Applying Automated Vocabulary Extraction and Word Sense Disambiguation in English-Learning Assistance

Chung-Chian Hsu **Chun-Ping Wu** **Hui-Chin Yen** **Yu-Fen Yang**
Nation Yunlin University of Science and Technology
{hsucc, g9823717, hyeh, yangy}@yuntech.edu.tw

Abstract

Learning English is an international trend and how to develop a learning-assistance system that supports effective English learning is an important issue in education. In the traditional way of English learning, learners have to read each word in a text to enhance their reading comprehension. However, in the current era of information technology, the costly and inefficient of learning way fail to meet the needs of learners. Therefore, this study presents an assistance system consisting of three major components: Automated Vocabulary Extraction, Word Sense Disambiguation and Ranking of Vocabulary Frequency, which is called Vocabulary Learning-Assistance System (VLAS). The core functions of the proposed system include three parts: First, it provides the translation of vocabulary based on a Word Sense Disambiguation technique. Second, the system can extract vocabulary in the articles automatically and assign level of the word based on learners' learned vocabulary and the predefined level of vocabulary. Finally, the system provides the Ranking of Vocabulary Frequency based on term frequency. Through the VLAS, efficiency and effectiveness of vocabulary learning are expected to improve. Experimental results indicate VLAS can significantly reduce cognitive load of the learners.

Keywords: Nature Language Process, Word Sense Disambiguation

1. Introduction

In recent years, with the accelerated growth in computer hardware technologies and network technologies, more and more information increasing through time, internet-based applications are bringing about. In the English-language education field, the traditional way of English learning has been unable to satisfy the requirements of learners, how to make use of the immediacy and convenience of internet to design a useful learning environment, which has become an important issue.

In the global village environment, it becomes increasingly important to equip people with foreign language skills; therefore, learning English has become an international trend. Generally, English learning can be divided into four issues including listening, speaking, reading and writing skills, however, Wilkins (Wilkins 1972) argued that "Without grammar very little can be conveyed, without vocabulary nothing can be conveyed". Learners with poor number of vocabulary usually misunderstanding content or have poor comprehension when reading English articles with poor number of vocabulary (Lin and Hsieh 2001). Therefore, vocabulary learning in English language learning is extremely important.

In addition, the existing online vocabulary learning tools have the following disadvantages:

(1) These systems do not provide automatic filtering of vocabulary. Learners must view the articles' vocabulary one by one, and select the vocabulary to their personal glossaries. The loading of learners is very high when learners are reading a new article.
(2) These systems only support a simple translation. In general, this type of online vocabulary learning tool will support the translation, but most of

these systems only support a simple translation. Namely, these systems fail to provide the most appropriate translation to the learners according to the original content of the article.

This study aims to present an assistance system based on Word Sense Disambiguation, Automated Vocabulary Extraction and Ranking of Vocabulary Frequency, which called Vocabulary Learning-Assistance System (VLAS).

The VLAS can assist learners in the amount of vocabulary learning and reading comprehension progress. The VLAS is constructed with the following three functions: the first one is Word Sense Disambiguation. We want to provide vocabulary translation to learners. It is a common problem that polysemous vocabulary appears in natural language processing tasks; therefore, how to determine the appropriate translation and provided to learners that is also the focus of this study. The second is Automated Vocabulary Extraction. We use data preprocessing approach and special rules to filter useful vocabulary from articles and make the vocabulary available for learners, which rules are based on the learners' personal glossaries and predefined parameter of the vocabulary level. Word Sense Disambiguation techniques are used to allow the learners getting the most appropriate meaning of vocabulary of an article from the candidate list. In the third function, we also provide the function of the ranking of vocabulary frequency; it can calculate the term frequency of article or paragraphs, which may help learners in reading comprehension.

In sum, this study presents an assistance system based on Word Sense Disambiguation, Automated Vocabulary Extraction and Ranking of Vocabulary Frequency, which is expected to assist learners to reduce the loading of vocabulary and improves their reading comprehension skill.

This study is divided into five sections. In Section 2, we briefly review related studies of vocabulary learning tools, Word Sense Disambiguation and WordNet. In Section 3, we propose an assistance system based on Word Sense Disambiguation, Automated Vocabulary Extraction of Vocabulary Frequency. In Section 4, the experimental results of Word Sense Disambiguation. In Section 5, conclusions are described.

