Computational Linguistics and Chinese Language Processing
Vol. 18, No. 2, June 2013, pp. 1-18

Assessing Chinese Readability using Term Frequency and Lexical Chain

Yu-Ta Chen*, Yaw-Huei Chen*, and Yu-Chih Cheng*

Abstract

This paper investigates the appropriateness of using lexical cohesion analysis to assess Chinese readability. In addition to term frequency features, we derive features from the result of lexical chaining to capture the lexical cohesive information, where E-HowNet lexical database is used to compute semantic similarity between nouns with high word frequency. Classification models for assessing readability of Chinese text are learned from the features using support vector machines. We select articles from textbooks of elementary schools to train and test the classification models. The experiments compare the prediction results of different sets of features.

Keywords: Readability, Chinese Text, Lexical Chain, TF-IDF, SVM.

1. Introduction

Readability of an article indicates its level in terms of reading comprehension of children in general. Readability assessment is a process that measures the reading level of a piece of text, which can help in finding reading materials suitable for children. Automatic readability assessment can significantly facilitate this process. There are other applications of automatic readability assessment such as the support of building a web search engine that can distinguish the reading levels of web pages (Eickhoff, Serdyukov, & de Vries, 2010; Miltsakaki & Troutt, 2008) and the incorporation into a text simplification system (Aluisio, Specia, Gasperin, & Scarton, 2010). Traditional measures of text readability focus on vocabulary and syntactic aspects of text difficulty, but recent work tries to discover the connections between text readability and the semantic or discourse structure of texts (Feng, Elhadad, & Huenerfauth, 2009; Pitler & Nenkova, 2008).

Most of the existing work on automatic readability assessment is conducted for English

* Department of Computer Science and Information Engineering, National Chiayi University, Chiayi, Taiwan, R.O.C.
E-mail: {s0960413, ychen, s0990413}@mail.ncyu.edu.tw

text. In contrast, research on readability assessment for Chinese text is still in its initial stage. This paper investigates the appropriateness of using lexical cohesion analysis to improve the performance of Chinese readability assessment. More specifically, we build lexical chains, which are sequences of semantically related terms, in an article to represent the lexical cohesive structure of texts, and then derive features from the result of lexical chaining to capture the lexical cohesive information. Consisting of term frequency features and lexical chain features, various combinations of features are evaluated for generating prediction models on Chinese readability using support vector machines (SVMs). The prediction models are trained and tested on articles selected from textbooks of elementary schools in Taiwan. The results are compared for different sets of features.

This paper is organized as follows. Section 2 introduces related work in readability assessment and lexical cohesion analysis. Section 3 discusses the research methodology of our analysis, including problem definition, text processing, feature deriving, and prediction model building. Section 4 presents the experiments and the experimental results. Section 5 gives conclusions and directions for future work.

2. Related Work

This section briefly surveys existing work in the areas of readability assessment and lexical cohesion analysis.

2.1 Readability Assessment

Traditional readability formulae for English are based on shallow features such as average sentence length and average number of syllables per word to approximate syntactic and vocabulary difficulty in text (Kincaid, Fishburne Jr., Rogers, & Chissom, 1975; McLaughlin, 1969). However, this kind of measure makes strong assumptions about text difficulty and may not be always reliable.

With the growth of computational power, researchers began to have the ability to use word frequency as a better measure of word difficulty (Chall & Dale, 1995; Stenner, 1996). Word frequency information can be used in two ways. One is to maintain lists of common and rare words and to use the percentage of words in the article that are present or absent in the lists as features to measure the reading difficulty of that article (Chall & Dale, 1995; Lin, Su, Lai, Yang, & Hsieh, 2009; Schwarm & Ostendorf, 2005). The other is to compute the numbers of occurrences of words from a corpus and to use the computed word frequencies as features to measure the reading difficulty (Stenner, 1996). The effects of both methods rely on careful choice of corpus used to generate the word lists and frequency information, however, the second method is more flexible in that it can be incorporated into other models such as the term frequency-inverse document frequency (TF-IDF) scheme.

Some researchers suggest that text readability can be measured by factors in semantic aspect in addition to vocabulary and syntactic ones. Aluisio *et al.* (2010) consider the ambiguity ratio of terms for each part-of-speech (POS) as a feature for assessing text readability in Portuguese. Feng, Jansche, Huenerfauth, & Elhadad (2010) use some features inspired by cognitive linguistics to measure text readability, such as the number of named entities and the distribution of lexical chains in an article.

