# GRAW+: A Two-View Graph Propagation Method With Word Coupling for Readability Assessment

**Zhiwei Jiang**
*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China.*
*E-mail: jiangzhiwei@outlook.com*

**Qing Gu**
*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China.*
*E-mail: guq@nju.edu.cn*

**Yafeng Yin**
*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China.*
*E-mail: yafeng@nju.edu.cn*

**Jianxiang Wang**
*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China.*
*E-mail: wjxnju@outlook.com*

**Daoxu Chen**
*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China.*
*E-mail: cdx@nju.edu.cn*

Existing methods for readability assessment usually construct inductive classification models to assess the readability of singular text documents based on extracted features, which have been demonstrated to be effective. However, they rarely make use of the interrelationship among documents on readability, which can help increase the accuracy of readability assessment. In this article, we adopt a graph-based classification method to model and utilize the relationship among documents using the coupled bag-of-words model. We propose a word coupling method to build the coupled bag-of-words model by estimating the correlation between words on reading difficulty. In addition, we propose a two-view graph propagation method to make use of both the coupled bag-of-words model and the linguistic features. Our method employs a graph merging operation to combine graphs built according to different views, and improves the label propagation by incorporating the ordinal relation among reading levels. Experiments were conducted on both English and Chinese data sets, and the results demonstrate both effectiveness and potential of the method.

Received March 26, 2017; revised June 20, 2018; accepted July 14, 2018

 DOI: 10.1002/asi.24123

## Introduction

Readability assessment evaluates the reading difficulties of text documents, which are normally represented as discrete reading levels. Automatic readability assessment is a challenging task, which has attracted researchers' attention from the beginning of the last century (Collins-Thompson, 2014). Traditionally, it can be used by educationists to choose appropriate reading materials for students of different education or grade levels. In modern times, it can be used by web search engines to do personalized searches based on web users' educational backgrounds.

Existing methods for readability assessment usually concentrate on feature engineering and then applying inductive classification models to utilize the features. In the early stages, researchers proposed readability formulas to measure the readability of texts (Zakaluk & Samuels, 1988). These formulas are usually attained by linear regression on several easy-to-compute text features relevant to reading difficulty. Recently, by employing machine-learning techniques, classification-based methods have been proposed and demonstrated to be more effective than readability formulas (Benjamin, 2012; Collins-Thompson, 2014). These methods

combine the rich representation of texts with sophisticated prediction models.

Most current methods assess the readability of text documents singularly, and ignore the interrelationship among documents on readability, which can be useful in assessing the readability of documents based on the labeled ones. For example, two documents can be of the same reading level, if they consist of words that have similar reading difficulty. Hence, we propose a graph propagation method for readability assessment, which can model and utilize the interrelationship among text documents.

To measure the relationship among documents, we use the bag-of-words (BoW) model, which is commonly used for text classification and clustering (Huang, 2008; Sebastiani, 2002). However, to measure the relationship on readability, the basic BoW model requires improvements, since it ignores the fact that different words may have similar reading difficulties. Figure 1 illustrates the improved use of the BoW model for readability assessment using a simple example. In Figure 1, the left matrix is built from the basic BoW model for three documents (that is, D1, D2, and D3) consisting of four tokens (that is, school, law, syllabus, and decree). Among the three documents, D1 and D2 are two relatively difficult documents both containing two easy words (school or law) and two difficult words (syllabus or decree), while D3 is an easy document that contains two easy words (school). By calculating the cosine similarities based on the basic BoW model (the bottom left subfigure), the result shows that D1 is more similar to D3 than to D2, which is inconsistent with their similarities on reading difficulty.

To overcome the shortcoming of the basic BoW model, we designed a word coupling method. As shown in Figure 1, the word coupling method first measures the similarities among words on reading difficulties (the word coupling matrix). Then the method makes the words of similar difficulties (for example, school and law) share their occurrence frequencies with each other (by matrix multiplication), which leads to the coupled BoW model (the coupled BoW matrix). In this way, the documents will be similar on readability if their words have similar distributions on reading difficulties.

To build the coupled BoW model, the key point is to construct the word coupling matrix. For this purpose, we first estimate the occurrence distributions of words in sentences of different reading difficulties, and then compute their similarities on reading difficulty based on the distributions.

Besides the coupled BoW model, the linguistic features can also be adopted by our method. On the one hand, we use the linguistic features as complementation of the coupled BoW model to construct graphs from multiple views. On the other, the linguistic features are used to reinforce the label propagation algorithm by providing the prior knowledge.

In this article, we propose a two-view graph propagation method with word coupling for readability assessment. Our contributions are as follows (a preliminary version of this work appeared in Jiang, Sun, Gu, Bai, & Chen, 2015). (i) We apply the graph-based method for readability assessment, which can make use of the interrelationship among documents to estimate their readability. (ii) We propose the coupled BoW model, which can be used to measure the similarity of documents on reading difficulty. (iii) We propose a two-view graph building strategy to make use of both the coupled BoW model and the linguistic features. (iv) We propose a reinforced label propagation algorithm, which can make use of the ordinal relation among reading levels. Extensive experiments were carried out on data sets of both English and Chinese. Compared with the state-of-art methods, the results demonstrate both effectiveness and the potential of our method.

## Background and Related Work

### *Readability Assessment*

Research on automatic readability assessment has spanned the last 70 years (Benjamin, 2012). Early research mainly focused on the designing of readability formulas (Zakaluk & Samuels, 1988). Many well-known readability formulas have been developed, such as the SMOG formula

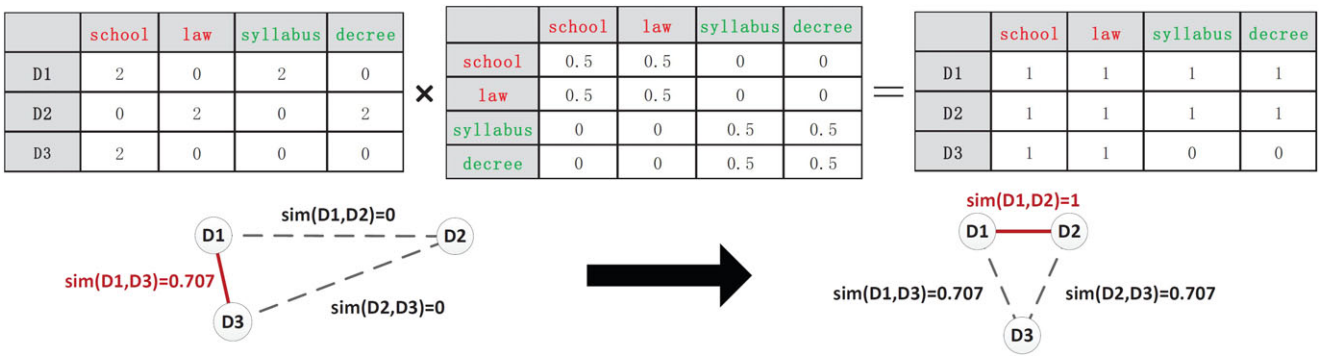



FIG. 1. A motivation example of the word coupling method. The left matrix is a basic BoW matrix. The central matrix is a word coupling matrix. The right matrix is the coupled BoW matrix. [Color figure can be viewed at wileyonlinelibrary.com]


DOI: 10.1002/asi

(McLaughlin, 1969) and the FK formula (Kincaid, Fishburne, Rogers, & Chissom, 1975). A key observation in these studies is that the vocabulary used in a document usually determines its readability (Pitler & Nenkova, 2008). A general way of using the vocabulary is the statistical language model (Collins-Thompson & Callan, 2004; Kidwell, Lebanon, & Collins-Thompson, 2009). More recently, researchers have explored complex linguistic features combined with classification models to obtain robust and effective readability prediction methods (Denning, Pera, & Ng, 2016; Feng, Jansche, Huenerfauth, & Elhadad, 2010; Schwarm & Ostendorf, 2005). While most studies are conducted for English, there are studies for other languages, such as French (François & Fairon, 2012), German (Hancke, Vajjala, & Meurers, 2012), and Bangla (Sinha, Dasgupta, & Basu, 2014). In addition, researchers have used the representation learning techniques for readability assessment (Cha, Gwon, & Kung, 2017; Tseng, Hung, Sung, & Chen, 2016).

### *The Bag-of-Words Model*

The BoW model has been widely used for document classification owing to its simplicity and general applicability. It constructs a feature space that contains all the distinct words of a language (or text corpus). Traditionally, it assumes that words are independent, while recently, capturing the word coupling relationship has attracted much attention (Cao, 2015). Billhardt, Borrajo, and Maojo (2002) studied the coupling relationship based on the co-occurrence of words in the same documents. Kalogeratos and Likas (2012) generalized the relationship by taking into account the distance among words in the sentences of each document. Cheng, Miao, Wang, and Cao (2013) estimated the co-occurrence of words by mining the transitive properties. Inspired by these studies, this article modifies the BoW model for readability assessment, and provides the coupled BoW model incorporating the co-occurrence of words in different levels of reading difficulty.

