社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－ＦＩ－ 76 (20)
2004－ＮＬ－163 (20)
2004／9／17

英文読解のためのコースウェアの作成

内山将夫† 谷村緑‡ 井佐原均†

英文読解教材作成のための素材は豊富である．しかし，それらの素材を取捨選択して，1つのコースウェアとしての教材を作成するのは，困難である．本稿では，そのようなコースウェアとしての教材を，学習対象とする語彙とコーパスとから自動的に作成する方法を提案する．その方法によれば，学習対象である語彙をなるべくコンパクトに網羅するような文書集合を選択することができるので，それをコースウェアとすることにより，読解を通した効率的な語彙の獲得ができると考える．提案手法により作成された読解教材の種々の統計量を，無作為抽出の場合と比べることにより，作成された教材が，コンパクトに語彙を網羅していることが確認された．

Constructing English Reading Courseware

Masao Utiyama†, Midori Tanimura‡, and Hitoshi Isahara†

English reading materials are abundant on the Internet. However, it is still difficult to select proper materials to construct a courseware that can be used through a semester English reading course. We proposed a method for constructing a courseware from a target vocabulary and corpus. The method was designed to extract a minimal set of articles (from the corpus) that contained the vocabulary. The constructed courseware consisted of articles that had dense occurrence of the target vocabulary. Thus, the courseware was efficient in presenting the vocabulary to students through reading. The statistics of the courseware compared with those of the randomly sampled articles indicated that the courseware is promising.

†情報通信研究機構．National Institute of Information and Communications Technology.
‡近畿大学．Kinki University.

1 はじめに

英文読解教材を作るための素材は豊富である．たとえば，ニュース記事であれば，CNN[1]やTIME[2]やBBC[3]などがある．しかし，これらの素材を適切に取捨選択して，1つのコースウェアとしての教材を作成するのは，困難である．ただし，ここでの教材とは，たとえば，大学の半期の授業，もしくは，それに付随する自習教材として使えるような，1つの一貫したものであり，それを学習することにより，英文読解の技能が修得できることを目的とするものを指す．

このようなコースウェアとしての教材を，自動作成するのが我々の目標である．特に，本稿においては，英文を読解するためには，ある程度の語彙を獲得することが必要不可欠であることから，作成された教材を学習することにより，そのような語彙が「効率的」に獲得可能となる教材を作成することを目的とする．ただし，ここでの「効率的」とは，以下のように定義する．

定義 1 なるべく少量の文章を読むだけで，与えられた語彙が学習できること．

定義 2 教材の最初の方を学習するだけで，必要な語彙の多くが学習できること．

定義1については，少量の文章で語彙を学習できるならば，多量に読むよりも良いであろうという考えから採用し，定義2については，途中で学習を中断したときにも，効果が高いようにという考えから採用した．

このような教材を作成するときの前提として，我々は，(1) 獲得目標としての語彙を任意に設定できることと，(2) 教材作成に必要な素材としての文章 (文書) の集合がコーパスとして与えられていることとを設定した．そして，これらが満されるとき，与えられた語彙をなるべく効率的に学習できる教材を作成することを目的とした．

このような教材が，もし，自動作成できれば，それは，特に，English for Special Purposes (ESP) としての英語教育に有効である．なぜなら，そのようなことができれば，ESPのための語彙とコーパスとを与えるだけで，工学や医学や経済などの，学習者が必要とする特定分野の英語学習のための読解教材が，容易に作成できるからである．それに対して，それらの教材を，各分野個別に人手で作成するとすると，それには多大なコストがかかる．

[1]http://www.cnn.com/
[2]http://www.time.com/time/
[3]http://www.bbc.co.uk/

以下では，効率的な語彙獲得を目的とした教材の作成法について述べ，次に，その効率性について調べる．

2 最適化としての教材作成

本節では，コースウェアとしての教材を自動作成するときの一般的な方針として,「最適化としての教材作成」を提案する．

まず，用語を定義する．複数の文書からなるコーパスを D とする．D 中の任意の文書の集合を $A(\subseteq D)$ とする．なお，A は順序付けされたリストであっても良い．A がリストである場合には，そのリストの先頭の文書から学習を開始し，A が集合である場合には，学習の順番は任意であるとする．次に，最適化の目的関数を $f(A)$ とし，$f(A)$ が大きいほど，A は「良い」教材であるとする．また，A についての制約を考え，それが満されているときには1，そうでないときには0を取る述語として，$\delta(A)$ を考える．

