

概念ベースを用いた算数問題 濃度問題 の意味理解方式

Method of Semantic Understanding of "Arithmetic Problems-Concentration Problems-" Using Concept-Base

松山 剛*1
MATSUYAMA Gou

渡部 広一*1
WATABE Hirokazu

河岡 司*1
KAWAOKA Tsukasa

*1 同志社大学大学院 工学研究科 知識工学専攻

Department of Knowledge Engineering and Computer Sciences, Graduate School of Engineering, Doshisha University

When we humans receive uncertain information, we interpret it suitably, to understand what the speaker is trying to say. This is possible because we have the word and the connection between concepts about "commonsense". In order to realize computers which can understand humans' various demands, a "Commonsense Judgment Mechanism" must be built in, to deal with such "commonsense". This paper proposes a method to understand "arithmetic problems-concentration problems" by observing the contents and composition of words of the problems, with the concept-base automatically built from dictionaries and the knowledge base which stored the knowledge about the problems.

1. はじめに

近年、我々の周りには様々なコンピュータが登場し、その発展は目覚しく、今や我々の生活には必要不可欠なものとなっている。しかし、コンピュータは人間と同じように柔軟で知的な判断を行うことができない。このような知的なコンピュータを実現するためには、人間の持つ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必要がある。常識にはものの単位や四則演算など数量に関するもの、時間や場所、感覚や感情に関するものなどがある。

本稿では、常識の中でも特に、数量に関する常識を用いて、数量に関係した自然言語文章を意味理解し、その判断を行うことを目的としている。そこで、概念ベース〔小島 2002〕などの常識知識や濃度問題に関する知識を集めた知識ベースを用い、算数問題の意味理解システム〔福業 2004〕の中でも、公式を使って解くことが可能な定式化された問題である濃度問題の意味理解を行う手法を提案する。

2. 算数問題の意味理解システム

算数問題は一般問題と定式化された問題(特殊問題と呼ぶ)に分けられる。一般問題とは、小学校の算数問題や日常会話における数量を扱う会話など、一般的な四則演算などを利用して解く問題である。例えば、「林檎が 5 個あります。蜜柑が 10 個あります。合わせていくつの果物がありますか。」というような問題である。特殊問題とは、SPI の非言語能力問題のように公式を利用して解くことのできる濃度問題、数列問題、N 進法問題や仕事算などである。

算数問題の意味理解システム(図 1)は、数量を含む自然言語文章の意味理解を行うシステムである。システムの処理は、問題のパターン分類と、個別演算システムによる意味理解に大きく分けられる。前者は、一般問題や特殊問題(濃度問題、数列問題など)のいずれかであるかを分類する処理である。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、解を導く処理である。