### *Graph-Based Label Propagation*

Graph-based label propagation is applied on a graph to propagate class labels from labeled nodes to unlabeled ones (Subramanya, Petrov, & Pereira, 2010). It has been successfully applied in various applications, such as dictionary construction (Kim, Verma, & Yeh, 2013), word segmentation and tagging (Zeng, Wong, Chao, & Trancoso, 2013), and sentiment classification (Ponomareva & Thelwall, 2012). Typically, a graph-based label propagation method consists of two main steps: graph construction and label propagation. During the first step, some forms of edge weight estimation and edge pruning are required to build an efficient graph (Jebara, Wang, & Chang, 2009; Ponomareva & Thelwall, 2012). In addition, nodes in the graph can be heterogeneous; for example, both instance nodes and class nodes can coexist in a birelational graph (Jiang, 2011), and nodes (words) of English can be linked to nodes (words) of the target language (Chinese) in a bilingual graph (Gao, Wei, Li, Liu, & Zhou, 2015). During the second step, propagation algorithms are required to propagate the label distributions to all the nodes (Kim et al., 2013; Subramanya et al., 2010) so that the classes of unlabeled nodes can be predicted.

## The Proposed Method

In this section, we first present the overview of the proposed method (GRAW+). Then we describe two main parts of the method in detail: feature representation and readability classification.

### *An Overview of the Method*

The framework of GRAW+ is depicted in Figure 2. GRAW+ takes an auxiliary text corpus and a target document set as inputs. The auxiliary text corpus contains unlabeled sentences used to construct the word coupling matrix. The target document set contains both labeled and unlabeled documents on readability. The objective of GRAW+ is to predict the reading levels of the unlabeled

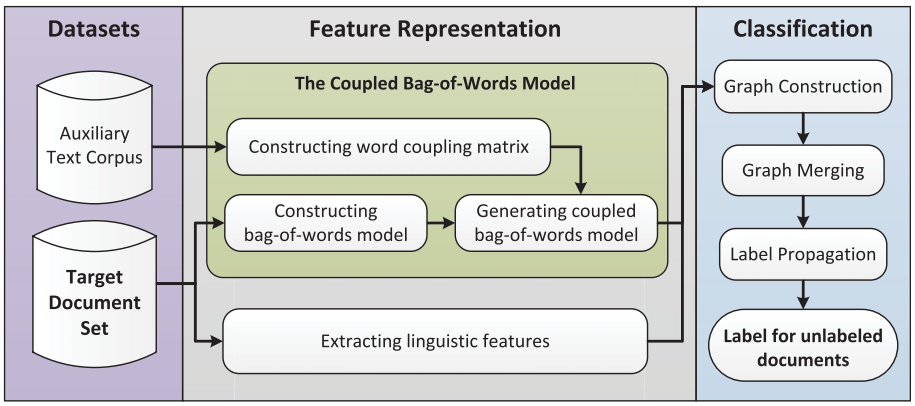



FIG. 2. The framework of GRAW+. [Color figure can be viewed at wileyonlinelibrary.com]


DOI: 10.1002/asi

documents in the target set based on the labeled ones. GRAW+ includes two main stages: feature representation and readability classification.

During the first stage, the documents in the data set are mapped into feature vectors from two distinct views: the cBoW (coupled BoW) view and the linguistic view. From the cBoW view, the coupled BoW model is required, and from the linguistic view, suitable linguistic features can be borrowed from previous studies. By representing documents from these two views, both word-level difficulty distribution and document-level readability features can be measured, and hence provide extensive information for readability assessment.

During the second stage, a two-view graph propagation method is proposed for readability classification, which consists of three main steps: graph construction, graph merging, and label propagation. In the first step, both labeled and unlabeled documents are used to build graphs, where each document is represented by a node, and their similarities on reading difficulty are represented by edge weights. In the second step, the intra-view merging operation is used to merge the homogeneous graphs within the same view, and the inter-view merging operation is used to merge the heterogeneous graphs from different views. In the third step, a label propagation algorithm is designed on the merged graph.

### *The Coupled Bag-of-Words Model*

To build the cBoW model, first we construct the word coupling matrix. Then we transform the basic BoW model to the coupled BoW model using the word coupling matrix.

*Constructing the word coupling matrix.* As stated before, the word coupling matrix is constructed to represent the similarities of word pairs on reading difficulty. For this purpose, we assume the simple fact that easy words tend to appear in easy sentences, while difficult words tend to appear in difficult sentences. Hence, we can estimate the reading difficulty of a word by its distributions of occurrence probabilities in sentences from different difficulty levels. The difficulty distributions can then be used to estimate the similarities among words on reading difficulty by computing the distribution divergence. Since sentences with labeled difficulty levels are hard to acquire, we use unlabeled sentences instead, and label the sentences by estimating their difficulty levels with heuristic functions.

Figure 3 presents the three steps required to construct the word coupling matrix: per-sentence reading difficulty estimation, per-word difficulty distribution estimation, and word coupling matrix construction. In the first step, each sentence in the text corpus is assigned a weak label, which is a reading score computed in a heuristic function. In the second step, based on the weak labels, the difficulty distribution of each word is estimated, according to their distributions of occurrences in these sentences. In the final step, the similarities among words are calculated using the distribution divergence.

**Step 1: Per-sentence reading difficulty estimation.** The precise estimation of sentence-level readability is a hard problem and has recently attracted the attention of many researchers (Pilán, Volodina, & Johansson, 2014; Schumacher, Eskenazi, Frishkoff, & Collins-Thompson, 2016; Vajjala & Meurers, 2014). For efficiency, we use heuristic functions to make a rough estimation. Specifically, we consider the linguistic features designed for readability assessment that have been demonstrated to be effective in previous studies (Feng et al., 2010; Schumacher et al., 2016), and choose the most used linguistic features that can be operated at the sentence level to build the heuristic functions. In total, eight heuristic functions $h \in \{len, ans, anc, lv, art, ntr, pth, anp\}$ corresponding to eight distinct features from three aspects are used to compute the reading score of a sentence, as shown in Table 1.

Let $S$ denote the set of all the sentences ready for constructing the word coupling matrix. Given a sentence $s \in S$, its reading difficulty can be quantified as a reading score $r(s) = h(s)$ by using one of the eight functions. The more difficult $s$ is, the greater $r(s)$ will be.

Considering that the reading score $r(s)$ may be continuous, we discretize $r(s)$ into several difficulty levels. Let $\eta$ denote the predetermined number of difficulty levels, $r^h_{max}$ and $r^h_{min}$, respectively, denote the maximum and minimum

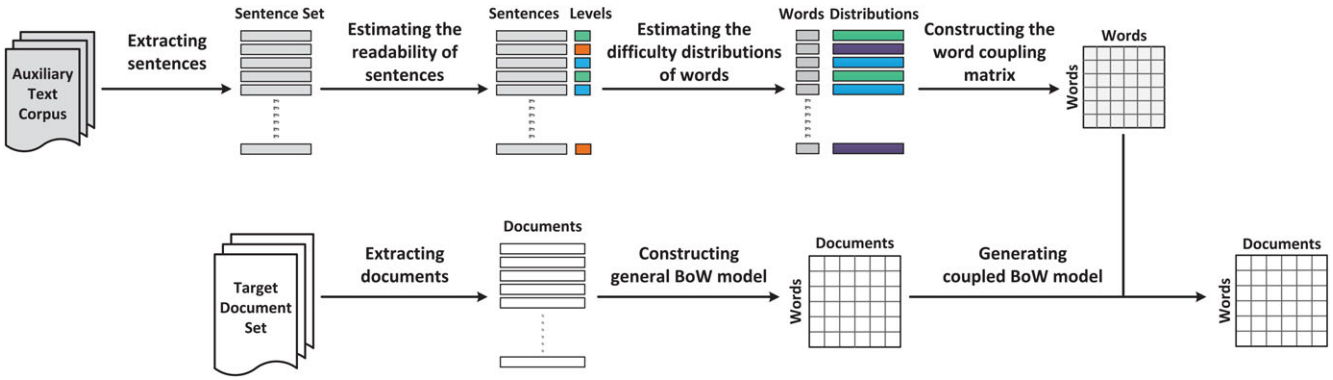



FIG. 3. The process of representing documents using the coupled bag-of-words model. [Color figure can be viewed at wileyonlinelibrary.com]


DOI: 10.1002/asi

TABLE 1. Three aspects of estimating reading difficulty of sentences using heuristic functions.

| Aspect | Function | Description |
|---|---|---|
| Surface | len(s) | the length of the sentence s. |
| | ans(s) | the average number of syllables (or strokes for Chinese) per word (or character for Chinese) in s. |
| | anc(s) | the average number of characters per word in s. |
| Lexical | lv(s) | the number of distinct types of POS, that is, part of speech, in s. |
| | atr(s) | the ratio of adjectives in s. |
| | ntr(s) | the ratio of nouns in s. |
| Syntactic | pth(s) | the height of the syntax parser tree of s. |
| | anp(s) | the average number of (noun, verb, and preposition) phrases in s. |

reading score of all the sentences in $S$, where $h$ refers to one of the eight functions. To determine the difficulty level $l^h(s)$ ($l^h(s) \in [1, \eta]$) of a sentence $s$, the range $\left[r^h_{min}, r^h_{max}\right]$ is evenly divided into $\eta$ intervals. $l^h(s)$ will be $i$, if the reading score $r(s)$ resides in the $i$-th interval. For each of the three aspects, we compute one $l^*(s)$ for a sentence $s$ by combining the heuristic functions using the following equations.

$$l^{\text{sur}}(s) = max\left(l^{len}(s), l^{ans}(s), l^{anc}(s)\right)$$
$$l^{\text{lex}}(s) = max\left(l^{lv}(s), l^{atr}(s), l^{ntr}(s)\right) \quad (1)$$
$$l^{\text{syn}}(s) = max\left(l^{pth}(s), l^{anp}(s)\right)$$

**Step 2: Per-word difficulty distribution estimation.** The difficulty distribution of each word is computed based on the sentence-level reading difficulty. Since each sentence contains many words and each word may appear in many sentences, we estimate the difficulty distributions of words according to their distributions of occurrences in sentences.