Some Chinese-specific factors, such as radical familiarity, number of strokes, geometry or shape of characters, are also considered (Lau, 2006). However, it is unclear whether these character-level features can truly benefit the readability assessment on Chinese text. Recently, machine learning based approaches also have been proposed for accessing Chinese readability (Chen, Tsai, & Chen, 2011; Sung, Chang, Chen, Cha, Huang, Hu, & Hsu, 2011).

2.2 Lexical Cohesion Analysis

Two properties of texts are widely used to indicate the quality of a text, coherence and cohesion. According to Morris and Hirst (1991), coherence refers to the fact that there is sense in a text, while cohesion refers to the fact that elements in a text tend to hang together. The former is an implicit quality within the text, whereas the latter is an explicit quality that can be observed through the text itself. Observing the interaction between textual units in terms of these properties is a way of analyzing the discourse structure of texts (Stokes, 2004). Discourse structure of a text is sometimes subjective and may require knowledge from the real world in order to truly understand the text coherence. However, according to Hasan (1984), analyzing the degree of interaction between cohesive chains in a text can help the reader indirectly measure the coherence of a text. Such cohesion analysis is more objective and less computationally expensive.

Halliday and Hasan (1976) classify cohesion into five types: (1) conjunction, (2) reference, (3) lexical cohesion, (4) substitution, and (5) ellipsis. Among these types, lexical cohesion is the most useful one and is the easiest to identify automatically since it requires less implicit information behind the text to be discovered (Hasan, 1984). Lexical cohesion is defined as the cohesion that arises from semantic relationships between words (Morris & Hirst, 1991). Halliday and Hasan (1976) further define five types of lexical cohesive ties in text: (1) repetition, (2) repetition through synonymy, (3) word association through specialization/ generalization, (4) word association through part-whole relationships, and (5) word association through collocation. All of the semantic relationships mentioned above except for collocation can be obtained from lexicographic resources such as a thesaurus. The collocation information can be obtained by computing word co-occurrences from a corpus or be captured using an n-gram language model with $n > 1$.

Lexical chaining is a technique that is widely used as a method to represent lexical cohesive structure of a text (Stokes, 2004). A lexical chain is a sequence of semantically related words in a passage, where the semantic relatedness between words is determined by the above-mentioned lexical cohesive ties usually with the help of a lexicographic resource such as a thesaurus. Lexical chains have been used to support a wide range of natural language processing tasks including word sense disambiguation, text segmentation, text summarization, topic detection, and malapropism detection.

Different lexicographic resources capture different subset of the lexical cohesive ties in text. Morris and Hirst (1991) use Roget's thesaurus to find cohesive ties between words in order to build lexical chains. WordNet (Fellbaum, 1998) is an online lexical database and has predominant use in information retrieval and natural language processing tasks, including lexical chaining. The major relationship between words in WordNet is synonymy, and other types of relationships such as hypernymy and hyponymy are defined among synsets, sets of synonymous words, forming a semantic network of concepts.

HowNet is a lexical database for Chinese words developed by Dong (n.d.). The idea of HowNet is to use a finite set of primitives to express concepts or senses in the world. The whole set of primitives are defined in a hierarchical structure based on their hypernymy and hyponymy relationships. Each sense of a word is defined in a dictionary of HowNet using a subset of the primitives. HowNet so far has two major versions: the 2000 version and the 2002 version. The 2000 version defines a word sense by a flat set of primitives with some relational symbols that determine the relation between the primitive and the target word sense. On the other hand, the 2002 version of HowNet uses a nesting grammar to define a word sense. A definition consists of primitives and a framework. The framework organizes the primitives into a complete definition. Dai, Liu, Xia, & Wu (2008) propose a method to compute lexical semantic similarity between Chinese words using the 2002 version of HowNet. For traditional Chinese, E-HowNet (Extended HowNet) is a lexical semantic representation system developed by Academia Sinica in Taiwan (CKIP Group, 2009). It is similar to the 2002 version of HowNet with the following major differences: (1) Word senses (concepts) are defined by not only primitives but also any well-defined concepts and conceptual relations, (2) Content words, function words, and phrases are represented uniformly, and (3) The incorporation of functions as a new type of primitive. An example of word sense definition is shown in Figure 1. Due to the first major difference mentioned above, a word sense definition may contain another well-defined word sense, such as "大學" (university, college) in the example. A bottom level expansion of the definition can be obtained by expanding all well-defined concepts in the top level definition, as shown in Figure 2.

```
教授    N    {老師:
                location={大學}}
```

Figure 1. Top level definition of a word sense in E-HowNet.

```
教授    N    {human|人:
                domain={education|教育},
                predication={teach|教:
                                agent={~}
                            },
                location={InstitutePlace|場所:
                                domain={education|教育},
                                telic={or({study|學習:
                                                location={~}
                                        },
                                        {teach|教:
                                                location={~}
                                        }
                                      )
                                    },
                                qualification={HighRank|高等}
                            }
        }
```

Figure 2. Bottom level expansion of the definition of a word sense in E-HowNet.