2. Literature Review

This section briefly reviews the related studies and tools, section 2.1 introduce vocabulary learning tools. Section 2.2 will introduce studies related to Word Sense Disambiguation and section 2.2 introduced WordNet in details.

2.1 Vocabulary learning tool

The well-known portals, such as Google, Yahoo and Microsoft, provided the function of vocabulary translation. In addition, Yahoo also launched kimo mini-pen tool in December 2007, which not only provides the function of general online dictionary, but also provides the learners' personal glossaries management, in addition to other well-known portals do not provide special vocabulary learning mechanism.

However, the proposed systems (Chen and Chung 2008) focused on item response theory and learning memory cycle. The literatures have some drawbacks.

(1) These systems cannot help the learners to automatically filter the required vocabulary so that learners must be select vocabulary one by one by themselves. These systems are unable to help learners reduce their cognitive load in learning.
(2) In general, the systems used the function of vocabulary translation to list all the meanings of the vocabulary, rather than provided the most suitable translation to the learners according to the context of the article.

2.2 Word Sense Disambiguation

In the natural language processing field, polysemy is a common phenomenon. How to correctly analyze and understand natural language is a problem to be solved. Through the context of articles, automatically exclude ambiguity, the term polysemy to determine the significance of articles is the Word Sense Disambiguation.

The approach that Word Sense Disambiguation used before is artificial rule (Wilks 1972; Small 1980), but the cost of artificial rules is too high, which can only deal with limited number of information. Systems that used these methods require a huge dictionary or corpora, which need manual disambiguation information. Therefore, it is an important issue to think about how to have Word Sense Disambiguation to be used from manual to automatic mode.

It has been common to use two kinds of resources: a dictionary and corpora. The first resource, a dictionary (Lesk 1986) used the number of common words among the sense definition of a polysemous word and the sense definitions of its context words. (Wilks, Fass et al. 1990) defined the related words as frequently co-occurring words with the words in a sense definition of a machine-readable dictionary. (Yarowsky 1995) extracted the decision list form corpora automatically using sense definitions of a machine-readable dictionary.

The second resource for WSD is corpus. Corpus-based approaches are divided into two types: supervised learning and unsupervised learning. The supervised learning type, which is use of machine learning and artificial labeled data generated classifier, which through a variety of different situations on the appropriate meaning. The classifier learning data set are usually composed of the information marked by hand, and the target word meaning as well as other information. Another type is unsupervised learning, which is based on unsupervised machine learning with corpora, this type of approach focuses on “one sense per discourse”.

2.3 WordNet

The WordNet is a large lexical database of English, which is developed by Cognitive Science Laboratory at Princeton University under the guidance of Professor George A. Miller. Since 1985, it has more than 25 years of history. The current version is WordNet 3.0. WordNet was not originally intended to have considerable impact on computational linguistics or natural language processing tasks. In the late 80s because of the need for semantic computing, computational linguists found WordNet, and applied to the field of natural language processing tasks.

The feature of WordNet is that it is based on the meaning of the word rather than on lexical grammar to organize messages. WordNet thought synonym set (Synset) to represent the concept. WordNet provides a brief summary for each of the definition of Synset and records the various semantic relations between Synsets.

WordNet has adequate amount of vocabulary. As of 2010, the database contains 155,287 words organized in 117,695 synsets for a total of 206,941 word-sense pairs; in a compressed form, it is about 12 megabytes in size (http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html).

Many studies have utilized WordNet to calculate the similarity. In this study, we also use WordNet to calculate the similarity between words in word sense disambiguation.

3. Method

This section describes system architecture and the details of Word Sense Disambiguation and Automated Vocabulary Extraction. First, an overview of system architecture is presented in Section 3.1. Next, the system components and details of Word Sense Disambiguation, Automated Vocabulary Extraction and Ranking of

Vocabulary Frequency will be introduced in Section 3.2.

3.1 System architecture

An English-learning assistance system based on Automated Extraction and Translation of Vocabulary by Word Sense Disambiguation is presented. Fig. 1 shows the details of system architecture.

