LF-004

自動言い換え技術を利用した三つの英語学習支援システム

Three Assistance Systems for Learners of English Using Automatic Paraphrasing Techniques

村田 真樹 †
Masaki Murata

井佐原 均 †
Hitoshi Isahara

Abstract

We have developed three English learner assistance systems based on automatic paraphrasing techniques to help learners of English and English-language beginners. One system extracts personal error patterns in the user's English usage. The second transforms English sentences containing the letters "l" and "r" into sentences containing fewer instances of "l" and "r". (Japanese people have trouble pronouncing "l" and "r" properly in English, so this system will be useful for, as an example, transforming a draft of a presentation to make it easier for a Japanese speaker to read to an audience.) The third is an annotation system that provides definition sentences to make difficult English words easier to understand. We believe that these systems will be useful both for learners of English and in studies on second-language acquisition.

1. はじめに

自動言い換えに関する研究 [1, 2, 3] は，自然言語処理の文生成，要約，質問応答 [4, 5, 6] と多岐にわたるもので近年重要視されつつある．自動言い換えの研究のためにいくつかの技術が開発されてきたが，本稿ではこの技術を利用して英語学習者を手助けすることを目的とする．そこで，自動言い換えの技術を利用して三つの英語学習支援システムを作成した．一つ目のシステムは各ユーザの英語運用における個人的な誤りのパターンを自動抽出するものである．二つ目のシステムは，日本人にとって正しく発音することが難しい R と L を含む英語の文章を，なるべく R と L を含まないような文章に書き換えるものである．三つ目のシステムは，難しい英語単語にその単語の定義文を与えることで文章の読解を支援するシステムである．これらのシステムは英語学習者のみならず第二言語獲得に関する研究 [7, 8, 9] においても役に立つものであると考えられる．

2. 個人的な英語運用の誤りパターンを抽出するシステム

我々はすでに，互いに関係のあるテキストの対から差分を自動的に取り出す，様々な言い換えに関する研究を行なってきた．例えば，話し言葉と書き言葉のテキスト対から差分を取り出し，書き言葉と話し言葉の間の言い換えパターンの抽出を行なった．この研究では話し言葉のテキストとして学会での講演発表を書き起こしたものを，書き言葉のテキストとしてはその講演発表に対応する論文原稿を利用した [10]．また，同じ特許の請求項と実施例を照合することで，特許テキストに関する書き換え規則や同義語表現の抽出も行なった [11]．この言い換えパターンを抽出する技術を利用して，個人的な英語運用の誤りパターンを抽出するシステムを作成した．このシステムでは，誤りパターンを抽出する対象としているその個人の英文校閲前のテキストと，それの英文校閲後のテキストを照合することで，その個人の英語運用の誤りパターンを抽出する．実際にこのシステムを使って，筆頭著者の書いた 80 個の英語論文 (これらの論文はネイティブの英文校閲を受けている) を使って誤りパターンを抽出してみた．それぞれの論文で英文校閲前のものと英文校閲後のものを UNIX の Diff コマンドで差分をとり，この差分を誤りパターンとした (この Diff コマンドで差分を取り出す方法についての詳細は文献 [12, 10, 11, 13] を参照のこと)．そして抽出された誤りパターンの頻度を計算した．その結果を表 1 に示す．表のφは空文字を意味し，"φ ⇒" と "⇒ φ" はそれぞれ挿入と削除を意味する．

結果から，筆頭著者の最も多い誤りの原因は "the" や "a" などの冠詞に関わるものであることがわかる．冠詞の用法は日本人にとって難しいものとされており [14]，この結果でもそれが裏付けられている．次に多いものとして"is" ⇒ "was" や"are" ⇒ "were" などの時制に関する誤りに気づく．ここで面白いことは，冠詞の誤りはだいたい対称的である (方向性がない) のに対して，時制に関する誤りは対称的でないことである．(例えば，"is" ⇒ "was" の頻度は大きいが"was" ⇒ "is" の頻度は小さい．) 時制の誤りが非対称であることの一つの理由は，"is" と "are" などの現在時制はデフォルトの表現形式であり，一般によく使われる表現であるので，深く考えずに表記すると現在時制を記述してしまうからではないかと思われる．

