センター試験「化学」正誤判定問題に対する自動解答システム

古瀬 弘樹 †　　松崎 拓也 ‡　　佐藤 理史 ‡
†名古屋大学 工学部 電気電子・情報工学科　‡名古屋大学大学院 工学研究科
{h_furuse, matuzaki, ssato}@nuee.nagoya-u.ac.jp

1 はじめに

本研究は、国立情報学研究所を中心とした、「ロボットは東大に入れるか」プロジェクトの一環として行ったものである。本研究の対象としたのは、大学入試センター試験「化学」のうち、図1のような、複数の選択肢から正しい（あるいは誤った）ものを一つ選ぶタイプの問題で、かつ、教科書などのテキストを検索することによって各選択肢の正誤を判定しうるタイプの問題である。ひとくちに「テキスト検索によって正誤を判定しうる」と言っても、問われる内容は多岐にわたり、例えば図1の選択肢③のように化学用語や概念の定義を問うものもある一方で、選択肢⑤のように日常生活における化学物質の用法を問うものもある。さらに、選択肢①②④のように反応など、物質の性質や変化を問うものもある。

本研究では、問題の選択肢及び知識源となる教科書データからキーワードを抽出し、両者を比較することによって正誤を判定することを試みた。さらに誤りであろう選択肢を事前に検出し、選択肢を絞るために、教科書の文や wikipedia の項目との比較によって原因となっているキーワードを特定することを試みた。

先行研究として、Kanayama ら [1] および Kobayashi ら [2] によるセンター試験「世界史」の自動解答システムがある。これらの研究では、選択肢中のキーワードを一つずつ隠し、隠した語を問う factoid 問題に対するシステムの解答を隠した語と比較することで誤りを検出する手法を用いている。本研究の選択肢絞り込みも同様のアイデアに基づいているが、隠した語と同じ系統の語と比較するだけでなく、隠した語の系統が不明な場合でも適用できる絞り込み手法を新たに考案し評価を行った。

2 正誤問題自動解答システムの構成

本研究で作成した解答システムの全体像を図2に示す。まず入力された自然言語文を形態素解析し、キーワードとなる語を抽出する。次に2つの手法を用いて選択肢を絞り込む。最後に、各選択肢と教科書のテキスト文から得られたキーワードとの一致の度合いを計算し、解答を選び出す。

問1　金属に関する記述として**誤りを含むもの**を，以下から一つ選べ。
① ナトリウムの単体は、常温で水と激しく反応する。
② 鉛の単体は、希塩酸や希硫酸によく溶ける。（誤り）
③ ブリキは、鋼板の表面にスズをメッキしたものである。
④ アルミニウムは、希塩酸にも濃い水酸化ナトリウム水溶液にも溶ける。
⑤ ステンレス鋼は鉄を主成分とする合金であり、腐食しにくいため台所の流し台に利用されている。

図 1: 「化学」正誤判定問題の例

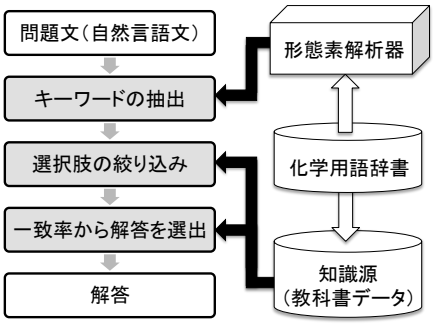

図 2: システムの流れ

2.1 用語辞書の作成

問題文および選択肢の形態素解析には MeCab [1] を使用した。化学の問題文には物質名など化学用語が多数含まれる。例えば「ノナン」という物質名は「ノ」「ナン」と分割されてしまい、正しい解析結果が得られない。このような解析誤りを防ぐため、予め化学における重要な用語をまとめ、MeCab の辞書に追加した。追加した用語には、例えば「水酸化ナトリウム」「塩基性」といった語が含まれている。用語辞書の作成のためには、センター形式模試の過去問と高校化学の教

[1] http://taku910.github.io/mecab/

科書である数研出版社の「化学」(2012) を使用した。

2.2 キーワードの抽出と解答を選出

上で述べた用語辞書に、化学で頻出する用言を加えたものをキーワードリストとした。キーワードリストには 1639 個のキーワードが登録されている。このうち、用言は 21 個である。これを基に、教科書データと問題それぞれからキーワードを抽出する。教科書データ T は一文ごとに区切り、各文 s から抽出したキーワード集合を $k(s)$ とする。問題文については、各選択肢 o から抽出したキーワード集合を $k(o)$ とし、さらに問題文のうち選択肢以外の部分から抽出したキーワードは全ての $k(o)$ に加えた。

