Vol.2016-NL-225 No.2
2016/1/22

大学入試化学の自動解答システムにおける格フレーム辞書を用いた係り受け解析誤りの訂正と省略の検出

吉田 達平[1,a)]　松崎 拓也[1]　佐藤 理史[1]

概要：高校化学の計算問題の自動解答システムの開発を行った．システムではまず問題文に対する言語処理を行い，問題の意味を表す中間表現を生成したのち，計算処理を行い解答を導出する．本稿では特に前半の言語処理に関して，係り受け解析誤りの修正と，ゼロ代名詞のような省略された要素の検出について報告する．これらは非文法的なパターンを検出するヒューリスティクスと，高校化学分野に対する簡易的なオントロジーに基づき選択制限を記述した格フレーム辞書を用いて行う．

キーワード：大学入試センター試験，化学，質問応答，形態素解析，係り受け解析，ゼロ代名詞，格フレーム

Tappei Yoshida[1,a)]　Takuya Matsuzaki[1]　Satoshi Sato[1]

1. はじめに

我々は大学入試化学の自動解答システムの開発を行っている．本稿では特にその言語処理部分について報告する．本研究は，国立情報学研究所を中心とする「ロボットは東大に入れるか」プロジェクト [1] の一環である．

従来の質問応答タスクは，知識源から知識を正しく検索することが中心的な課題であるが，入試化学の自動解答はそうではない．従来の研究対象にない，以下の 3 つの課題がある．**（1）問題文の理解**：従来の質問応答タスクと違い，質問文と知識源の表層的な一致や類似では解けない．化学実験のような具体的な手続きの記述を，抽象的な意味表現にマッピングための深い理解が必要が必要となる．**（2）問題文の不完全性**：表層的な情報だけでは計算を行うのに十分でない．各段階での足りない情報を，言語知識および分野知識を用いて推定する必要がある．**（3）処理の複雑性**：(1)（2）で挙げたように，問題文の理解や，知識の検索・利用が求められ，さらに計算のための数学的推論などを統合的に行う必要がある．

本稿で対象とするのは，「(1) 問題文の理解」のうちの，基本的な言語解析である．システムでは，問題文の係り受け構造に従ってその意味表示を導出するため，精確な言語解析は重要な課題である．本稿では特に，係り受け解析誤りの修正と，ゼロ代名詞の検出に注目した．既存の言語解析システムを化学の問題文にそのまま適用すると，化学用語に関して適切な形態素区切りと品詞情報を与えたうえでも，3 割程度の文において誤りが発生する．これに対しては化学の知識を用いて，言語解析の精度向上を試みる余地がある．本稿では，化学のドメイン固有の知識に基づいたヒューリスティクスと格フレーム辞書を用いて，それらの問題の解決に取り組んだ．

[1] 名古屋大学大学院工学研究科
Graduate School of Engineering, Nagoya University
[a)] tappei_y@nuee.nagoya-u.ac.jp

2. 背景

この節では，本稿で取り組む言語処理の課題の背景となっている，化学の入試問題の自動解答システムについて概説する．詳細については文献 [2] を参照されたい．

2.1 システムの全体像

システムは言語処理部と計算処理部と 2 つで構成される．言語処理部は問題文から，問題の意味を形式的に表した中間表現にマッピングする．言語処理部は，問題中の各文に対する意味表現を得る**文解析器**と，個々の文の意味表現を組み合わせ，問題全体に対する意味表現へと変換する**文脈処理器**に分けられる．システムの全体像を図 1 に示す．本

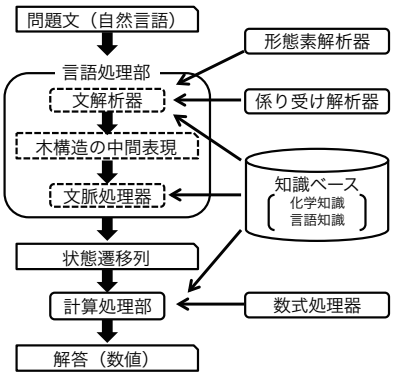

図 1 システムの全体像

ラベル	単語
SUBNAME	塩酸，水酸化ナトリウム水溶液，アルミニウム，…
BODY	気体，固体，沈殿，…
EVE	加える，加熱（する），溶解（する），…
HEAT	熱量，生成熱，燃焼熱，…
QUA	物質量，質量，濃度，…

表 1 用語分類ラベルの例

稿で取り組むのは，このうち言語処理部のうち文解析器の改良である．

言語処理部：文解析器

文解析器は係り受け構造に沿って各文の意味表現を組み立てる（図 2）．文の意味表現は，物質名や化学反応などのイベント名をノードとする木構造で，物質の量や，イベントに関わる物質といった親ノードの属性を子ノードが表す．文解析器の処理は以下の 3 つのステップからなる．① **用語へのラベル付**: 化学物質名，燃焼・混合といったイベントを表す述語，また物理量など，化学の入試問題において重要な役割を果たす語に対しラベル付けする．「一酸化炭素 => SUBNAME（物質名）」のように，その上位概念にあたるラベルを付与する．ラベルの例を表 1 に載せた．② **係り受け解析**: 形態素解析および係り受け解析を行う．本稿の実験では，形態素解析には mecab[3]，係り受け解析には cabocha[4] を用いた．③ **木構造の生成**: 例えば「一酸化炭素が生成する」というフレーズに対しては，「一酸化炭素が→生成する」という係り受け構造に従って，「生成する」のイベントの@substance（生成する物質）属性として「一酸化炭素」を設定する．

