Organizing English Reading Materials for Vocabulary Learning

Masao Utiyama, Midori Tanimura and **Hitoshi Isahara**
National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Souraku-gun, Kyoto 619-0289 Japan
{mutiyama,mtanimura,isahara}@nict.go.jp

Abstract

We propose a method of organizing reading materials for vocabulary learning. It enables us to select a concise set of reading texts (from a target corpus) that contains all the target vocabulary to be learned. We used a specialized vocabulary for an English certification test as the target vocabulary and used English Wikipedia, a free-content encyclopedia, as the target corpus. The organized reading materials would enable learners not only to study the target vocabulary efficiently but also to gain a variety of knowledge through reading. The reading materials are available on our web site.

1 Introduction

EFL (English as a foreign language) learners and teachers can easily access a wide range of English reading materials on the Internet. For example, current news stories can be read on web sites such as those for CNN,[1] TIME,[2] or the BBC.[3] Specialized reading materials for EFL learners are also provided on web sites like EFL Reading.[4]

This situation, however, does not mean that EFL learners and teachers can easily select proper texts suited to their specific purposes, for example, learning vocabulary through reading. On the contrary, EFL teachers have to carefully select texts, if they want their students to learn a specialized vocabulary through reading in a particular discipline such as medicine, engineering, or economics. However, it is problematic for teachers to select materials for learning a target vocabulary with short authentic texts.

It is possible to automate this selection process given the target vocabulary to be learned and the target corpus from which texts are gathered (Utiyama et al., 2004). In this research (Utiyama et al., 2004), we used a specialized vocabulary for an English certification test as the target vocabulary and used newspaper articles from *The Daily Yomiuri* as the target corpus. We then organized a set of reading materials, which we called *courseware*[5], using the algorithm in Section 2. The courseware consisted of 116 articles and contained all the target vocabulary. We used the courseware in university English classes from May 2004 to January 2005. We found that the courseware was effective in learning vocabulary (Tanimura and Utiyama, in preparation).

Based on the promising results, our next goal is to distribute courseware (produced with our algorithm) to EFL teachers and learners so that we can receive wider feedback. To this end, the courseware we constructed (Utiyama et al., 2004) is inadequate because it was prepared from The Daily Yomiuri, which is copyrighted. We therefore replaced The Daily Yomiuri with English Wikipedia,[6] a free-content encyclopedia, and developed new course-

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

[5] Courseware usually includes software in addition to other materials. However, in this paper, the term *courseware* is used to refer to the reading materials only.
[6] http://en.wikipedia.org/wiki/Main_Page

ware. It is available on our web site.[7]

In the following, will we first summarize our algorithm and then describe details on the courseware we constructed from English Wikipedia.

2 Algorithm

We want to prepare *efficient* courseware for learning a target vocabulary. We defined *efficiency* in terms of the amount of reading materials that must be read to learn a required vocabulary. That is, efficient courseware is as short as possible, while containing the required vocabulary. We used a greedy method to develop the efficient courseware (Utiyama et al., 2004).

Let C be the courseware under development and V be the target vocabulary to be learned. We iteratively select a document (from the target corpus) that has the largest number of new types[8] (types contained in V but not in C) and put it into C until C covering all of V. "C covers all of V" means that each word in V occurs at least once in a document in C.

More concretely, let V_{todo} be the part of V not covered by C, and let V_{done} be $V - V_{\text{todo}}$. We iteratively put document d into C that maximizes $G(\cdot)$,

$$\begin{aligned} & G(d|\alpha, V_{\text{todo}}, V_{\text{done}}) \\ = \; & \alpha g(d|V_{\text{todo}}) + (1-\alpha) g(d|V_{\text{done}}), \end{aligned} \quad (1)$$

until C covers all of V. We then define $g(\cdot)$ as

$$\begin{aligned} & g(d|V_x) \\ = \; & \frac{k_1+1}{k_1((1-b)+b\frac{|W(d)|}{E(|W(\cdot)|)})+1} |W(d) \cap V_x|, \end{aligned} \quad (2)$$

where $W(d)$ is the set of types in d, $E(|W(\cdot)|)$ is the average for $|W(\cdot)|$ over the whole corpus, and k_1 and b are parameters that depend on the corpus. We set k_1 as 1.5 and b as 0.75. $g(d|V_x)$ takes a large value when there is a large number of common types between $W(d)$ and V_x and d is short. These effects are due to $|W(d) \cap V_x|$ and $\frac{|W(d)|}{E(|W(\cdot)|)}$ respectively. As $g(\cdot)$ is based on the Okapi BM25 function (Robertson and Walker, 2000), which has been shown to be quite efficient in information retrieval,[9] we expected $g(\cdot)$ to be effective in retrieving documents relevant to the target vocabulary.