Let $\mathcal{V}$ denote the set of all the words appearing in $S$, $p_{\text{t}}$ denotes the difficulty distribution of a word (term) $t \in \mathcal{V}$. $p_t$ is a vector containing $\eta$ (that is, the number of difficulty levels) values, the $i$-th part of which can be calculated by Equation 2.

$$p_{\text{t}}(i) = \frac{1}{n_t} \cdot \sum_{s \in S} \delta(t \in s) \cdot \delta(l(s) = i) \quad (2)$$

where $n_t$ refers to the number of sentences containing $t$. The indicator function $\delta(x)$ returns 1 if $x$ is true and 0 otherwise.

**Step 3: Word coupling matrix construction.** Given the set of words $\mathcal{V}$, a word coupling matrix is defined as $C \in R^{|\mathcal{V}| \times |\mathcal{V}|}$, the element of which reflects the correlation between two words (that is, terms). The correlation between each pair of words can be computed according to the similarity measure of their difficulty distributions.

Given two words (terms) $t_1$ and $t_2$, whose difficulty distributions are $p_{t_1}$ and $p_{t_2}$, respectively, we use a symmetric version of Kullback–Leibler divergence (Kullback & Leibler, 1951) to measure their distribution difference, which averages the values of the divergence computed from both directions. The equation is:

$$c_{KL}(t_1, t_2) = \frac{1}{2}\left(KL(p_{t_1} \| p_{t_2}) + KL(p_{t_2} \| p_{t_1})\right) \quad (3)$$

where $KL(p \,||\, q) = \sum_i p(i) \log \frac{p(i)}{q(i)}$ and $i$ is the element index. After that, the logistic function is applied to get the normalized distribution similarity, that is:

$$sim(t_1, t_2) = \frac{2}{1 + e^{c_{KL}(t_1, t_2)}} \quad (4)$$

Given a word $t_i$, only $\lambda$ other words with highest correlation (similarity) are selected to build the neighbor set of $t_i$, denoted as $\mathcal{N}(t_i)$. If a word $t_j$ is not selected (that is, $t_j \notin \mathcal{N}(t_i)$), the corresponding $sim(t_i, t_j)$ will be assigned 0. After that, the word coupling matrix ($C^*$) with $sim(t_i, t_j)$, as its elements are normalized along the rows so that the sum of each row is 1. Based on three different $l^*(s)$, we construct three distinct word coupling matrices $C^{sur}$, $C^{lex}$, and $C^{syn}$.

While a large volume of vocabulary will make the construction of the word coupling matrix time-consuming, we provide a strategy to filter out less informative words based on their distributions on reading difficulty. The filtering measure is the entropy of the words, which can be calculated by Equation 5. By sorting the words ascendingly according to entropy, the last $\alpha \in [0, 1]$ proportion will be filtered out.

$$Ent(t) = H(p_t) = -\sum_{i=1}^{\eta} p_t(i) \log p_t(i) \quad (5)$$

*Generating the Coupled Bag-Of-Words Model.* In the basic BoW model, words are treated as being independent of each other, and the corresponding BoW matrix is sparse and ignores the similarity among words on reading difficulty. For readability assessment, the coupled BoW model can be implemented by multiplying the word coupling matrix and the basic BoW matrix, and the resulting coupled BoW matrix will be dense and focus on similarities on reading difficulty.


DOI: 10.1002/asi

One of the popular schemes of the BoW model is TF-IDF (Term Frequency and Inverse Document Frequency). Given the set of documents $\mathcal{D}$ and the set of words $\mathcal{V}$, the TF-IDF matrix is defined as $M^{bow} \in R^{|\mathcal{V}|\times|\mathcal{D}|}$, which can be calculated based on the logarithmically scaled term (that is, word) frequency (Salton & Buckley, 1988) as follows:

$$M^{bow}_{t,d} = tf_{t,d} \cdot idf_{t,d} = (1 + \log f(t,d)) \cdot \log\left(1 + \frac{|\mathcal{D}|}{|\{d|t \in d\}|}\right) \tag{6}$$

where $f(t, d)$ is the number of times that a term (word) $t \in \mathcal{V}$ occurs in a document $d \in \mathcal{D}$.

By adopting the TF-IDF matrix $M^{bow}$ from the basic BoW model, the coupled TF-IDF matrix $M^*$ can be generated by the following equation:

$$M^* = C^* \cdot M^{bow} \tag{7}$$

Technically, three coupled TF-IDF matrices $M^{sur}$, $M^{lex}$, and $M^{syn}$ can be built according to the three word coupling matrices $C^*$, developed in the previous section.

### The Linguistic Features

The readability of documents is influenced by many factors, such as vocabulary, composition, syntax, semantics, and so on. Since vocabulary factors have been incorporated in the cBoW model, we integrate the other three factors into the linguistic view as complementation of the cBoW view. From the linguistic view, we build $M^l \in R^{n_l\times|\mathcal{D}|}$ ($n_l$ refers to the number of features) for the documents in $\mathcal{D}$. Based on the recent work on document-level readability assessment (Feng et al., 2010; Jiang, Sun, Gu, & Chen, 2014; Vajjala & Meurers, 2012), we select three groups of linguistic features: surface features, lexical features, and syntactic features, which are described as follows. Since our proposed method aims for language-independency, we select mostly the language-independent features and add some popular language-dependent features adapted from English to Chinese.

*Surface Features.* Surface features are the kind of features that can be directly acquired by counting the grammatical units in a document, such as the average number of characters per word, syllables per word, and words per sentence (Vajjala & Meurers, 2012). We adopt them in our method and add extra features used in Jiang et al. (2014). The extra features include the average number of polysyllabic words (for example, the number of syllables is greater than 3) per sentence, the average number of difficult words (for example, the number of characters is greater than 10) per sentence, the ratio of distinct words (that is, without repetition), and the ratio of unique words.

*Lexical Features.* Lexical features are relevant to the lexical types (for example, part of speech), which can be acquired by lexical analysis. Vajjala and Meurers (2012) employed lexical richness measures for readability assessment, which include 15 SLA (Second Language Acquisition) measures and two extra designed measures. Jiang et al. (2014) developed counterparts of all the measures for Chinese documents. We adopt the two sets of lexical features for both English and Chinese. In addition, we also add the five features proposed by Feng et al. (2010) to compensate for the missed lexical types.

*Syntactic Features.* Syntactic features are features that are computed based on the syntactic structures of sentences that may require the parse tree analysis. Following Vajjala and Meurers (2012), we adopt features computed on units of three levels: sentence, clause, and T-unit. In addition, we add the features designed in Jiang et al. (2014), which count the relative ratios of different types of parse tree nodes and phrases (that is, subtrees). Examples include the average number of noun phrases per sentence, the average number of parse tree nodes per words, and the ratio of the extra high tree.

### Two-View Graph Propagation

Based on the cBoW and linguistic views, we propose a two-view graph propagation method for readability classification. While the general graph-based label propagation (Zhu & Ghahramani, 2002) contains two steps (that is, graph construction and label propagation), our method adds an extra graph merging step to make use of multiple graphs. In addition, since grade levels are in ordinal scale, we further propose a reinforced label propagation algorithm.

*Graph Construction.* Given a feature representation $X \in \{M^{sur}, M^{lex}, M^{syn}, M^l\}$, we can build a directed graph $G^*$ to represent the interrelationship on readability among the documents, where the node set $\mathcal{D}$ contains all the documents. Given the similarity function, we link node $d_i \in \mathcal{D}$ to $d_j \in \mathcal{D}$ with an edge of weight $G^*_{i,j}$, defined as:

$$G^*_{i,j} = \begin{cases} sim(d_i, d_j) & \text{if } d_j \in \mathcal{K}(d_i) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $\mathcal{K}(d_i)$ is the set of $k$-nearest neighbors of $d_i$ with top-$k$ similarities. The similarity function $sim(d_i, d_j)$ can be defined by the Euclidean distance as follows:

$$sim(d_i, d_j) = \frac{1}{\sqrt{\sum_{v=1}^{n}\left(X_{v,i} - X_{v,j}\right)^2} + \epsilon} \tag{9}$$

where $\epsilon$ is a small constant to avoid zero denominators.


DOI: 10.1002/asi

*Graph Merging.* By constructing graphs from two views, we get four graphs, where three graphs correspond to the three coupled TF-IDF matrices (denoted $G^{sur}$, $G^{lex}$, and $G^{syn}$, respectively), and one graph corresponds to the linguistic features (denoted $G^{l}$). The former three graphs are homogeneous, since they share the same view, while the last graph is different. For the three homogeneous graphs, we provide an intra-view homogeneous graph merging strategy to merge them into one (denoted $G^{c}$). To combine the graphs from both views, we provide an inter-view heterogeneous graph merging strategy to merge graph $G^{c}$ and graph $G^{l}$ into the final graph $G^{cl}$.

**Intra-view homogeneous graph merging.** Given the three graphs (that is, $G^{sur}$, $G^{lex}$ and $G^{syn}$), where each node has $k$ neighbors, we merge them into the graph $G^{c}$ where each node still has $k$ neighbors. The basic idea is keeping the common edges while removing edges containing redundant information, as shown in Figure 4. Given a node $v$, firstly we reserve its neighbors which are common in all three graphs. Secondly, for the candidate nodes, which are neighbors of $v$ in at least one graph, we select one by one the node which possesses the least number of common neighbors (that is, the nodes that are already selected in $\mathcal{K}^{c}(v)$). The objective is to keep the number of triangles in $G^{c}$ to a minimum. The edge weights of $G^{c}$ are averaged on the corresponding edges appeared in the three graphs.

