中国語 MC-FIB 問題の AQG システムの開発

黄　魯南*　　星野　綾子**　　加藤　恒昭*
*東京大学大学院総合文化研究科〒153-8902 東京都目黒区駒場3-8-1
huang.lunan@gmail.com　　kato@boz.c.u-tokyo.ac.jp
**NEC共通基盤ソフトウエア研究所〒211-8666神奈川県川崎市中原区下沼部1753
a-hoshino@cj.jp.nec.com

1 はじめに

AQG（Automatic Question Generation）は練習問題を自動的に作成することである。CALL（Computer-Assisted Language Learning）分野のトピックの一つとして、近年多くの関心が寄せられている。

MC-FIB（Multiple-choice Fill-In-the-Blank）とは、問題文（文章）に空所を作り、そこに埋めるために最も適切なものを選択肢の中から選択させる形式の問題である。また以下のように、錯乱肢によって、A）語彙問題、B）文法問題、C）混合問題、三つのタイプに分けられる。

A) I want to [ ] to school.
1. go 2. make 3. take 4. do
B) I want [ ] to school.
1. to go 2. go 3. going 4. gone
C) I want [ ] to school.
1. to go 2. to make 3. gone 4. made

図 1　MC-FIB問題の類型

MC-FIBの自動生成には、空所の決定、錯乱肢の作成など、複数の課題があるが、本研究は英語の先行研究を参考し、中国語の性質に合わせて、中国語の語彙問題を中心にしたMC-FIB問題の錯乱肢を作成する手法を提案する。そして、被験者実験を用いて、他の手法で作成した錯乱肢との比較を通して、本手法の有効性を検証する。そして、手法の改善と将来の課題を検討していく。

2 先行研究

2.1 Coniam（1997）の手法（参考文献[1]）

Coniam は錯乱肢生成に、削除された単語、すなわち問題の正解となる単語（以下、本稿ではWと称す）の頻度と品詞の情報を用いた手法を提案した。その手順は以下の通りである。

1. 文章のすべての単語に品詞タグをつける。
2. 単語を文章から削除し空所を作る。
3. W の品詞と頻度を得る。
4. W と同頻度帯にあり、かつ品詞が同じの単語を錯乱肢とする。

2.2 Sumitaら（2005）の手法（参考文献[5]）

Sumita らはシソーラスを利用して錯乱肢を生成する手法を提案した。そして、別解を生成しないように、サーチエンジンのヒット数を用いた錯乱肢候補のフィルタリングを採用している。その手順は以下の通りである。

1. コーパスから問題文を抽出する。
2. 空所位置を選択する。
3. Wのすべての類義語をシソーラスから抽出し、候補集とする。
4. 候補集から、単語を一つずつ取り出し、問題文の空所に代入する。
5. 上記単語を含んだ問題文の一部をサーチエンジンにかけ、ヒット数を観察する。
6. ヒット数がゼロより大きい場合、その単語を候補集から取り除く。
7. 最後に、候補集に残った単語を錯乱肢とする。

Sumita らは、英語母語話者と日本人受験者を利用し、二つの実験を行い、手法の有効性を検証した。第一の実験では、まず母語話者に生成した問題を見せる。受け入れられる問題なら解かせ、そうでなければエラーが発生したとする。その結果、母語話者の正解率は、93.5%の正解率であった。第二の実験では、TOEIC スコアが 400 点から 900 点までの間の学習者を100名集め、提案手法で作成した320問を解かせた。その結果、スコア間の相関係数は 0.8 であったという。

2.3 先行研究の分析

先行研究では、錯乱肢生成について、頻度情報、品詞情報、意味情報と三つの情報を利用した手法が提案され、また、別解を生成しないように、サーチエンジンヒット数を利用して、錯乱肢を選別する手法が提案されている。

以下は本研究と関係のある、錯乱肢生成において、頻度情報、品詞情報、サーチエンジンヒット数それぞれを用いる有用性を分析する。

頻度情報：単語の難しさを測る一つの重要な基準はその頻度である。一つの頻度帯にある単語

は、学習者にとって同等の難しさをもつと見なすことができる。学習者のレベルはがWの頻度までの単語が分かるものとしたら、Wより頻度が高すぎるものは、学習者にとって簡単すぎると推測される。逆に、Wより頻度が大幅に低いと、学習者にとって難しすぎる。

