二級自動車整備士試験問題からの キーワード自動抽出の試み

青　木　恒　夫

1　はじめに

二級自動車整備士の資格試験(検定および認定)は例年，ガソリン自動車整備士試験が 3，7，10 月に，ジーゼル自動車整備士試験が 3，10，12月の計6回実施される。本学のような自動車整備士を養成する教育機関では，その教育効果を高める目的で学生に過去の試験問題を教材として提供する。従来，ここで提供される教材は，試験問題コピーの切り貼りによって，用途に応じた編集を施したのちに利用されていたが，10年ほど前からはワード・プロセッサ専用機の普及により，電子的な編集へと変化している。また，パーソナル・コンピュータの普及により，より汎用的なテキスト形式によるデータの提供も一般化してきた。しかし，提供される問題はデータベースとして利用できるほど詳細な整理分類がなされているわけでなく，多くは5分野(工学一般，エンジン，シャシ，電気，法規)程度の大まかなカテゴリーに手作業で分類して利用されているに過ぎない。

今回，テキスト形式で記述された二級自動車整備士試験問題から最小単位の問題(以下，単位問題と称す)を自動的に切り出し，各単位問題の出題情報，カテゴリー，問題形式，正解情報等をインデックスとして付加するプログラム abskey(*abstract the keyword*)を開発した。本プログラムは，日本語形態素解析システムと共に動作し，問題文中に含まれる自動車技術関連用語(以下，自動車用語と称す)をキーワードとして自動抽出する機能を有する。

2　abskey の動作アルゴリズム

abskey は次の例のように，二級自動車整備士試験問題のファイル名を引数として起動する。

```
abskey mondai.txt > result
```

起動後は引数に指定された問題文を解析し，単位問題ごとにインデックスを付加したテキスト・データを出力する。結果は，CSV(*Comma Separated Value*)形式で「標準出力」[1]に出力されるが，この例のようにOS(*Operating System*)のリダイレクション(*redirection*)を利用してファイルに格納するのが一般的な使用法である。以下の節では，abskey の動作アルゴリズムを説明する。

[1]通常はディスプレイが標準出力に割り当てられている。

2.1　出題情報の取得と単位問題の切り出し

ガソリン／ジーゼルの別，出題年月などの出題情報は問題文のヘッダから取得する。以下に問題文冒頭の例を示すが，行頭の番号とコロン(:)は筆者が説明のために付加したものである。

```
1:２ｄ１９９２０３
2:２級ジーゼル　１９９２年（平成４年）３月
3:
4:〔１〕次の各々について，適切なものには○を，適切でないものには×を記入しなさい。
5:１．エンジンの容積効率は，過給すると向上する。（正解○）
```

1行目の「２ｄ１９９２０３」は，試験の種別(二級ガソリン／ジーゼル)と実施年月を示す。2行目は利用者のための識別用見出しである。3行目の空行をはさんで4行目以降が問題の本文となる。問題本文からは，単位問題ごとに大問題番号(〔１〕)，小問題番号(１．)および問題本文を別々に抽出する。単位問題には，○×や法令記述問題のように1行で完結するものと選択問題のように大問題番号の範囲全体で一つの問題を構成するものに分けられる。筆者の知る限りでは，二級自動車整備士試験問題の問題構成は固定化されているので，単位問題を切り出したり，出題分野を特定することは比較的容易である。各問題の末尾には，「(正解○)」や「(正解イ－７，ロ－８，ハ－７，ニ－２，ホ－１)」のような正解となる情報が付加されているので，正解だけを抽出することもできる。

2.2　問題形式の推定

abskeyでは問題形式を以下の7つに分類し，プログラム内部で形式を推定したのちにインデックスの一つとして出力している。

- **単○**：1行で完結する○×問題
- **記述**：1行で完結する法令記述問題で，正解は主に数値による記述が要求される
- **計算**：計算問題
- **選択**：選択穴埋め問題
- **総○**：複数の○×問題から構成される総合問題で，1行で完結する○×問題に分割できない
- **複合**：選択，○×，計算等の複数形式の問題を含むもの
- **不明**：abskeyで推定できなかったもの

二級自動車整備士試験問題の問題構成から，「単○」および「記述」は特定の大問題番号に属するため，機械的に決定できる。また，大問題〔２〕も伝統的に計算問題であり，問題形式の特定

は容易である。ただし，計算問題はシャシや電気の分野で時々出題されることがあり，問題文から推定する必要がある。

問題文から問題形式を推定するには，問題形式の特徴を表す複数のキーワードを問題本文から検索することにより行う。例えば「選んで」というキーワードがあれば，ほぼ選択問題と推測できるが，更に「記号で」というキーワードがあれば選択問題と推定できる。abskeyでは，このように複数のキーワードを問題本文から検索することにより，過去十年間の試験問題では，100%の精度で問題形式を推定することができる。