It has been suggested that coherent texts are easier to read (Feng *et al*., 2010), and some previous studies have used lexical-chain-based features to assist in readability assessment of English text (Feng *et al*., 2009; Feng *et al*., 2010). Some other ways of modeling text coherence are also used for readability assessment, such as the entity-grid representation of discourse structure and coreference chains (Barzilay & Lapata, 2008; Feng *et al*., 2009; Pitler & Nenkova, 2008). However, none of these discourse-based factors are tested on Chinese text for estimating readability. In this paper, we evaluate a combination of term frequency features and lexical chain features for generating classification models on Chinese readability.

3. Assessing Readability using SVM

This section presents the methodology adopted for assessing readability of Chinese text using SVM. We first explain the problem of readability assessment, basic concepts of SVM classification, and the system design. Then we describe how we conduct the text processing

step, followed by the features we use for representing each article in the corpus. Finally, we discuss the performance measures used in the experiments.

3.1 Problem Definition

Various types of prediction models have been tested on the task of readability assessment in previous research (Aluisio *et al*., 2010; Heilman, Collins-Thompson, & Eskenazi, 2008), including classification and regression models. Since several studies obtain better results when using SVM classification than regression models (Feng *et al*., 2010; Petersen & Ostendorf, 2009; Schwarm & Ostendorf, 2005), in this paper we treat the problem of Chinese readability assessment as a classification task where SVM is used to build classifiers that predict the reading levels of given texts.

Readability can be classified according to grade levels, but the difference between adjacent grades may be insignificant, which makes the classification result less accurate. More importantly, grade-level readability is too fine for many applications and a broader range of readability level is more practical. For example, the U.S. government surveyed over 26,000 individuals aged 16 and older and reported data with only five levels of literacy skills (National Center for Education Statistics, 2002). Therefore, we divide reading skills of elementary school students into three levels: lower grade, middle grade, and higher grade, where lower grade corresponds to the first and second grade levels, middle grade corresponds to the third and fourth grade levels, and higher grade corresponds to the fifth and sixth grade levels.

In this paper, we try to evaluate different combinations of features for predicting the reading level of a text written in traditional Chinese as suitable for lower grade or middle grade. We will build one prediction model for lower grade level and another prediction model for middle grade level. These binary SVM classifiers can be combined to solve the multiclass problem of predicting the reading level of an article (Duan & Keerthi, 2005; Hsu & Lin, 2002).

While most studies on readability assessment view the reading levels as discrete classes, we think readability is continuous. That is, an article that is suitable for students of a certain level must also be comprehensible for students of higher levels. Similarly, if a student can understand an article of a certain reading level, he/she must also be able to understand any article of a lower reading level. Therefore, when building classifiers for lower grade, we use articles of grades 1 and 2 as positive data, while the others are negative data. When building classifiers for middle grade, articles of grade 1 through grade 4 are used altogether as positive data, while those of higher grade levels are used as negative data.

3.2 Text Processing

After the data set is collected, each article is undergone a word segmentation process as a pre-processing step before deriving features from the texts. Word segmentation is done using a word segmentation system provided by Academia Sinica (CKIP Group; n.d.). The segmentation result is stored in XML format, where POS-tags are attached to all words and sentence boundaries are marked.

It is reported by Yang and Petersen (1997) that chi-square test (χ^2) performs better than other feature selection methods such as mutual information and information gain in automatic text classification. Therefore, we use chi-square test to evaluate the importance of terms in the corpus with respect to their discriminative power among reading levels. The chi-square test is used to test the independence of two events, which, in feature selection, are the occurrence of the term and the occurrence of the class. Higher chi-square test value indicates higher discriminative power of the term to the classes. For each prediction model, we compute chi-square test value for each term in the corpus. Such information will benefit our feature derivation process described below. We do not perform stop word removal and stemming because Collins-Thompson and Callan (2005) report that these processes may harm the performance of classifier on lower grade levels.