文解析が終わった段階で，個々の文に関して完全な意味表現を得るためには，(1) 親ノード-子ノードの関係からなる木構造が正しい，(2) 各ノードに対して，必要な属性情報がすべてそろっている，という 2 点が必要となる．よって文解析器では，(1) のために正確な係り受け解析が必要となり，(2) のために正しい省略（ゼロ代名詞）検出が必要となる．

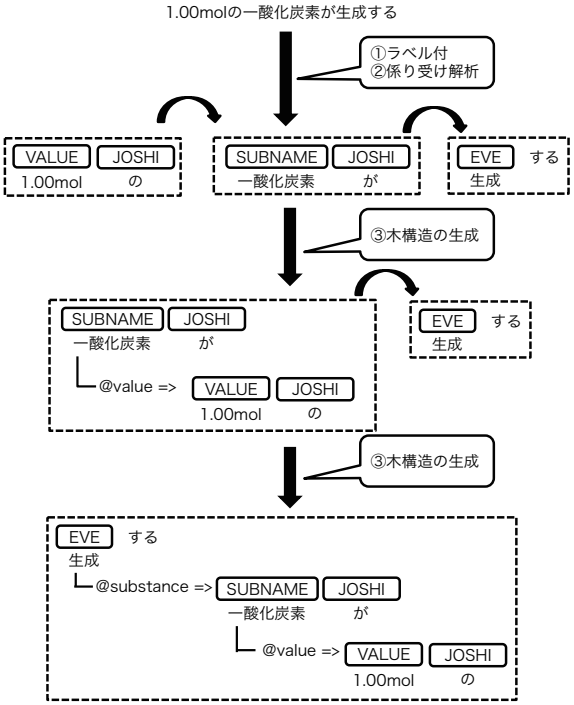

図 2 あるフレーズを言語解析する例

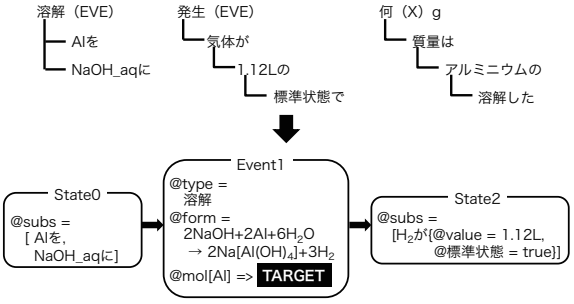

図 3 変換処理の例

言語解析部：文脈処理器

文脈処理器は文解析器で得られた解析結果を，問題文が記述する状況を状態遷移列としてあらわす中間表現へとマップする．この中間表現は，状態を表すノードと，生成といったイベントを表すノードが交互に現れる構造となっている．状態を表すノードは，どんな物質がどんな状態であるか，という情報の集合である．図 3 に言語処理で得られた「溶解」「発生」「何 g」を root とする 3 つの木構造の言語処理結果から，状態遷移列を生成する例を示す．

計算処理部

言語処理部で得られた状態遷移列を入力として計算処理を行う．まず問題文に与えられていない数値をすべて変数に置き換える．例えば図 3 の問題文において，発生した気体の体積は 1.12L と定数として与えられているが，水酸化ナトリウム水溶液の体積は与えられていないため，これを表す変数を用意する．次に変数同士の関係を表す方程式を，化学反応式や物理量間の関係を定める公式から立式する最後にそれらを連立方程式として解くことで解答を得る．連立方程式の求解には maple を利用している．

[

ラベル	単語
(1) 言語処理の課題	12/22
(2) 特定の化学知識の不足	13/22
(3) 問題表現の拡張	9/22

表 2 システムの課題

ラベル	品詞の制約
物質名	名詞-一般
数値	名詞-数
単位	名詞-接尾-助数詞

表 3 ラベルに対応した品詞の制約の例

2.2 システムの現状

開発データに対しシステムを動かしたところ，大学入試センター過去問 26 問のうち 4 問，benesse 模試 52 問のうち 11 問について自然言語から解答の数値を導出することができた．開発データについては節 6.1 で詳しく述べる．大学入試センター過去問の解けない 22 問について問題を分析すると，大きく分けて 3 つに課題が分類でき，その数を表 2 にまとめた（重複あり）．

(1) 言語処理の課題：言語解析誤りや，複雑な言語表現の理解など基本的な言語処理の問題
(2) 特定の化学知識の不足：熱化学や電気分解などは特定ドメインの知識が必要
(3) 問題表現の拡張：状態遷移列では表現しきれない問題が一部存在する