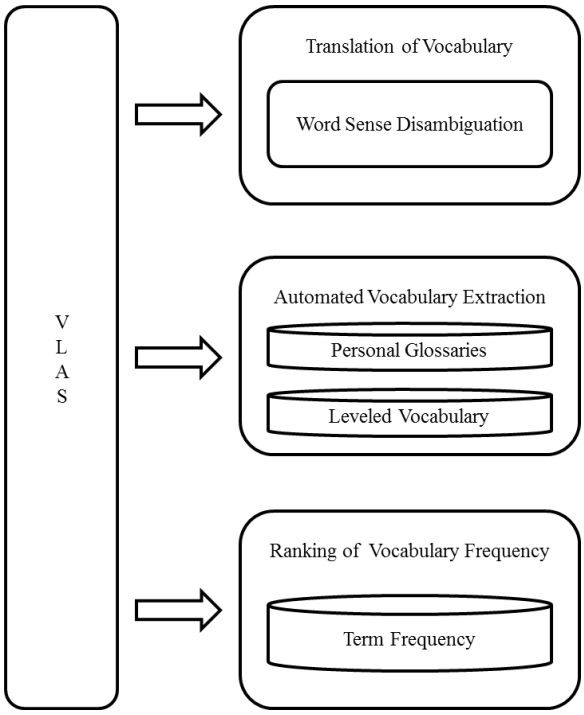

Fig. 1. The system architecture of VLAS

This system has three major components: Translation of Vocabulary, Automated Vocabulary Extraction and Ranking of Vocabulary Frequency. The Translation of Vocabulary mechanism is based on Word Sense Disambiguation technology. The Automated Vocabulary Extraction component can extract vocabulary based on predefined levels of vocabulary and learners' personal glossaries. The Ranking of Vocabulary Frequency is based on occurrence in the article or paragraphs.

3.2 Components of the system

This section describes the components of the system. The main components are divided into three parts: Translation of Vocabulary, Automated Vocabulary Extraction and Ranking of Vocabulary Frequency. The system components of VLAS are shown in Fig. 2.

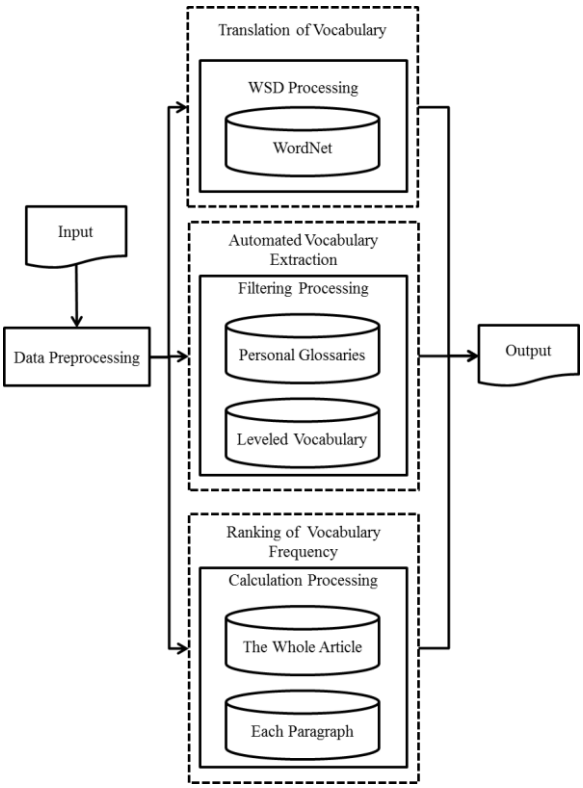

Fig. 2. The system components of VLAS

3.2.1 Processing Step

In Fig. 2, the data preprocessing is the first step, because articles have unstructured formats which include many useless items to learners in terms of English learning, such as stop words, numbers and tags. The articles' unstructured formats also affect the results of extraction. Therefore, we use data preprocessing to help learners collect meaningful and useful vocabulary to learn. The processing steps are shown in Fig. 3.