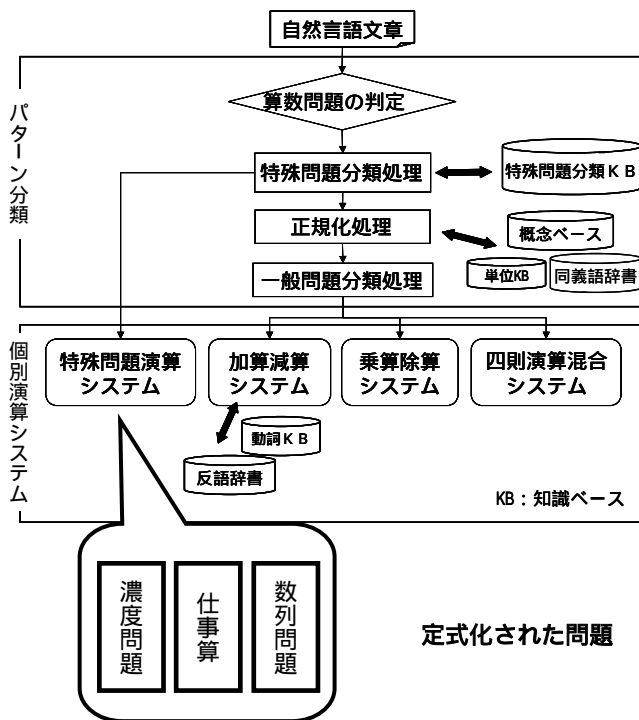


図 1 算数問題の意味理解システム

3. 濃度問題の意味理解

濃度問題の意味理解とは、算数問題の意味理解システムにおいて、特殊問題に分類される濃度問題を対象に、その意味理解を行い、適切な解を返すことである。

3.1 濃度問題とは

濃度問題とは、溶質(溶かすもの)、溶媒(ものを溶かす液体)、溶液(溶質と溶媒からなる混合物)に関する情報が与えられ、質問された答えを求める問題である。濃度問題の例としては、「食塩 80g を何 g の水に溶かすと、20%の食塩水ができるか。」のようなものが挙げられる。本稿では、特別な操作無し 蒸発

連絡先: 同志社大学大学院工学研究科 知識情報処理研究室, 〒610-0394 京都府京田辺市多々羅都谷 1-3, Tel: 0774-65-6944

操作 混合操作, のように濃度問題を操作の種類によって分類し(表 1), これらの濃度問題を意味理解の対象とする.

表 1 濃度問題の種類

濃度問題の種類(問題例)	操作の名称
100gの水に20gの砂糖を溶かして砂糖水を作った. この砂糖水の濃度は何%か.	特別な操作無し
5%の食塩水が200gある. この食塩水から何gの水を蒸発させると, 8%の食塩水になるか.	蒸発操作
4%の食塩水100gと, 10%の食塩水400gを混ぜると何%の食塩水ができるか.	混合操作

3.2 処理の流れ

以下に濃度問題の意味理解の流れを示す(図 2).

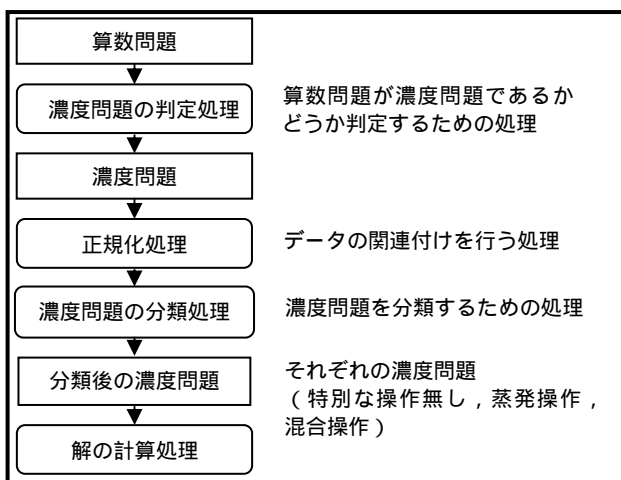


図 2 濃度問題の意味理解の流れ

以下, それぞれの処理を説明する.

3.3 濃度問題の判定処理

概念ベースや関連度計算^[渡部 2001]などの連想システムとその応用である記事関連度計算^[松田 2001], また判断知識として濃度問題知識ベースを用い, 算数問題が濃度問題であるかどうかの判定を行う.

(1) 概念ベースと関連度

概念ベースとは, ある単語(概念)とその意味特徴を表す属性と重みの集合で構成されたものである. 概念ベースには, 約 9 万語の概念が格納されており, 総属性数は約 254 万個である. 属性数は概念によって異なるが, 1 概念あたりの平均属性数は約 29 個である. ある概念 A に対して, その語の i 番目の属性を a_i , 重みを w_i , 概念 A の属性数を N 個とすると, 概念 A は以下のように表される.

$$A = \{(a_1, w_1), (a_2, w_2), \dots, (a_N, w_N)\}$$

関連度とは, 概念間の関連の強さを定量的に評価するものであり, 0 と 1 の間の数値で表す.