本研究は次の段階として，これらを処理する枠組みをシステムに追加し，より頑健なシステムを目指していく．本稿ではその第一歩として，精確な言語解析に取り組む．

3. 化学問題の言語処理で解決すべき課題

この節では本稿で取り扱う言語処理上の課題について説明する．

3.1 言語解析誤り

3.1.1 形態素解析誤り

問題文を既存の形態素解析エンジンに入力すると，化学式や物理量など，化学特有の表現の周囲で誤りが起きやすいことが観察された．具体的には，開発データ 94 文に対して形態素解析を行うと，5 文，6 形態素では誤りが生じた．物質名などの化学用語が未知語であることによる誤りを防ぐため，化学用語に関しては形態素区切りと品詞情報の制約を mecab （-p オプション付き）への入力として与えた．品詞情報の制約については表 3 に例をまとめた．

誤りのパターンを全て表 4 にまとめた．「名詞 + (並列助詞 | 格助詞)」という形のフレーズで誤りが多く生じていることがわかる．誤りの内容としては，名詞の直後に接続詞がくるような文法的に不適切な文や，フィラーや終助詞「や」といった非文法的な形態素並びが多い．特定のパターンに誤りが集中しているため，ヒューリスティックルールによる修正が有効であると期待できる．

誤りが生じた形態素（太字）とその周辺文脈	誤り出力	正解
塩酸**と**水酸化ナトリウム水溶液	接続助詞	並列助詞
アルケン A1.4g **と**アルカン B6.0g	フィラー	並列助詞
Cl^- **や** SO_4^{2-}	終助詞	並列助詞
SO_4^{2-} **と**反応して	接続助詞	格助詞
アセチレン 1mol **が**生成	接続詞	格助詞
7.2mL **で**変化した	接続詞	格助詞
ある濃度の	動詞	連体詞

表 4 開発データ 94 文中の形態素解析誤り

文節の種類	正解	誤り出力
名詞→述語	**銅に→加え**	**銅に→加熱した**
名詞→名詞	**100cm^3 に→含まれる**	**100cm^3 に→メタノール分子の**
その他	それぞれ→反応させた	それぞれ→十分に

表 5 文脈と無関係に文法・語法的に誤っている係り受け関係

文節の種類	正解	誤り出力
名詞→述語	100g を→加熱して	100g を→融解させ
名詞→名詞	メタンと→エタン	メタンと→混合気体を
述語→述語	融解させ→加熱して	融解させ→上昇させて
述語→名詞	占める→混合気体	占める→メタンと

表 6 文法的にはあり得るが，意味的に誤っている係り受け関係

文節の種類	(1) 文法的・語法的な誤り	(2) 意味的な誤り
名詞→述語	**20**	2
名詞→名詞	**4**	1
副詞→副詞	1	0
述語→述語	0	8
述語→名詞	0	1

表 7 係り受け解析誤りの種類と数

3.1.2 係り受け解析誤り

cabocha を用いて実験データに対して係り受け解析を行うと，文単位では 3 割程度が正しく解析できない．新聞等のテキストに対する精度に比べて低い値とは言えない．しかし前節で述べたようにシステムは正確な係り受け解析を前提にしており，一つの問題の中でただ一箇所の係り受け解析誤りが原因で解答が得られないことが，たびたび発生する．一般に，誤った係り受け関係は以下の 2 種類に分けられる．

(1) 文脈と無関係に文法・語法的に誤っているもの
(2) 文法的にはあり得る係り受けだが，意味を考えると誤っているもの

(1) の例を表 5 に，(2) の例を 6 にまとめ，それぞれの数を表 7 にまとめる．(1) を解決するためには，文法的・語法的な誤りを検出するための文法的知識が必要である．(2) を解決するためには，図 3 のような状態遷移列に自然言語文を変換した後，化学的な非整合性を検出する知識が必要である．

本稿では（1）のタイプの誤りのうち，特に数が多い「名詞→述語」と「名詞→名詞」について取り組んだ．これは図 2 の文解析結果を得た時点での誤り修正に集中するためである．

本稿では修正の対象としないが，（2）のタイプの誤りとしては以下のような例がある．

- 112L を**占める**メタンとエタンの混合気体

「**占める**」の係り受け先は「混合気体」であるが，これを cabocha で解析すると「占める→メタン」と出力する．これは文法的に誤りとは言えない．この解析誤りは，この出力結果に従って 112L のメタンと未知の量のエタンが存在するとして処理を進め，連立方程式を解く段階で，解が不定あるいは存在しないことが判明するまでは検出できない．

3.2 省略

入試化学の問題文では，述語の格要素が省略（ゼロ代名詞化）されることが多い．典型的な例を示す．

- ある濃度の HCl10.0mL を 0.50mol/L の水酸化バリウム Ba(OH)2 水溶液で中和滴定したところ，10.0mL **加えた**ときに過不足なく反応が完了した．

化学の実験操作としての「加える」という動詞には，「加える物質」と「加えられた物質」の 2 つの要素が関わる．つまり，「物質 1 に物質 2 を加えた」「物質 1 を物質 2 へ加えた」などの格要素が意味的には存在するが，ここではそれら 2 つの格要素が両方とも省略されている．この場合，省略を用いずに表現した場合の一例は「**水酸化バリウム Ba(OH)2 水溶液**に**ある濃度の HCl10.0mL** を加えた」となる．