以上の準備の下で，最適化としての教材作成においては，次のような $\hat{A}$ を求めたい．

$$\hat{A} = \arg \max_{A \subseteq D, \delta(A)=1} f(A) \tag{1}$$

このような $\hat{A}$ は，制約が満された「最良」な文書集合である．

次に，$\hat{A}$ を探索する算法について述べると，これは一般には，組み合せ最適化の問題である．つまり，コーパス中の全ての文書の組み合せのなかから，$\delta(A)$ を満し，かつ，$f(A)$ が最大のものを求める必要がある．この問題は，一般的には，効率的な探索方法がない問題であるので，次節においては，貪欲 (greedy) な算法により，近似解を求めている．

3 効率的な語彙獲得のための読解教材の作成算法

本節では，前節で導入した目的関数 $f(A)$ と制約 $\delta(A)$ の具体的な表現を述べ，次に，実際に利用した貪欲算法について述べる．

まず，我々は，コーパス D に加えて，語彙 V が与えられていることを前提とし，そこから V 全体を網羅するような教材 $A(\subseteq D)$ を作成したい．そのため，制約としては，以下を考える．

$$\delta(A) = \begin{cases} 1 & (\forall w \in V)(\exists d \in A)(\text{文書 } d \text{は単語 } w \text{を含む}) \\ 0 & \text{otherwise} \end{cases}$$

この制約は，語彙V中の全ての単語wについて，それが最低1回は，いずれかの文書dに出現することを表現している．次に，$f(A)$については，1節における効率性の定義1から，A 中の文書長の和が小さいほど大きい値を取るような関数を利用することが考えられる．これらの$\delta(A)$と$f(A)$とから，(1)式の最適解を得るには，たとえば，整数計画法による集合被覆問題に対する解法(ウィリアムス 1995)を適用することができるが，本稿では，以下で述べる算法，すなわち，これまでに網羅されていないようなV中の単語を，なるべく多く含むような文書を順次選択するという貪欲算法により，近似解として$\hat{A}$を得た．この算法によれば，なるべく少ない文書数で与えられた語彙を網羅することができるので，結果として，$f(A)$が大きい文書集合を近似解として得ることができる．また，選択された最初の方の文書は，後の方のものと比べて，V中の語彙をより多く含むので，効率性の定義2に沿った教材が作成できる．

より具体的には，Aを現在作成中の教材とし，$V_{\rm todo}$を，まだA中の文書に網羅されて(含まれて)いないVの単語とし，$V_{\rm done}$を既に網羅されている単語としたとき，以下のスコア

$$G(d|\alpha, V_{\rm todo}, V_{\rm done}) = \alpha g(d|V_{\rm todo}) + (1-\alpha) g(d|V_{\rm done})$$

が最大のdをAに追加するということをV中の単語全てがAに網羅されるまで繰り返す．ここで，

$$g(d|V_x) = \frac{k_1+1}{k_1((1-b)+b\frac{|W(d)|}{E(|W(d)|)})+1}|W(d)\cap V_x| \quad (2)$$

ただし，k_1とbとは，経験的に定める定数であり，本稿の実験においては，$k_1 = 1.5, b = 0.75$である．また，$|W(d)\cap V_x|$は，文書d中の単語集合$W(d)$とV_xとの共通単語数であるので，$g(d|V_x)$は，共通単語数が大きいときに大きい値となる．更に，$|W(d)|$は，文書dにおける異なり単語の数であり，$E(|W(d)|)$は，それのコーパス全体における平均値である．そのため，$g(d|V_x)$は，$\frac{|W(d)|}{E(|W(d)|)}$の影響により，異なり単語の数が少ない(文書長が短かい)文書において大きい値となる．すなわち，$g(d|V_x)$は，V_xと共通単語数が多く，かつ，文書長が短かい文書において大きい値となる．