(2) 記事関連度

記事間の関連の強さを定量化したものを記事関連度と定義する. 記事間の関連の強さを定量的に表現することで, 記事間の関連の強弱をコンピュータに判断させることができる. 記事を形態素解析し, 固有名詞などを除いたものを索引語と呼ぶ. 記事を 1 つの概念, 記事の索引語をそれぞ

れ概念の 1 次属性と見なし, 索引語の意味を考慮するために, 概念ベースを用いて索引語からその属性を取得し, 単語間の関連度と同様の方法で記事間の関連度を計算する.

(3) 濃度問題知識ベース

濃度問題知識ベースとは, 様々な濃度問題の形式のいずれかに一致するかどうかを理解できる知識として, 濃度問題をそれぞれ問題形式で集めたものであり, 全 10 問で構成されている.

(4) 処理の流れ

以下に濃度問題の判定処理の流れを示す(図 3).

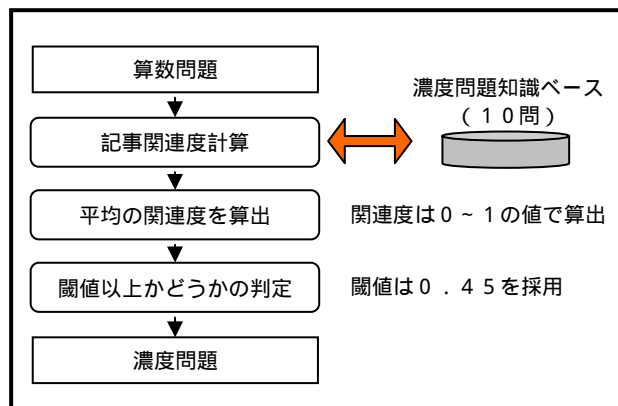


図 3 濃度問題の判定処理の流れ

まず入力文と濃度問題知識ベースの問題全てと記事関連度計算を行う. そして, 0~1 の値で算出された記事関連度から平均の関連度を算出し, その値が閾値(0.45)以上かどうかを調べる. この閾値はあらかじめ取得した濃度問題(30 問)と濃度問題知識ベースとの記事関連度計算の最小値(0.47)を参考に設定した値である. 記事関連度計算により求めた平均の関連度が閾値以上であれば, その問題は濃度問題であると判定する.

(5) 評価

SPI 問題集や Web から, 濃度問題を 50 問, 濃度問題ではない算数問題を 50 問, 計 100 問を収集し, 濃度問題の判定処理の評価を行った. その評価結果を図 4 に示す.

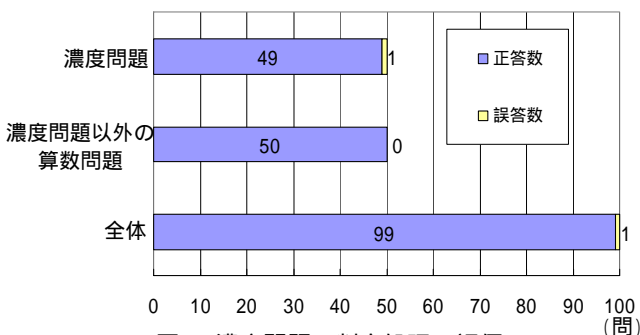


図 4 濃度問題の判定処理の評価

全体で 99%の正答率を得ることができた. 正しく判定されなかった問題は, 複雑な混合や分離操作を含む濃度問題である. 記事関連度が低くなったのは, 記事内のそれぞれの索引語の重み付けにおいて, 特に索引語の出現頻度を考慮していなかったためではないかと思われる. 索引語の頻度 TF や, 索引語が記事内に特徴的に現れているかという特定性を表すための尺度 IDF を用いた索引語の重み付けを考慮して記事関連度計算を行い, 新たに閾値を設定することで, この問題を解決できるのではないと思われる.