抽出したキーワード集合 $k(s)$ と $k(o)$ を比較しキーワード一致率を算出する。一致率の計算の際は、キーワードごとに異なる重みをつけた。キーワード x の重み $w(x)$ は $w(x) = N/C(x)$ と定義する。ここで、N は教科書の全単語数、$C(x)$ は教科書における x の出現回数を表す。選択肢 o と教科書の文 s のキーワード一致率 $A(o, s)$ は以下のように定義する。

$$A(o,s) = 100 \cdot \frac{\sum_{x \in k(o) \cap k(s)} w(x)}{\sum_{x \in k(o)} w(x)}$$

この $A(o, s)$ を教科書のすべての文に対して計算し、その最大値を選択肢のスコア $A(o)$ とする：$A(o) = \max_{s \in T} A(o, s)$ 。これを全ての選択肢に対して行い、正しいことを述べている選択肢を探す問題ならキーワード一致率 $A(o)$ の最も高いものを、誤りを含む選択肢を探す問題なら、$A(o)$ の最も低いものを出力する。

2.3 選択肢の絞り込み

開発データを用いて $A(o)$ を実際に計算し、結果を観察したところ、正しい内容の書かれた選択肢は比較的 $A(o)$ が高くなるものが多く、著しく低いものは存在しなかった。しかし、誤りを含む選択肢については、$A(o)$ が低いものが多数ある一方で、高い値になるものも多く、正しい内容の選択肢と比較すると、$A(o)$ の値がより広範囲に分布していた。誤りの選択肢の中で $A(o)$ の高いものを分析してみると、一語だけその内容に適さない語にすり替えられて作られた誤りの選択肢が多数見られた。このような選択肢では、すり替えられていた語を除くと正しい内容を表すキーワード集合が残るため、教科書とのキーワード一致率が比較的高くなってしまうと考えられる。この対策として二つの手法であらかじめ選択肢を絞ることを考えた。

【問題の選択肢 o 】
五円玉に使用されている黄銅は、銅とスズの合金である
$k(o)$ = {"黄銅", "銅", "スズ", "合金"}
【教科書データに存在する文 s 】
黄銅は、銅と亜鉛で作られる合金で、真ちゅうともいう。
$k(s)$ = {"黄銅", "銅", "亜鉛", "合金", "真ちゅう"}

除外したキーワード v	$k(o) \setminus \{v\}$	一致個数 $(k(o) \setminus \{v\}) \cap k(s)$
"黄銅"	{"銅", "スズ", "合金"}	2
"銅"	{"黄銅", "スズ", "合金"}	2
"スズ"	{"黄銅", "銅", "合金"}	3
"合金"	{"黄銅", "銅", "スズ"}	2

図 3: キーワード消去を用いた選択肢の絞り込み

2.3.1 キーワード消去を用いた選択肢の絞り込み

正しい内容を表す文から一語だけすり替えられて作られた誤りの選択肢ならば、すり替えられたキーワード v を $k(o)$ から除いた場合、教科書の文と一致するキーワードの個数は、v 以外のキーワードを除いた場合の一致個数よりも多くなるはずである。

例として、一語だけ誤った選択肢をこの手法によって検出する例を図３に示す。この例では「五円玉に使用されている黄銅は、銅とスズの合金である」という選択肢について正誤を判定するわけだが、この選択肢は、「亜鉛」が「スズ」にすり替えられた誤りの選択肢である。教科書データには「黄銅は、銅と亜鉛で作られる合金で、真ちゅうともいう。」という文 s が存在したとする。このとき、選択肢のキーワードを一つずつ除いて教科書データのキーワードと比較していくと、「スズ」を除いた場合のみ一致個数が大きくなる。これは残り全てが、正しい内容を持つ教科書の文 s に含まれているためである。

この考えをもとに、選択肢 o それぞれに対して以下の手続きを行う。まず、抽出されたキーワード $k(o)$ からひとつのキーワード v を除いて、残ったキーワード $k(o) \setminus \{v\}$ と教科書のキーワード $k(s)$ を比較し、v を除いた時の一致個数 $E(o, v)$ を計算する。

$$E(o,v) = max_{s \in T} |(k(o) \setminus \{v\}) \cap k(s)|$$

$E(o, v)$ を比較してある v に対してだけ $E(o, v)$ が高く、残りのキーワードに対しては $E(o, v)$ が等しい場合、その選択肢は誤りの可能性が高いと判断する。また、$E(o, v)$ が全て同じ場合は、その選択肢は正しい可能性が高いと判断する。それ以外の場合は、この手法では判断できないため絞り込みの対象とはしない。