**Inter-view heterogeneous graph merging.** Considering that edges of either $G^{c}$ or $G^{l}$ describe the relationship of documents on readability from a certain perspective, we reserve all the edges (that is, some nodes will have $>k$ neighbors) and use the factor $\beta$ to balance their weights on the final graph $G^{cl}$. The following equation defines the strategy:

$$G^{cl}_{i,j} = \beta G^{c}_{i,j} + (1-\beta) G^{l}_{i,j} \tag{10}$$

The parameter $\beta \in [0, 1]$ is used to specify the relative importance of the two graphs. Clearly, the case of $\beta = 0$ or 1 reduces the graph to a single view.

*Label propagation.* Given a graph $G$ constructed in previous sections, its nodes are divided into two sets: the labeled set $\mathcal{D}_l = \{d_1, d_2, \cdots, d_x\}$ and the unlabeled set $\mathcal{D}_u = \{d_{x+1}, d_{x+2}, \cdots, d_n\}$. The goal of label propagation is to propagate class labels from the labeled nodes (that is, documents with known reading levels) to the entire graph. We use a simplified label propagation algorithm presented in Subramanya et al. (2010), which has been proved effective (Kim et al., 2013). The algorithm iteratively updates the label distribution on a document node using the following equation:

$$q^{(i)}_d(y) = \frac{1}{\kappa_d}\left( q^{0}_d(y)\delta(d \in \mathcal{D}_l) + \sum_{v \in \mathcal{K}(d)} G_{d,v} q^{(i-1)}_v(y) \right) \tag{11}$$

At the left side of Equation 11, $q^{(i)}_d(y)$ is the afterward probability of $y$ (that is, the reading level $y$) on a node $d$ at the $i$-th iteration. At the right side, $\kappa_d$ is the normalizing constant to make sure the sum of the level probabilities is 1, and $q^{0}_d(y)$ is the initial probability of $y$ on $d$ if $d$ is initially labeled (1 if $d$ is labeled as level $y$ or 0 otherwise). $\delta(x)$ is the indicator function. $\mathcal{K}(d)$ denotes the set of neighbors of $d$. The iteration stops when the changes in $q^{(i)}_d(y)$ for all the nodes and reading levels are small enough (for example, less than $e^{-3}$), or $i$ exceeds a predefined number (for example, greater than 30). For each unlabeled document $\mathrm{d_i}$, its predicted reading level is y if the element $\mathrm{q_{d_i}(y)}$ is greatest in the latest label distribution $\mathrm{q_{d_i}}$.

**Reinforced label propagation.** The above label propagation algorithm (Subramanya et al., 2010) treats the class labels as in nominal scale, and ignores the ordinal relation among reading levels. We develop a reinforced label propagation algorithm to utilize the ordinal relationship. Preclassification is required using the linguistic features, which can provide extra information to amplify the edge weights.

Let documents have $m$ reading levels in ordinal scale. Given the $x$ labeled documents with reading levels $\{y_1, \cdots, y_x\}$, a preclassifier can be trained and the prelabels can then be predicted for the unlabeled documents, denoted as $\{\hat{y}_{x+1}, \cdots, \hat{y}_n\}$. Thus, we have the a priori label of all the documents $\{z_1, \cdots, z_x, z_{x+1}, \cdots, z_n\}$, where $z_i = y_i$ ($i \in \{1, \cdots, x\}$) for labeled documents and $z_j = \hat{y}_j$ ($j \in \{x+1, \cdots, n\}$) for unlabeled documents.

Based on the a priori labels, we amplify the weights of the edges which connect two nodes of similar a priori labels. The reinforced label propagation iteratively updates the label distribution on a document node using the following enhanced equation:

$$q^{(i)}_d(y) = \frac{1}{\kappa_d}\left( q^{0}_d(y)\delta(d \in \mathcal{D}_l) + \sum_{v \in \mathcal{K}(d)} G'_{d,v} q^{(i-1)}_v(y) \right)$$

$$G'_{d,v} = (m - |z_d - z_v|) G_{d,v} \tag{12}$$

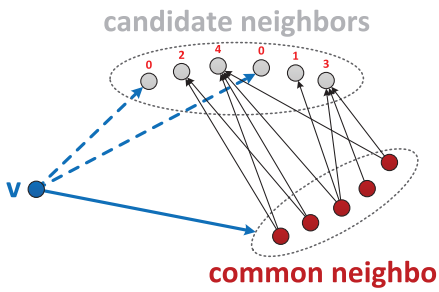



FIG. 4. Illustration of the intra-view homogeneous graph merging strategy. [Color figure can be viewed at wileyonlinelibrary.com]


DOI: 10.1002/asi

where $G'$ is a weighted graph the edge weights of which are modified from $G_{d,\ v}$ according to the difference in current reading levels of their end nodes. For each unlabeled document $d_i$, its final reading level is y if $q_{d_i}(y)$ is greatest in the final label distribution $q_{d_i}$.

## Experiments

In this section, we conduct experiments on data sets of both English and Chinese to investigate the following four research questions:

**RQ1:** Whether the proposed method (that is, GRAW+) outperforms the state-of-the-art methods for readability assessment?

**RQ2:** What are the effects of the word coupling matrix on the performance of the coupled bag-of-words model?

**RQ3:** How effective is the two-view graph propagation method, including the graph merging strategies and reinforced label propagation algorithm?

**RQ4:** Whether introducing the external text corpus can improve the quality of the word coupling matrix?

### *Corpus and Performance Measures*

To evaluate our proposed method, we used two data sets. The first is CPT (Chinese primary textbook) (Jiang et al., 2014), which contains Chinese documents of six reading levels corresponding to six grades. The second is ENCT (English New Concept textbook) (Jiang et al., 2015), which contains English documents of four reading levels. Both data sets (shown in Table 2) are built from well-known textbooks where documents are organized into grades by credible educationists. For the documents in CPT, we use the ICTCLAS tool (Zhang, 2013; Zhang, Yu, Xiong, & Liu, 2003) to do the word segmentation.

We conducted experiments on both CPT and ENCT using the hold-out validation, which randomly divides a data set into labeled (training) and unlabeled (test) sets by stratified sampling. The labeling proportion is varied to investigate the performance of a method under different circumstances. To reduce variability, under each case, 100 rounds of hold-out validation are performed, and the validation results are averaged over all the rounds. To tune the hyperparameters, we randomly choose one partition from the training set as the development set. We chose the precision (P), recall (R), and F1-measure (F1) as the performance measures.

### *Comparisons to the State-of-the-Art Methods*

To address RQ1, we implement the following readability assessment methods and compare our method GRAW+ with them:

- SMOG (McLaughlin, 1969) and FK (Kincaid et al., 1975) are two widely used readability formulas. We reserve their features and refine the coefficients on both data sets to befit the reading (grade) levels.
- SUM (Collins-Thompson & Callan, 2004) is a word-based method, which trains one unigram model for each reading level, and applies model smoothing among the reading levels.
- V&M (Vajjala & Meurers, 2012) is one of the current best readability assessment methods for English, which adopts three groups of features for classification. As majority of the features are designed specifically for English, we run V&M on ENCT only.
- Jiang (Jiang et al., 2014) is a readability assessment method for Chinese. It adopts five groups of features and designs an ordinal multiclass classification with voting for classification. We run Jiang on CPT only.
- SG-NN is a word embedding-based readability assessment method proposed by Tseng et al. (2016). In SG-NN, the representation of a document is generated by adding up the word embedding of all words in the document. The word embedding model used is Skip-Gram. The classification model used is the regularized neural network with one hidden layer.
- SG-KM-SVM is a word embedding-based readability assessment method proposed by Cha et al. (2017). In SG-KM-SVM, the representation of a document is generated by applying average pooling on the word embedding and cluster membership of all words in the document. The word embedding model used is Skip-Gram. The cluster membership is generated by K-means. SVM (Support Vector Machine) is used to predict the reading level of a document.
- SVM (Support Vector Machine) and LR (Logistic Regression) are two classification models that have widely been used for readability assessment in previous studies (Feng et al., 2010; Jiang et al., 2014).
- TSVM (Transductive SVM) (Joachims, 1998) is a classical transductive method, which has not been applied in readability assessment. Since GRAW+ is also a transductive method, we run TSVM here as a baseline.
- OLR (Ordinal LR) (Mccullagh, 1980) is a variant of LR that can predict in ordinal scale. As the reinforced label propagation in GRAW+ also exploits the ordinal relation among reading levels, we run OLR here as another baseline.
- Bi-LP is a graph propagation method that applies label propagation on a complex graph (Gao et al., 2015; Jiang, 2011). Bi-LP builds two separate subgraphs from cBoW view and linguistic view, and connects them using the bipartite subgraph. The label propagation algorithm is performed on the integrated graph and leads to two distributions for each document. A

TABLE 2. Statistics of the English and Chinese data sets.

| Data set | Language | #Grade | #Doc | #Sent | #Word |
|---|---|---|---|---|---|
| CPT | Chinese | 6 | 637 | 16,145 | 234,372 |
| ENCT | English | 4 | 279 | 4,671 | 62,921 |


DOI: 10.1002/asi

TABLE 3. The average precision, recall, and F1-measures (%) of the 11 methods for readability assessment on either data set (the labeling proportion is 0.7).