3.3 Feature Deriving

The use of term frequencies as the primary information for assessing Chinese readability has been investigated (Chen, Tsai, & Chen, 2011), where TF-IDF values of the terms with high discriminative power are used as features for SVM classification. This paper investigates the appropriateness of using lexical cohesion analysis to improve the performance of Chinese readability assessment. Therefore, we build lexical chains for both the training and testing documents and deriving features from the lexical chains to capture the lexical cohesive aspect of the texts.

A general algorithm for generating lexical chains is shown in Figure 3, which is a simplified version of that proposed by Morris and Hirst (1991) as described in (Stokes, 2004). The chaining constraints in the algorithm are highly customizable and are the key to the quality of the generated lexical chains. The allowable word distance constraint is based on the assumption that relationships between words are best disambiguated with respect to the words that lie nearest to each other in the text. The semantic similarity is the most important factor that determines term relatedness and is generally based on any subset of the lexical cohesive ties mentioned above. Figure 4 shows an example of the lexical chaining result.

Choose a set of highly informative terms for chaining, t_1, t_2, …, t_n.
The first candidate term in the text, t_1, becomes the head of the first chain, c_1.
For each remaining term t_i do
    For each chain c_m do
        If the chain is most strongly related to t_i with respect to allowable word distance and semantic similarity
        Then t_i becomes a member of c_m,
        Else t_i becomes the head of a new chain.
    End for
End for

Figure 3. A general lexical chaining algorithm.

Original Text:

在都會區房價飆高之時，銀行業整體壞帳率創下歷史新低，業界人士對此相當擔心，房價看來將持續疲軟到明年第 1 季，近 2 年承作的房貸物件將無上漲空間，尤其是泛公股行庫的整批房貸，多數是在「升緩、跌快」的市郊區，亦是銀拍屋的集中地，若明年房貸呆帳湧現，恐成爲銀行業系統性風險爆發的最大來源。

爲避免壞帳爆增牽連銀行的獲利，國銀和消金外銀間正默默建立共識，絕對不能以「衝業務」的理由，在房貸市場推銷低利產品，不單純是業者配合央行的特別監管，主要是台灣金融業再也經不起龐大虧損。

銀行業的「壞年」定義，意指容易發生貸款壞帳的條件氛圍，最常見的狀況即現階段的房價漲、投資氣氛濃；其相反即是「好年」，如全球金融海嘯期間或 SARS 期間，雖然房市冷、價格縮，有能力消費或擔保融貸的消費者卻都是「百中選一」的信用良好者，從銀行取得房貸的標的，對成數不奢望，還款時間卻往往超前計畫的 50%以上，銀行鮮少因此發生壞帳。

根據金管會統計，目前本國銀行的壞帳持續改善，整體銀行平均逾放比在歷史低點的 0.96%，完全擺脫多年前動輒 4 個百分點的可怕記錄。銀行業者認爲，融貸逾放的來源有 2 大項目，信用卡和房地產，前者經常維持在 2-3%之間，後者則因貸出利率僅 2%，銀行能夠獲利的空間很小，長期以來平均逾放率在 1%，實在經不起任何房價超跌的折損衝擊。

房貸圈目前存有一種默認的共識，泛公股行庫的三商銀和民營銀行的中信銀，不該帶頭促銷房貸產品，而消金外銀則繼續強化個人徵信，從風險管控著手，多管道降低「壞年」留下的逾放壓力。

Derived Lexical Chains:

lexical chain 1: (1)銀行業-3 (2)銀行業-25 (3)業務-34 (4)銀行業-45 (5)金融-59
lexical chain 2: (1)整體-4 (2)物件-14 (3)期間-61 (4)期間-62 (5)時間-74 (6)目前-79 (7)整體-82 (8)前者-94 (9)後者-95 (10)目前-104
lexical chain 3: (1)成數-73 (2)百分點-86 (3)利率-96
lexical chain 4: (1)系統性-26 (2)來源-28 (3)條件-50 (4)氛圍-51 (5)狀況-52 (6)氣氛-56 (7)價格-64 (8)能力-65 (9)信用-68 (10)標的-72 (11)來源-90 (12)壓力-118

Figure 4. An example of lexical chaining result.