次に多いものとしては，φ ⇒ "of" や"of" ⇒ "for" や"in" ⇒ "of" などの前置詞に関する誤りである．その他，"which" ⇒ "that" と"having" ⇒ "with" の誤りにも気づく．以上示した通り，われわれの言い換え技術を利用したシステムを用いることで，非常に簡単に，筆頭著者の個人的な英語の誤りパターンを抽出することができた．

特定のユーザに応じた教育システムは特に有益である．個人的な英語の誤りパターンを抽出することのできるわれわれのシステムは，そういう特定のユーザに応じた教育システムに役に立つものである．また，このシステムは，特定の個人だけでなく，特定の集団，例えば，ある英語教師が受け持つ学生の集団でどういう英語の誤りパターンが多いかを調べることにも利用できる．

3. RL 発音回避システム

我々は文章の自動言い換えの研究を数多く行なってきた．例えば，書き言葉を話し言葉に自動で言い換える研

†独立行政法人情報通信研究機構．{murata,isahara}@nict.go.jp
National Institute of Information and Communications Technology.

表 **1.** 個人的な英語誤りパターンの例

英文校閲前の表現	英文校閲後の表現	頻度
”the”	⇒ φ	393
φ	⇒ ”the”	357
”a”	⇒ ”the”	146
”a”	⇒ φ	122
φ	⇒ ”a”	120
”the”	⇒ ”a”	116
”is”	⇒ ”was”	68
φ	⇒ ”of”	56
”of”	⇒ ”for”	49
”in”	⇒ ”of”	36
”are”	⇒ ”were”	36
”of”	⇒ φ	34
”of”	⇒ ”in”	32
”in”	⇒ ”for”	32
”an”	⇒ ”the”	32
”which”	⇒ ”that”	31
”is”	⇒ ”are”	28
”a”	⇒ ”an”	28
φ	⇒ ”by”	28
φ	⇒ ”an”	27
”having”	⇒ ”with”	25
φ	⇒ ”that”	25
φ	⇒ ”in”	24
”the”	⇒ ”an”	23
φ	⇒ ”used”	23
”result”	⇒ ”results”	22
”and”	⇒ ”or”	22
φ	⇒ ”also”	22
φ	⇒ ”to”	21
φ	⇒ ”thus”	21
φ	⇒ ”only”	20
φ	⇒ ”and”	20
”we”	⇒ φ	19
”that”	⇒ φ	19
φ	⇒ ”as”	18
”metonymy”	⇒ ”metonymic”	17
”cases”	⇒ ”case particles”	17
”as”	⇒ φ	17
”are”	⇒ ”is”	17
φ	⇒ ”have”	17
”use”	⇒ ”used”	16
”short term”	⇒ ”short-term”	16
”only”	⇒ φ	16
”in”	⇒ ”at”	16
”have”	⇒ φ	16
”by”	⇒ φ	15
”When a”	⇒ ”A”	15
φ	⇒ ”method”	15

究をした [10]．また，文章をよりなめらかに言い換える推敲システムの研究もした．また，要約のようになるべく意味を変えずに文章を短く言い換える研究もした．また，質問応答システムでは質問文と知識データの文の照合を適切に行なえるように，質問文と知識データの文を互いになるべく類似した形の文になるように言い換える研究もした．これらのシステムはある一つの言い換えの統一的モデルで構築することができ，また，その言い換えの統一的モデルに関する研究も行なってきた [1, 3]．このような言い換え技術を使って，日本人にとって正しく発音することが難しい R と L を含む英語の文章 [15, 16] を，なるべく R と L を含まないような文章に書き換えるシステムを作成した．本節ではこのシステムについて述べる．本稿ではこのシステムのことを RL 発音回避システムと呼ぶ．このシステムは，日本人が国際会議などで英語で演説する際，その演説の原稿での文章をなるべく R と L を含まないような文章に書き換え，R と L を含む単語の発音が苦手な日本人にとって話しやすい英語文章に書き換える時に役に立つ．