品詞情報： 品詞情報は錯乱肢を作るというより、錯乱肢候補を絞る条件の一つとして、英語を対象とした頻度情報を用いた手法で、補助的に使われている。英語の場合は単語のスペルで、品詞が推測できることがあるため、品詞まで違った単語を正解と間違える学習者が少ないと考えられるので、品詞情報で錯乱肢候補を選別する必要がある。

サーチエンジンヒット数:ヒット数には二つの役割が考えられる。一つの重要な役割は別解を排除することである。もう一つは、よい選択肢を選別することである。文のヒット数が少ないのは四つの可能性がある。第一は、正しい文ではあるが、古典文や、方言などで、よく使われる文ではない。第二は、正しい文ではあるが、普通はそう言わなくて、それを表現するにもっと別のよく使う表現がある。第三は、多くの人がその文は正しくないと思うが、正しいと思う人もいる。第四は、正しい文ではないが、入力のミス、外国人が書いたなどの理由で、インターネットに載せられた。第二と第三の可能性はよい錯乱肢を作成することを示唆している。

3 本研究の手法

3.1 利用する情報

中国語の漢字は、音より、意味を表わすものである。中国語の単語の中では、複数の形態素からなる複合語が多い。劉ら(参考文献[2])によれば、二音節単語 32711 の中、31958 語が複合語であり、97.7%を占める。複合語を構成する形態素はさらに実在した意味をもつ語根とそうでない接辞とに分けることができる。31958 語の二音節複合語の中、接頭辞を含んだ構造をもつ単語 44 語, 接尾辞を含んだ構造をもつ単語 1098 語, 重複式単語は 118 語, 短縮式単語は 112 語、残り 30586 語すべて、二つの語根によって構成され、全体の 95.7%を占める。

中国語では語根によって構成される複合語が語彙の中で比率が高いという点に基づいて、語根を利用した錯乱肢を作成する手法を提案する。本手法は二音節単語を中心に実験を行う。

まず、以下の三つの表が必要でなる。

1. 同字単語表

中国語の常用字が 3500 収録された漢字表(参考文献[8])から漢字を一つずつ取り出して、LCMC 中国語コーパス(参考文献[3])によって頻度順に配列された単語表(参考文献[7])を利用して、その漢字を含んだすべての二音節単語を頻度順に並べ、同字単語表を作成する。作成された項目の例を以下に示す。

土 ″土地″,″领土″,″国土″,″土司″,″土壤″‥‥

図 2 同字単語表の一例

2. 接辞表

本手法は、語根の意味を利用するもので、接辞からは候補を作らない方針であるため、接辞表によって単語に含まれた接辞を識別し、その接辞の字からは候補を作らないようにする。

3. 排除単語表

本手法は、一般的な複合語に基づくものなので、専門用語、人名、地名、単純語、短縮語などを、錯乱肢にしない方針であるため、排除単語表によって、上記のような単語を候補から削除する。

3.2 錯乱肢の決定方法

錯乱肢を取得するための手順は以下の通りである。

1. 入力した二音節単語(W)を二つの漢字に分けて、接辞表を走査し、接辞表にないものを語根(S、二つの字とも語根ならば S1、S2 と表記する)と認定する。
2. 同字単語表を利用し、Sの単語リストから、 Wの近傍、閾値(Lw)個数個以内にある単語を候補集に入れる。そして、排除単語表を利用して、その中にある単語を候補集から削除する。
3. 候補集から単語を、MC-FIB 問題の空所に一つずつ入れ、前後の閾値(Ls)内の単語と組んだ文の断片、すなわち、空所後(Ls＋1)番目の単語から前方の(Ls＊2＋1)-gram のフレーズをグーグルにかけて、フレーズで検索して、ヒット数が1以上のものであれば、候補集から削除する。こうして、最終候補集が作成される。
4. 最終候補集から、ランダムに三つの候補を取り出して、錯乱肢とする。また、教師が MC-FIB 問題を作成する際の補助として本システムを使う場合、このステップで、人手で候補を選出し、錯乱肢とすることもできる。

3.3 閾値の決定

閾値 Lw は、候補となる単語の頻度をコントロールしている。本研究の実験では、まず、Lwを3と設定する。

閾値 Ls は、グーグルにかける空所位置前後に含める単語の数をコントロールする。閾値を大きく設定すれば、インターネットというコーパスも有限であるため、別解になる単語もヒット数がゼロになり、選択肢に別解が入る恐れがある。閾値を低く設定すれば、候補が過剰に排除され、結果的に、十分な数の錯乱肢が残らない恐れがある。Lsを1とする場合で、20 個の文に対して、調査を行った結果、19 個(グーグルはアルゴリズムの改変、また検索、クリック履歴によって検索結果が異なることがある。本稿が引用した結果はすべて検討当時のものである)で過剰な排除が起こった。そのうち、三つ以上の単語を過剰に排除したのは15問であった。Lsを2とした場合、その 6 文の候補にある別解を、すべて排除できた。Lsを3とした場合、候補に別解がある6文のうち4文しか候補から別解を排除できなかった。つまり 67%の成功率しかなかった。以上より、Ls を 2 と設定することが最も適切だと考えられる。

4 評価

提案した手法の有効性を検証するために、被験者実験によって、1)本手法で作成した錯乱肢(以下、同字選択肢と呼ぶ)と 2)頻度情報を用いた錯乱肢(以下、頻度選択肢と呼ぶ)と3)品詞情報を用いた錯乱肢(以下、品詞選択肢と呼ぶ)を比較する。

実験はテストの形式で、中国語検定試験二級(参考文献[4])の第二問の MC-FIB 問題の問題文から、二音節単語を考査対象にしたものを 20 個抽出し、実験問題文とした。一問に、選択肢を 10 個用意する。そのうち、同字選択肢と頻度選択肢と品詞選択肢が三つずつ、正解が一つとした。同字選択肢は、本手法で作った候補集から、無作為に三つを抽出したものである。頻度選択肢は、二音節単語表から、Wの前後各三語(二音節単語表は頻度順で並んでいるため)を抽出してから、その 6 語から、無作為に三つを抽出するものである。品詞選択肢は、台湾の中央研究院の現代漢語平衡コーパス(参考文献[6])を利用する。コーパスでWの品詞と同じ品詞の語を含んだ文を検索して、ヒットした文から無作為に、三つの文を抽出する。その文に含まれたWの品詞と同じ品詞の語を錯乱肢とする。作成された各問題の選択肢に別解がないことを確認した。以下は、一例である。選択肢の中、B)、E)、G)は同字選択肢で、A)、D)、I)は品詞選択肢で、C)、H)、F)は頻度選択肢で、J)は正解である。

例 他有一个非常幸福____的家庭。
A)值得 B)美化 C)宅院 D)深入 E)美感
F)赞许 G)满怀 H)转而 I)随意 J)美满

実験対象は、中国語学習歴1年半から2年までの中級レベル中国語学習者6名(左下図では1番～6番)、中国語学習歴 3～10 年の上級レベル中国語学習者 3 名(左下図では 7 番～9 番)である。各学習者が誤った問題の、錯乱肢の種類ごとの個数とその合計は以下の通りである。

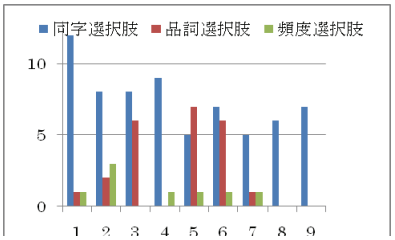

図 3 学習者ごとの不正解数

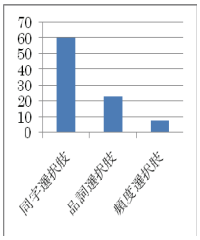

図 4 合計

実験結果、5 番以外の受験者は同字選択肢を選択して誤った問題が最も多かった。同字選択肢は品詞選択肢や頻度選択肢より紛らわしいと分かる。そして、中級グループの受験者は、品詞選択肢でよく間違っているに対し、上級グループの受験者はほとんど同字選択肢だけで間違っていることが分かる。つまり、上級レベルの中国語学習者をレベル分けするには、他の錯乱肢生成方法より同字選択肢を用いた手法が有効だと考えられる。また、中級レベルの受験者の中で、同字選択肢で間違う問題数のバラつきが大きいという点から、中級レベルの中国語学習者をレベル分けするにも、同字選択肢を用いた問題が有効だと考えられる。

その他、実験の結果から、二つの示唆を見出すことができる。第一に、中級グループの受験者は、品詞選択肢での間違いが多いのに対し、上級グループの受験者はほとんど同字選択肢だけで間違えているという点から、上級レベルと中級レベルの学習者を分けるには、品詞選択肢を用いた問題が有効であると考えられる。第二に中級グループの中では、品詞選択肢で間違う問題数のバラつきが大きいという点から、中級レベルの中国語学習者をレベル分けするには、品詞選択肢を用いた問題も有効であると考えられる。

5 改善にむけての考案

5.1 過剰な排除を防ぐには

本手法で作成した選択肢に別解はなかったが、過剰な排除の問題は依然として、存在する。言い

換えれば、その問題はインターネットの中の情報の信頼度の問題である。つまり、もともと間違った文がヒットする。ヒット数から見れば、別解になれる単語を空所に入れた文の断片で検索すれば、ヒット数が 10 万以上であるのが普通であることに対し、そういった誤った文の断片のヒット数は、大半の場合が一万以下に止まるが、一部の文には、空所に別解を入れても、錯乱肢となる候補を入れても、検索した結果のヒット数ともに 10 万を超える場合と、1 万を下回る場合がある。このため、場合に応じて、ヒット数に如何に適切な閾値を設定するかは一つの将来の課題である。

5.2 高品質の錯乱肢を生成するには

よりよい錯乱肢を作りだすため、錯乱肢を絞る二つのアルゴリズムを検討する。

錯乱肢を絞るアルゴリズムに、Ls＝1 の時の候補の集合（以下は 3-gram モデルと呼ぶ）と Ls＝2 の時の候補の集合（以下は 5-gram モデルと呼ぶ）との差集合を利用して、候補を絞ることができる（以下、当手法を差集合手法と呼ぶ）。この手法では、Ls＝2 の場合でヒット数ゼロの単語は、正解ではない単語と考え、Ls＝1 の場合でヒット数ゼロの単語は、この文脈とまったく無関係な単語をと考える。5-gram モデルと 3-gram モデルの差集合は、空所に入れて、その前後各一語からなる三つの語の連鎖が文脈によって起こりえるが、空所に入れて、その前後各二語からなる五つの語の連鎖は非文となる単語の集合である。

3.2 節の閾値 Ls の設定についての議論では、20 問の中 5 問で過剰な排除が 3 個以下しか起こらなかった。一方、四択問題には、少なくとも三つの錯乱肢が必要なので、差集合の手法を利用すれば、20 問のうち、5 問も錯乱肢がそろわない問題が起こることとなる。つまり、差集合法は 25%という高い比率の問題に適用できないということである。このため、差集合は一部の問題に良い錯乱肢を作る一方で、不十分な手法でもあり、さらなる考案が必要である。

差集合を利用した手法の改善として、3-gram の起点を変えて、それぞれ、空所前の一番目、二番目、三番目の単語から数える 3-gram モデルで、ヒット数ゼロとなった単語の積集合を取ることで、3-gram モデルの下で残る候補の数を絞ることができる。5-gram モデルで得た集合とこの積集合の差集合は、その語を空所に入れて、前後各一語、前方二語、後方二語のどちら一方からなる三つの語の連鎖を見たときに、少なくとも、そのうちの一つは文脈により起こりえるような語であって、しかし、前後各 2 語となる 5 語の連鎖は非文となるような語の集合である。単純な 3-gram モデル（mid3-gram）を利用した手法より、緩めた共起関係を利用している。差集合改善手法は錯乱肢の数を確保した上での紛らわしさにおいては、最も良いと考えられる。

ただし、差集合改善手法はすべての場合に適用できるわけではない。緩めた条件なので、すべての候補がいずれかの連鎖でヒット数がゼロではなく、結局 5-gram モデルと同じ結果になる可能性が高い。このため、本節で議論した三つの手法を如何に場合に応じて利用するかは、今後の課題である。

6 結論

本研究は、中国語 MC-FIB 問題 AQG の語彙問題の錯乱肢作成について、語根を利用する手法を提案して、単語の頻度と品詞による錯乱肢との比較を通して、その有効性を検証した。

最後に、手法の改善を巡って、ヒット数と差集合を利用する手法を提案したが、その実証は将来の課題である。

参考資料

[1] Coniam, David. A preliminary inquiry into using corpus word frequency data in the automatic generation of english language cloze tests. *CALICO Journal*, 16(2-4):15-33, 1997.

[2] 劉云、俞士汶、朱学峰、現代漢語合成詞結構数据庫、2000.

[3] McEnery, Tony. Richard Xiao, The Corpus of Mandarin Chinese. 2004.

[4] 日本中国語検定協会、中国語検定試験第 59 回～第 67 回二級試験問題、 2006-2009.

[5] Sumita, Eiichiro. Fumiaki Sugaya, and Seiichi Yamamoto. Measuring non-native speakers' proficiency of english by using a test with automatically-generated fillin-the-blank questions. Ann Arbor, Michigan, U.S., June 2005.

[6] http://dbo.sinica.edu.tw/SinicaCorpus/

[7] http://www.ling.lancs.ac.uk/

[8] http://www.moe.edu.cn/