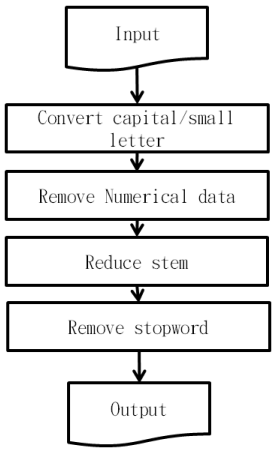

Fig. 3. The steps of data preprocessing.

The steps described in detail as below:

(1) Convert capital/small letter:
To consider the same word but use upper or lower case, all the vocabulary are converted into lower case.

(2) Remove numerical data:
Numerical data, such as date, time, year … etc, is useless to vocabulary learning, so they can be removed.

(3) Stem words:
In this step, we would correct the verb tense into the present tense. We use Martin Porter's Porter Stemming Algorithm (Porter 1980) to reach the goal.

(4) Remove stopwords:
Stopwords, such as "i", "you", "he", "am", "are", "is" … etc, appear frequently in an article, but they are often meaningful and unimportant in an article.

3.2.2 Translation of Vocabulary

The problem of deciding which sense of the word was intended by the writer is an important problem in Word Sense Disambiguation field. As mentioned in the section 2, WSD system usually uses two kinds of resources: a dictionary and corpora. We consider the use of WordNet to achieve this function.

How to identify the most appropriate translation of target word, which has been the purpose of WSD system. The processing steps are shown in Fig. 4.

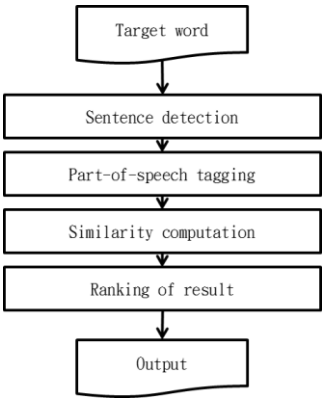

Fig. 4 The steps of WSD

The WSD steps are elaborated as below:

(1) Sentence detection
The purpose of this step is to detect the sentence of the target word. The sentence detection processing is the preprocessing for WSD. Then we use the sentence for the process followed. The algorithms of sentence detection are shown as follows:

```
function GetSentence (w,d):
input: w, the target word
       d, the source document
returns: S, the sentence containing target word
1.  d = source document
2.  S = null
3.  p = getPosition(w)
4.  startFlag = 0
5.  endFlag = 0
6.  While true
7.  If(getWord(p) == ".")
8.     startFlag  = p+1
9.     break
10. Else
11.    P = p-1
12. End while
13.
14. While true
15. If(getWord(p) == ".")
16.    endFlag = p
17.    break
18. Else
19.    p = p+1
20. End while
21. S = getWords(startFlag, endFlag)
22.
23. Return S
```

Fig. 5 The algorithm of GetSentence

(2) Part-of-Speech tagging

This step is intended to mark the Part-of-Speech of the target word in the text. A word may have multiple Part-of-Speeches in WordNet and every part of speech may also have multiple senses. If we can determine the speech of the target word in advance, we need to deal with similarity calculation of the single part of speech rather than all parts of the speech.

We use LingPipe (http://alias-i.com/lingpipe) to tag the part-of-speech. LingPipe is a java-based natural language processing toolkit distributed with source code by Alias-i. The Part-of-Speech tagging result is shown in Figure 6.

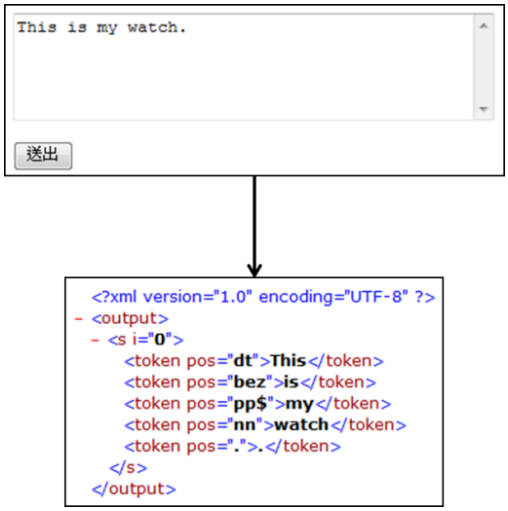

Fig. 6 The result of Part-of-Speech tagging

In the WordNet, there have only the information of noun, verb, adjective and adverb, so this step needs to identify only the target word belonging to one of the four Part-of-Speeches.