The algorithm is adopted in this paper for the construction of lexical chains. We select nouns in the balanced corpus created by Academia Sinica (CKIP Group, 2010) with word frequency higher than a given threshold as candidate terms for lexical chaining. We apply the method proposed by Dai *et al*. (2008) to compute semantic similarity between words using E-HowNet instead of HowNet as the lexical database. The difference is that the primitives of function type are treated as descriptors. Let P and Q be two word senses and the number of modifying primitives of P is less than that of Q. The semantic similarity between P and Q is computed by Equation 1,

$$\begin{aligned} Sim(P,Q) &= \alpha \times Sim(P',Q') \\ &+ \beta \times \frac{\sum_{0 \le i < |P|} \max_{0 \le j < |Q|} (Sim(P_i, Q_j))}{|P|} \\ &+ \gamma \times \frac{|S \cap T|}{|S| + |T|} \end{aligned} \tag{1}$$

where P' and Q' are the primary primitives of P and Q, respectively, $|P|$ and $|Q|$ are the numbers of modifying primitives in their respective word senses, S and T are the sets of descriptors of frameworks of P and Q, respectively, $|S \cap T|$ is the number of common descriptors of S and T, $|S|$ and $|T|$ are the numbers of descriptors in S and T, and α, β, and γ are the relative weights of the three parts.

After constructing lexical chains, we derive five features from the lexical chains for each article. The five features are the number of lexical chains, the average length of lexical chains, the average span of lexical chains, the number of lexical chains with span longer than the half length of the article, and the average number of active chains per word. The features are normalized by dividing the article length. Table 1 shows the lexical chain features and their representing codes used in this paper.

Table 1. List of lexical chain features.

Code	Feature
lc-1	Number of lexical chains
lc-2	Average length of lexical chains
lc-3	Average span of lexical chains
lc-4	Number of long lexical chains
lc-5	Average number of active chains per word

3.4 SVM Classification

We apply support vector machines (SVM) as the modeling technique for our classification problem. The goal of an SVM, which is a vector-space-based large margin classifier, is to find

a decision surface that is maximally far away from any data point in the two classes. When data in the input space (X) cannot be linearly separated, we transform the data into a high-dimensional space called the feature space (*F*) using a function $\phi: X \rightarrow F$ so that the data are now linearly separable. Then in the feature space we find a linear decision function that best separates the data into two classes. An SVM toolkit, LIBSVM (Chang & Lin, n.d.), is used for building prediction models. When training the prediction model for each reading level, texts belonging to that reading level are used as positive data, while the rest of the texts are used as negative data. We follow the procedure suggested by Hsu, Chang, & Lin (2010) including the use of radial basis function kernel, scaling, and cross-validation.

3.5 Evaluation

In this paper, we use precision, recall, F-measure, and accuracy to evaluate the learned prediction models. For the test data, we use the same procedure for text processing and feature deriving. Correct prediction refers to the agreement between the predicted reading level and the original reading level. We compute the following quantities: true positive (*TP*) is the number of articles correctly classified as positive, false negative (*FN*) is the number of positive articles incorrectly classified as negative, true negative (*TN*) stands for the number of articles correctly classified as negative, and false positive (*FP*) refers to the number of negative articles incorrectly classified as positive. Precision, recall, F-measure, and accuracy are defined as follows.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F} - \text{measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \tag{4}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

We will test on different sets of features to find the best feature combination for training the prediction models.

4. Experiments

In this section we present our experiment setup and the results of the experiments on the textbooks corpus using different feature combinations.

4.1 Experiment Environment

The program modules for the experiments are written in Java programming language running on a PC with Microsoft Windows environment, Intel Core 2 Quad CPU, and 2GB of RAM. The corpus used as empirical data is stored in a Microsoft Access database. The lexicographic resources used for lexical semantic similarity computation in the experiments are stored as pure-text files in CSV format. LIBSVM is used for learning and testing SVM prediction models.

4.2 Empirical Data

The corpus used as empirical data consists of articles selected from the textbooks of elementary schools in Taiwan. We collect the digital versions of the textbooks of three subjects, Mandarin, Social Studies, and Life Science, for all of the six grade levels from publishers Nan I and Han Lin, resulting in a total number of 740 articles. Table 2 shows details of the collected data set.

Table 2. Summary of the textbooks corpus.

Reading Level	Grade Level	Mandarin	Social Studies	Life Science	No. of Articles
lower	1st grade	42	0	73	115
	2nd grade	56	0	55	111
middle	3rd grade	61	53	0	114
	4th grade	67	50	0	117
higher	5th grade	83	58	0	141
	6th grade	88	54	0	142
Total		397	215	128	740

4.3 Experiment Design

In each experiment, we use one set of features with a fixed parameter setting and target a certain grade level. We equally divide the corpus into five data sets to support 5-fold cross validation, and we present the average precision, recall, F-measure, and accuracy of the five folds.

Since the textbooks corpus does not contain articles beyond elementary school levels, we only build prediction models for lower grade and middle grade. For convenience, we denote feature sets by a string with special syntax. Feature types are indicated in the string by the abbreviation of that feature type. For example, “lc” refers to the lexical chain feature type and “tf” refers to the TF-IDF feature type. Options of a feature type are indicated in the string by a

dash followed by the code name for that option, attached to the end of the feature type indicator.