3.4 正規化処理

人間のようにコンピュータに問題の文章を理解させるためには、ここで述べる正規化処理が必要になる。正規化処理とは具体的に、濃度問題と判定された算数問題において、濃度問題の分類処理、解の計算処理に必要な情報を取得し、それぞれのデータの関連付けを行い、問題文の意味的特徴を捉えるための処理である。必要な情報は、問題文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り、(数量 Xa , 概念 A [その種類], 数量 Xb) や (概念 A [その種類], 数量 Xb) のような正規化表現である。例を図 5 に示す。

この処理過程では、溶質、溶媒、溶液、数値データを判断するために各種知識ベースやシソーラス [NTTコミュニケーション科学研究所 1997] を用いており、知識ベースに存在しない未知語に対しては未知語処理を行うことで対応する。また、一般的な濃度問題において、例えば金属などは溶質としてふさわしくないという事実から、溶質が金属であるかどうかの判断を行うためにシソーラスを用いた。

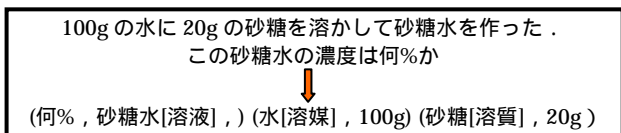


図 5 濃度問題における正規化処理の例

(1) 知識ベース

- 溶質判断知識ベース
溶質として多く出現する「食塩」、「砂糖」などの具体的な物質や、溶質であるかどうかを判断するための「固体」、「金属」などの語を集めたもので、全 6 語で構成されている。
- 溶媒判断知識ベース
溶媒として多く出現する「水」、溶媒であるかどうかを判断するための「溶媒」などの語を集めたもので、全 6 語で構成されている。
- 溶液判断知識ベース
溶液として多く出現する「食塩水」、「砂糖水」などの具体的な物質や、溶液であるかどうかを判断するための「溶液」、「水溶液」などの語を集めたものであり、全 5 語で構成されている。
- 単位知識ベース
数値を表す語であるかどうかの判断や、その語を数値に変換する時に、基準となる単位に統一するために必要となる単位(%, g, kg など)を集めたものを単位知識ベースと呼び、5 語で構成されている。本稿では、質量の基準の単位に「g」を用いている。

(2) 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 2710 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(図 6)。

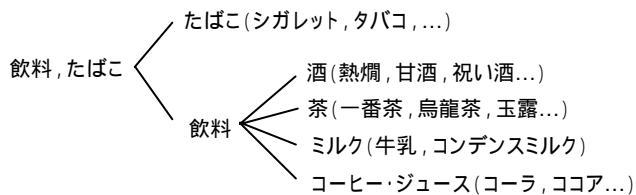


図 6 シソーラス

(3) 未知語処理

知識ベースには代表的な語のみ格納している。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るため、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、未知語処理を行う。

その方法として以下の 2 つを提案する。概念ベースを用いて未知語の 1 次属性を調べることにより知識ベースの中に一致する語がないかを調べて置換する手法の未知語処理(図 7)と、関連度計算を用いて未知語と知識ベースの中で最も関連が強いものを調べて置換する手法の未知語処理(図 8)である。