| Methods | | CPT | | | ENCT | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1-measure | Precision | Recall | F1-measure |
| SMOG | | 28.48 | 25.38 | 21.12 | 55.00 | 41.41 | 41.20 |
| FK | | 34.48 | 24.68 | 18.73 | 60.78 | 45.65 | 46.16 |
| SUM | | 37.39 | 33.35 | 33.78 | 71.18 | 71.45 | 67.44 |
| V&M | | — | — | — | 86.98 | 85.08 | 85.63 |
| Jiang | | 48.07 | 47.53 | 47.30 | — | — | — |
| SG-NN | | 45.58 | 45.87 | 44.96 | 88.93 | 88.17 | 88.24 |
| SG-KM-SVM | | 40.96 | 40.77 | 39.98 | 80.87 | 79.92 | 80.33 |
| SVM | Linguistic | 48.43 | 48.01 | 47.83 | 87.95 | 87.08 | 86.85 |
| | TF-IDF | 43.10 | 44.64 | 42.47 | 92.26 | 90.03 | 90.66 |
| | Linguistic+TF-IDF | 51.04 | 51.26 | 50.74 | 88.02 | 86.00 | 85.96 |
| LR | Linguistic | 46.53 | 46.43 | 46.16 | 88.05 | 87.16 | 87.28 |
| | TF-IDF | 33.10 | 34.00 | 30.65 | 88.26 | 86.18 | 86.69 |
| | Linguistic+TF-IDF | 46.64 | 46.78 | 46.22 | 90.80 | 89.17 | 89.67 |
| TSVM | Linguistic | 53.32 | 51.28 | 48.95 | 90.57 | 87.72 | 88.31 |
| | TF-IDF | 37.55 | 38.92 | 30.93 | 28.54 | 46.28 | 35.30 |
| | Linguistic+TF-IDF | 44.35 | 41.12 | 33.47 | 27.83 | 45.14 | 34.43 |
| OLR | Linguistic | 51.02 | 47.93 | 47.98 | 51.53 | 54.40 | 49.93 |
| | TF-IDF | 34.25 | 30.38 | 27.44 | 55.01 | 55.35 | 53.18 |
| | Linguistic+TF-IDF | 51.90 | 50.51 | 49.27 | 62.62 | 62.33 | 60.27 |
| Bi-LP | LP | 46.52 | 46.50 | 44.92 | 87.53 | 83.19 | 83.63 |
| | Reinforced LP | 47.41 | 46.87 | 45.90 | 88.71 | 84.51 | 85.04 |
| GRAW | $G^c$ | 50.41 | 50.60 | 49.67 | 90.27 | 88.16 | 88.67 |
| | $G^f$ | 32.99 | 38.98 | 32.95 | 86.16 | 78.80 | 78.55 |
| | $G^{cf}$ | 51.43 | 53.26 | 50.86 | 92.39 | 90.68 | 91.15 |
| GRAW+ | $G^c$ | 53.70 | 53.64 | 53.19 | 91.86 | 90.55 | 90.88 |
| | $G^l$ | 40.99 | 43.99 | 39.97 | 87.34 | 80.90 | 81.00 |
| | $G^{cl}$ | **54.21** | **55.16** | **54.07** | **93.33** | **92.02** | **92.38** |

simple average is used to determine the final class label of each document.

- GRAW (Jiang et al., 2015) is the previous version of our method. We verify if GRAW+ can improve the performance of GRAW by adding new facilities.

For the four baseline classification models (that is, LR, SVM, TSVM, and OLR), documents are represented as feature vectors. To make a fair comparison, we build the feature vector by concatenating two sets of features: the linguistic features used in this article (denoted Linguistic), and the TF-IDF features (denoted TF-IDF), since GRAW+ incorporates both.

For GRAW+, we apply the reinforced label propagation on each of the three graphs: $G^c$, $G^l$, and $G^{cl}$, denoted as $\text{GRAW}^{\text{c}}_{+}$, $\text{GRAW}^{\text{l}}_{+}$, $\text{GRAW}^{\text{cl}}_{+}$, respectively. Unless otherwise specified, we fixed $\eta$ to 3, $\alpha$ to 0, and $\beta$ to 0.5. $\lambda$ and $k$ are tuned on the development set. Sentences from the whole data set are used as the auxiliary text corpus. LR is used to build the preclassifier for the reinforced label propagation.

Table 3 gives the average performance of each method on both data sets where the proportion of the labeled (training) set is 0.7. Specifically, the precision, recall, and F1-measure of all the methods are calculated by averaging the results on all reading (grade) levels on either English or Chinese data sets. The values marked in bold in each column refer to the maximum (best) measures gained by the methods.

From Table 3, the readability formulas (SMOG and FK) perform poorly in both the precision and recall measures, so that their F1-measures are generally the poorest. However, SMOG and FK still have acceptable performance on the English data set ENCT. The unigram model (SUM) performs a little better than the readability formulas. On ENCT, it has relatively good performance, while on the Chinese data set CPT, its performance is not satisfactory. Both V&M on ENCT and Jiang on CPT perform well, which means both the linguistic features developed and the classifiers trained are useful. The two word embedding based methods (SG-NN and SG-KM-SVM) achieve better performance measures than SUM on both data sets. On ENCT, SG-NN performs better than V&M. By adopting features from our proposed two views, the two commonly used models (SVM and LR) perform a little better than V&M and Jiang, which demonstrates the usefulness of the two views. The transductive method (TSVM) slightly outperforms SVM and LR, which suggests that the unlabeled documents can provide valuable information for readability classification. In addition, by adding the TF-IDF features into the feature set, the performance of both SVM and LR have improved, but the performance of TSVM becomes worse. This may be due to the heterogeneous spaces of the two views, which should not be roughly combined through


DOI: 10.1002/asi

concatenation. By employing the ordinal relation among the reading levels, OLR achieves a good performance on CPT. But on ENCT its performance is poor, which suggests the instability of OLR. On both data sets, both GRAW and GRAW+ can outperform Bi-LP with either LP or reinforced LP, which demonstrates that our graph merging strategy is more effective than the graph integrating strategy in Bi-LP. By comparing Bi-LP with $G^c$ and $G^l$ from GRAW+, which roughly correspond to the two subgraphs in Bi-LP, it can be observed that the performance of Bi-LP is better than $G^l$ but worse than $G^c$. By simply linking the two subgraphs through the bipartite graph, Bi-LP may not take full advantage of the two subgraphs. In general, $GRAW^{cf}$ (applying the general label propagation on $G^{cf}$) performs better than all the above baselines, which demonstrates the effectiveness of our method. And last, by applying the reinforced label propagation algorithm, $GRAW^{cl}_{+}$ performs the best in all the three measures on both data sets.

We studied the effect of the labeling proportion on the performance of these methods on both data sets. The F1-measure averaged over the reading levels was used, since it is a good representative of the three measures according to Table 3. Figure 5 depicts the performance trends of the five baseline methods and the two versions of our method (that is, GRAW and GRAW+) by varying the labeling proportion from 0.1 to 0.9 step by 0.1.

From Figure 5, neither SMOG nor FK benefits from the enlarged labeled set. This suggests that the performance of the readability formulas can hardly be improved by accumulating training data. The other methods achieve better performance on larger labeled sets, and outperform the two readability formulas even if the labeling proportion is small. Both Jiang on CPT and V&M on ENCT perform better than SUM. GRAW outperforms the baseline methods over all the labeling proportions on both data sets and performs well even when the labeling proportion is small. Again, as the enhanced version of GRAW, the performance of GRAW+ is consistently improved over the labeling proportions.

*Effects of the Word Coupling Matrix*

For RQ2, we first compare the coupled BoW model to the basic BoW model for graph construction. Three graphs are built by using each of the three coupled BoW matrices (that is, $M^{sur}$, $M^{lex}$, and $M^{syn}$) generated from the three word coupling matrices (that is, $C^{sur}$, $C^{lex}$, and $C^{syn}$), respectively, and one graph is built by using the TF-IDF (that is, basic BoW) matrix for comparison. The general label propagation is applied on each graph to measure the performance of readability classification. The labeling proportion is varied from 0.1 to 0.9 on both the English and Chinese data sets. Figure 6a depicts the averaged F1-measures resulting from the four graphs. From Figure 6a, the three coupled BoW matrices greatly outperform the TF-IDF matrix, especially on the Chinese data set CPT. This demonstrates that the word coupling matrices are effective in improving the performance of the basic BoW model for readability assessment.

Second, we investigate which parts of GRAW+ (the parameters, the word filtering, and the size of auxiliary sentence set) take effect on the performance of the word coupling matrices.