4.4 Experiments on Lexical Chain Features

To test the capability of lexical chain features on Chinese readability assessment, the lexical chain features listed in Table 1 are used and the results are shown in Table 3 and Table 4.

Table 3. Result of classifier for lower grade using lexical chain only.

Feature set	Precision	Recall	F-measure	Accuracy
lc-1-2-3-4-5	0.76	0.57	0.65	0.81

Table 4. Result of classifier for middle grade using lexical chain only.

Feature set	Precision	Recall	F-measure	Accuracy
lc-1-2-3-4-5	0.70	0.83	0.76	0.68

4.5 Comparison with TF-IDF Features

It is interesting to see whether incorporating a small number of TF-IDF features into lexical chain features can produce the same or even better results. We first use TF-IDF features generated from top 50 to top 500 terms to produce classifiers for lower grade. The precision, recall, F-measure, and accuracy of the classifiers using different number of TF-IDF features are shown in Table 5. Then, we add the five lexical chain features to the TF-IDF feature sets and repeat the same experiments. Their precision, recall, F-measure, and accuracy values are shown in Table 6. Figure 5 illustrates line graphs generated from F-measure values of the two tables, from which we find that the overall performance is improved for lower grade classifiers when using a combination of TF-IDF features and lexical chain features.

Table 5. Result of classifier for lower grade using TF-IDF features only.

Feature set	Precision	Recall	F-measure	Accuracy
tf-top50	0.78	0.87	0.82	0.88
tf-top100	0.81	0.86	0.83	0.89
tf-top200	0.80	0.89	0.84	0.90
tf-top300	0.82	0.89	0.85	0.90
tf-top400	0.86	0.89	0.87	0.92
tf-top500	0.84	0.89	0.87	0.92

Table 6. Result of classifier for lower grade using lexical chain and TF-IDF.

Feature set	Precision	Recall	F-measure	Accuracy
lc-1-2-3-4-5 + tf-top50	0.85	0.85	0.85	0.91
lc-1-2-3-4-5 + tf-top100	0.83	0.87	0.85	0.91
lc-1-2-3-4-5 + tf-top200	0.90	0.83	0.86	0.92
lc-1-2-3-4-5 + tf-top300	0.95	0.91	0.93	0.95
lc-1-2-3-4-5 + tf-top400	0.93	0.93	0.93	0.96
lc-1-2-3-4-5 + tf-top500	0.93	0.89	0.91	0.95

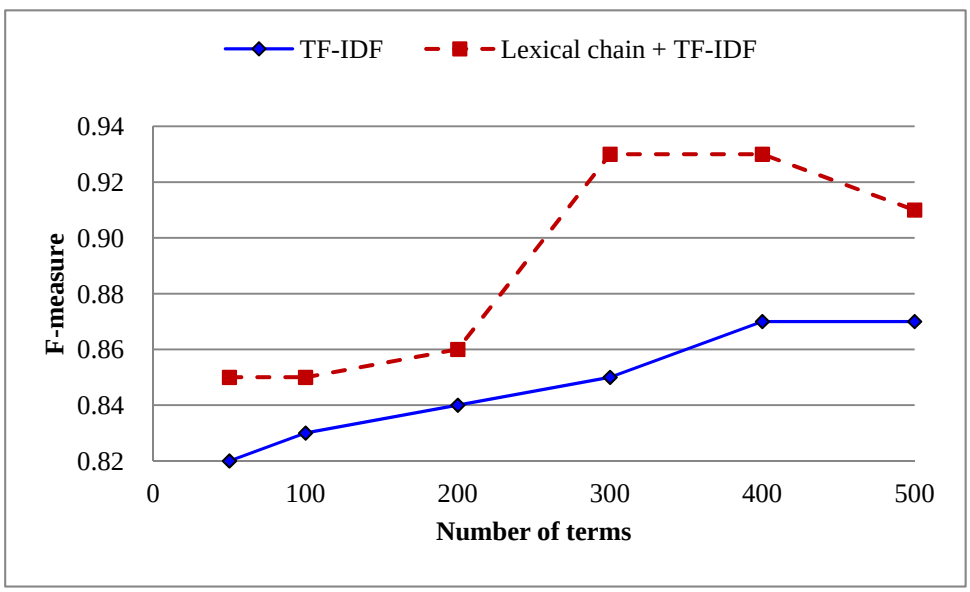

Figure 5. Result of classifier for lower grade.