2.3 形態素解析

1節で述べたように，abskeyは日本語形態素解析システムをプログラム内部で起動することにより，問題文中に含まれる自動車用語を自動的に抽出し，キーワードとして出力する機能を持っている。形態素解析とは「自然言語に対して，意味を持つ最小の言語単位(形態素)に分解すること」であるが，詳細については文献[1]を参照していただきたい。

abskey内部では，切り出された単位問題を一時的なファイルとしてディスク上に書き出し，その問題文ファイルを形態素解析プログラム Breakfast 4.0.4.f [3]に処理させている。処理結果は品詞タグを付けた別の一時ファイルとして出力させ，そのファイルからキーワードを抽出している。Breakfastの形態素解析アルゴリズムについては公開されていないが，JUMANの説明[2]をそのまま引用すると，

- ある特定の位置からはじまるすべての可能な形態素を辞書引きによって得る．
- 辞書引きによって得られた個々の形態素に対して，その直前の位置に存在するすべての形態素との連接可能性のチェック，および，コストの計算を行なう．
  - 連接可能性のチェックによって連接不可能とわかった形態素間の接続は行われない．また，その位置で連接可能なもののうち最良(最小)のコストと比較して `.jumanrc` の(コスト幅 ??)によって定義される数値以上のコストの差を持つ形態素の接続は行われない．

とある。BreakfastもJUMANの辞書を形態素解析に用いていることから，解析結果の優先度を決定する「コスト」の概念はほぼ同様と思われる。

品詞情報を含む形態素解析結果から自動車用語のみを抽出する方法はいくつか考えられる。当初，日本語形態素解析システムの標準仕様で形態素解析を実施し，「名詞類」[2]と解析された形態素について，別途準備した自動車用語辞書から該当する用語の有無を調べるアプローチを試みた。

[2]使用した形態素解析システムには普通名詞，サ変名詞，固有名詞など6種類の名詞が定義されている。

しかし，次の例で示すように「自動車用語」として意味を持つ範囲の形態素を独自に抽出することはできなかった。

「燃料噴射時期」→
「燃料(普通名詞)」＋「噴射(サ変名詞)」＋「時(名詞性名詞助数辞)」＋「期(普通名詞)」

そこで，日本語解析システムの解析辞書に自動車用語を含ませ，更に自動車用語が優先して抽出されるために，その単語の「品詞コスト」を他の名詞類より小さく設定することにした。具体的には，自動車用語は「固有名詞辞書」(Noun.koyuu.dic)に含ませることにし，品詞コストを「10」[3]に設定して優先度を高めた。しかし，自動車用語以外で問題文中に含まれる「ときの」や「厚」などの人名と思われる単語まで抽出されることから，従来の固有名詞辞書(約37,000語)をすべて破棄し，新たに自動車用語のみを登録した固有名詞辞書を作成した。その結果，自動車用語は他の名詞に比べ優先的に解析され，固有名詞として抽出された形態素はすべて自動車用語であることが確認された。

2.4　キーワードの抽出

自動車用語[4]のみを固有名詞辞書に登録し，Breakfastに品詞タグ付で処理させると次のような解析結果が得られる。例文は，2.1節で示した問題文である。なお，解析結果中の「␣」は空白を明示的に表したものである。

エンジン␣固有名詞
の␣名詞接続助詞
容積効率␣固有名詞
は␣副助詞
，␣読点
過給␣固有名詞
する␣動詞␣␣␣␣␣␣␣␣␣␣サ変動詞␣␣␣␣␣␣␣基本形
と␣述語接続助詞
向上␣サ変名詞
する␣動詞␣␣␣␣␣␣␣␣␣␣サ変動詞␣␣␣␣␣␣␣基本形
。␣句点

abskeyは，この形態素解析結果から，品詞タグに「固有名詞」が含まれるすべての行の形態素を自動車用語のキーワードとして抽出している。なお，同名のキーワードが複数抽出されること

[3]他の名詞類のうち普通名詞，副詞的名詞，サ変名詞，数詞，時相名詞は100に，形式名詞は10に設定されている。
[4]実際の運用では出題内容の特徴を表す用語を登録しており，自動車用語には限定していない。

が予想されるが，抽出されたすべてのキーワードをプログラム内部で一旦ソーティングし，重複した出力が行われないよう配慮している。

3　自動車用語辞書(形態素解析辞書)の作成

Breakfastで用いるJUMAN形式の品詞辞書は，次に示すような1形態素1行で構成される。

```
(名詞 (固有名詞 ((見出し語 潤滑装置) (読み じゅんかつそうち))))
(名詞 (固有名詞 ((見出し語 小型四輪貨物) (読み こがたよんりんかもつ))))
```