解析誤り同様に，語彙的情報を用いて省略を検出・補完したい．このためには意味的に必須の（無ければ省略されている）格要素をまとめた格フレームの知識が必要である．このような知識があれば，係り受け解析結果に基いて省略を検出することができる．例えば，ヲ格が「加える」に必須であることが分かっていれば，先の例では省略を検出できる．またこのような処理をするためにも，正しい係り受け解析は必須である．

格フレームを用いた日本語ゼロ代名詞の検出に関する先行研究としては，例えば Seki et al. (2002) [5] や笹野と黒橋（2011）[6] によるものがある．これらの研究ではいずれも新聞記事や WEB テキストを対象として評価を行っており，先行詞の同定までを含め 40%〜50%程度の精度が得られたことを報告している．化学問題を解くという目的のためには，この程度の精度で十分とは考えにくい．しかし化学問題のように，より限定的なドメインでは比較的小規模な格フレーム辞書を用いて，より高い効果を挙げられる可能性がある．

id	述語	が	を	で	と
1	混合する	×	物質, 1+	×	物質, 1+
2	混合する	×	物質, **2**+	×	×
3	**混合される**	物質, 1+	×	×	物質, 1+
4	発生する	×	**熱**, 1	×	×
5	中和する	×	物質, **1**	物質, **1**	×
6	希釈する	×	物質, 1	物質, **0**+	×

表 8 格フレーム辞書の一部

4. 格フレーム辞書の設計と開発

3 節でのべた課題を解決するための基本的な知識源として格フレーム辞書を開発した．辞書のフォーマットは表 8 に示す形をとる．1 行は 1 フレームとに対応している．各フレームは，ある述語が持ちうる格要素の組み合わせ，および，それぞれの格要素の意味クラスとその数（単数・複数・いずれも可）を指定する．ID にそって各エントリを順に説明する．

(1) 物質 1 を物質 2 と混合する
(2) （複数の）**物質を**混合する：ヲ格の名詞句は並列句や指示詞「これら」などで表される，複数の物質である必要がある．
(3) 物質 1 が物質 2 へ**混合される**：「混合する」と「混合される」のように同一の述語の異なる態は別エントリとして扱う．
(4) **熱**が発生する：格要素の意味クラスは物質が中心だが，それ以外にも色々なラベルが入りうる
(5) 物質 1 を物質 2 で中和する：「混合する」は意味の性質上何個でも名詞が入りうるが，「中和する」のヲ格要素は単一の物質のみを含む．
(6) 物質を希釈する / 物質 1 を**物質 2 で**希釈する：この場合，デ格の物質はあってもなくてもよい．水溶液を希釈するならデ格は水であることは自明であるからである．

この節では，以下，格フレームの設計の詳細について述べる．

この格フレーム辞書のエントリは動詞に限らず，入試化学問題の解析で重要となる述語的表現一般に対して格フレームを付与した．具体的には以下の 3 種類である

(1) 「混合する」「燃やす」「(体積を）占める」などのように，中間表現の生成に関わる動詞
(2) 「何 mol か」のような「何 + ... + 終助詞か」という形の質問を表す表現
(3) 判定詞（コピュラ）「だ」「である」．

以降，「名詞 + 格助詞 → 述語」の組み合わせを「格要素-述語ペア」と呼ぶ．

4.1 格要素に対する制約

格フレーム中の格要素は 3 種類に分類できる．

(1) 必ず存在する（無ければ省略されていると考えられる）格要素：表 8 では必要なエンティティの数として「1」「1+」「2」…と表現されている

(2) あってもなくてもよい格要素：表 8 では 0 個以上，つまり 0+と表現されている．

(3) その述語が取り得ない格要素：表 8 では「×」として表現されている．

開発する格フレーム辞書も，「格要素-述語ペア」をこの 3 値で評価する．

「名詞+助詞」の文節は，単一のエンティティ（物質など）だけを示しているとは限らない．代表的な例は並列である．例えば「物質 1 と物質 2 を混合する」というフレーズにおいて，「物質 2 を」という文節は，「物質 1 と物質 2」という並列名詞句を形成したあとで「混合する」に係っている．並列以外には「これら」等，そもそも複数のエンティティを指す表現がある．辞書の設計において，このような現象は重要である．例えば「物質 1 を物質 2 と混合する」と「物質 1 と物質 2 を混合する」は意味としては同じである．しかし前者のフレームは「ヲ格とト格に 1 つ以上の物質が必要」，後者のフレームは「ヲ格に 2 つ以上の物質が必要」という制約として表すことができる．このような，格要素に含まれているべきエンティティの数に関する制約を，「1」(1 つだけ)，「1+」(1 つ以上いくつでも)，「2」(ちょうど 2 つ）などのラベルで表す．