In Eq. (1), α is used to combine the scores of document d, which are obtained by using V_{todo} and V_{done}. It is defined as

$$\alpha = \frac{|V_{\text{done}}|}{1 + |V_{\text{done}}|} \quad (3)$$

This implies that even if $|W(d) \cap V_{\text{todo}}|$ is 1, it is as important as $|W(d) \cap V_{\text{done}}| = |V_{\text{done}}|$. Consequently, $G(\cdot)$ uses documents that have new types of the given vocabulary in preference to documents that have covered types.

To summarize, efficient courseware is constructed by putting document d with maximum $G(\cdot)$ into C until C covers all of V. This allows us to construct efficient courseware because $G(\cdot)$ takes a large value when a document has a large number of new types and is short.

3 Experiment

This section describes how the courseware was constructed by applying the method described in the previous section. We will first describe the vocabulary and corpus used to construct the courseware and then present the statistics for the courseware.

3.1 Vocabulary

We used the specialized vocabulary used in the Test of English for International Communication (TOEIC) because it is one of the most popular English certification tests in Japan. The vocabulary was compiled by Chujo (2003) and Chujo et al. (2004), who confirmed that the vocabulary was useful in preparing for the TOEIC test. The vocabulary had 640 entries and we used 638 words from it that occurred at least once in the corpus as the target vocabulary.

3.2 Corpus

We used articles from English Wikipedia as the target corpus, which is a free-content encyclopedia that anyone can edit. The version we used in this study had 478,611 articles. From these, we first discarded stub and other non-normal articles. We also discarded short articles of less than 150 words. We then selected 60,498 articles that were referred to (linked) by more than 15 articles. This 15-link threshold was

[7] http://www.kotonoba.net/~mutiyama/vocabridge/

[8] A *type* refers to a unique word, while a *token* refers to each occurrence of a type.

[9] BM25 and its variants have been proven to be quite efficient in information retrieval. Readers are referred to papers by the Text REtrieval Conference (TREC, http://trec.nist.gov/), for example.

set empirically to screen out noisy articles. Finally, we extracted a 150-word excerpt from the lead part of each of these 60,498 articles to prepare the target corpus. We set 150-word limit on an empirical basis to reduce the burden imposed on learners. In short, the target corpus consisted of 60,498 excerpts from the English Wikipedia. In the rest of the paper, we will use the term *an article* to refer to *an excerpt* that was extracted according to this procedure.

3.3 Example article

Figure 1 has an example of the articles in the courseware. It was the first article obtained with the algorithm. It shares 27 types and 49 tokens with the target vocabulary. These words are printed in **bold**.

Corporate finance

Corporate finance is the specific **area** of **finance dealing** with the **financial decisions corporations** make, and the tools and **analysis** used to make the **decisions**. The discipline as a whole may be divided between **long-term** and short-term **decisions** and techniques. Both share the same goal of enhancing **firm** value by **ensuring** that return on **capital exceeds cost** of **capital**. **Capital investment decisions** comprise the **long-term** choices about which **projects receive investment**, whether to **finance** that **investment** with equity or debt, and when or whether to pay dividends to shareholders. Short-term **corporate finance decisions** are called working **capital management** and **deal** with balance of **current** assets and **current** liabilities by **managing cash**, **inventories**, and short-term borrowing and lending (e.g., the **credit** terms **extended** to **customers**). **Corporate finance** is closely related to managerial **finance**, which is slightly broader in scope, describing the **financial** techniques **available** to all **forms** of business ... (more)

Figure 1: Example article

3.4 Courseware statistics

3.4.1 Basic courseware statistics

Table 1 lists basic statistics for the courseware constructed from the target vocabulary and corpus.[10] The courseware consisted of 131 articles. Each article was 150 words long because only excerpts were used. The average number of tokens per article shared with the vocabulary ("num. of common tokens" in the Table) was 18.4 and that of types ("num. of common types") was 12.4. About $12.3\%(= \frac{18.4}{150} \times 100)$ of the tokens in each article were covered by the vocabulary. Each article in the courseware was referred to by 70.7 articles on average as can be seen from the bottom row. Table 1 indicates that articles in the courseware included many target words and were heavily referred to by other articles.

[10] On our web site, we prepared 10 sets of article sets called *course-1* to *course-10*. These 10 courses were obtained by repeatedly applying our algorithm to the English Wikipedia removing articles included in earlier courses. The statistics presented in this paper were calculated from the first courseware, *course-1*.

3.4.2 Distribution of covered types

Figure 2 plots the increase in the number of covered types against the order (ranking) of articles that were put into the courseware. The horizontal axis represents the ranking of articles. The vertical axis indicates the number of covered types. The increase was sharpest when the ranking value was lowest (left of figure). The dotted horizontal lines indicate 50% and 90% of the target vocabulary. These lines cross the curved solid line at the 22nd and 83rd articles, i.e., 16.8% and 63.4% of the courseware, respectively. This means that learners can learn most of the target vocabulary from the beginning of the courseware. This is desirable because learners sometimes do not have enough time to read all the courseware.