*The effect of the parameters $\eta$ and $\lambda$.* To investigate the effects of $\eta$ and $\lambda$ on the performance of the word coupling matrices, we vary the values and compute the average F1-measures on the two data sets, as shown in Figure 6b. The graph was built using $M^{syn}$, and the other two graphs present similar trends. From Figure 6b, a small $\eta$ (for example, 2 or 3) is good on the Chinese data set CPT. However, on the English data set ENCT, $\eta = 2$ leads to the poorest performance. It shows that the increasing of $\eta$ causes vibrated performance, and the trend is further complicated when involving $\lambda$. Above all, $\eta = 3$ gives a preferable option on both data sets. For $\lambda$, most matrices exhibit similar trends that rise first and then keep stable on both data sets, while some may drop when $\lambda$ is too great. This suggests that making a relatively large number of neighbors for each word (that is, $\lambda$ = 2,800 on CPT and

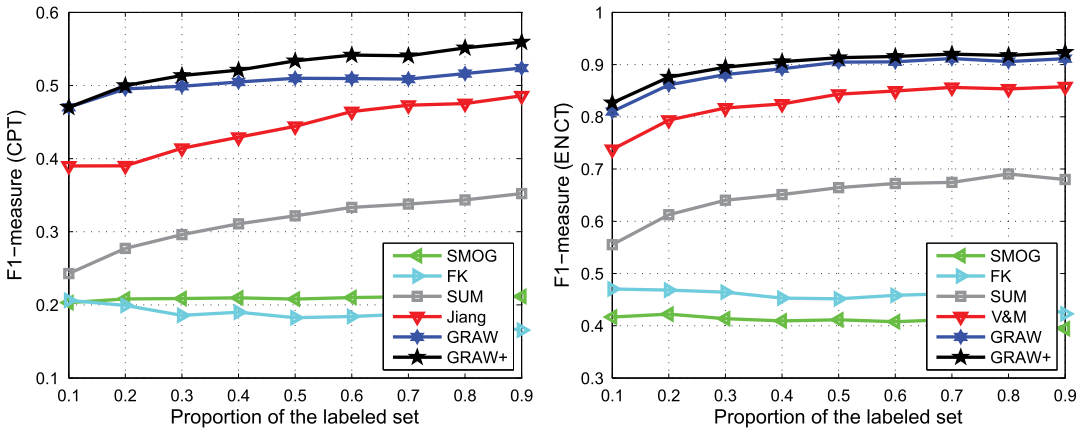
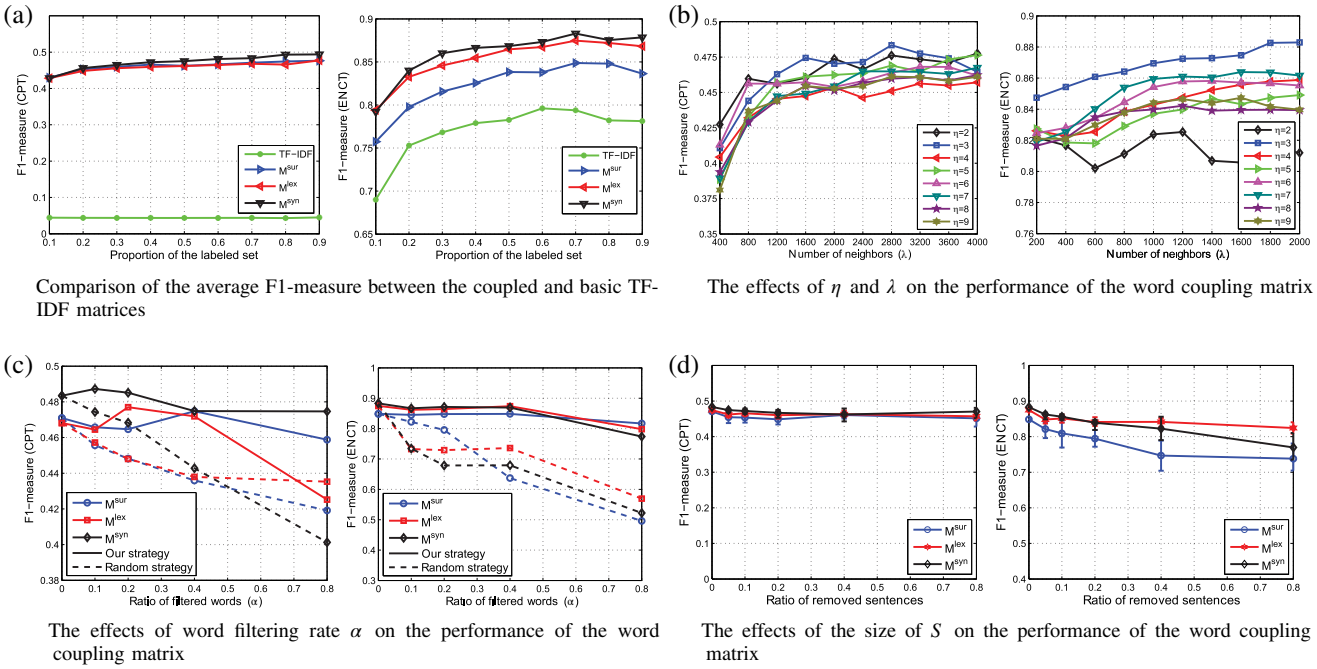



FIG. 5. The average F1-measures of the seven methods on both data sets (Jiang is running on CPT only, while V&M is running on ENCT only) with the labeling proportion varied from 0.1 to 0.9 step by 0.1. [Color figure can be viewed at wileyonlinelibrary.com]


DOI: 10.1002/asi




Comparison of the average F1-measure between the coupled and basic TF-IDF matrices

The effects of $\eta$ and $\lambda$ on the performance of the word coupling matrix

The effects of word filtering rate $\alpha$ on the performance of the word coupling matrix

The effects of the size of $S$ on the performance of the word coupling matrix

FIG. 6. The performance comparison among the word coupling matrices constructed from different perspectives. [Color figure can be viewed at wileyonlinelibrary.com]

$\lambda$ = 2,000 on ENCT) will result in an effective word coupling matrix.

*The effect of word filtering.* To investigate the effect of the word filtering strategy on the performance of the coupled BoW model, we vary the ratio $\alpha$ of filtered words, and compute the average F1-measures resulting from the three coupled BoW matrices (that is, $M^{sur}$, $M^{lex}$, and $M^{syn}$). The random filtering is depicted for comparison, which filters out words from the vocabulary randomly. From Figure 6c, we find that the random filtering performs worse than our word filtering strategy on both data sets. By employing our word filtering strategy, a stable performance can be attained for all three coupled BoW matrices on both data sets when no more than 40% words are filtered out.

*The effect of the size of S.* To investigate if the size of $S$ (that is, the sentence set) takes effect on the performance of GRAW+, we vary the size of $S$ by randomly removing sentences from it. Figure 6d depicts the average F1-measures resulting from the three coupled BoW matrices. From Figure 6d, on the Chinese data set CPT, the performance of GRAW+ suffers little from removing sentences, even if only 20% of sentences are left for building the word coupling matrices. However, on the English data set ENCT, the mean performance drops evidently and the deviation increases evidently when too many sentences are removed. This suggests that cumulating a sufficient text corpus is required for building a suitable word coupling matrix for the coupled BoW model, and factors other than the number of sentences may influence the corpus quality, which will be discussed later.

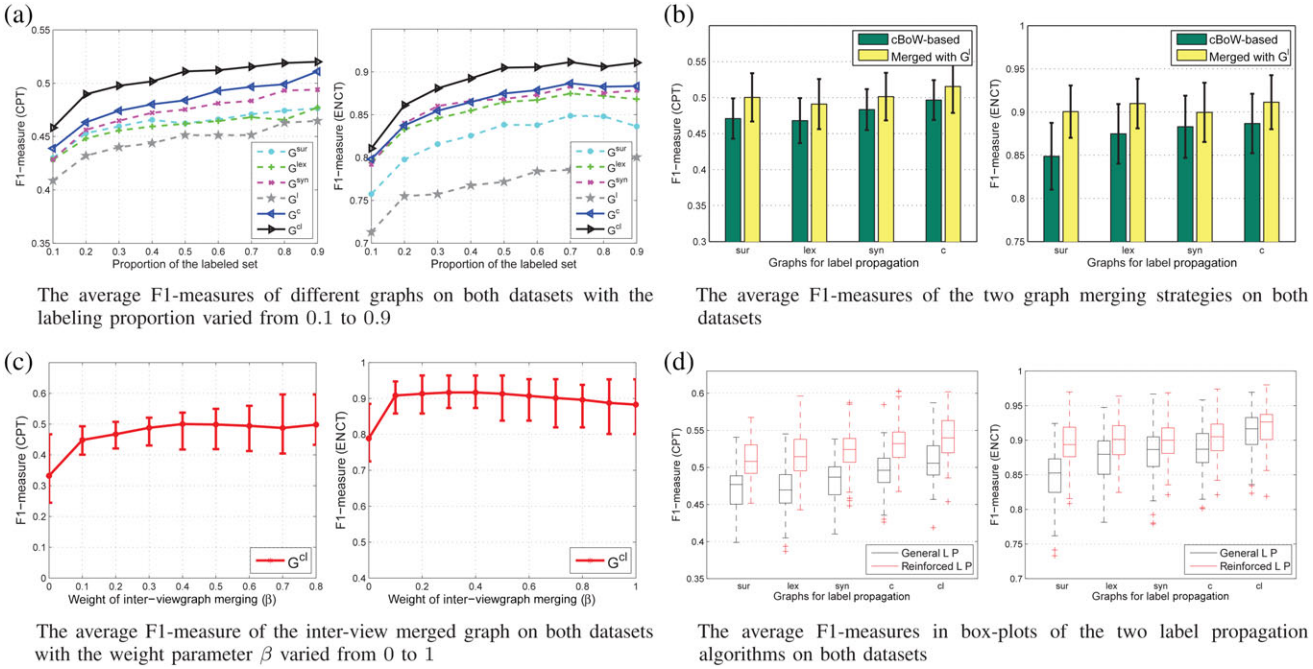
*Effectiveness of Two-View Graph Propagation*

For RQ3, we conducted experiments to validate the effectiveness of the graph merging strategies and the reinforced label propagation algorithm.