格要素に対する選択制限は，主として以下の 3 種類の意味クラスに格要素を分類することで表現する．現在は以下の 3 種類である

(1) 物質（酸素，塩酸，気体，など）

(2) 物理量名（物質量，質量，など）

(3) 数値（1.00mol，0.49%，など）

ただし一部のフレームでは，上記のもの以外の意味クラス，意味クラスでなく具体的な名詞の表層形など，さまざまな粒度で制約をかける．上記以外の意味クラスの例としては，例えば「占める」という動詞は，「1L の体積を占める」という表現はあるが，「1mol/L の濃度を占める」という表現は意味をなさない．よって格フレーム辞書としては「体積を占める」は可であるが，「[物理量] を占める」は不可となる．このように，述語によっては（1）〜（3）の意味クラスだけでは選択制限を十分に表現できない．

4.2 格の種類

格フレーム辞書で定義する格は「が」「を」「に」など格助詞に 1 対 1 に対応するものに限らない．具体的には以下のものである．

(1) 格助詞：「が」「の」「を」「に」「へ」「と」「から」「より」「で」「や」

述語	nil	を	に	で
加える	体積	物質	物質	×

表 9 格フレーム辞書の一部

foreach x **in** 入力文の N-best 形態素解析結果
  $x' \leftarrow$ ルールによる x の修正
  foreach y **in** x' に対する N-best 係り受け解析結果
    if y が名詞-名詞関係誤りを含む **then** 次の候補へ
    if y が格フレーム制約に違反する **then** 次の候補へ
    return y // 誤りを含まない解析が見つかった
  end
end

図 4 形態素・係り受け解析誤りの修正アルゴリズム

(2) その他の助詞：「は」(係助詞)，「まで」(副助詞)

(3) 格助詞相当句：「を用いて」

(4) 仮想的な助詞：nil

係助詞「は」と格助詞「が」は書き換えれる場合と，そうでない場合があるので，運用のしやすさのために辞書の設計としては区別する．例えば「水**は**気体である」というフレーズを，「水**が**気体である」とは書き換えられない．格助詞相当句も格助詞と同様に扱う．

仮想的な助詞 nil は，名詞と述語が助詞を介さず係り受け関係を持ちうる場合に用いる．この場合，仮想的な助詞「nil」が存在すると考え，他の助詞と全く同様に表 9 の形式で扱う．これによって，例えば「16mL →加える」といった係り受けを正解と判定し，「塩酸→加える」という係り受け解析誤りを検出することができる．

5. 形態素・係り受け解析誤りおよび省略の検出

5.1 形態素・係り受け解析誤りの検出・修正

形態素解析および係り受け解析において，それぞれ n-best 解析結果を得て，誤り修正のルールを適用する．アルゴリズムの概要を図 4 に示す．形態素解析の n-best 出力は mecab-ipadic の-N オプションを用いて得た．係り受け解析の n-best 出力は，cabocha に実装されている Sassano (2004) [7] の Shift-reduce 法による解析アルゴリズムの各段階で，スコア上位 n 個の解析候補を保持するよう改変することで得た．

形態素誤り修正は mecab-ipadic の n-best 解析結果をそれぞれ書き換えることで行う．書き換えには，パターンに基づいたヒューリスティックルールを適用する．形態素解析結果が表 10 のパターンに当てはまった場合に，表 10 に従って書き換える．「名詞+**と**」の場合は，並列助詞の場合と格助詞の場合について，それぞれ書き換えた解析結果を 2 つ出力する．そして，図 4 に従って，書き換えた n-best 解析結果から，スコアが最も高い順に cabocha へ入力する．

係り受け解析誤り修正は，図 4 に従って，cabocha の出力結果から誤りでない最も高いスコアのものを選ぶことで行う．誤りの検出は格フレーム辞書との矛盾の有無を調べるのに加え，名詞を主辞とする文節どうしの関係の誤りを

パターン	正解の品詞
名詞+**と**	並列助詞 または 格助詞
名詞+**や**	並列助詞
名詞+**が**	格助詞
名詞+**で**	格助詞
ある+［物理量］	連体詞

表 10 係り受け解析の実験結果

ヒューリスティックルールの適用でチェックすることで行う．前処理として，格要素中のエンティティの数に関する制約をチェックするために，並列の検出が必要である．係り受け関係のある 2 つの文節が次の条件を両方満たしている時に並列とした．

(1) どちらの文節も主辞は名詞であり，同じ意味クラスに属する（物質名，物理量，数値，など）
(2) 係り受け元の文節に，並列助詞「と」，並列助詞「や」，格助詞「から」，接続詞「および」，または当該の名詞で文節を終える．

5.1.1 誤った名詞-名詞関係の検出

以下の 2 つのルールで，名詞を主辞とする文節動詞の誤った係り受け関係を検出する．

ルール 1

これは「2 つの名詞に係り受け関係がある時，係り受け元の文節の助詞に基いて誤りを検出する」ルールである．例としては「塩酸が→質量を」のような係り受け解析誤りを検出する．この例での「塩酸が」という文節は，原則として述語の文節にしか係り得ない．具体的には，係助詞「は」あるいは格助詞「が」「を」「に」「へ」「から」「より」「で」のいずれかが名詞の後に現れる文節が，名詞を主辞とする文節に係っている場合に誤りとする．ただし，名詞+判定詞の形の文節に係る場合，あるいは「〜から〜までの」の形の句を作る場合は誤りとしない．