なお，(2)式の$g(d|V_x)$は，情報検索で使われているBM25 (Robertson and Walker 2000)という尺度を簡略化したものである．$g(d|V_x)$がBM25に基づいた理由は，BM25は，情報検索において，高精度に質問と関連した文書を検索できることが実証されているので，それに基づいた$g(d|V_x)$を利用することにより，V_xと良く関連した文書が得られると考えたためである．

次に，重みαは，

$$\alpha = \frac{|V_{\rm done}|}{1+|V_{\rm done}|}$$

とした．この重みを利用することで，$V_{\rm todo}$での1単語の共通が，$V_{\rm done}$全体での共通と同程度に勘案される．そのため，これまでに網羅されていないような単語を，なるべく多く含むような文書，および，既に網羅されている単語との共通単語数も大きいような文書が優先して教材に採用されるため，効率性の定義1，2に沿うような教材が作成できる．

4 実験

本節では，上記の貪欲算法を，実際の語彙とコーパスに対して適用した結果として作成された教材(以下では作成教材と呼ぶ)の性質を調べるとともに，実際の授業における利用例を述べる．

4.1 語彙

学習対象とする語彙としては，(中條 2003)により作成されたTOEIC[4]学習用語彙「レベル1(補習向け)」「レベル2(初級向け)」「レベル3(中級向け)」を用いた[5]．この語彙は全部で640項目からなり，それぞれのレベルの項目数は，レベル1が200, レベル2が200, レベル3が240である．これらの選定の詳細は(中條 2003)に記述されている．これらの語彙は3レベルに分かれているが，本実験においては，特に，それらを区別することなく，全語彙640項目を学習対象の語彙とした．なお，これら項目において，単語を複数含むもの(たとえば「広告」という項目に対する「ad」と「advertisement」等)については，その全ての単語を学習対象とし，また，後述のコーパスに出現しない単語については，それを語彙から除去した．その結果，全部で642単語を学習対象の語彙とした．

4.2 コーパス

教材作成のためのコーパスとしては,「読売新聞記事データ」における「The Daily Yomiuri」の1989年から

[4] Test of English for International Communication (http://www.toeic.or.jp/toeic/index.html)

[5] これらの語彙はhttp://www5d.biglobe.ne.jp/~chujo/で公開されている．

2001年までの約11万記事を元データとした．これらの記事から，読解のときの負担が軽くなるように，300単語以下のものを選ぶと約4万1千記事である．ここで，The Daily Yomiuri の記事の特徴として，それと内容が対応するような読売新聞記事が，記事によっては存在するというものがある．そのため，そのような特徴を生かした教材を作成する可能性も考慮して，上述の300単語以下の記事から，更に，読売新聞において対応記事が存在する可能性が高いもののみを選ぶと約2万5千記事である．ただし，対応記事については，(内山井佐原 2003) において対応付けされた記事対のうちで，対応付けの信頼性が高いと報告されている，対応付けスコア0.111106681以上の記事対を利用した．

この約2万5千記事に対して，以下の処理を適用した結果のコーパスに対して，3節の算法を適用した．その処理は，(1) 各記事を1文単位に整形したあとで，(2) 各文の各単語に品詞をタグ付けし，基本形を得るというものである[6]．なお，ソフトウェア上の制限として，基本形は，名詞・動詞・形容詞・副詞のいずれかについてしか求めることができなかった．

4.3 作成された教材の例

本節では，作成教材における記事の例として，その最初の記事の一部を図1に示す．図1の記事の全文では，異なり語数では43語，延べ語数では61語が学習対象の語彙と共通している．それら共通単語については，初出のものを太字，それ以外を斜体で示す．

図1の記事は，3節の算法を，上述の学習対象語彙とコーパスに対して適用した結果として得られる最初の記事であるから，その算法における，最良の記事である．したがって，2番目以降に教材として採用される記事については，共通単語数は減少していく．しかし，減少していったとしても，ただ無目的に記事を集めるのに比べれば，学習対象語彙中の単語を多く含む記事を抽出できる．そのことについては，次節でより詳しく調べる．