(3) Similarity computation

This step is to calculate sentence similarity between the sentence of the target word and the description of each sense in WordNet. We calculate the sentence similarity based on word similarity of each word. Each sense of which the description is most similar to the sentence containing the word is identified as the sense of the word. The word similarity method is based on Pirró's algorithm (Pirró 2009). The example of word similarity is that the word similarity of "dog" and "cat" is greater than the similarity of "dog" and "chair". The samples are shown as in Table 1.

Table 1 The sample of word similarity.

sample	similarity
"dog" and "cat"	0.7420
"dog" and "chair"	0.3439

(4) Ranking of the results

The final step is to sort the results of similarity comparison, and provides to the users. The results allow the users to select the most appropriate sense. If all of the results are 0, we provide the default which is the first sense in the dictionary to the users. Because the first sense in WordNet has the most frequent usage, a screenshot of word senses in WordNet is shown in Fig. 7.

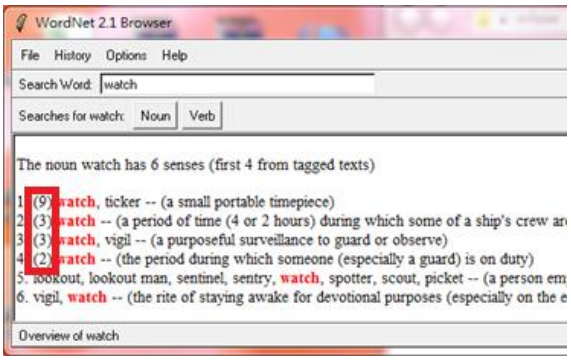

Fig. 7 The screenshot of WordNet Browser

The number in red box is obtained from the corpus. The higher the number is, the higher the probability of the sense appears in sentences.

3.2.3 Automated Vocabulary Extraction

We used data preprocessing and filtering rules to implement the Automated Vocabulary Extraction function. Fig. 2 presents detailed information of Automated Vocabulary Extraction processing.

Based on Fig. 2, there are two major processes in Automated Vocabulary Extraction: preprocessing step and filtering step. In the preprocessing stage, we use data preprocessing in order to filter out unimportant or meaningless vocabulary for learners. In the filtering rules, we use learners' personal glossaries and the predefined level of vocabulary to extract vocabulary for learners after preprocessing. The details of preprocessing stage and vocabulary filter rules are described as the following sections.

3.2.3.1 Filtering rules of vocabulary

The system uses several filtering rules to extract vocabulary to learners. Not all words in the article are useful to the learners. Some of the words are already learned by the learners. Some of them are just symbols or numbers. Therefore, this study conducted two ways to filter useless words. The system provides the learners helpful vocabulary in order to reduce the learners' learning loading. The filtering ways are divided into two parts and described below.

(1) Filtering based on the vocabulary of a predefined level.
English words have been categorized according to the English ability of the learners, such as GEPT, TOEIC vocabulary and other related information. Therefore, this study will use the predefined level of vocabulary as the basis, and provide to learners the vocabulary with higher level than the predefined level of vocabulary.

(2) Filtering based on personal glossaries.
After learners undertake a number of tasks, they will accumulate personal glossaries. When learners do more learning tasks, they will gradually increase their understanding of vocabulary of new tasks. Therefore, this rule is based on the learners' personal glossaries, allowing learners to organize the unknown vocabulary.

3.2.4 Ranking of Vocabulary Frequency

According to the calculation of the term frequency in the article or paragraphs, Ranking of Vocabulary Frequency function is used to help learners find the main idea of the article and the main idea of individual paragraphs so as to improve learners' reading comprehension. Because the article after data preprocessing step, frequent terms may be representative of the article or the paragraph and have a degree of significance.

In the calculation rules, we can get the ranking of the article main idea and the paragraph main idea after data preprocessing according to the following formula.