ルール 2

これは「同じ意味クラスに属する名詞を含む文節どうしの係り受け関係は，並列句を作るものだけを許容する」ルールである．例えば同じ意味クラス【物質名】に属する「塩酸」と「硫酸」は，「塩酸と→硫酸」という係り受け関係はあり得るが，「塩酸の→硫酸」という関係はない．同じラベルが続く場合，係り受け元の文節は並列助詞「と」を含むなど，前述の並列句の条件を満たしていなければならない．この条件を満たさない，同じ意味クラスの名詞が係り受け関係を持つが，並列ではない場合に誤りとして検出する．

5.1.2 格フレーム辞書との比較

格フレーム辞書との比較は，名詞の文節から述語の文節への係り受けの正しさを調べる．4 章で，格要素-述語ペアを 3 種類に分類したうちの，「ありえない」フレームとラベル付けした格要素が，係り受け解析結果にあった場合，誤りとする．格フレーム辞書で，並列区など複数のエンティティを想定している格要素が，単数のエンティティとなっている場合，現状はすべて省略が起きているとして誤りとして検出しない（後述）．同様にに，単数のエンティティが想定される格要素が，複数のエンティティとなっている場合は誤りではないとした．例えば，「中和」という反応は「中和されるもの（ヲ格）」と，「中和するもの（デ格）」は意味的にはそれぞれ絶対に単数なので，格フレーム辞書もそれに従って単数としている．しかし「硫酸と塩酸を中和する」のようにヲ格に複数の要素がきても語法的には問題がない．ただしこの時，2 節の文脈処理器において，「硫酸の中和」と「塩酸の中和」という，二つの状態遷移列が生成される．

5.2 省略の検出

省略の検出も名詞から述語への係り受けと，格フレーム辞書を比較することで行う．文中の述語に対して，(a) 格フレーム辞書の対応するエントリ，(b) 文中で係っている格要素，の 2 つを比較する．(a) において「必須の格要素」とされたものが，(b) で満たされてなければそこに省略が発生していると検出できる．複数のエンティティが要求される格要素には，並列名詞句や「これら」がくる必要がある．この条件が満たされない場合は，差分の要素数だけ省略が存在すると検出する．

「塩酸を反応させた」というフレーズで例を示す．「反応させる」について，格フレーム辞書に以下の 3 つのエントリがあるとする（実際はさらに多くのエントリがあり，複雑な制約がある）．

(1) ［物質］を［物質］と反応させる（{"を"=>1，"と"=>1}）
(2) ［物質］と［物質］を反応させる（{"を"=>2}）
(3) ［物質］を反応させる（{"を"=>1}）

() 内は格助詞と格要素の持つエンティティの数を，{"格助詞"=>数}という表記で表したものである．(b) としては{"を"=>塩酸}となる．ここで (a) と (b) を比較すると，本来の格フレームが (1) であった場合，省略された要素は{"と"=>1}である．同様に (2) の場合は，ヲ格が複数なので差分をとって，省略として{"を"=>1}を検出する．(3) の場合は省略は起こっていない．

今回は省略を言語現象としてまとめ補完に向けて考察をするが，その補完はしない．補完を行う場合，前述の例のように複数の格フレームの候補からの絞込が必要である．例の場合，次のような推論が行われる．

- (1) と (2) は化学的には同値なのでどちらでもよい
- 高校化学の範囲で塩酸が単独で起こる化学変化は出てこないので (3) はあり得ない

このように，語彙的情報だけでない，化学的な推論を伴うので，今回は対象外とした．

問題の例

アルミニウムを水酸化ナトリウム水溶液に溶解させたところ，
標準状態で1.12Lの気体が発生した．
溶解したアルミニウムは何gか．

図 5 計算問題の例

モデル	数学 500 文	化学 94 文
(1) のみで訓練	0.845	0.93
(1) + (2) で訓練	0.937	0.94

表 11 使用する学習モデルの係り受け精度（文節単位）

6. 係り受け解析誤りの自動修正に関する評価実験

6.1 実験条件・実験データ

開発データは benesse 模試 18 回分から抽出した計算問題 53 問とし，実験データは Z 会模試 [8] と駿台模試 [9] から抽出した計算問題 19+11 問とする．本稿で言う計算問題とは，図 5 の例のように「アルミニウム質量は何 g か」と問題文中で求めるべき数値が明示されているものを指している．つまり，計算を含まない問題や，計算を用いた上で正誤判定する問題は対象外とする．

評価に用いるのは中間表現の生成に必要な文のみとし，それぞれ開発データで 93 文，実験データで 84 文が該当した．中間表現の生成に不要な文とは，例えば次の問題における太字で示した文である．

不要な文の例

エタン C2H6 が完全燃焼すると，水と二酸化炭素が生じる．1mol のエタンが完全燃焼する時，過不足なく反応する酸素の質量は何 g か．**最も適当な数値を下の ①〜⑥ のうちから一つ選べ．g**