2.3.2 用語集を用いた選択肢の絞りこみ

一語だけすり替えられたキーワードに注目すると、同じ系統の語にすり替えられている場合が多い。例え

表 1: 2 種の絞り込みによる判定

絞り込み手法	選択肢に対する判定					
2.4.1 の絞り込み	×	×	○	○	不明	不明
2.4.2 の絞り込み	×	不明	×	不明	×	不明
最終的な判断	×	×	不明	○	×	不明

ば「五円玉に使用されている黄銅は、銅とスズの合金である」という選択肢では、「亜鉛」が同じ金属元素である「スズ」に書き換えられている。そこで選択肢に、ある系統に属するキーワード、例えば金属元素名が出現した場合、それをその他の金属元素に置き換えたときに正しい内容を表すキーワード集合となるかどうかを判定し、誤った選択肢を検出する手法を考えた。

この手法では、wikipedia から化学に関するページを選出し、タイトルとテキスト（以下「説明文」と呼ぶ）を対応付けたデータを用いた。タイトル t の説明文を $T(t)$ で表す。まず選択肢から抽出されたキーワード $k(o)$ から、金属元素の集合 M に含まれるキーワード y を抜き取る。残ったキーワード $k(o) \setminus \{y\}$ をそれぞれの金属元素 $m \in M$ の説明文 $T(m)$ と比較して、一致したキーワードの重みの和 $S(o, y, m)$ を計算する。

$$S(o, y, m) = \sum_{x \in (k(o) \setminus \{y\}) \cap k(T(m))} w(x)$$

ここから最大値を与える金属元素名 $\hat{m}$ を求める。

$$\hat{m} = argmax_{m \in M} S(o, y, m)$$

$\hat{m} \neq y$ なら元々のキーワードは不適切であると判断し、その選択肢は誤りの可能性が高いとする。

なお、今回は誤りを含む選択肢を観察した際、同系統のキーワードとして金属元素が多く見受けられたため、金属元素のみを対象としてこの手法を取り入れた。手法としては金属元素以外にも用いることができる。

2.3.3 2 種の絞り込みによる判定

2.4.1 と 2.4.2 で述べた手法を併用して、選択肢の候補を絞り込む。2 種の絞り込み結果の併合は表 1 のように行う。×は絞り込みによって誤りを含む選択肢と判断された場合、○は絞り込みによって正しい内容が書かれた選択肢だと判断された場合、不明は判断不能であった場合を示す。正しい選択肢を選ぶ問題の場合は、最終的に○または不明と判断された選択肢から一致率計算を行って解答を選び、誤りを含む選択肢を選ぶ問題の場合は、最終的に×または不明と判断された選択肢から一致率計算を行って解答を選ぶ。

表 2: 正誤判定問題出力結果

	正解数	選択肢毎
一致率計算のみ	35/100	318/491(64.78%)
2.4.1 の絞り込みのみ	32/100	310/491(63.14%)
2.4.2 の絞り込みのみ	35/100	316/491(64.36%)
2 種の絞り込みを使用	34/100	305/491(62.19%)
開発データ	12/27	94/134(70.15%)

表 3: 絞り込みの結果

種類	TP	TN	FN	FP	不明
2.4.1 の絞り込みのみ	83	43	49	29	287
2.4.2 の絞り込みのみ	0	28	29	0	434
2 種の絞り込みを使用	77	52	65	24	273

3 結果・考察

センター形式模試の過去問を用いて、システムの評価を行った。教科書データとして、東京書籍「化学 1」(2007)、東京書籍「化学 2」(2009)、岩波書店「理化学辞典第 5 版」(1999)、の 3 つを合わせて用いた。また、システムの評価に用いる問題データとして、進研マーク模試、代ゼミ センター試験模試、駿台 センター試験実戦問題集、河合 マーク式総合問題集から問題を抜粋し、そのうち、5 択の問題 26 問、4 択の 1 問からなる 27 問（合計 134 選択肢）を開発用データ、5 択の問題 91 問、4 択の 8 問、6 択の 1 問からなる 100 問（合計 491 選択肢）を評価用データとした。

3.1 実験結果

評価結果を表 2 に示す。開発データに対する結果は 2 種の絞り込みを使用したもののみ示す。さらに、個々の選択肢を、1 問ずつの二択問題とみた場合の評価も行った。この場合は開発データでの結果を元に、キーワード一致率 $A(o)$ が 70 %以上なら正しい、70 %未満なら誤っていると判断して、結果を算出した。多肢選択問題としてみると 20.36 %、二択の問題としてみた場合は 50.0 %がベースラインとなるので、提案手法はランダムな解答よりは明らかによい性能と言える。