*Effectiveness of graph merging.* We compared graphs built on singular coupled BoW matrix (that is, $G^{sur}$, $G^{lex}$, and $G^{syn}$) to the intra-view merged graph (that is, $G^{c}$) and the inter-view merged graph (that is, $G^{cl}$). Figure 7a depicts the averaged F1-measures resulting after applying the general label propagation on these graphs. From Figure 7a, the merged graph $G^{c}$ outperforms the three basic graphs on both data sets in most cases. Within the three singular matrices, $G^{syn}$ performs best, especially on the English data set ENCT, where it can outperform $G^{c}$ slightly when the labeling proportion is small (0.2−0.4). By combining the graph from the linguistic view, $G^{cl}$ performs evidently better than $G^{c}$ on both data sets, while $G^{l}$ always performs the poorest. Figure 7b further validates the effectiveness of the graph merging strategies. By merging the graph from the linguistic view, all the cBoW-based graphs (green bars) get consistently improved performance (yellow bars). The intra-view merging strategy provides a stable improvement for all the graphs built from the cBoW view.

To study the effect of $\beta$ (in Equation 10) on the performance of the merged graph $G^{cl}$, we present the results of applying the general label propagation on $G^{cl}$ with varied $\beta$ in Figure 7c. From Figure 7c, $G^{cl}$ performs well when $\beta$ is in range [0.4, 0.8] on CPT and [0.2, 0.4] on ENCT. This means that the cBoW view requires more weight on CPT than on ENCT. Note that the graph with $\beta$ = 0 equals $G^{l}$,


DOI: 10.1002/asi




The average F1-measures of different graphs on both datasets with the labeling proportion varied from 0.1 to 0.9

The average F1-measures of the two graph merging strategies on both datasets

The average F1-measure of the inter-view merged graph on both datasets with the weight parameter $\beta$ varied from 0 to 1

The average F1-measures in box-plots of the two label propagation algorithms on both datasets

FIG. 7. The performance comparison among the graph merging strategies and the reinforced label propagation algorithm. [Color figure can be viewed at wileyonlinelibrary.com]

which stands for the linguistic view. The sharp rises at 0.1 on both data sets indicate the necessity of the cBoW view.

To verify that graph merging is superior to matrix concatenation, we present the averaged F1-measures resulting after applying the general label propagation algorithm on graphs built by matrix concatenation in Table 4. "sur·lex·syn" refers to the graph built by concatenating the three coupled TF-IDF matrices and "sur·lex·syn·lin" refers to the graph built by concatenating the three coupled TF-IDF matrices and the linguistic matrix. The former is compared to the intra-view graph merging strategy, while the latter is compared to the inter-view graph merging strategy. As shown in Table 4, the two graph merging strategies always outperform the matrix concatenation on both data sets. Besides, "sur·lex·syn·lin" performs worse than "sur·lex·syn," which implies that matrix concatenation is not a good choice in integrating heterogeneous vector space models.

*Effectiveness of reinforced label propagation.* To study the effectiveness of the reinforced label propagation, we compared the general label propagation algorithm to the reinforced label propagation algorithm. Figure 7d depicts the boxplots of applying the two label propagation algorithms on the three singular and two merged graphs.

Figure 7d, shows that the reinforced label propagation algorithm outperforms the general label propagation algorithm on both data sets no matter which of the five graphs is used, which means that our enhancement to the general label propagation algorithm is effective, and the ordinal relation among reading levels shall be utilized. Since preclassification is required to get the a priori labels, the reinforced label propagation provides a way to combine two weak classifiers into a stronger one.

### *External Corpus for Constructing Word Coupling Matrix*

For RQ4, we investigated the effects of using external corpus on constructing the word coupling matrix. We collect two external corpora for both languages: the Chinese Wikipedia (denoted Cwiki) and the English Wikipedia (denoted Ewiki), as shown in Table 5. For the documents in Cwiki, we use the ICTCLAS tool (Zhang, 2013; Zhang et al., 2003) to do the word segmentation.

Different from previous experiments, which a construct word coupling matrix based on the target data set (CPT or ENCT) itself, we conducted experiments to verify if we can use the external text corpus for constructing universal word coupling matrices. Figure 8 depicts the performance

TABLE 4. Comparison between the graph merging strategies and the matrix concatenation strategies.

| Strategy | | CPT | ENCT |
|---|---|---|---|
| Matrix concatenation | sur•lex•syn | 47.45 | 88.43 |
| | sur•lex•syn•lin | 45.82 | 86.44 |
| The intra-view merged graph $G^c$ | | 49.67 | 88.67 |
| The inter-view merged graph $G^{cl}$ | | 51.54 | 91.16 |


DOI: 10.1002/asi

TABLE 5. Statistics of the external corpora for both languages.

| External corpus | Language | #Sent | #Word |
|---|---|---|---|
| Cwiki | Chinese | 200,000 | 4,788,173 |
| Ewiki | English | 40,000 | 948,755 |

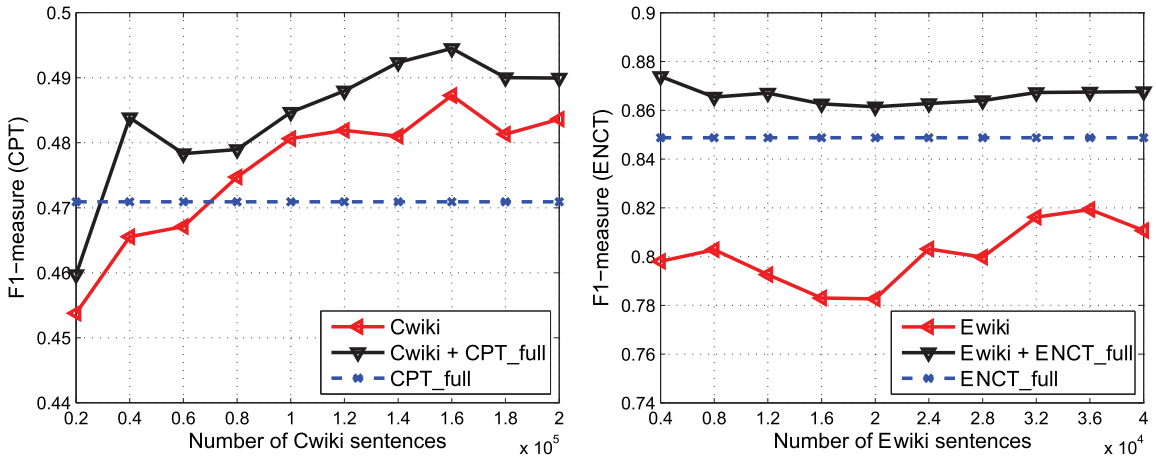



FIG. 8. The effects of using external text corpus on the performance of the word coupling matrix. [Color figure can be viewed at wileyonlinelibrary.com]

of applying the general label propagation algorithm on $G^{sur}$ in line charts, where the trends in different colors correspond to the word coupling matrices built using a different corpus. From Figure 8, it shows that the external corpus (that is, Cwiki) alone can achieve good performance for Chinese when the size is large enough (for example, larger than $8 \times 10^4$). However, for English, only using the external corpus is not enough. The reason may be that there is a great difference between ENCT and the Ewiki corpus, and the word matrices built on Ewiki alone is insufficient. In addition, by combining the target data set with the external corpus (that is, the black line), the performance can be consistently improved. This suggests that the external text corpus is beneficial in enhancing the quality of the word coupling matrices.

To further study the effect of the external text corpus on GRAW+, we applied the two label propagation algorithms on the two merged graphs (that is, $G^c$ and $G^{cl}$). The results are listed in Table 6. From Table 6, on the Chinese data set, as compared with the internal corpus, the performance of Cwiki is a little worse with $G^c$, but becomes comparable to $G^{cl}$. This suggests that GRAW+ can make the Cwiki corpus a suitable substitution in constructing effective word coupling matrices. On the English data set, the observation is similar, and by mixing the external and internal corpus, the performance can be further improved.

## Discussion

During the experiments, we mainly studied our method GRAW+ from two perspectives. In the first perspective, we verified whether GRAW+ is effective compared to the current well-known readability assessment methods. We compared GRAW+ with the readability formulas, the language model-based method, the linguistic feature-based methods, and the state-of-the-art word embedding-based methods. The experiment results show that GRAW+ performs the best among all these methods on both English and Chinese data sets. The reason may be that the baseline methods only evaluate the readability of documents from either the vocabulary view or linguistic view, while our method employs both views. In addition, it can be found that the existing word embedding-based methods are rather preliminary for readability assessment, and do not take the sequence of words into consideration, which will ignore the syntactic difficulty and discourse difficulty of documents. In our future work, we plan to capture the word sequences and sentence structures, and obtain the high-level representation of documents using the deep neural networks.

In the second perspective, we studied which factors will mostly affect the performance of GRAW+. Factors such as the GRAW parameters, the word filtering strategies, and

TABLE 6. Constructing the word coupling matrix by using external text corpus.

| GRAW+ | | Chinese | | | English | | |
|---|---|---|---|---|---|---|---|
| | | CPT | Cwiki | Cwiki+CPT | ENCT | Ewiki | Ewiki+ENCT |
| General LP | $G^c$ | 49.67 | 47.70 | 49.55 | 88.67 | 83.05 | 89.45 |
| | $G^{cl}$ | 51.54 | 50.72 | 51.13 | 91.16 | 89.59 | 91.30 |
| Reinforced LP | $G^c$ | 53.19 | 51.82 | 52.01 | 90.88 | 86.97 | 90.87 |
| | $G^{cl}$ | 54.07 | 54.25 | 53.94 | 92.38 | 91.08 | 92.39 |


DOI: 10.1002/asi

the size of the auxiliary sentence set were studied in the experiments. There are still others we have not considered, such as the word sense ambiguity, which may affect the reading difficulties of words under different contexts. For example, words with multiple meanings may bring an extra burden to readers in identifying the proper sense within the contexts, and thus affect the readability. In this article, we do not explicitly take the word sense ambiguity into consideration when building the BoW model. However, the reading difficulty of a word is estimated by its difficulty distribution, which may imply whether the word has multiple meanings in a degree. If a word has multiple meanings of different reading difficulties, its difficulty distribution will reflect such properties. The explicit exploration of word sense ambiguity and its effect on readability is planned in our future work.