上述の言い換えの統一的モデルは変形部と評価部の二つのモジュールからなる．まず，言い換えの対象となる文がシステムに入力される．変形部においていくつかの可能な言い換え候補が生成され，評価部においてこれらの候補の中から最も適切なものを選択し，この選択した言い換え候補を使って言い換えを行ない，言い換えた文を出力する．

RL 発音回避システムの作成のために，変形部には WordNet 2.0 [17] の同義語を，また，評価部には以下の条件を利用した．

- 英語文で発音しにくい R+母音や L+母音の表現を含む個数が小さいほど良い言い換えとする．
- R+母音や L+母音の表現を含む個数が同じ場合は入力された元の表現の方が良い言い換えとし，また，変形後の表現同士の比較では，入力されたデータの各部分単語列の英語テキストでの生起確率がより大きくなるような言い換えを良いものとする．
- 英語テキストでの各表現の出現が前二単語，後ろ二単語の文脈を含めて 1 個以上があることを条件とする．(文としての適切性の判定)

英語テキストとしては BNC コーパス [18] を用いた．本稿では，母音の判定には文字を利用し，a,i,u,e,o,y を後ろにくっつけて持つ r,l の表現を，R+母音，L+母音の表現とした．また，変形規則に用いる同義語表現には動詞の変化形，名詞の複数形なども追加して用いた．

この条件で著者が今まで国際会議で口頭発表してきた発表の原稿を入力として与え，R や L をあまり含まない英文への変形の実験を行なった．この実験の結果の一例を表 2 に示す．表の縦線で囲った部分が言い換えられた表現で，上の表現が下の表現に言い換えられている．それぞれ R+母音，L+母音が少なくなる表現に書き換えられている．“approach” を “way” に書き換えたり，“length” を “size” に書き換えたりして，発音しやすい語へ正しく言い換えているものがあった．しかし，今のところこのシステムでは，言い換えると微妙に意味が異なってしま

表 **2. RL** 発音回避システムの出力例

正しい言い換え
We think a good | approach way | is to construct it using “X *no* Y”.
The criteria used to | select determine | the most appropriate transformation type must be predefined.
This figure shows the | structure composition | of the thesaurus.
$length$ d is the | length size | of a document d.
誤った言い換え
This is the | title name | of the query.
P of d and t is the | location determination | of the first occurrence of a term *t* in the document *d*.
This term is for weighting terms which are | followed used | by the Japanese-language particle “nado”.

う誤りもあった．このシステムの今後の応用としては，今のところまだ性能がそれほど高いというわけではないので，言い換えた結果のみを出力するのではなく，言い換えの候補をいくつかの評価部で用いた値 (R+母音や L+母音の表現を含む個数など) とともに値の順に提示し，そこでユーザに言い換えに適切な表現を選ばせるという支援システムのような形の利用が良いと思われる．また，本稿では R と L に限ったが，F や V などの他の発音の難しい文字についても同様の機能を有するシステムの開発をしたいと思っている．

4. 難しい語に定義文を付与するシステム

黒橋ら [19] は辞書の定義文を使って文章をわかりやすいものに言い換える研究を行なっている．辞書の定義文は単語の説明であるので，それらは文をわかりやすいものに言い換えるときに用いることができるのである．この考え方にしたがって，本節では，わかりやすくなるように，難しい語に自動で定義文を付与するシステムを作成した．

