベースにない場合もある。この場合には未知語処理を行う。この処理より、未知語を未知語と知識ベース内の最高関連度をとった語と置き換えるのだが、閾値は未知語が名詞の場合には 0.23、動詞の場合には 0.07 とする。尚、これらの数値は実験的に算出したものである。

この正規化の例を図 4 に示す。

ある仕事の途中で、 <u>和子</u> さんと <u>4日</u> 、 <u>啓子</u> さんと <u>8日</u> 休んだ。
(和子, 4日, 休む), (啓子, 8日, 休む)
↓
(和子, 4日, 休暇), (啓子, 8日, 休暇)

図 4 仕事算における正規化の例

(4) 意味理解モデル

5 章 (3) で得られた全ての正規化表現を仕事表 (表 4) に格納する事により、答えを導く事ができる。

表 4 仕事表

	能力	休暇	実働①	実働②
A	a_A	r_A	$w1_A$	$w2_A$
B	a_B	r_B	$w1_B$	$w2_B$
全員	S_a	S_r	S_{w1}	S_{w2}

能力: A (B) が一人で仕事を終わらすのにかかる期間

休暇: A (B) が休んだ期間

実働①: A (B) だけで働いた期間

実働②: A (B) が働いた累計の期間

(5) 数式表現

表 4 から、以下の数式が成り立つ。

$$1/a_A \times w2_A + 1/a_B \times w2_B = 1$$

6. 加算減算システム

(1) 加算減算問題

収集した加減算問題の中でも、以下の 2 つの種類の一般的な文章表現による加算減算問題を対象とする。

(a) 時系列変化問題

例: 「紙が 80 枚あります。36 枚配りました。何枚残りましたか。」

(b) 部分全体・包含関係問題

例: 「赤い花が 15 本、黄色い花が 17 本咲いています。花はいくつ咲いていますか。」

(2) 判断知識

判断知識は加算減算問題を理解するために必要な知識を集めたものである。その知識として単数状態動詞知識ベース、変化明示語知識ベースを作成した。

(a) 単数状態動詞知識ベース

対象とする状態が一つである場合、動詞の意味を一意に決めることができる。例えば、「買う」はその状態に入ってくる意味をもつので「+の動詞」、「売る」はその状態から出ていく意味をもつので「-の動詞」となる。このようにある状態に対して入るか出ていくかという観点から分類した動詞を単数状態動詞知識ベースに格納した (表 5)。知識ベースにない語は未知語として扱い、未知語処理を行うことにより対処する。

表 5 単数状態動詞知識ベース

動 詞	演 算	動 詞	演 算
ある	+	使う	-
もらう	+	食べる	-
乗る	+	降りる	-

(b) 変化明示語知識ベース

ある概念の数量が変化することを示す語を集めたものを変化明示語知識ベースという。例えば、「初め、...であった」「残っている」などは、数量が何らかの変化をし、その前後でどのような状態であるかを表す動詞や語である。変化前と変化後を示す語を格納した。

(3) 意味理解モデル

加減算問題は動詞の意味によって+か-の演算が決定する。そこで、パターン分類で得られた正規化表現にさらに動詞を追加したものを加減算の内部表現形式とする。

(a) 内部表現形式

内部表現形式は（概念 A、数量 a、動詞 Va）となる。この例を示す。「紙が 80 枚あります。36 枚配りました。何枚残りましたか。」という文章から正規化表現（紙, 80 枚）、（紙, 36 枚）、（紙, x 枚）が得られ

る。これと文章を照らし合わせて動詞を追加する。動詞の対応は各概念の数量と動詞の距離を見て、近いものを取得する。これにより、(紙, 80 枚, ある), (紙, 36 枚, 配る), (紙, x 枚, 残る) の内部表現形式が得られる。

(c) 表による内部表現形式

得られた内部表現形式を以下の表に格納して整理する。表は変化前, + の変化量, - の変化量, 変化後の 4 つに分けて考える。内部表現形式の動詞が単数状態動詞知識ベースに存在する場合は, その演算の意味により + の変化量または - の変化量に格納される。また, 動詞が変化明示語知識ベースに存在する場合は, 変化前または変化後に格納される (表 6)。

表 6 加減算の意味理解表

変化前	変化量 $\Delta +$	変化量 $\Delta -$	変化後
(概念 A, a_0, V_0)	(概念 A, a_+, V_+)	(概念 A, a_-, V_-)	(概念 A, a, V)