4.4 作成された教材の性質

本節では，作成教材が，1節で述べた効率性の定義1, 2を満すことを示す．そのときに，効率性の比較の対象として，無作為に記事を抽出したときと比べての有効性を示す．

[6] `http://www2.nict.go.jp/jt/a132/members/mutiyama/software.html` で公開されているソフトウェアを利用した．

Streamlining to cost NTT over 1.4 tril. yen

NTT Corp's restructuring plan, which aims to **transfer** 110,000 workers to subsidiaries, will **cost** the telecom giant a hefty 1.4 trillion yen to 1.5 trillion yen, The Yomiuri Shimbun learned Thursday.

The plan is **expected** to be so **expensive** because of ballooning **retirement** and other **compensation allowances** that will be paid to about 55,000 workers.

NTT will earmark lump-sum **expenses** in its **fiscal** 2001 **account** settlement ending in March to make up for the *costs* of the large-scale streamlining plan scheduled to be **implemented** in spring.

The nation's largest **telecommunications company**, which originally **forecast** after-tax **profits** of 3 billion yen for the **current** *fiscal* year, is **predicting** a loss of hundreds of billions of yen.

Under the restructuring plan, NTT will *transfer* a **total** of 110,000 of its 210,000 workers, mostly from its two **regional** phone **operators**–NTT East Corp. and NTT West Corp.–to other group *companies* to be set up. Among those *transferred*, 55,000 workers aged 51 and above will be **retired** and rehired at **salaries** as much as 30 percent lower than those they are currently **receiving**.

図 1: 作成教材における記事の一部

まず，作成教材における基本的な統計量を調べる．まず，その教材を構成する記事数は116である．次に，各記事の基礎統計量を表1に示す．

表 1: 各記事の基礎統計量 (記事数=116)

	平均	標準偏差	最小値	最大値
記事長	180.2	65.2	44	296
共通延べ語数	25.3	14.6	1	65
共通異なり語数	17.4	8.8	1	43
共通新出異なり語数	5.5	7.2	1	43

表1において,「記事長」とは，各記事における延べ語数である．次に,「共通延べ語数」とは，上述したTOEIC学習用語彙Vについて，各記事において，Vに含まれる単語の延べ語数のことである．ただし，ある単語がVに含まれるとは，その単語自体，あるいは，その単語の可能な基本形のいずれかがVに含まれることと定義する．また,「共通異なり語数」とは，計数の方法を延べ語数の場合と同様にした場合における，異なり語数のことである．最後に,「共通新出異なり語数」とは3節で述べたV_{todo}と各記事との共通異なり語数のことである．

次に，これらの統計量が，無作為抽出の場合と比較して，どの程度の量であるのかを調べる．そのために，作成教材の延べ語数20900語と同じだけの延べ語数と

表 2: 統計量の比較

	傾向	教材	平均	SD	上側	下側	超	以下	以上	未満
記事数	多い	116	111.1	3.4	0.076	0.924	62	938	99	901
平均記事長	短い	180.2	188.3	5.9	0.918	0.082	901	99	938	62
平均共通延べ語数	多い	25.3	19.3	1.1	0.0	1.0	0	1000	0	1000
平均共通異なり語数	多い	17.4	12.8	0.6	0.0	1.0	0	1000	0	1000
平均共通新出異なり語数	多い	5.5	3.6	0.1	0.0	1.0	0	1000	0	1000
網羅率	高い	1.0	0.616	0.016	0.0	1.0	0	1000	0	1000

なるように無作為に記事を抽出して1つの記事セット(これを無作為記事セットと呼ぶ)とするということを1000回繰り返し，1000個の無作為記事セットを得た．そして，それらについて，種々の統計量を計算し，作成教材と比較した．その一覧を表2に示す．