このシステムでは，まず最初に難しい単語を抽出し，その後でその抽出した難しい単語に定義文を与える．現在のシステムでは難しい単語を特定するために，二つの方法を用意している．一つはユーザが今までに書いたことや読んだことがない単語を難しい単語とする方法で，もう一つは英語のコーパスでの出現頻度が低いものを難しい単語とする方法である．(言語学習の分野では英語のコーパスでの単語の頻度は，簡単な単語や難しい単語を特定することによく用いられる [9, 20]．)

このシステムに実際に WordNet 2.0 のマニュアルを入力として与えた結果の例を表 3 に示す．この結果は，BNC のコーパスでの頻度が 1000 以下のものを難しい単語とする方法を使っている．定義文は EDR 辞書のもの [21] を使った．単語の語幹化には WordNet 2.0 を使った．“[Notes: ...]” の部分は本システムによって付与された定義文である．現在のシステムはすべての定義文を出力するようになっている．“|” の記号は複数の定義文が辞書にある場合その複数の定義文は “|” で区切って表示される．表 3 では英語文の定義文を表示しているが，日本人のユーザのために日本語の定義文を表示するようにしてもよい．“[Caution!]” は，難しい単語と判定されたがその語が EDR 辞書になかった場合に表示される．“[Above!]” は，すでに前方でその語の定義文が付与されていることを意味する．表 3 の結果では，“corpus” と “consortium” と “lexicography” が正しく難しい単語として抽出され，それらには定義文が与えられている．

さらに，難しい単語の抽出にわれわれの論文に出現していない単語を難しいとする方法を，BNC のコーパスでの頻度が 1000 以下のものを難しい単語とする方法に併せて利用した．そうすると，われわれのよく知る “corpus” は難しい単語として抽出されなくなった．(言語処理研究者である著書らは “corpus” という語をよく知っている．) それぞれのユーザはそれぞれのユーザの得意な専門的な知識を持っている．われわれのシステムはそういう個々のユーザの特徴を利用して，そのユーザがよく知っている単語には定義文を付与しないという，ユーザ依存の処理ができるのである．

また，出力結果の例では “encoded” に “[Caution!]” が付与されている．このため，ユーザは “encoded” が難しい単語であると認識できる．このことは英語学習者にとっては学習の手助けとなる．また，“[Caution!]” には別の興味深い効果がある．“recordeded” にも “[Caution!]” が与えられている．“recordeded” はミススペルでそういうミススペルを検出するのにも役に立つ．

将来的には，個々のユーザの読んだり書いたりした文章をすべて格納し，ある新しい文が入力された時に，そのユーザに対して，そのユーザの全生涯で初めて出現した語がどれであるのかということを示すシステムも作りたいと考えている．(文書の初出の表現を強調表示する研究は文献 [11] でも述べている．) また，英語の教科書において初出の語をなんらかの強調表示をするのも面白いのではないかと考えている．訳語をふって読者の手助けをするシステムはいくつかあるが [22]，ユーザ依存の処理や初出の表現を強調する考え方は珍しいと思う．今後はこのあたりを発展させた研究をしたいと思っている．

表 **3.** 難しい語に定義文を付与するシステムの出力例

The British National Corpus [Notes: a collection of all the works of a special type, on a special subject|a gland of an insect, called corpus allatum|a total assemblage of law in a country|an extract of corpus luteum of a pig or a cow|the glassy fluid from the eye|a set of material or data for use during study, especially for linguistic analysis] is a very large (over 100 million words) corpus [Above!] of modern English, both spoken and written.
The project was carried out and is managed by an industrial/academic consortium [Notes: an association of creditor nations] lead by Oxford University Press, ...
The spoken part (10%) includes a large amount of unscripted [Notes: (of a speech or discussion that is broadcast) spoken naturally or without previous arrangement] informal conversation, recordeded [Caution!] by volunteers selected from different age, region and social classes in a demographically [Notes: in a demographical way] balanced way, ...
... will be useful for a very wide variety of research purposes, in fields as distinct as lexicography, [Notes: the creating and printing of dictionaries] artificial intelligence, speech recognition and synthesis, literary studies, and all varieties of linguistics.
The corpus [Above!] is encoded [Caution!] according to the Guidelines ...