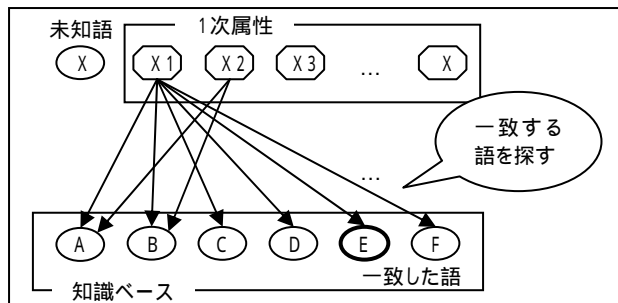


図 7 未知語処理 のイメージ

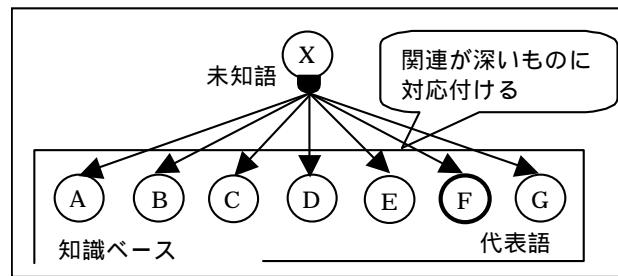


図 8 未知語処理 のイメージ

(4) 処理の流れ

以下、正規化処理の流れを示す。

問題文を形態素解析ツールで形態素解析する(図 9)。名詞と名詞をつなげるといった複合語の処理を行う。例:「食塩」+「水」 「食塩水」など。

の結果を参考に、問題文の名詞に着目し、その名詞が溶質、溶媒、溶液のいずれかに該当するかを、各判断知識ベースや未知語処理を用いることにより判断する。

() 溶質、溶媒、溶液のいずれかに該当すれば、その前後に数値を表す語があるかどうかを、単位知識ベースを用いて判断する。数値を表す語があれば、着目した名詞と数値を表す語を保持する。数値を表す語がなければ次の名詞を調べる。

() 溶質、溶媒、溶液のいずれにも該当しない場合、次に未知語処理を用いて該当するものがあれば、() の処理を行う。未知語処理を用いても該当するものがなければ次の名詞を調べる。

を問題文の最後まで繰り返し、得られた名詞と数値を表す語を処理結果とする。

同じ情報を意味する正規化表現がある場合は一方を省略する。

以下、正規化処理の具体例を示す。

例: 25%の食塩水を作るには、食塩 125g を何 g の水に溶かせばよいか。

形態素解析を行う準備として、問題文の英数記号を半角文字に、「,」と「.」を「,」と「.」に変換する処理を行う。

問題文はただの文字列であるので、まず形態素解析を行う(図9)。

25% / の / 食塩 / 水 / を / 作る / に / は / 、 / 食塩 / 125g / を / 何 g / の / 水 / に / 溶かせ / ば / よい / か / 。

図9 問題文の形態素解析

形態素解析を行うと、図9に示したように「食塩水」は「食塩 / 水」のような処理結果になるので、連続する名詞を複合語とみなす複合語処理を行う。複合語は名詞として扱う。

名詞「食塩水」に着目し、溶質、溶媒、溶液判断知識ベースを参照する。この例では、溶液判断知識ベースに「食塩水」という語があるので、「食塩水」は溶液であることがわかる。

「食塩水」が溶液であるとわかったので、その前後に数値を表す語があるかどうかを、単位知識ベースを参照しながら判断する。すると、「25%」という数値を表す語があるので、「食塩水」と「25%」という語を保持する。ここで、着目語の前後をどこまで調べるかであるが、着目語の前を調べる範囲を決める条件として、句読点「,」「.」や助詞「から」「が」「に」「へ」や他の名詞が表れたことを用いる。それ以降は調べない。また、着目語の後ろを調べる範囲を決める条件として、句読点「,」「.」や助詞「から」「に」「へ」や他の名詞が表れたことを用い、こちらも同様でそれ以降は調べない。図10に数値を表す語の保持と、調べる範囲の例を示す。

25% / の / 食塩水 / を / 作る / に / は / 、 / 食塩 / 125g / を / 何 g / の / 水 / に / 溶かせ / ば / よい / か / 。

図10 数値を表す語の保持と、調べる範囲

25% / の / 食塩水 / を / 作る / に / は / 、 / 食塩 / 125g / を / 何 g / の / 水 / に / 溶かせ / ば / よい / か / 。

図11 繰り返し結果

~ の処理により、問題文から以下の結果が得られる。

(25%,食塩水[溶液],) (水[溶媒],何g) (食塩[溶質],125g)

3.5 濃度問題の分類処理

濃度問題を3.1節で示した「操作の種類により区別した濃度問題」に分類する。濃度問題とその正規化処理の結果を参照し、溶液の状態数(ある濃度の溶液の数)の特徴や、問題文の動詞に着目することで、濃度問題の分類が行えるのではないかと考え、その条件を以下のように定義した(図12)。