表2において，各行は，各種統計量を示す．各列については,「傾向」は，作成教材の当該統計量が，無作為記事セットと比べてどういう傾向にあるかであり,「教材」の数値は，作成教材における当該統計量の値である．次に,「平均」と「SD」は，当該統計量の無作為記事セット全体における平均と標準偏差である．また,「上側」と「下側」は，当該統計量が，無作為記事セットの平均と標準偏差による正規分布をすると仮定したときに， 作成教材の数値を超える値となる確率(上側確率)と，それ以下の値となる確率(下側確率)である．なお，上側確率が「0.0」となっている欄における確率は，10^{-7}未満である．最後に,「超」「以下」「以上」「未満」にある数値は，それぞれ，作成教材の数値「を超える」「以下」「以上」「未満」の数値であるような無作為記事セットの数である．そのため,「超」と「以下」の和，および,「以上」と「未満」の和は1000である．

表2においては，前述の通り，各記事セットの延べ語数は20900語と一定である．そのため，記事数が多いということは，平均記事長が短いということと同じことである．これら記事数と平均記事長について，上側確率と下側確率は，それぞれ，0.076 と 0.082 であるので，これらの傾向は，統計的には有意というほどではないが，記事数は多い，あるいは，平均記事長は短いといえる．このことは(2)式での説明を裏付けている．

次に，表2の「平均共通延べ語数」「平均共通異なり語数」「平均共通新出異なり語数」の行には，それぞれ，表1の対応する統計量の各記事あたりの平均値についての数値がある．これらについては，上側確率等を見ると分かるように，無作為記事セットが作成教材よりも大きい数値をとる可能性は，非常に小さいと言える．これより，作成教材は，学習対象語彙を多く含む記事集合を選択していると言える．

最後に,「網羅率」とは，学習対象語彙Vのなかで，実際に使われた単語の，V全体に占める割合である．これは，作成教材については，算法の性質上，かならず1になる．しかし，無作為記事セットについては，そうなるとは限らないし，実際に，網羅率の平均は0.616である．すなわち，無作為記事セットにおいては，Vの約40%が出現しない．また，上側確率も 0.0 である．そのため，作成教材は，無作為記事セットと比べて，定義1の意味で効率良くVを網羅していると言える．

以上においては，無作為記事セットとの比較により，効率性の定義1の観点から，作成教材の性質を述べた．

次に，図2に，全語彙Vについて，網羅された異なり語数の増加の様子を実線で示す．

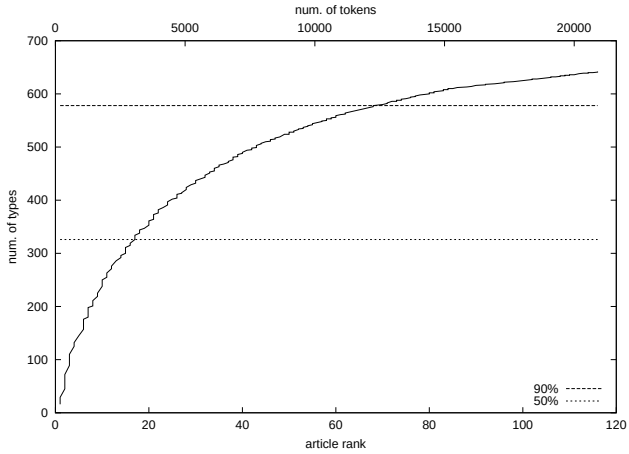

図 2: 記事の順位や延べ語数と異なり語数(全体)

図2の横軸が記事の順位(article rank)や延べ語数(num. of tokens)で，縦軸が，Vで網羅された異なり語数である．これより，最初の方における，異なり語数の増加が大きいことがわかる．図2では，延べ語数が100増えるごとに，Vでの異なり語数を数えたときに，初めてVの50%および90%を超えるところに横線を引いてある．それらは，それぞれ，異なり語数が326語と578語のところである．そして，このときの延べ語数は，3700語と13900語であり，記事の順位は，17位と68位である．そしてこれらは，延べ語においては，18%および67%に相当する．つまり，作成教材の最初の18%や67%で，全