5. おわりに

本研究では，自動言い換えの技術を利用して三つの英語学習支援システムを構築した．これらのシステムが，英語学習者や第二言語獲得の研究に役に立つことを願っている．

謝辞

独立行政法人情報通信研究機構の和泉絵美氏と小谷克則氏にはこの研究に対して有益なコメントと手助けをしていただきました．ここに感謝します．

参考文献

[1] 村田真樹, 井佐原均, 言い換えの統一的モデル — 尺度に基づく変形の利用 —, 言語処理学会第 7 回年次大会ワークショップ論文集, (2001).

[2] IWPT, *NLPRS'2001 Workshop on Automatic Paraphrasing: Theories and Applications*, (2001).

[3] Masaki Murata and Hitoshi Isahara, Universal model for paraphrasing — using transformation based on a defined criteria —, *NLPRS'2001 Workshop on Automatic Paraphrasing: Theories and Applications*, (2001).

[4] 加藤直人, 浦谷則好, 局所的要約知識の自動獲得手法, 言語処理学会誌, Vol. 6, No. 7, (1999).

[5] 村田真樹, 内山将夫, 井佐原均, 類似度に基づく推論を用いた質問応答システム, 自然言語処理研究会 2000-NL-135, (2000), pp. 181–188.

[6] Tetsuro Takahashi, Kozo Nawata, Kentaro Inui, and Yuji Matsumoto, Effect of structural matching and paraphrasing in question answering, *IEICE Transactions on Information and Systems*, Vol. E86–D, No. 9, (2003), pp. 1677–1685.

[7] Heidi Dulay, Marina Burt, and Stephen Krashen, *Language Two*, (Oxford University Press, 1982).

[8] Diane Larsen-Freeman and Michael H. Long, *An Introduction to second language acquisition research*, (Longman, 1991).

[9] Sylviane Granger, *Learner English on Computer*, (Longman, 1998).

[10] Masaki Murata and Hitoshi Isahara, Automatic extraction of differences between spoken and written languages, and automatic translation from the written to the spoken language, *LERC 2002*, (2002).

[11] Masaki Murata and Hitoshi Isahara, Using the diff command in patent documents, *Proceedings of the Third NTCIR Workshop (PATENT)*, (2002).

[12] 村田真樹, 井佐原均, 同義テキストの照合に基づくパラフレーズに関する知識の自動獲得, 情報処理学会自然言語処理研究会 2001-NL-142, (2001).

[13] 村田真樹, 井佐原均, diff を用いた言語処理 — 便利な差分検出ツール mdiff の利用—, 言語処理学会誌, Vol. 9, No. 2, (2002).

[14] 村田真樹, 長尾真, 名詞の指示性を利用した日本語文章における名詞の指示対象の推定, 言語処理学会誌, Vol. 3, No. 1, (1996).

[15] 小池生夫, 第二言語習得研究に基づく最新の英語教育, (大修館書店, 1994).

[16] 垣田直巳, 小篠敏明, 英語の誤答分析, (大修館書店, 1983).

[17] Princeton University, Wordnet 2.0, (2003).

[18] Oxford University Computing Services, British national corpus, (1995).

[19] 黒橋禎夫, 酒井康行, 鍛冶伸裕, 国語辞典に基づく文章理解とパラフレーズ, 言語処理学会第 7 回年次大会ワークショップ論文集, (2001).

[20] 大学英語教育学会 (JACET) 語彙研究会, JACET8000, (2003), http://members.at.infoseek.co.jp/jacetvoc/.

[21] 日本電子化辞書研究所, EDR 電子化辞書仕様説明書, (1993).

[22] Victor Poznanski, Pete Whitelock, Jan IJdens, and Steffan Corley, Practical glossing by prioritised tiling, *COLING-ACL '98*, (1998), pp. 1060–1066.