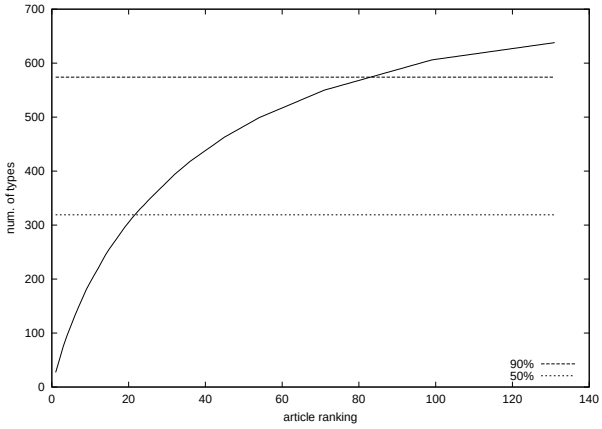

Figure 2: Increase in the number of covered types

3.4.3 Document frequency distribution

Figure 3 has target words that occurred in eight articles or more. The numbers in parentheses indicate the document frequencies (DFs) of the words, where the *DF* of a word is the number of articles in which the word occurred. These words were the most basic words in the target vocabulary with respect to the courseware.

Table 2 lists the distribution of DFs. The first column lists the different DFs of the target words. The values in the "#DF" column are the numbers of

Table 1: Basic courseware statistics (number of articles: 131, length of each article: 150 words)

	Average	SD	Min	Median	Max
Num. of common tokens	18.4	10.8	1	16	55
Num. of common types	12.4	5.5	1	12	27
Num. of incoming links	70.7	145.3	16	32	1056

SD means standard deviation.

words that occurred in the corresponding DF articles. The "CUM" and "CUM%" columns show the cumulative numbers and percentages of words calculated from the values in the second column. As we can see from Table 2, more than 50% of the target words occurred in multiple articles. Consequently, learners were likely to be sufficiently exposed to efficiently learn the target vocabulary.

service (19), form (17), information (12), feature (12), operation (11), cost (11), individual (10), department (10), consumer (9), company (9), product (9), complete (9), range (9), law (9), associate (9), cause (9), consider (9), offer (9), provide (9), present (8), activity (8), due (8), area (8), bill (8), require (8), order (8)

Figure 3: Target words and their DFs.

Table 2: Document frequency distribution

DF	#DF	CUM	CUM%
19	1	1	0.2
17	1	2	0.3
12	2	4	0.6
11	2	6	0.9
10	2	8	1.3
9	11	19	3.0
8	7	26	4.1
7	20	46	7.2
6	25	71	11.1
5	35	106	16.6
4	36	142	22.3
3	71	213	33.4
2	118	331	51.9
1	307	638	100.0

4 Conclusion

While many teachers agree that vocabulary learning can be fostered by presenting words in context rather than isolating them from this, it is very difficult to prepare reading materials that contain the specialized vocabulary to be learned. We have proposed a method of automating this preparation process (Utiyama et al., 2004). We have found that our reading materials prepared from The Daily Yomiuri were effective in vocabulary learning (Tanimura and Utiyama, in preparation).

Our next goal is to distribute courseware (produced with our algorithm) to EFL teachers and learners so that we can receive wider feedback. To this end, we replaced The Daily Yomiuri, which is copyrighted, with the English Wikipedia, which is a free-content encyclopedia, and developed new courseware whose statistics were presented and discussed in this paper. This courseware, which is available on our web site, can be used to supplement classroom learning activities as well as self-study. We hope it will help EFL learners to learn and teachers to teach a broader range of vocabulary.

References

K. Chujo, T. Ushida, A. Yamazaki, M. Genung, A. Uchibori, and C. Nishigaki. 2004. Bijuaru beishikku niyoru TOEIC-yoo goiryoku yoosei sofutowuea no shisaku (3) [The development of English CD-ROM material to teach vocabulary for the TOEIC test (utilizing Visual Basic): Part 3]. Journal of the College of Industrial Technology, Nihon University, 37, 29-43.

K. Chujo. 2003. Eigo shokyuushamuke TOEIC Goi 1 & 2 no sentei to sono kouka [Selecting TOEIC vocabulary 1 & 2 for beginning-level students and measuring its effect on a sample TOEIC test]. Journal of the College of Industrial Technology Nihon University, 36: 27-42.

S. E. Robertson and S. Walker. 2000. Okapi/Keenbow at TREC-8. In *Proc. of TREC 8*, pages 151–162.

Midori Tanimura and Masao Utiyama. in preparation. Reading materials for learning TOEIC vocabulary based on corpus data.

Masao Utiyama, Midori Tanimura, and Hitoshi Isahara. 2004. Constructing English reading courseware. In *PACLIC-18*, pages 173–179.