不要な文は，この例で示した 2 つに分類することができる．(1) 問題の背景：「エタン C2H6 が…生じる」は，ヒントとして一般的な事実を述べたものであり，中間表現の生成のためには不要である．(2) 指示のための定型文：「最も適当な…選べ」は，解答の形式を指示する定型文であり言語処理する必要はない．

cabocha の学習モデルは 2 つ用意した．一つは cabocha とともに配布されている，京大テキストコーパスで訓練を行ったモデル，もう一方は次の 2 つのデータから学習している．(1) 京大コーパス 24283 文，(2) センター試験数学 IA，IIB および国立大 2 次試験数学問題のテキストに係り受けアノテーションを施したもの計 9086 文．このモデルを用いた数学の問題文 500 文，化学 94 文でのテストの結果を表 11 に示す．

6.2 実験結果とその分析

ベースラインシステムと 2 つの提案システムの係り受け解析の精度を評価した．

システム	純正モデル	数学モデル
baseline	64/94	59/94
システム 1	68/94	67/94
システム 2	72/94	76/94

表 12 開発データでの係り受け解析結果

システム	純正モデル	数学モデル
baseline	61/84	59/84
システム 1	64/84	60/84
システム 2	62/84	61/84

表 13 実験データでの係り受け解析結果

- baseline：mecab+ipadic と cabocha をそのまま適用したシステム
- システム 1：baseline に，形態素解析誤り修正を追加したシステム
- システム 2：システム 1 に，係り受け解析誤り修正を追加したシステム

正解率は文単位で評価し，一文中のすべての文節境界および係り受け関係が正しい時に正解とした．評価した結果を表 12 と表 13 にまとめる．提案手法であるシステム 2 として，開発データ内の非文法的な係り受け解析誤りをすべてカバーできるよう，格フレーム辞書とヒューリスティックルールを開発したところ，開発データに関しては約 70%から約 80%へ向上した．しかし，その過程で得られたシステム 2 を実験データに対して適用したところ，大きな性能の向上は見られなかった．一方で，baseline に形態素解析誤り修正のみを追加したシステム 1 では，開発データ，実験データともに，形態素修正ルールが悪影響を及ぼすことはなかった．現在の修正ルール（表 10）による悪影響がないことは，今後データを増やしてさらなる検証を行いたい．

システム 2 の開発データに結果の内訳を表 14 に示した．全て文の数である．

(1)「修正された誤り」：形態素・係り受け解析誤り修正によって狙い通り誤りが修正された誤り
(2)「修正されなかった誤り」：テストでは修正に失敗したが，辞書に適切な格フレームがあれば，理想的には修正することができた誤り．意味的な処理をしないと修正されないものは含まれていない．
(3)「誤りと検出された正解」：baseline システムで出力した正しい解析結果が誤りとして検出された誤り

なお，(1) については，2 つのモデルの一方だけで修正された場合も 1 文として数えている．(2) と (3) の原因は全て格フレーム辞書の不足であった．名詞を主辞とする文節同士の係り受け解析誤りを検出するヒューリスティックルールは悪影響を及ぼさなかった．このように，(1) 〜 (3) の結果，全体としての性能の向上は見られなかった．

原理的にはできる誤りの修正に失敗した原因は，想定外の格フレームの出現である．本手法は解析結果と矛盾しな

現象	数
(1) 修正された誤り	6
(2) 修正されなかった誤り	8
(3) 誤りと検出された正解	3

表 14 実験データでの係り受け解析テストの結果の内訳

タイプ	数
(1) 述語の被修飾語	27
(2) 先行する述語の格要素	22
(3) 言い換え元の述語の格要素	9
(4) 問題文中には出てこない	19
計	77

表 15 開発データ 94 文に観測された省略

い格フレームが，辞書内のエントリに一つも存在しない場合に誤りと出力する．よって述語ごとに格フレームを十分に書ききらないといけないという課題がある．実験の結果 94 文の開発データでは足りないことが分かったので，今後データを追加してさらなる実験を行う必要がある．

7. 省略された格要素の自動検出および先行詞の出現パターンに関する調査

5 節で述べた格フレームを用いたゼロ代名詞の検出手法について，開発データを用いた評価を行った．結果として 77 箇所でゼロ代名詞が検出された．以下で述べる，述語による被修飾語と共参照の関係にあるものを含めれば，これらはすべて省略と考えるのが妥当なものであった．本稿では取り上げなかったが，ゼロ代名詞の先行詞を同定する処理は解答システムの言語処理部で重要な位置を占める．本節の残りでは，この処理の高精度な実現に向け，開発データで観察されたゼロ代名詞と先行詞の関係について分類し考察を行う．

7.1 開発データに観測された省略のパターン

省略を開発データ 94 文に関して調査し，どんなパターンがあるか言語現象としてまとめた．述語と，ゼロ代名詞の先行詞との関係は，以下の 種類に分類できることが分かった．

(1) 述語の被修飾語
(2) 一連の化学的変化・操作が連続する場合に，先行する述語の格要素である
(3) 述語自体が先行する述語の言い換えである場合に，先行する述語の格要素と共参照の関係にある
(4) 問題文中には出てこない

開発データ 94 文に観測された（1）〜（4）の数を表 15 にまとめる．(1) の例としては，「含まれている酸」というフレーズが典型的である．この時「含まれる」の省略されたガ格と「酸」が共参照の関係にある．