学習対象語彙の50%や90%を網羅している．以上より，定義2の意味での効率性があることが言える．

4.5 作成された教材の利用例

作成された教材は，関西地区のある大学の授業で，3クラスにおいて，利用されている．これらクラスの英語の能力は，初級から中級程度である．そして，作成教材は，授業の補助教材として利用されている．その利用の形態は，各授業の前に，課題として与えられた記事を各自が自習しておき，授業においては，確認テストとして，課題記事におけるTOEIC語彙についての日本語訳を書かせるというものである．

このような形態での利用をした場合における，現時点(2004年8月)での感触として，まず，学生に与える教材の動機付けという点から述べると，これは非常に高いものである．その理由の1つとしては，教材に利用したTOEIC用の語彙もコーパスも，実際の言語資料から作成したものであり，それらが現実の言語使用に密接に結びついていることがあると考えている．

次に，作成教材と学生の英語能力との関係であるが，作成教材の一部には，学習者の現時点での英語能力を超えている部分もある．しかし，単語の学習方法として，単語のリストではなく，文脈のなかで単語を学習することにより，文脈により単語の意味が変化することを実感すると同時に，読解自体に対する抵抗感も減少するという効果があるようである．

5 関連研究

音声言語処理技術を教育に利用しようという試みはe-Learningとして盛んである．たとえば，最近のワークショップとして，(Burstein and Leacock 2003)には，英語のエッセイの自動採点とか，文法とか発音のチェックとか，テスト問題の自動作成とかに，音声言語処理技術を利用する研究がある．

これらの研究と本稿での研究の主要な相異点は，本稿での研究が，たとえば，大学の半期の授業，もしくは，それに付随する自習教材として使えるようなコースウェアとしての教材を自動的に作成することを目的としているのに対して，これらの研究では，コースウェアというよりは，ある授業計画のなかでの個々の技能に関する授業や自習における自動化を目的としていることである．そのため，我々の研究は，これまでの研究と相補的なものであると言える．

次に，本稿においては，学習対象の語彙とコーパスとが与えられていることを前提としたが，学習対象の語彙については，その選定を補助するために，学習対象のコーパスから特徴的な単語を抽出する研究がある(中條 内山 2004; 内山, 中條, 山本, 井佐原 2004)．つまり，学習対象のコーパスがあれば，そこから，語彙は抽出可能である．したがって，これらの研究と本稿での研究とを組み合わせることにより，学習対象のコーパスがあれば，比較的容易に，語彙と読解教材とを作成できることが期待できる．

6 おわりに

本稿では，与えられた学習対象語彙とコーパスとから，その語彙を効率的に獲得できるような読解教材を作成することを目的とする算法を提案し，それにより作成された教材について，その性質および利用例を述べた．作成された教材について，そこに含まれる学習対象語彙の数を調べたところ，それは，無作為抽出された記事集合に比べて，統計的に極めて有意に大きく，作成された教材の有効性が示された．今後の課題は，作成された教材の実際の授業における有効性を検証することである．更に，単語だけでなく，熟語や他の文法事項なども考慮した教材の作成も必要である．

参考文献

Burstein, J. and Leacock, C. (Eds.) (2003). *HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing.*

中條清美 (2003). “英語初級者向け「TOEIC語彙1,2」の選定と効果.” 日本大学生産工学部研究報告 Vol.36, pp.27–42.

中條清美 内山将夫 (2004). “統計的指標を利用した特徴語抽出に関する研究.” 関東甲信越英語教育学会研究紀要第18号 pp. 99–108.

Robertson, S. E. and Walker, S. (2000). “Okapi/Keenbow at TREC-8.” In *Proc. of TREC 8*, pp. 151–162.

内山将夫 井佐原均 (2003). “日英新聞の記事および文を対応付けるための高信頼性尺度.” 自然言語処理, **10** (4), 201–220.

内山将夫, 中條清美, 山本英子, 井佐原均 (2004). “英語教育のための分野特徴単語の選定尺度の比較.” 自然言語処理, **11** (3), 1–33.

ウィリアムスH.P. (1995). 数理計画モデルの作成法. 産業図書.