The Design of a Lexical Difficulty Filter for Language Learning on the Internet

Chin-Hwa Kuo*, David Wible**, Chih-Chiang Wang*, and Feng-yi Chien**

**Computers and Networking (CAN) Laboratory, Department of Computer Science and Information Engineering, Tamkang University, chkuo@mail.tku.edu.tw*

***Graduate Institute of Western Languages and Literature, Tamkang University, dwible@mail.tku.edu.tw*

Abstract

In this paper, we describe the design and implementation of a tool called the Lexical Difficulty Filter (LDF) intended to help learners or teachers in selecting sentences appropriate to the level of students for English vocabulary learning. The LDF serves as a bridge between learner and information resource bank. It selects suitable sentences from the output of standard English corpus concordancers. The core technology in the design of LDF is a fuzzy expert system. The results obtained are very encouraging in this pioneering work. We achieve 94.33% accuracy rate in imitating the judgments of a human expert in determining the degree of difficulty of a sentence for a given learner level.

1. Introduction

The Internet is an important resource for information acquisition. This information can benefit to teachers as well as learns. However, the information obtained from the Internet directly may not be suitable to learners, and further processing is often required. In this paper, a *Lexical Difficulty Filter* (LDF) for language learning on the Internet is described. The main purpose of the designed LDF is to determine the *degree of difficulty* of a given English sentence or paragraph. After this filtering process, the learning materials that are suitable to learner's level are delivered. As a result, learners are in a better-prepared learning environment.

Text processing, such as retrieval, understanding, summarization, clustering and classification, has been a lively field of research for several decades. The design concept of the LDF in our work is similar to text classification. The core technology used in this paper is based on a fuzzy rule generation system, an extended development based on the work by Hong and Lee in [1]. The LDF is able to achieve 94.33% accuracy in determining the degree of difficulty of the corresponding text.

2. System Overview

Using the Internet and/or electronic documents is an important trend in delivering learning materials. However, the materials obtained directly from such an environment may not be suitable for learners. Therefore, pre-processing may be required. As shown in Fig. 1, the LDF determines the degree of the difficulty of a given electronic text. This can be applied to offer learners numerous sentences that contain a new vocabulary word in context, but with the difficult sentences filtered out by the LDF. Or it can be used by teachers to check the level of lexical difficulty of a certain text for his or her students.

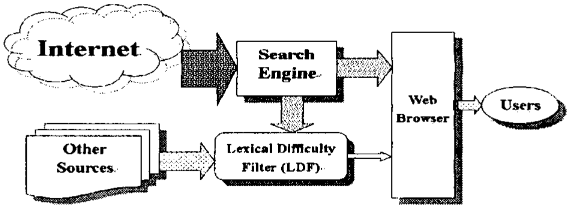

Figure 1. System overview

3. Design of LDF

Sentence examples provided by British National Corpus Online Concordancing are used as our sample data set. Each sentence is translated into four representative parameters:

1. Threshold Value

British National Corpus maintains a database called "vocabulary frequency list". Each user determines a limit value based on this database. Words that are within this limit value are assumed to be included in the user's vocabulary.

2. Words

The number of words contained in the sentence.

3. Percentage (PofW)

The percentage of words that are within the given threshold.

0-7695-1013-2/01 $10.00 © 2001 IEEE

4. Difficulty

The degree of difficulty should be provided by a human expert for each sentence in the training sample. After training, the LDF will then determine the difficulty levels for sentences in the test sample.

The system architecture of the LDF is illustrated in figure 2. The functions of each component are as follows.

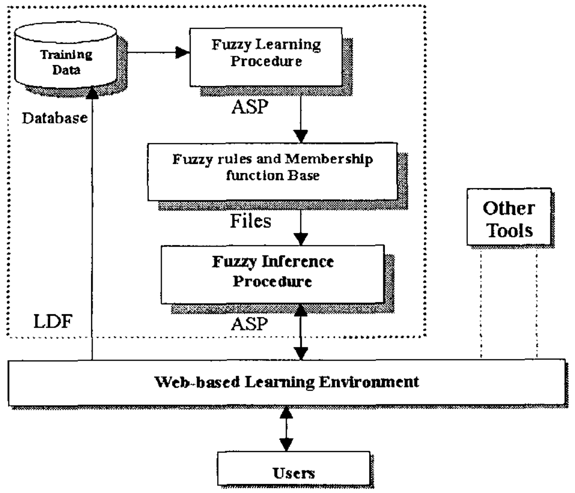
Training Data
Database
Fuzzy Learning Procedure
ASP
Fuzzy rules and Membership function Base
Files
Fuzzy Inference Procedure
ASP
LDF
Other Tools
Web-based Learning Environment
Users

Figure 2. Architecture of the LDF

- Web-Based Learning Environment

  This is a collection of web-based GUI. It includes many tools to facilitate the use of the learning environment. The LDF can be viewed as one of these tools.
- Training Data

  A database contains training sentences with user dependent degree of difficulty judged by a human expert.
- Fuzzy Learning Procedure

  After the inputs to the learning algorithm are the training examples. And the corresponding outputs are the fuzzy rules, and membership functions.
- Fuzzy Rules and Membership Functions Base

  Files that encapsulate fuzzy rules and membership functions are generated by the fuzzy learning procedure.
- Fuzzy Inference Procedure

  This procedure classifies sentences into difficulty levels by applying fuzzy rules and membership functions on parameters of the sentences.

4. Results

A large data set of 260 sentences was analyzed to ascertain the practicality of our system. Professional English teachers were asked to classify the data. Half of the data set was then used as the training sample for our algorithm. The rest was used as the testing sample. The fuzzy learning algorithm generates 167 rules. The computations of the degree of difficulty are based on the concept of a contingency table [2].

The statistical results are shown in Table 1. The overall accuracy achieves 94.33%.

Table 1. The error rate of the LDF

	Error Rate
Easy	8.46 %
Normal	0.1 %
Hard	8.46 %
Avg.	**5.67 %**

The average error rate of our algorithm on the testing sample is 5.67%. Furthermore, a scrutiny of the inaccurately classified sentences shows that the errors are never off by more than a single level. Thus, an "easy" classification by the system will at most be a "normal" classification and not a "hard" even if the classification is incorrect.

5. Conclusions

An integration of the lexical difficulty filter into web-based environments will allow teachers to provide students with materials of appropriate difficulty levels from the web without spending precious human capital to ascertain the difficulty levels of the materials [3].

The core technology in the designing of LDF is based on fuzzy expert systems. The results obtained are very encouraging in this pioneer work. We achieve a 94.33% accuracy rate in the determination of the degree of difficulty for a given learner level. There are several areas of potential extension to this research. The analysis of vocabulary only includes each word's value in the vocabulary frequency list. No information about the actual word is used. The additional information may enable the algorithm to classify difficulty levels more accurately. Another extension is the analysis of optimal sample sizes of the training data set.

6. References

[1] T. P. Hong and C. Y. Lee, "Learning fuzzy knowledge from training examples", In Proceedings of *ACM-CIKM98*, page 161-166, 1998.

[2] D. D. Lewis, "Evaluation and optimizing autonomous text classification systems", In Proceedings of *SIGIR*, page 246-254, 1995.

[3] IWiLL group, http://www.iwillnow.org, 2001.