溶液の状態数 1	→	特別な操作無し
溶液の状態数 2かつ「蒸発」という語が含まれる	→	蒸発操作
溶液の状態数 2または3	→	混合操作

図12 濃度問題の分類の条件

3.6 解の計算処理

濃度問題の解を求めるための基本となる公式は、

$$\text{濃度}(\%) = (\text{溶質} / \text{溶液}) \times 100 \dots ()$$

$$\text{溶質} = \text{溶液} \times (\text{濃度} / 100) \dots () \quad \text{溶液} = \text{溶質} + \text{溶媒}$$

であり、蒸発操作や混合操作の濃度問題を解くために、

$$(\text{溶質 A} + \text{溶質 B}) / ((\text{溶質 A} + \text{溶媒 A}) \pm (\text{溶質 B} + \text{溶媒 B})) = \text{溶質 C} / (\text{溶質 C} + \text{溶媒 C}) \dots ()$$

の公式を用いる。正規化処理の結果と上の公式から解を求める。

3.7 評価

濃度問題の判定処理によって濃度問題であると正しく判定された問題計147問に対して、本稿で提案した濃度問題の意味理解の手法の評価を行った。その結果を図13に示す。

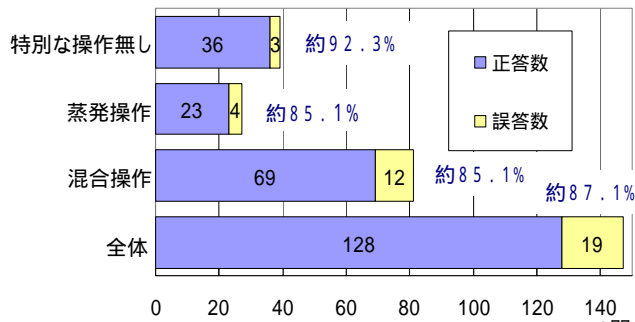


図13 濃度問題の意味理解における提案手法の評価

提案手法の正答率は約87.1%であった。各操作における正答率も全て85%以上と高い精度が得られた。誤った問題を調べてみると、質問の問われ方に「いくらか」などの語が用いられた場合や、着目語(名詞「食塩水」など)が省略された場合など、正規化処理の段階で正しく関連付けが行われていないことがわかった。今後はこれらの問題点の改善方法を考える必要がある。

4. おわりに

本稿では、算数問題の意味理解システムの特長問題に分類される濃度問題の内容に着目し、その特徴を調べることで、多種多様な濃度問題の意味理解を行う手法を提案した。これにより、これまでできなかった濃度問題に対する意味理解が可能となった。今後は、濃度問題以外の算数問題にも着目し、その意味理解を行うことで、より幅広い算数問題の意味理解の実現、そして、数量を含む自然な会話文を対象とした柔軟な判断の実現が可能になると思われる。

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「知能情報科学とその応用」における研究の一環として行った。

参考文献

[小島 2002] 小島 一秀, 渡部 広一, 河岡 司: 連想システムのための概念ベース構成法 - 属性信頼度の考え方に基づく属性重みの決定, 言語処理学会ジャーナル, 自然言語処理, Vol.9, No.5, pp.93-110, 2002.

[稲葉 2004] 稲葉 栄美子, 渡部 広一, 河岡 司: 常識知識を用いた算数問題の意味理解, 情報処理学会自然言語処理研究会資料, 2004-NL-159, pp.147-154, 2004.

[渡部 2001] 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.

[松田 2004] 松田 全弘: 記事関連度計算を用いた記事分類方式, 同志社大学工学部知識工学科卒業論文, 2004.

[NTT コミュニケーション科学研究所 1997] NTT コミュニケーション科学研究所: 日本語語彙体系, 岩波書店, 1997.