絞り込みの精度は表 3 に示す結果となった。491 選択肢のうち、TP は実際に正しいものを正しいと判断した場合、TN は実際に誤っているものを誤っていると判断した場合、FN は実際には正しいものを誤っていると判断した場合、FP は実際には誤っているものを正しいと判断した場合の数を示す。ただし、FN、FP でも直ちに不正解となるとは限らない。例えば誤りを見つける問題で正しい選択肢を FN と判断した場合で

表 4: 誤った解答の原因

原因	該当問題数
教科書情報不足・複数文に渡る記述	41
キーワード不足	22
絞り込みに問題	12
解析ミス	9
化学特有でない語が多い	8
キーワード数、問題文長に偏り	7

も、誤りの選択肢の候補が増えるだけである。

表 2 の通り、選択肢の絞り込みを行っても結果は改善しなかった。特に 2.4.1 の手法では、誤りの原因となるキーワードを除いた時に一致個数が多くなる事で誤りが検出できると期待したが、それ以外のキーワードを除いた場合に、誤りの原因となるキーワードにすり替える前の選択肢の内容とは無関係な教科書文で一致個数が多いものが存在し、一致個数に差が出なかったために、誤りの検出に失敗した場合が多く見られた。

3.2 考察

2 種の絞り込みを利用した際に誤った解答を出力した 66 問に対して、原因を分析し、表 4 にまとめた。なお、複数の原因があるものはそれぞれカウントしている。「教科書情報不足・複数文に渡る記述」は、そもそも教科書データに直接は記載されていない内容を問う問題や、選択肢の内容が教科書では複数の文に分けて記述されていた問題を指す。例えば「硝酸は、光に対して不安定なため、褐色のビンに保存される」という選択肢は正しい内容を示しているが、教科書データでは 2 文に分けてこの内容が記載されていたため、この選択肢から抽出したキーワードを多く含む文が存在せず、低い一致率となってしまった。

「キーワード不足」は正誤を判断する上でキーワードが十分に抽出できていなかった問題を指す。例えば「黄リンは空気中で自然発火するので、水中に保存する」という選択肢において、本来「自然発火」「水中」などもキーワードとして抽出すべきものであるが、現在のキーワードリストにはこれらが含まれていなかった。また、類似する原因にはなるが、「化学特有でない語が多い」は化学用語をほとんど含まない選択肢から、キーワード抽出が十分にできなかったケースを指す。例えば「ビニール袋に密封されていた使い捨てカイロを袋から取り出し，空気中でよく揉んだところ，熱が発生した」という選択肢からはキーワードとして「熱」しか抽出できなかった。

「絞り込みに問題」は選択肢の絞り込みの際、誤った絞り込みをした場合や、本来絞り込んで欲しいものを「真偽不明」として逃した場合を指す。これは 3.1 でも述べた通りである。

「キーワード数、問題文長に偏り」は選択肢を比べた時、文の長さに差があったり抽出したキーワード数が大きく異なったため、長い文あるいは多数のキーワードを持つ選択肢が不利になってしまった問題を指す。

表 4 の通り、最も多い原因は教科書データの不足や複数の文をうまく利用できなかった場合であり、次いでキーワード不足によるものが多く見られた。教科書に存在しない情報に関しては、形式化した知識データを網羅的に作成し利用するのが有効だと思われる。また、複数文に渡る情報も、組み合わせて利用することを考えると、形式化したデータとして持つのが利用しやすいと考えられる。一方でキーワード不足が原因のもの、特に化学特有でない語が多数使われる問題は、テキスト検索以外の方法で対応するのが難しい問題が多く、化学用語以外の名詞等もキーワードに加えて補充し、検索の精度を上げるのが有効だと思われる。

4 おわりに

本研究では、高校化学において重要な物質名などの用語を辞書として整理し、これを利用したセンター試験「化学」の正誤問題に対する自動解答システムの開発を行った。ランダムベースラインと比較すると成果はあったものの、絶対的な正答率はまだまだ低い。

今後は、同じ方式で全部の問題を解くのではなく、問題ごとに中間表現に翻訳し、化学物質や化学反応式などのデータベースから必要な知識を検索し、個々の問題に対応していく方法を考えている。

謝辞 教科書データおよび模試データを提供いただいた「ロボットは東大に入れるか」プロジェクトおよび、東京書籍様、ベネッセコーポレーション様、学校法人高宮学園代々木ゼミナール様に感謝いたします。

参考文献

[1] Hiroshi Kanayama, Yusuke Miyao, and John Prager. Answering yes/no questions via question inversion. *Proc. COLING*, pp. 1377–1391, 2012.

[2] Mio Kobayashi, Hiroshi Miyashita, Ai Ishii, and Chikara Hoshino. Nul system at qa lab-2 task. *Proc. NTCIR-12*, pp. 414–420, 2016.