The same set of experiments is conducted for the middle grade classifiers. Precision, recall, F-measure, and accuracy values of classifiers generated from TF-IDF features and the combination of TF-IDF and lexical chain features are shown in Table 7 and Table 8, respectively. The line graphs of F-measure values are shown in Figure 6, where the combined TF-IDF and lexical chain features generate the same or better F-measure in all cases. Therefore, incorporating a small number of TF-IDF features into lexical chain features is recommended for middle grade classifiers.

Table 7. Result of classifier for middle grade using TF-IDF features only.

Feature set	Precision	Recall	F-measure	Accuracy
tf-top50	0.81	0.88	0.84	0.79
tf-top100	0.81	0.90	0.85	0.81
tf-top200	0.83	0.92	0.87	0.83
tf-top300	0.86	0.90	0.88	0.84
tf-top400	0.82	0.92	0.87	0.83
tf-top500	0.82	0.95	0.88	0.84

Table 8. Result of classifier for middle grade using lexical chain and TF-IDF.

Feature set	Precision	Recall	F-measure	Accuracy
lc-1-2-3-4-5 + tf-top50	0.82	0.87	0.84	0.80
lc-1-2-3-4-5 + tf-top100	0.84	0.89	0.86	0.82
lc-1-2-3-4-5 + tf-top200	0.87	0.88	0.88	0.84
lc-1-2-3-4-5 + tf-top300	0.89	0.87	0.88	0.85
lc-1-2-3-4-5 + tf-top400	0.83	0.93	0.88	0.84
lc-1-2-3-4-5 + tf-top500	0.83	0.93	0.88	0.84

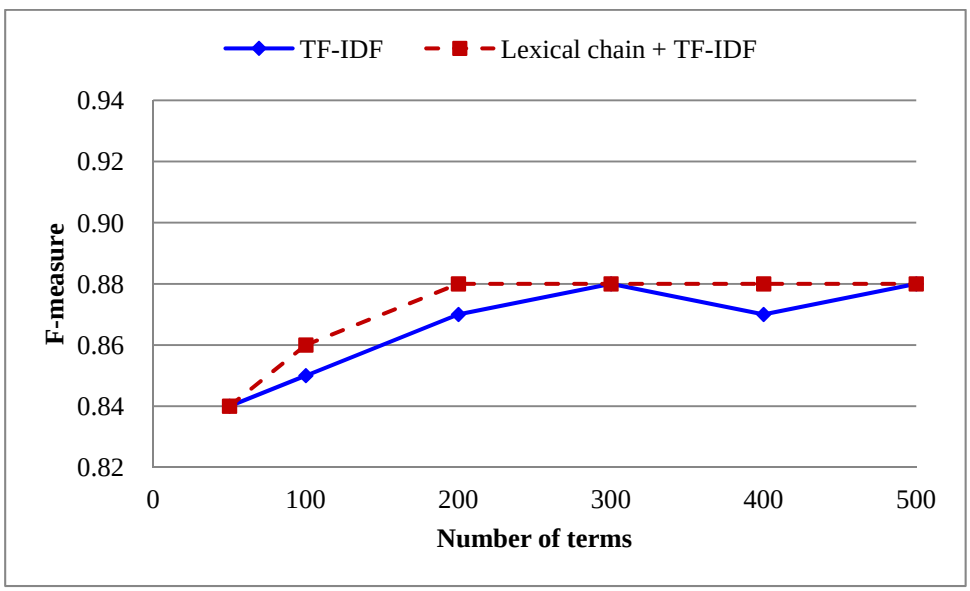

Figure 6. Result of classifier for middle grade.

5. Conclusions

This paper focuses on evaluating the effect of lexical cohesion analysis, more specifically, the effect of features based on lexical chains and term frequency, on the performance of readability assessment for Chinese text. The experiments produce satisfactory results on the textbooks corpus. Combining lexical chain and TF-IDF features usually produces better results, suggesting that both term frequency and lexical chain are useful features in Chinese readability assessment.

Future work can be done to have more articles annotated with reading levels or resort to other types of corpora where reading levels are inherent. On the other hand, lexical cohesion is only one of several aspects of text cohesion, and other aspects of text cohesion may also have some impact on the task of readability assessment. Several existing models of text cohesion, such as Coh-metrix and entity grid representation, try to model other aspect of text cohesion and have been extensively used in other natural language processing tasks such as writing quality assessment. Future work can be done to verify whether these models can benefit the task of readability assessment for Chinese text.