(d) 数式表現

表 6 から, 各概念において以下の数式が成り立つ。

$$a_0 + a_+ - a_- = a$$

(a_0 : 変化前の数量, a_+ : + の変化量,

a_- : - の変化量, a : 変化後の数量)

7. 乗算除算システム

(1) 意味理解モデル

乗算除算問題は得られた正規化表現の概念同士の論理関係を取得すれば答えを導くことができる。よって, 内部表現形式は正規化処理で得られた概念と数量のセットとなる。

(a) 表による内部表現形式

7 章 (2) で述べる概念同士の論理関係の取得方法により内部表現形式を以下の表に格納する (表 7)。

- ・ 1 台の自動車に 4 人乗れます。
- ・ 3 では何人乗れますか。

表 7 乗除算の意味理解表

自動車[台]	人[人]
(a1) 1	(b1) 4
(a2) 3	(b2) x

(b) 数式表現

表 7 から, 以下の数式が成り立つ。

$$a_1 b_2 = a_2 b_1$$

(a_1, a_2 : 概念 A の数量, b_1, b_2 : 概念 B の数量)

この数式を解いて答えを導くことができる。

(2) 論理関係の取得

どの概念の数量が論理的に対応しているのかを文章から判断し, 論理関係を取得する。取得方法は以下の 2 通りである。

(a) 論理関係を特定する語が含まれる場合

一般に, 「～ずつ」という語はもう一方の概念の数量を 1 とした場合の数量を表す。例えば, 「箱に林檎を 3 個ずつ入れる」というと, 「箱 1: 林檎 3」という論理関係を示している。そこで, 内部表現形式のうち「ずつ」という語がついているものは, もう一方の概念の数量が 1 であるものと組み合わせ, 残りの 2 つも組み合わせると論理関係を 2 つ得る。

(b) 論理関係を特定する語が含まれない場合

論理関係を特定する語がない場合は, 文章中出现する数量同士の距離の近さで判断する (図 5)。

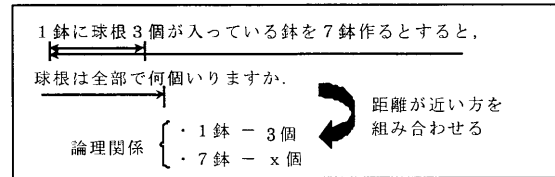


図 5 論理関係取得の例

8. 意味理解システム全体の評価

特殊問題計 150 問, 加減算問題計 200 問, 乗除算問題計 150 問の計 500 問を評価データとして意味理解システムにかけ, パターン分類, 各個別演算システムを経て最終的に正しい答えを出力したものを正解として評価を行った (図 6)。

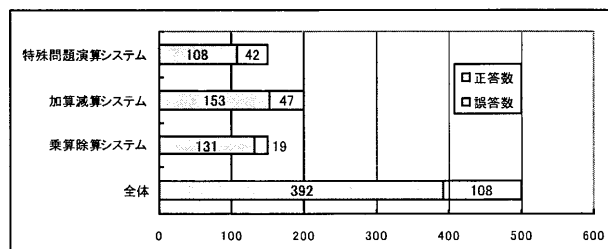


図 6 意味理解システム全体の評価

正答率は, 特殊問題演算システムが 72.0%, 加減算システムが 76.5%, 乗除算システムが 87.3%, システム全体が 81.1%となった。

9. おわりに

本研究では文章の表現形式ではなく算数問題の内容に着目して文章中から演算に必要な情報を取得し, 算数問題を解く手法を提案した。また, 概念ベースや関連度計算により知識の拡張を行った。これらにより, 極めて多様な表現および語に対応することが可能となった。

今後, 個別演算システムなどを追加することでより幅広い算数問題の意味理解を実現できると考えられる。

謝辞

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行ったものである。

参考文献

1. NTT コミュニケーション科学研究所監修, 「日本語語彙体系」, 岩波書店, 東京, 1997.
2. 笠原 要, 松澤 和光, 石川 勉, “国語辞書を利用した日常語の類似性判別”, 情報処理学会論文誌, Vol.38, No.7, pp.1272-1283, 1997.
3. 広瀬 幹規, 渡部 広一, 河岡 司, “概念間ルールと属性としての出現頻度を考慮した概念ベースの自動精錬手法”, 信学技報, NLC2001-93, pp.109-116, 2002.
4. 渡部 広一, 河岡 司 “常識的判断のための概念間の関連度評価モデル”, 自然言語処理, Vol.8, No.2, pp.39-54, 2001.