## Conclusion

In this article we proposed a two-view graph propagation method with word coupling for readability assessment. The coupled bag-of-words model was designed to model the relationship among text documents on readability, which can improve the accuracy of readability assessment. The model can be used with the linguistic features in the graph-based classification framework, which includes graph construction, merging, and label propagation. The reinforced label propagation algorithm was developed to make use of the ordinal relation among reading levels. Experiments were conducted on both Chinese and English data sets. The results show that our method can outperform the state-of-art methods for readability assessment. In addition, the separate experiments demonstrate the usefulness of the coupled bag-of-words model, the graph merging strategies, and the reinforced label propagation algorithm respectively.

In future work, we plan to strengthen our method from the following perspectives: (i) We will explore the effects of semantic factors including the word sense ambiguity on readability assessment of texts. (ii) We will extend the set of linguistic features by considering the coherence features and domain-concept features. (iii) We will test our method on extra data sets of different languages, especially Chinese, and adapt the method to improve its effectiveness on specific languages. (iv) We will design the end-to-end neural networks for readability assessment, in order to express the reading difficulties from different levels, for example, vocabulary, syntactic, semantic, and others.

## Acknowledgments

This work was supported by the National Key R&D Program of China under Grant No. 2018YFB1003800; National Natural Science Foundation of China under Grant Nos. 61373012, 61321491, 91218302; the Fundamental Research Funds for the Central Universities under Grant No. 020214380049. This work is partially supported by the Collaborative Innovation Center of Novel Software Technology and Industrialization.

## References

Benjamin, R.G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. Educational Psychology Review, 24(1), 63–88.

Billhardt, H., Borrajo, D., & Maojo, V. (2002). A context vector model for information retrieval. Journal of the Association for Information Science and Technology, 53(3), 236–249.

Cao, L. (2015). Coupling learning of complex interactions. Information Processing and Management, 51(2), 167–186.

Cha, M., Youngjune G., & H.T. Kung. (2017). Language modeling by clustering with word embeddings for text readability assessment. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp.2003–2006). Singapore: ACM.

Cheng, X., Miao, D., Wang, C., & Cao, L. (2013). Coupled term-term relation analysis for document clustering. In Proceedings of the 2013 International Joint Conference on Neural Networks (pp. 1–8). Dallas, TX, USA: IEEE.

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. ITL – International Journal of Applied Linguistics, 165(2), 97–135.

Collins-Thompson, K., & Callan, J.P. (2004). A language modeling approach to predicting reading difficulty. In Proceedings of the 2004 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 193–200). Boston, Massachusetts, USA: Association for Computational Linguistics.

Denning, J., Pera, M.S., & Ng, Y.K. (2016). A readability level prediction tool for k-12 books. Journal of the Association for Information Science and Technology, 67(3), 550–565.

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A comparison of features for automatic readability assessment. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 276–284). Beijing, China: Chinese Information Processing Society of China.

François, T., & Fairon, C. (2012). An AI readability formula for French as a foreign language. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 466–477). Jeju Island, Korea: Association for Computational Linguistics.

Gao, D., Wei, F., Li, W., Liu, X., & Zhou, M. (2015). Cross-lingual sentiment lexicon learning with bilingual word graph label propagation. Computational Linguistics, 41(1), 21–40.

Hancke, J., Vajjala, S., & Meurers, D. (2012). Readability classification for German using lexical, syntactic, and morphological features. In Proceedings of the 24th International Conference on Computational Linguistics (pp. 1063–1080). Mumbai, India: Indian Institute of Technology Bombay.

Huang, A. (2008). Similarity measures for text document clustering. In Proceedings of the Sixth New Zealand Computer Science Research Student Conference (pp. 49–56). Christchurch, New Zealand: New Zealand Computer Science Research Student Conference.

Jebara, T., Wang, J., & Chang, S.-F. (2009). Graph construction and b-matching for semi-supervised learning. In Proceedings of the 26th Annual International Conference on Machine Learning (pp. 441–448).

Jiang, J.Q. (2011). Learning protein functions from bi-relational graph of proteins and function annotations. In International Conference on Algorithms in Bioinformatics (pp.128–138). Saarbrücken, Germany: Springer.

Jiang, Z., Sun, G., Gu, Q., Bai, T., & Chen, D. (2015). A graph-based readability assessment method using word coupling. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 411–420).

Jiang, Z., Sun, G., Gu, Q., & Chen, D. (2014). An ordinal multi-class classification method for readability assessment of Chinese documents. In Knowledge Science, Engineering and Management (pp. 61–72). Sibiu, Romania: Springer.

Joachims, T. (1998). Making large-scale SVM learning practical. In Bernhar Schölkopf, Christophe J. C. Burges, & Alexande J. Smola (Eds.), Advances in Kernel Methods-Support Vector Learning. Cambridge, MA: MIT Press.

Kalogeratos, A., & Likas, A. (2012). Text document clustering using global term context vectors. Knowledge and Information Systems, 31(3), 455–474.

Kidwell, P., Lebanon, G., & Collins-Thompson, K. (2009). Statistical estimation of word acquisition with application to readability prediction. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (pp. 900–909). Singapore: Association for Computational Linguistics.

Kim, D.S., Verma, K., & Yeh, P.Z. (2013). Joint extraction and labeling via graph propagation for dictionary construction. In Proceedings of the 27th AAAI Conference on Artificial Intelligence (pp. 510–517). Bellevue, Washington, USA: AAAI Press.

Kincaid, J.P., Fishburne, R.P., 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* (Tech. Rep.). Memphis, TN: Naval Air Station.

Kullback, S., & Leibler, R.A. (1951). On information and sufficiency. The Annals of Mathematical Statistics, 22(1), 79–86.

Mccullagh, P. (1980). Regression models for ordinal data. Journal of the Royal Statistical Society, 42(2), 109–142.

McLaughlin, G.H. (1969). Smog grading: A new readability formula. Journal of Reading, 12(8), 639–646.

Pilán, I., Volodina, E., & Johansson, R. (2014). Rule-based and machine learning approaches for second language sentence-level readability. In Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 174–184). Baltimore, Maryland, USA: Association for Computational Linguistics.

Pitler, E., & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In 2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference (pp. 186–195). Honolulu, Hawaii, USA: Association for Computational Linguistics.

Ponomareva, N., & Thelwall, M. (2012). Do neighbours help?: An exploration of graph-based algorithms for cross-domain sentiment classification. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 655–665). Jeju Island, Korea: Association for Computational Linguistics.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5), 513–523.

Schumacher, E., Eskenazi, M., Frishkoff, G., & Collins-Thompson, K. (2016). Predicting the relative difficulty of single sentences with and without surrounding context. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 1871–1881). Austin, Texas, USA: Association for Computational Linguistics.

Schwarm, S.E., & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 523–530).

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47.

Sinha, M., Dasgupta, T., & Basu, A. (2014). Influence of target reader background and text features on text readability in Bangla: A computational approach. In Proceedings of the 25th International Conference on Computational Linguistics (pp. 345–354). Dublin, Ireland: Association for Computational Linguistics.

Subramanya, A., Petrov, S., & Pereira, F. (2010). Efficient graph-based semi-supervised learning of structured tagging models. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 167–176). MIT Stata Center, Massachusetts, USA: Association for Computational Linguistics.

Tseng, H.C., Hung, H.T., Sung, Y.T., & Chen, B. (2016). Classification of text readability based on deep neural network and representation learning techniques [In Chinese]. In Proceedings of the 28th Conference on Computational Linguistics and Speech Processing (ROCLING 2016) (pp. 255–270). National Cheng Kung University, Tainan, Taiwan: Association for Computational Linguistics and Chinese Language Processing.

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP (pp. 163–173). Montréal, Canada: Association for Computational Linguistics.

Vajjala, S., & Meurers, D. (2014). Assessing the relative reading level of sentence pairs for text simplification. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014 (pp. 288–297). Gothenburg, Sweden: Association for Computational Linguistics.

Zakaluk, B.L., & Samuels, S.J. (1988). Readability: Its past, present, and future. Newark, DE: International Reading Association.

Zeng, X., Wong, D.F., Chao, L.S., & Trancoso, I. (2013). Graph-based semi-supervised model for joint Chinese word segmentation and part-of-speech tagging. In Proceeding of the 51st Annual Meeting of the Association for Computational Linguistics (pp. 770–779). Sofia, Bulgaria: Association for Computational Linguistics.

Zhang, H.P. (2013). ICTCLAS [Computer software]. Retrieved from http://ictclas.nlpir.org

Zhang, H.P., Yu, H.K., Xiong, D.Y., & Liu, Q. (2003). HHMM-based Chinese lexical analyzer ICTCLAS. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing (Vol. 17, pp. 184–187).

Zhu, X., & Ghahramani, Z. (2002). *Learning from labeled and unlabeled data with label propagation* (Tech. Rep. No. CMU-CALD-02-107). Center for Automated Learning and Discovery, CMU: Carnegie Mellon University, USA.