Acknowledgments

This research was supported in part by the National Science Council of Taiwan under Grant NSC99-2511-S-415-007-MY2.

Reference

Aluisio, S., Specia, L., Gasperin, C., & Scarton, C. (2010). Readability assessment for text simplification. In *NAACL-HLT 2010: The 5th Workshop on Innovative Use of NLP for Building Educational Applications*.

Barzilay, R., & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, *34*(1), 1-34.

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Cambridge, MA: Brookline Books.

Chang, C.-C., & Lin, C.-J. (n.d.). LIBSVM: A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chen, Y.-H., Tsai, Y.-H., & Chen, Y.-T. (2011). Chinese readability assessment using tf-idf and svm. In *International Conference on Machine Learning and Cybernetics (ICMLC2011)*, Guilin, China.

CKIP Group. (n.d.). A Chinese word segmentation system, http://ckipsvr.iis.sinica.edu.tw/

CKIP Group. (2009). *Lexical semantic representation and semantic composition - An introduction to E-HowNet*. (Technical Report), Institute of Information Science, Academia Sinica.

CKIP Group. (2010). *Academia Sinica Balanced Corpus (Version 3.1)*. Institute of Information Science, Academia Sinica.

Collins-Thompson, K., & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology (JASIST)*, *56*(13), 1448-1462.

Dai, L., Liu, B., Xia, Y., & Wu, S. (2008). Measuring semantic similarity between words using HowNet. In *International Conference on Computer Science and Information Technology 2008*, 601-605.

Dong, Z. (n.d.). HowNet knowledge database. http://www.keenage.com/

Duan, K.-B., & Keerthi, S. S. (2005). Which is the best multiclass SVM method? An empirical study. In *Proceedings of the Sixth International Workshop on Multiple Classifier Systems*.

Eickhoff, C., Serdyukov, P., & de Vries, A. P. (2010). Web page classification on child suitability. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database and some of its applications*. Cambridge, MA: MIT Press.

Feng, L., Elhadad, N., & Huenerfauth, M. (2009). Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, 229-237.

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A comparison of features for automatic readability assessment. In *The 23rd International Conference on Computational Linguistics (COLING 2010): Poster Volume*, 276-284.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Hasan, R. (1984). Coherence and cohesive harmony. In J. Flood (Ed.), *Understanding reading comprehension: Cognition, language and the structure of prose* (pp. 184-219), Newark, DE: International Reading Association.

Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the 3rd ACL Workshop on Innovative Use of NLP for Building Educational Applications*, 71-79.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2010). *A practical guide to support vector classification*. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Hsu, C.-W. & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, *13*(2), 415-425.

Kincaid, J. P., Fishburne R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated readability index, Fog count and Flesch reading ease formula) for Navy enlisted personnel*. Research Branch Report 8-75. Chief of Naval Technical Training: Naval Air Station Memphis.

Lau, T. P. (2006). *Chinese readability analysis and its applications on the Internet*. (Master Thesis), Computer Science and Engineering, The Chinese University of Hong Kong.

Lin, S.-Y., Su, C.-C., Lai, Y.-D., Yang, L.-C., & Hsieh, S.-K. (2009). Assessing text readability using hierarchical lexical relations retrieved from WordNet. *Computational Linguistics and Chinese Language Processing*, *14*(1), 45-84.

McLaughlin, G. H. (1969). SMOG grading - A new readability formula. *Journal of Reading*, *12*(8), 639-646.

Miltsakaki, E., & Troutt, A. (2008). Real time Web text classification and analysis of reading difficulty. In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications*.

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, *17*(1), 21-48.

National Center for Education Statistics. (2002). *Adult literacy in America* (3rd ed.). Washington, D. C.: U.S. Dept. of Education.

Petersen, S. E., & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech and Language*, *23*, 89-106.

Pitler, E., & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 186-195.

Schwarm, S. E., & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the ACL*, 523-530.

Stenner, A. J. (1996). Measuring reading comprehension with the Lexile framework. In *Fourth North American Conference on Adolescent/Adult Literacy*.

Stokes, N. (2004). *Applications of lexical cohesion analysis in the topic detection and tracking domain*. (Ph.D. Thesis), Department of Computer Science, National University of Ireland, Dublin.

Sung, Y.-T., Chang, T.H., Chen, J.-L., Cha, J.-H., Huang, C.-H., Hu, M.-K., & Hsu, F.-Y. (2011). The construction of Chinese readability index explorer and the analysis of text readability. In *21th Annual Meeting of Society for Text and Discourse Process*, Poitiers, France.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 412-420.