The formula for counting term frequency for main idea can be specified as

$$tf_{w_{i,a}} = \sum_{j=1}^{k} I(w_i, w_j) \qquad (1)$$

where w_i is the input word, k is the length of article a and $I(w_i,w_j)$ is an indication function which returns 1 if $w_i = w_j$ and otherwise 0.

The formula for counting the term frequency in paragraphs can be specified as

$$tf_{w_{i,p}} = \sum_{j=1}^{k} I(w_i, w_j) \qquad (2)$$

where w_i is the input word, k is the length of paragraph p, and $I(w_i,w_j)$is an indication function which returns 1 if $w_i = w_j$ and otherwise 0.

According to the above formulas, we can get the word frequency in the article or each paragraph. By Eq. (1), we sort the words in the article, and list the most frequent occurrences words to the learners as the main idea. By Eq. (2), we sort the words in each paragraph, and list the most frequent occurrences words to the learners as the paragraph idea. The algorithms are shown in Fig. 8.

function GetRankList (a):
input: a, the article
returns: L_a, the ranking list of article
L_p, the ranking list of each paragraph
1. L_a = {}
2. L_p = {}
3. P = {}
4. j = 1
5.
6. For each token $c_i \in a$:
7. $p_j = p_j \cup c_i$
8. If ((c_i == newline) And (c_{i-1} == "."))
9. p_j = removeStopword(p_j)
10. p_j = stemming(p_j)
11. For each term $t_i \in p_j$:
12. For each term $t_j \in p_j$:
13. $tfp_{ti} = tfp_{ti} + I(t_i,t_j)$
14. $tfa_{ti} = tfa_{ti} + I(t_i,t_j)$
15.
16. $j = j+1$
17. End for
18.
19. L_a = sorted tfa by descending
20. L_p = sorted tfp by descending
21.
22. Return L_a and L_p

Fig. 8. The algorithm of GetRankList

4. Experimental results

In this section, we focus on two directions, first is contribution of Part-of-Speech tagging, second is ranking of Word Sense Disambiguation.

We selected five articles from an English magazine. After automated vocabulary extraction according to the intermediate level of GEPT, we got all the matching words for the experiment, the information are shown in Table 2.

Table 2 The number of vocabulary before and after filtering according to the intermediate level of GEPT

Article	1	2	3	4	5
Before filtering	323	352	572	549	865
After filtering	21	13	38	33	44

4.1 The contribution of POS tagging

In this section, we analyze the contribution of Part-of-Speech tagging. We observe the differences before and after Part-of-Speech tagging, and the difference is the contribution of this processing.

For example, a word may have multiple parts of speech (e.g. verb, noun, adjective, adverb, etc.), if we can get the correct part of speech of the word from its original, then we need to deal with only these candidate senses from a particular part of speech and ignore the other parts so as to reduce the computation load.

The contribution of POS tagging is show in Table 3, depicting a reduction of 24% compared to that without using POS tagging.

Table 3 The contribution of POS tagging

sense number before POS tagging	sense number after POS tagging	reduction of computational load
638	485	24%

4.2 Translation of vocabulary

In this section, we perform word sense disambiguation experiments. We use the words from the previous extraction based on intermediate level of GEPT and calculate sentence similarity between source sentence and each sense definition.

The calculation of sentence similarity is based on word similarity. According to the similarity score, we sort it and provide to learners. The accuracy is shown in Table **4**.

Table 4 The accuracy of WSD

	One POS with one sense	One POS with multiple senses	Multiple POS with multiple senses
Number of vocabulary	21	78	50
Number of sense	21	287	330
Random accuracy (%)	100.00	0.43	0.50
First sense accuracy (%)	100.00	46.15	42.00
Accuracy of top1 by VLAS (%)	**100.00**	**58.97**	**54.00**
Accuracy of top3 by VLAS (%)	**100.00**	**92.31**	**88.00**

In Table 4, we compared three methods, random accuracy, first sense accuracy and

VLAS. There have three types of word structure, one POS with one sense, one POS with multiple senses and multiple POS with multiple senses. According to the results, the random accuracy is the lowest, and the first sense accuracy is only a little lower than accuracy of top1 by VLAS, because the dictionary usually put the most common sense in the top. Our method is the best in each type, because we consider the POS tagging and sentence similarity. The accuracy of top3 by VLAS did not achieve one hundred per cent. We analyzed the data and found major reasons as follows.