キーワード抽出を目的とした自動車用語辞書の作成には，一般に公開されている自動車用語辞典[4][5]などの見出しを利用する方法が考えられる。しかし，「摩擦力」などの自動車技術用語に限定されない用語は含まれないことが多く，出題内容の特徴を表すキーワードの抽出という要求を満たすことができなかった。そのため，実際の試験問題テキストから出題内容の特徴を表す単語のみを残す方法により，例に示すような「見出し語,読み」から構成される辞書ファイルを独自に作成することにした。

```
潤滑装置,じゅんかつそうち
小型四輪貨物,こがたよんりんかもつ
```

1回の試験問題から平均250個ほどの見出し語が抽出できる。複数の試験問題においては当然であるが，見出し語の抽出を行っている試験問題内においても重複した抽出が予想される。しかし，以下の作業で見出し語の重複をさける処理を行うので，作業者が重複した抽出を行っても構わない。なお，見出し語の右に続くコンマと読みは手入力による。

抽出した辞書ファイルは試験問題ごとに分割されているので，OSのコマンドやテキスト・エディタを利用して1個の辞書ファイルにまとめる。次に，別途作成したmkbfdic(*Making the Breakfast dictionary*)というプログラムを使用してJUMAN形式の辞書に変換する。mkbfdicは，重複した見出し語を削除し，JUMAN形式の辞書に変換するプログラムで，見出し語ファイルの異常部を検出する機能も持っている。

以上の方法により，10年間の試験問題から約1,000個の見出し語を持つ，二級自動車整備士試験問題に対応する自動車用語辞書が作成できた。

4　abskeyによる処理と結果の評価

abskeyは，Pentium 90MHz程度のコンピュータの場合，漢字6000文字ほどの試験問題を約6分で処理する。処理時間で最も大きいウェイトを占めるのは形態素解析であるので，速度の速いコンピュータを用いれば処理速度はさらに向上する[5]。

[5]Pentium II 450MHz Dual, Windows NT 4.0 Serverでは，約30秒で処理が終わる。

出力は単位問題ごとに，10個のカラムが半角のコンマで区切られたテキスト(CSV形式)として「標準出力」へ出力される。各カラムはダブル・コーテーションで前後を囲まれて出力される。複数のキーワードが出現する場合には，半角のスペースが区切りとして挿入される。実際の出力は次の例のように，カラムの先頭から 1.試験の種別(２ｄ，２ｇ)，2.出題年(西暦)，3.出題月(2桁)，4.大問題番号，5.小問題番号(小問題に分けられない場合は０)，6.分野(工学一般，エンジン，シャシ，電気，法令)，7.問題形式(単○，記述，計算，選択，総○，複合，不明)，8. キーワード(複数の場合は半角スペース区切り)，9.問題本文，10.正解 となる。

”２ｄ”,”１９９２”,”０３”,”１”,”１”,”工学一般”,”単○”,”エンジン 過給 容積効率 ”,”エンジンの容積効率は，過給すると向上する。”,”○”

出力されるCSV形式は，表計算やデータベースに標準で読み込み可能な形式であるので，複数のデータ検索を組み合わせたアプリケーション，特に教材用試験問題自動編纂システムや学生用の自習システムなどの作成に応用できる。

出力結果について，出題内容の特徴を表すキーワードの抽出が大きな課題であったが，実際の試験問題から辞書を作成しているため，ほぼ満足する出力が得られた。ただ，自動車用語辞書によって性能が大きく左右されるので，更に慎重な辞書メンテナンスが必要と思われる。

5　おわりに

本学では，1997年から学内LAN(*Local Area Network*)において二級自動車整備士試験問題(全文テキスト)の提供を積極的に行っている。教員は各回の試験問題をワープロ・ソフトウェア上で編集して利用されているようだが，適切なデータベースとして提供されていないので，勘に頼る編集にならざるを得なかった。しかし，abskey によって体系的に整理されたデータを提供することが可能となる。より積極的に多方面の応用に供していただければ幸いである。

参考文献

[1] 長尾 真ら： 自然言語処理, 岩波講座 ソフトウェア科学, 岩波書店 1996.

[2] 松本裕治ら： 日本語形態素解析システム JUMAN version 2.0 使用説明書, 京都大学工学部 長尾研究室, 奈良先端科学技術大学院大学 松本研究室 1994.

[3] 株式会社 富士通研究所： Breakfast 4.0.4f User's Manual, 株式会社 富士通研究所 1997.

[4] 大須賀和美： 新自動車用語辞典〔改訂増補〕, 精文館書店 1985.

[5] 社団法人 自動車技術会： 自動車用語和英辞典, 社団法人 自動車技術会 1997.