(2) の例を「**これら**を混合したのち純水を**加え**て 100mL をとり…」というフレーズで示す．この場合「加える」はニ格が省略されているが，照応先は「混合する」のヲ格にあたる「これら」である．このフレーズは「混合したのち…加えて…とり…」という一連の操作を描写している．このような時間的に連続した操作や化学変化の記述において，操作の中心となっている格要素はたびたび省略される．

(3) の例を「**HCl** を **Ba(OH)2 水溶液**で中和滴定したところ，10.0mL **加え**たときに…」というフレーズで示す．この場合，「加える」のヲ格とニ格が省略されているが，その先行詞は，それぞれ「Ba(OH)2 水溶液」，「HCl」である．ここでの「加える」とは「中和滴定」と同一のイベント（あるいはその一部）を指す．よって「加える」の省略されている格要素は，省略前の 2 つの格要素が助詞を変えてそのまま当てはまる．このように言い換えられたり，あるいは全く同一の表現で同じイベントが指示されるときに省略が起きやすい．

(4) の例を「生じた PbCl2 の沈殿を**ろ過し**て全て**除去し**た．」というフレーズで示す．この時，「ろ過」，「除去」，に省略されカラ格の格要素が支持する対象は，ともに前後の文中には出てこない．省略されているのは「今操作している物質」である．このような場合，ゼロ代名詞の指示対象を同定する処理は複雑になる（現在確認しているのは「今操作している物質」のみである）．

7.2 照応問題の解決に対する考察

照応の解決には，イベント表現（実験操作・化学変化に関する述語）に関するオントロジーを作成する必要がある．例えば（3）の例を処理するには，

(a) 概念的には「加える」は「中和滴定」の上位にあり，後者は前者で言い換えられる
(b)「中和滴定」を「加える」で言い換える場合，前者のヲ格・デ格が，それぞれ後者のニ格・ヲ格に対応しているという 2 つの手がかりが必要である．

このように，イベント表現間の概念的関係，格要素のイベント表現における役割（中和する物質 / される物質のような），が少なくとも必要である．他にも格フレーム辞書により化学的に詳細な制約を表現することも必要である．(4) の例では「ろ過する」のカラ格は，現在の格フレーム辞書では単に「物質」としているが，「ろ過」の意味を考えると「少なくとも液体と沈殿物が混ざった混合物」と絞り込める．

他には問題文中のイベント表現の間の関係性を，より正確に把握する必要を認識している．例えば (2) の例の「混合」と「加える」は連続して起こる 2 つのイベントで，(3) の例の「中和滴定」と「加える」は同一のものである．このような処理を行うためにはイベント表現や物質名のような内容語だけでなく，接続詞や非自立語に注目しなければならない．実際 (2) と (3) の例ではそれぞれ「混合した**のち**」「中和滴定した**ところ**」というように，ともに非自立語が構造を読み取る鍵となる．

8. まとめ

センター試験「化学」の計算問題における言語処理の問題に取り組んだ．今回は，形態素解析・係り受け解析の誤りの修正と，ゼロ代名詞の検出への検討を行った．形態素・係り受け解析誤りの修正は，質問応答のドメインが化学に閉じていることを利用した，格フレーム辞書とヒューリスティックルールを適用したシステムを開発した．しかし係り受け解析誤りの修正は，開発データの不足から格フレーム辞書が十分な規模得られず，オープンテストでは大きな向上が見られなかった．ゼロ代名詞の検出に対する検討では，言語知識と化学知識を組み合わせたオントロジーが必要な事がわかった．引き継き，格フレーム辞書やオントロジーの強化を続け，入試化学の問題文の言語処理の精度向上を目指す．

参考文献

[1] 新井紀子, 松崎拓也. ロボットは東大に入れるか?—国立情報学研究所「人工頭脳」プロジェクト—. 人工知能学会誌, Vol. 27, No. 5, pp. 463–469, 2012.
[2] 吉田達平, 松崎拓也, 佐藤理史. 大学入試化学の計算問題の自動解答システム. 2015 年度人工知能学会全国大会, 2015.
[3] 工藤拓. Mecab: Yet another part-of-speech and morphological analyzer. `http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html`.
[4] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. Vol. 43, No. 6, pp. 1834–1842, 2002.
[5] Kazuhiro Seki, Atsushi Fujii, and Tetsuya Ishikawa. A probabilistic method for analyzing japanese anaphora integrating zero pronoun detection and resolution. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pp. 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
[6] 遼平笹野, 禎夫黒橋. 大規模格フレームを用いた識別モデルに基づく日本語ゼロ照応解析. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3328–3337, dec 2011.
[7] Manabu Sassano. Linear-time dependency analysis for japanese. In *Proceedings of Coling 2004*, pp. 8–14, Geneva, Switzerland, Aug 23–Aug 27 2004. COLING.
[8] 大学入試模試センター. 2010-駿台 大学入試完全対策シリーズ 大学入試センター試験実戦問題集化学 I. 駿台文庫株式会社, 2009.
[9] Z 会出版編集部. 平成 22 年用 センター試験 実戦模試 6 化学 I. 株式会社 Z 会, 2009.