(1) The word is a multi-word or phrase.
After automated vocabulary extraction, the multi-word or phrase was cut to single words for example, "*flock* to", "*tone* down". Therefore, a multi-word or phrase after automated vocabulary extraction, it lost its original meaning.

(2) The word is a person name or terminology.
In the experiment, some words are person names or terminology in the original article, for example, " *van* paasschen says.". Therefore, in the WSD, process exception raised since person names and terminology do not exist in a regular English dictionary.

(3) The original sentence in the article is too short.
In the process of obtaining the sentence, some of the sentences are too short, the information of sentence is not enough to express the original meaning, for instance, "i kind of go into my *shell*" As a result, WSD fail to identify the correct sense.

4.3 Ranking of Extracted Vocabulary

In this section, we perform ranking of extracted vocabulary experiments. We selected the previous five articles, and identify the top-3 frequent keywords. The ranking results are shown in Table 5.

Table 5 The top 3 keywords for each article

Article title	Top-3 frequent keywords
Sport Stacking	sport, stack, player
Hannah Montana: "Tween" Queen	show, miley, hannah
Traveling Through Texas	cowboy, dinosaur, texas
Workplace Personalities	introvert, people, personality
Luxury Hotels	luxury, city, hotel

In Table 5, the top-3 frequent keywords for each article are consistent with the theme of each article. The article "Sport Stacking" is talking about the promotion of sport stacking in school. The article "Hannah Montana: "Tween" Queen" is talking about the introduction of a famous female singer in the United States. The article "Traveling Through Texas" is talking about traveling through texas. The article "Workplace Personalities" is talking about various personalities in workplace. The article "Luxury Hotels" is talking about expensive and luxury hotels, for example Dubai's Burj Al-Arab. All the keywords are related to the topics of the articles. In other words, the extracted frequent keywords can represent the main idea of the article.

5. Conclusion

In this study, we have proposed a learning-assistance system based on Word Sense Disambiguation, Automated Vocabulary Extraction and Ranking of Vocabulary Frequency, which can assist learners to reduce the loading of vocabulary learning and improve their reading comprehension skill.

The proposed learning-assistance system based on Automated Vocabulary Extraction and WSD of this study aims to achieve the following contribution.

(1) To reduce the vocabulary loading of learners. Through Automated Vocabulary Extraction, this system helps learners reduce the vocabulary

loading when reading, thereby increasing their learning motivation.

(2) To strengthen the learners' reading comprehension. Through Translation of Vocabulary and Ranking of Vocabulary Frequency, the system not only provides the translation of vocabulary from articles, but also the main idea and the paragraph idea of an article to the learners. As a result, for learners, the effectiveness in reading comprehension will increase.

Acknowledgment

This research was supported in part by National Science Council, Taiwan, under grant NSC 98-2410-H-224-010-MY2.

References

[1] Chen, C.-M. and C.-J. Chung (2008). "Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle." Computers & Education **51**(2): 624-645.

[2] Lesk, M. (1986). Automatic sense disambiguation: How to tell a pine cone from an ice cream cone. In Preceedings of the 1986 SIGDOC Conference, New York, Association for Computing Machinery.

[3] Lin, B. and C.-t. Hsieh (2001). "Web-based teaching and learner control: a research review." Computers & Education **37**(3-4): 377-386.

[4] Pirró, G. (2009). "A semantic similarity metric combining features and intrinsic information content." Data & Knowledge Engineering **68**(11): 1289-1308.

[5] Porter, M. (1980). "An algorithm for suffix stripping." Program: Electronic Library & Information Systems **40**(3): 211-218.

[6] Small, S. (1980). Word expert parsing: a theory of distributed word-based natural language understanding, University of Maryland. **Doctoral dissertation**.

[7] Wilkins, D. A. (1972). Linguistics in language teaching. Cambridge,, MIT Press.

[8] Wilks, Y. (1972). Grammar, meaning and the machine analysis of language. London,, Routledge and K. Paul.

[9] Wilks, Y., D. Fass, et al. (1990). "Providing machine tractable dictionary tools." Machine Translation **5**(2): 99-154.

[10] Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics.