A Graph-based Readability Assessment Method using Word Coupling

Zhiwei Jiang, Gang Sun, Qing Gu*, Tao Bai, Daoxu Chen
State Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
jiangzhiwei@outlook.com, sungangnju@163.com,
guq@nju.edu.cn, bt@xjau.edu.cn, cdx@nju.edu.cn

Abstract

This paper proposes a graph-based readability assessment method using word coupling. Compared to the state-of-the-art methods such as the readability formulae, the word-based and feature-based methods, our method develops a coupled bag-of-words model which combines the merits of word frequencies and text features. Unlike the general bag-of-words model which assumes words are independent, our model correlates the words based on their similarities on readability. By applying TF-IDF (Term Frequency and Inverse Document Frequency), the coupled TF-IDF matrix is built, and used in the graph-based classification framework, which involves graph building, merging and label propagation. Experiments are conducted on both English and Chinese datasets. The results demonstrate both effectiveness and potential of the method.

1 Introduction

Readability assessment is a task that aims to evaluate the reading difficulty or comprehending easiness of text documents. It is helpful for educationists to select texts appropriate to the reading/grade levels of the students, and for web designers to organize texts on web pages for the users doing personalized searches for information retrieval.

Research on readability assessment starts from the early 20th century (Dale and Chall, 1948). Many useful readability formulae have been developed since then (Dale and Chall, 1948; McLaughlin, 1969; Kincaid et al., 1975). Currently, due to the development of natural language processing, the methods on readability assessment have made a great progress (Zakaluk and Samuels, 1988; Benjamin, 2012; Gonzalez-Dios et al., 2014). The word-based methods compute word frequencies in documents to estimate their readability (Collins-Thompson and Callan, 2004; Kidwell et al., 2009). The feature-based methods extract text features from documents and train classification models to classify the readability (Schwarm and Ostendorf, 2005; Feng et al., 2010; François and Fairon, 2012; Hancke et al., 2012).

In this paper, we propose a graph-based method using word coupling, which combines the merits of both word frequencies and text features for readability assessment. We design a coupled bag-of-words model, which correlates words based on their similarities on sentence-level readability computed using text features. The model is used in a graph-based classification framework, which involves graph building, graph merging/combination, and label propagation. We perform experiments on datasets of both English and Chinese. The results demonstrate both effectiveness and potential of our method.

The rest of this paper is organized as follows: Section 2 introduces backgrounds of our work. Section 3 presents the details of the method. Section 4 designs the experiments and explains the results. Finally, Section 5 concludes the paper with planned future work.

2 Background

In this section, we introduce briefly three research topics relevant to our work: readability assessment, the bag-of-words model and the graph-based label propagation method.

2.1 Readability Assessment

Research on readability assessment has developed three types of methods: the readability formula, the word-based methods and the feature-based methods (Kincaid et al., 1975; Collins-Thompson and Callan, 2004; Schwarm and Os-

*Corresponding author.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 411–420,
Lisbon, Portugal, 17-21 September 2015.

tendorf, 2005). During the early time, many well-known readability formulae have been developed to assess the readability of text documents (Dale and Chall, 1948; McLaughlin, 1969; Kincaid et al., 1975). Surface text features are defined in these formulae to measure both lexical and grammatical complexities of a document. The word-based methods focus on words and their frequencies in a document to assess its readability, which mainly include the unigram/bigram/n-gram models (Collins-Thompson and Callan, 2004; Schwarm and Ostendorf, 2005) and the word acquisition model (Kidwell et al., 2009). The feature-based methods focus on extracting text features from a document and training a classification model to classify its readability (Feng et al., 2010; François and Fairon, 2012; Hancke et al., 2012). Suitable text features are usually essential to the success of these methods. The Support vector machine and logistic regression model are two classification models commonly used in these methods.

2.2 The Bag-of-Words Model

The bag-of-words model is mostly used for document classification. It constructs a feature space that contains all the distinct words in a language (or the document set). A document is represented by a vector, whose components reflect the weight of every distinct word contained in the document. Normally, it assumes the words are independent. Now the capturing of the relationship among words has attracted considerable attention (Wong et al., 1985; Cheng et al., 2013). Inspired by these works, this paper adopts the bag-of-words model in readability assessment, and refines the model by computing similarity among words on reading difficulty.

2.3 The Graph-based Label Propagation Method

Graph-based label propagation is applied on a graph to propagate class labels from labeled nodes to unlabeled ones (Kim et al., 2013). It has been successfully applied in various applications, such as dictionary construction (Kim et al., 2013), word segmentation and tagging (Zeng et al., 2013), and sentiment classification (Ponomareva and Thelwall, 2012). Typically, a graph-based label propagation method consists of two main steps: graph construction and label propagation (Zeng et al., 2013). During the first step, a similarity function is required to build edges and compute weights between pairs of the nodes (Daitch et al., 2009). Some form of edge pruning is required to refine the graph (Jebara et al., 2009). After that, effective algorithms have been developed to propagate the label distributions to all the nodes (Subramanya et al., 2010; Kim et al., 2013).

3 The Proposed Method

In this section, we present GRAW (Graph-based Readability Assessment method using Word coupling), which constructs a coupled bag-of-words model by exploiting the correlation of readability among the words. Unlike the general bag-of-words model which models document relationship on topic, the coupled bag-of-words model extends it to model the relationship among documents on readability. In the following sections, we describe in detail how to build the coupled bag-of-words model. The model is then used in the graph-based classification framework for readability assessment.

3.1 The General Bag-of-Words Model

TF-IDF (Term Frequency and Inverse Document Frequency) is the most popular scheme of the bag-of-words model. Given the set of documents $\mathcal{D}$, the TF-IDF matrix M can be calculated based on the logarithmically scaled term (i.e. word) frequency (Salton and Buckley, 1988) as follows.

$$\begin{aligned} M_{t,d} &= tf_{t,d} \cdot idf_{t,d} \\ &= (1 + \log f(t,d)) \cdot \log \frac{|\mathcal{D}|}{|\{d|t \in d\}|} \end{aligned} \quad (1)$$

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

3.2 The Coupled Bag-of-Words Model

As shown in Figure 1, three main stages are required to construct the coupled bag-of-words model: per-sentence readability estimation, word coupling matrix construction and coupled TF-IDF matrix calculation. The following sections describe the details of these stages.

3.2.1 Per-Sentence Readability Estimation

Two steps are required for the per-sentence readability estimation. The first is to compute a reading score of a sentence by heuristic functions. The second is to determine the difficulty level of the sentence by discretizing the score.

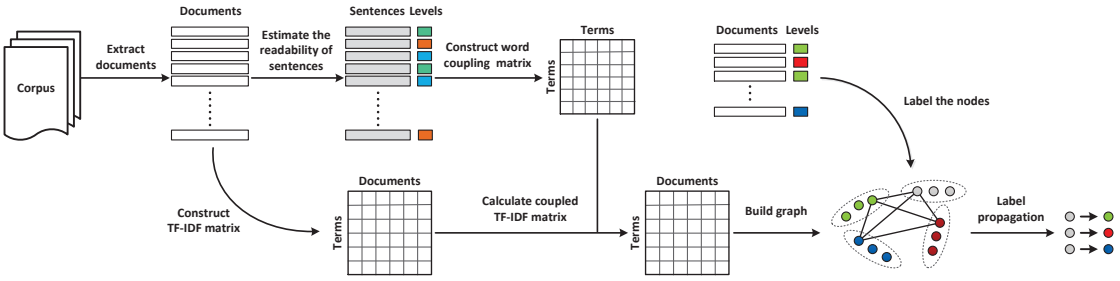


Figure 1: The Framework of GRAW

Step 1. Given a sentence s, its reading difficulty can be quantified as a reading score which is a continuous variable denoted by $r(s)$. The more difficult s is, the greater $r(s)$ will be. Based on text features of s, $r(s)$ can be computed by one of the eight heuristic functions listed in Table 1 which are grouped into three aspects.

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, i.e. part of speech, in s.
	$atr(s)$	the ratio of adjectives in s.
	$ntr(s)$	the ratio of nouns in s.
Syntatic	$pth(s)$	the height of the syntax parser tree of s.
	$anp(s)$	the average number of (noun, verb, and preposition) phrases in s.

Table 1: Three aspects of estimating reading difficulty of sentences using heuristic functions

Step 2. Let η denote the pre-determined number of difficulty levels, r_{max} and r_{min} denote the maximum and minimum reading score respectively of all the sentences in $\mathcal{D}$. To determine the difficulty level $l^*(s)$ ($l^*(s) \in [1, \eta]$) of a sentence s, the range $[r_{min}, r_{max}]$ is divided into η intervals, so that each interval contains the reading scores of $\frac{1}{\eta}$ of all the sentences. The assumption is that all the sentences are equally distributed among the difficulty levels. $l^*(s)$ 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. The assumption is that the reading difficulty of a sentence may be determined by the maximum measure on the text features.

$$\begin{aligned} l^{sur}(s) &= \max\left[l^{len}(s), l^{ans}(s), l^{anc}(s)\right] \\ l^{lex}(s) &= \max\left[l^{lv}(s), l^{atr}(s), l^{ntr}(s)\right] \\ l^{syn}(s) &= \max\left[l^{pth}(s), l^{anp}(s)\right] \end{aligned} \tag{2}$$

3.2.2 Word Coupling Matrix Construction

Let $\mathcal{V}$ denote the set of all the words, a word coupling matrix is defined as $C^* \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$, the element of which reflects the correlation between two words (i.e. terms). Two steps are required to construct this matrix. The first is to count the difficulty distributions of words, and the second is to compute the correlation between each pair of words according to the similarity of their difficulty distributions.

Step 1. Let $\mathcal{S}$ denote the set of all the sentences, p_t denote the difficulty distribution of a word (term) t. p_t is a vector containing η (i.e. the number of difficulty levels) values, the i-th part of which can be calculated by the following formula.

$$p_t(i) = \frac{1}{n_t} \cdot \sum_{s \in \mathcal{S}} \delta(t \in s) \cdot \delta(l^*(s) = i) \tag{3}$$

where n_t refers to the number of sentences in which t appears. The indicator function $\delta(x)$ returns 1 if x is true and 0 otherwise. $l^*(s)$ refers to one of the functions $l^{sur}(s)$, $l^{lex}(s)$ or $l^{syn}(s)$.

Step 2. Given two words (terms) t_1 and t_2, whose level distributions are p_{t_1} and p_{t_2} respectively, we measure the distribution difference $c_{KL}(t_1, t_2)$ using the Kullback-Leibler divergence (Kullback and Leibler, 1951), computed by the following formula.

$$c_{KL}(t_1, t_2) = \frac{1}{2}\left(KL(p_{t_1}||p_{t_2}) + KL(p_{t_2}||p_{t_1})\right) \tag{4}$$

where $KL(p||q) = \sum_i p(i) \log \frac{p(i)}{q(i)}$. After that, the logistic function is applied on the computed difference to get the normalized distribution similarity, i.e.

$$sim(t_1, t_2) = \frac{2}{1 + e^{c_{KL}(t_1, t_2)}} \tag{5}$$

Given a word t_i, only λ 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 (i.e. $t_j \notin \mathcal{N}(t_i)$), the corresponding $sim(t_i, t_j)$ will be assigned 0. After that, the word coupling matrix (i.e. C^*) with $sim(t_i, t_j)$ as elements is normalized along the rows so that the sum of each row is 1. Based on three different $l^*(s)$, we construct three word coupling matrices C^{sur}, C^{lex} and C^{syn}.

3.2.3 Coupled TF-IDF Matrix Calculation

In the general bag-of-words model, the words are treated as independent of each other. However, for readability assessment, words may be correlated according to the similarity of their difficulty distributions. To improve the TF-IDF matrix M described in Section 3.1, we multiply it by the word coupling matrix C^*, so that the term frequencies are shared among the highly correlated (coupled) words. We denote the coupled TF-IDF matrix as M^*, obtained by the following formula.

$$M^* = C^* \cdot M \tag{6}$$

Specifically, three homogenous coupled TF-IDF matrices M^{sur}, M^{lex} and M^{syn} can be built according to the three word coupling matrices C^*.

3.3 Graph-based Readability Assessment

We employ the coupled bag-of-words model for readability assessment under the graph-based classification framework as described in the previous work (Zhu and Ghahramani, 2002). Firstly, we construct a graph representing the readability relationship among documents by using the coupled bag-of-words model to compute the relations among these documents. Secondly, we estimate reading levels of documents by applying label propagation on the graph.

3.3.1 Graph Construction

We build a directed graph G^* to represent the readability relation among documents, where nodes represent documents, and edges are weighted by the similarities between pairs of documents. Given a similarity function, we link documents d_i to d_j with an edge of weight G^*_{ij}, defined as:

$$G^*_{i,j} = \begin{cases} sim(d_i, d_j) & \text{if } d_j \in \mathcal{N}(d_i) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $\mathcal{N}(d_i)$ is the set of k-nearest neighbors of d_i determined by the similarity function.

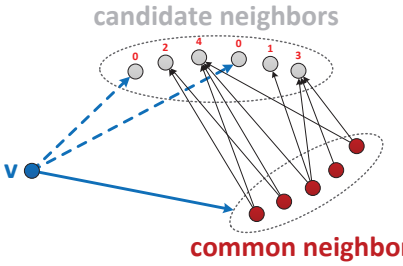


Figure 2: Illustration of the graph merging strategy

Given the coupled matrix $M^* \in \mathbb{R}^{m \times |\mathcal{D}|}$ which maps each document into a m-dimension feature space, the similarity function $sim(d_i, d_j)$ can be defined by the Euclidean distance as follows.

$$sim(d_i, d_j) = \frac{1}{\sqrt{\sum_{k=1}^{m} (M_{k,i} - M_{k,j})^2 + \epsilon}} \tag{8}$$

where ϵ is a small constant to avoid zero denominators.

Merge the three graphs Refer to Section 3.2, the three coupled TF-IDF matrices will lead to three different document graphs, denoted as G^{sur}, G^{lex} and G^{syn} respectively. To take advantage of the three aspects at one time, we need to merge the three graphs into one, denoted as G^c.

In G^c, each node also keeps k neighbors, and some edges shall be filtered out from the three graphs. The basic idea is to remove edges containing redundant information, as shown in Figure 2. For each node v, we firstly select the neighbors which are common in all the three graphs (i.e. $\mathcal{N}^{sur}(v) \cap \mathcal{N}^{lex}(v) \cap \mathcal{N}^{syn}(v)$). Secondly, for the rest candidate nodes, which are the neighbors of v in at least one graph, we select one by one the node which possesses the least number of common neighbors (from all the three graphs) with the nodes that are already selected in $\mathcal{N}^c(v)$. 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.

Combine with the feature-based graph Previous studies usually extract text features from documents to assess the readability using classification models. Here, we also take into consideration the feature-based graph, where similarities among documents are computed on text features. We use the features defined in (Jiang et al., 2014), where the model based features are eliminated since the

computation depends on pre-assigned class labels, and represent a document as a vector of the feature values. We compute the similarity between any pair of documents using the Euclidean distance, and built the feature-based graph (denoted as G^f) in the same way as above.

Additionally, to take advantage of both graphs, we combine them into one (denoted as G^{cf}) using the following formula.

$$G^{cf}_{i,j} = \max\left[G^{c}_{i,j}, G^{f}_{i,j}\right] \quad (9)$$

3.3.2 Graph Propagation

Given a graph G^* constructed in previous sections, its nodes are divided into two sets: the labeled set V_l and the unlabeled set V_u. The goal of label propagation is to propagate class labels from the labeled nodes (i.e. documents) to the entire graph. Here, we use a simplified version of the label propagation method presented in (Subramanya et al., 2010), which has been proved effective (Kim et al., 2013). The method iteratively updates the label distribution on a document node using the following equation.

$$p_d^{(i)}(l) = \frac{1}{\kappa_d}\left(p_d^0(l)\delta(d \in V_l) + \sum_{v \in \mathcal{N}(d)} G_{d,v} p_v^{(i-1)}(l)\right) \quad (10)$$

At the left side of Eq.10, $p_d^{(i)}(l)$ is the afterward probability of l (i.e. the class label) on a node d at the i-th iteration. At the right side, κ_d is the normalizing constant to make sure the sum of all the probabilities is 1, and $p_d^0(l)$ is the initial probability of l on d if d is initially labeled (i.e. belonging to the labeled set V_l). $\delta(x)$ is the indicator function. $\mathcal{N}(d)$ denotes the set of neighbors of d. The iteration stops when the changes in $p_d^{(i)}(l)$ for all the nodes and label values are small enough (e.g. less than e^{-3}), or i exceeds a predefined number (e.g. greater than 30).

4 Empirical Studies

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

RQ1: Whether the proposed method (i.e. GRAW) outperforms the state-of-the-art methods for readability assessment?

RQ2: What are the effects of adding the word coupling matrix to the general bag-of-words model?

RQ3: Whether the graph merging strategy is effective, and whether the performance can be further improved by combining the feature-based graph.

4.1 Corpus and Metrics

To evaluate our proposed method, we collected two datasets. The first is CPT (*Chinese primary textbook*) (Jiang et al., 2014), which contains Chinese documents of six reading levels. The second is ENCT (*English New Concept textbook*) which contains English documents of four reading levels. Both datasets are built from well-known textbooks where documents are labeled as grade levels by credible educationists. The details of the datasets are listed in Table 2.

Dataset	Language	#Grade	#Doc	#Sent	#Word
CPT	Chinese	6	637	16145	234372
ENCT	English	4	279	4671	62921

Table 2: Statistics of the datasets on both English and Chinese

We conduct experiments on both datasets using the cross-validation which randomly divides a dataset into labeled (training) and unlabeled (test) sets. The labeling proportion is varied to investigate the performance of GRAW under different circumstances. To reduce variability, given certain labeling proportion, 100 rounds of cross-validation are performed, and the validation results are averaged over all the rounds. We choose the precision (P), recall (R) and F1-measure (F1) as the performance metrics.

4.2 Comparison to the State-of-the-Art Methods

To address RQ1, we implement the following readability assessment methods and compare GRAW to them: (1) SMOG (McLaughlin, 1969) and FK (Kincaid et al., 1975) are two widely used readability formulae. We reserve their core measures (i.e. text features, and number of strokes for Chinese instead of number of syllables), and refine the coefficients on both datasets to befit the reading (grade) levels. (2) SUM (Collins-Thompson and Callan, 2004) is a word-based method, which trains one unigram model for each grade level, and applies model smoothing both inter and intra the grade levels. (3) LR and SVM refer to two feature-based methods which incorporate text features defined in (Jiang et al., 2014) to represent documents as instances. The logistic regression model and

Dataset	Level	Metric	Methods						
			SMOG	FK	SUM	LR	SVM	GRAW$_c$	GRAW$_{cf}$
CPT (Chinese)	Gr.1	P	57.48	74.07	71.76	71.87	73.18	73.26	**75.29**
		R	17.31	9.69	36.31	71.17	67.28	73.28	**83.17**
		F1	26.14	15.25	47.67	71.23	69.70	72.98	**78.89**
	Gr.2	P	34.73	31.44	37.94	51.62	50.78	52.05	**55.83**
		R	31.06	28.00	29.73	56.48	59.45	57.36	**66.06**
		F1	32.66	29.42	33.13	53.67	54.53	54.40	**60.37**
	Gr.3	P	20.05	20.79	28.12	44.15	48.89	46.33	**51.72**
		R	58.84	75.53	25.06	43.94	49.94	58.59	**68.41**
		F1	29.89	32.40	26.35	43.72	49.04	51.57	**58.74**
	Gr.4	P	25.06	28.94	25.60	33.35	33.92	39.90	**44.15**
		R	**41.06**	31.82	28.76	31.82	33.64	35.42	28.88
		F1	31.03	29.69	26.91	32.24	33.58	**37.32**	34.57
	Gr.5	P	33.57	45.00	28.71	37.70	37.30	**45.02**	37.33
		R	4.00	2.71	34.41	**36.12**	34.29	27.12	19.00
		F1	7.02	4.95	31.10	**36.61**	35.47	33.35	24.45
	Gr.6	P	0.00	6.67	32.21	40.47	**46.53**	45.91	44.24
		R	0.00	0.35	45.81	39.03	43.48	51.81	**54.06**
		F1	0.00	0.67	37.55	39.48	44.65	**48.38**	48.15
	Avg.	P	28.48	34.48	37.39	46.53	48.43	50.41	**51.43**
		R	25.38	24.68	33.35	46.43	48.01	50.60	**53.26**
		F1	21.12	18.73	33.78	46.16	47.83	49.67	**50.86**
ENCT (English)	Gr.1	P	54.65	60.79	**96.59**	88.60	90.74	95.42	95.53
		R	67.50	73.50	84.77	**89.32**	85.45	83.77	83.95
		F1	60.18	66.36	**90.09**	88.64	87.76	89.01	89.18
	Gr.2	P	50.11	56.23	78.30	85.51	**90.80**	88.60	89.03
		R	59.28	63.93	35.07	86.07	92.86	96.76	**96.86**
		F1	54.17	59.69	48.11	85.54	91.68	92.42	**92.70**
	Gr.3	P	29.49	32.09	40.53	88.31	89.08	85.36	**89.73**
		R	24.22	26.94	68.33	86.17	84.78	94.17	**96.56**
		F1	26.40	29.15	50.77	86.94	86.16	89.40	**92.92**
	Gr.4	P	85.73	94.00	69.30	89.79	81.20	91.70	**95.26**
		R	14.64	18.21	**97.64**	87.07	85.21	77.93	85.36
		F1	24.06	29.46	80.81	88.02	81.79	83.84	**89.81**
	Avg.	P	55.00	60.78	71.18	88.05	87.95	90.27	**92.39**
		R	41.41	45.65	71.45	87.16	87.08	88.16	**90.68**
		F1	41.20	46.16	67.44	87.28	86.85	88.67	**91.15**

Table 3: The average Precision, Recall and F1-measure (%) per reading level of the seven methods for readability assessment on both datasets when the labeling proportion is 0.7

support vector machine are used as the classifiers respectively.

For GRAW, we implement label propagation on both the merged graph G^c and the final graph G^{cf} (Section 3.3), denoted as GRAW$_c$ and GRAW$_{cf}$ respectively. Table 3 gives the average performance measure per reading level resulted by the implemented methods on both datasets. Unless otherwise specified, we fixed η to 3, and λ to 2800 for CPT and 2000 for ENCT. The proportion of the labeled (training) set is set to 0.7.

In Table 3, the precision, recall and F1-measure of all the seven methods are calculated per reading (grade) level on both English and Chinese datasets. The values marked in bold in each row refer to the maximum (best) measure gained by the methods.

From Table 3, the readability formulae (SMOG and FK) perform poorly on either the precision or recall measure, and their F1-measure values are generally the poorest. Both SMOG and FK are designed for English, and have acceptable performance on the English dataset ENCT. The unigram model (SUM) performs a little better than the readability formulae. On ENCT, It has relatively good performance on grade levels 1 and 4, while on the Chinese dataset CPT, the performance is not satisfactory. The feature-based methods (LR and SVM) perform well on both ENCT and CPT, which means both the text features developed and the classifiers trained are useful. In general, GRAW$_c$ performs better than both LR and SVM, which demonstrates the effectiveness of our method. In addition, by combining the feature-based graph (GRAW$_{cf}$), GRAW can be improved, and performs the best on all the three metrics over the majority of reading levels on both datasets. The only exception is on level 5 in CPT, which suggests the requirement of further improvements.

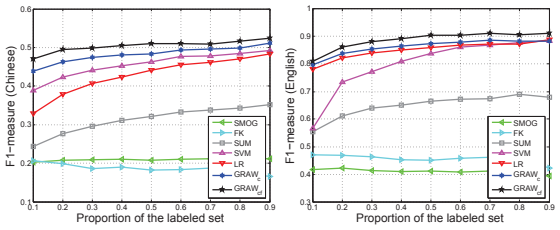
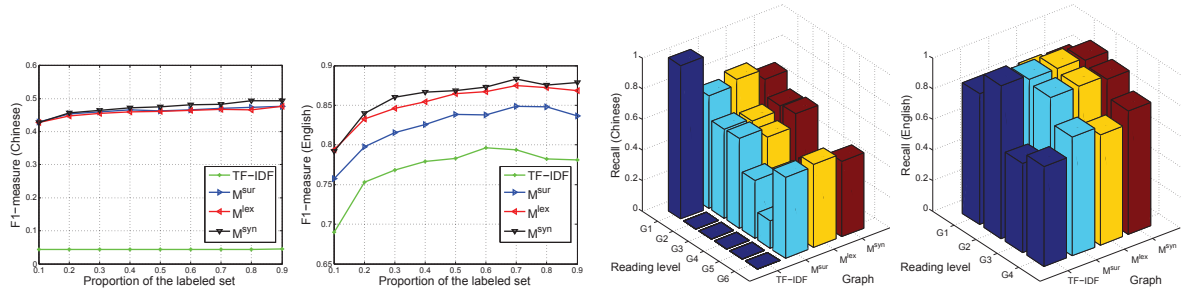


Figure 3: The average F1-measure of the seven methods on both datasets with the labeling proportion varied from 0.1 to 0.9

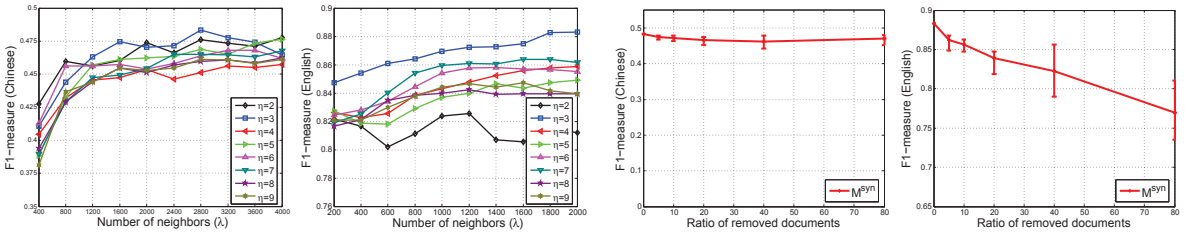
We study the effect of labeling proportion on the performance of these methods on both datasets. The F1-measure averaged over the reading levels is used, since it is a good representative of the three metrics according to Table 3. Figure 3 depicts the performance trends of all the methods.

From Figure 3, neither SMOG nor FK benefits



(a) Comparison of the average F1-measure between the coupling and general TF-IDF matrices

(b) Comparison of the Recall rate per reading level between the coupling and general TF-IDF matrices



(c) The effects of η and λ on the performance of the word coupling matrix

(d) The effect of corpus size on the performance of the word coupling matrix

Figure 4: Four perspectives on the effectiveness of the word coupling matrices

from the increasing size of the labeled set. This suggests that the performance of the readability formulae can hardly be improved by accumulating training data. The other 5 methods achieve better performance on larger labeled set, and outperform the two formulae even if the labeling proportion is small. Both LR and SVM perform better than SUM, but the performance is not good when the labeling proportion is less than 0.3, especially on the Chinese dataset. On the Chinese dataset, SVM performs better than LR, while on the English dataset, the situation is reversed. Both versions of GRAW outperform the other methods over the labeling ranges on both datasets. In addition, GRAW performs well when the labeling proportion is still small. Again, by combining the feature-based graph, the performance of GRAW is consistently improved.

In summary, GRAW can outperform the state-of-the-art methods for readability assessment on both English and Chinese datasets. By combining the feature-based graph, the performance of GRAW can be further improved.

4.3 Effects of the Word Coupling Matrix

For RQ2, we firstly compare the coupled bag-of-words model to the general model in the process of graph construction. Four graphs are built by using each of the three word coupling matrices (i.e. M^{sur}, M^{lex} and M^{syn}) and the TF-IDF matrix respectively. Label propagation is applied on each graph to predict reading levels of unlabeled documents. The labeling proportion is varied from 0.1 to 0.9 on both the English and Chinese datasets. Figure 4(a) depicts the average F1-measure resulted from the four graphs.

From Figure 4(a), the three word coupling matrices greatly outperform the TF-IDF matrix, especially on the Chinese dataset. This demonstrates that the word coupling matrices are very effective in improving the performance of the general bag-of-words model for readability assessment.

Secondly, we investigate the performance of the four matrices per reading level. Figure 4(b) depicts the recall rate per reading level of the four corresponding graphs in bar charts. The labeling proportion is set to 0.7. The recall rate is used because it makes the reason evident that the TF-IDF matrix performs poorly. From Figure 4(b), on the Chinese dataset, nearly all the unlabeled documents are classified as level 1 by the TF-IDF matrix, in which the word frequencies are too few to make meaningful discrimination among the reading levels. On the English dataset, the TF-IDF matrix performs better, but still prefers to classify documents into lower levels.

As described in Section 3.2.2, η (the number of difficulty levels of sentences) and λ (the number of neighbors pertained for each document node) are two parameters in building the word coupling

matrices. To investigate their effects on the performance of the built matrices, we vary the values of both η and λ, and compute the average F1-measure on the two datasets. Figure 4(c) depicts the results in line charts, where η varies from 2 to 9 step by 1, while λ varies from 400 to 4000 step by 400 on Chinese and from 200 to 2000 step by 200 on English (the difference is due to the dissimilar number of documents between the two datasets). The three word coupling matrices exhibit similar behavior during experiments, hence, only M^{syn} is depicted.

From Figure 4(c), a small η (e.g. 2 or 3) is good on the Chinese dataset. However, on the English dataset, $\eta = 2$ leads to the poorest performance. It seems the increasing of η causes vibrated performance, and the trend is further complicated when involving λ. Above all, $\eta = 3$ gives a preferable option on both datasets. For λ, most of the lines exhibit a similar trend that rises first and then keeps stable on both datasets, although some may drop when λ is too large. This suggests that making a relatively large number of the other words as the neighbors of one (i.e. $\lambda = 2800$ on the Chinese dataset and $\lambda = 2000$ on the English dataset) will make an effective word coupling matrix.

The word coupling matrix constructed in GRAW uses the whole corpus on either English or Chinese. To investigate if the corpus size takes effects on the performance of GRAW, we vary the proportion of the corpus used by randomly removing documents from each reading level. Figure 4(d) depicts the average F1-measure resulted by M^{syn}. The removing ratio is selected from $\{0, 0.05, 0.1, 0.2, 0.4, 0.8\}$. Both the mean values and deviations are shown on the line chart.

From Figure 4(d), on the Chinese dataset, the performance of GRAW suffers little from removing documents, even if only 20% documents are left for building the word coupling matrix. However, on the English dataset, the mean performance drops sharply and the deviation increases evidently. This suggests that cumulating sufficient corpus is required for building a suitable word coupling matrix in GRAW, and factors other than number of documents may influence the corpus quality, which deserves further study.

In summary, the word coupling matrix plays an essential role in GRAW. For building a suitable word coupling matrix, the number of difficulty levels of sentences (η) can be set to 3, and a relatively large number of the other words should be selected as the neighbors of a word. A sufficient corpus is required for refining the matrix.

4.4 Effectiveness of Graph Combination

For RQ3, we compare graphs built on each singular word coupling matrix (i.e. M^{sur}, M^{lex} and M^{syn}) to the merged graph (i.e. GRAW$_c$) and the combined graph (i.e. GRAW$_{cf}$). Figure 5 depicts the average F1-measure resulted after applying label propagation on these graphs with labeling proportion varied from 0.1 to 0.9. The feature-based graph (i.e. G^f) is also depicted for comparison.

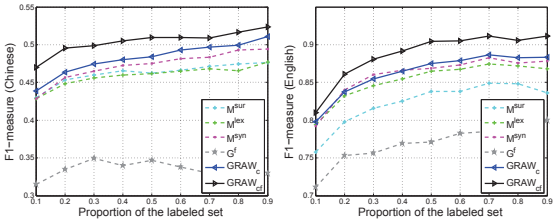


Figure 5: The average F1-measure of different types of graphs on the English and Chinese datasets

From Figure 5, the merged graph GRAW$_c$ outperforms the three basic graphs on both datasets in most cases. Within the three, M^{syn} performs best, especially on the English dataset, where it can outperform GRAW$_c$ slightly when the labeling proportion is small $(0.2 - 0.4)$. By combining the feature-based graph, GRAW$_{cf}$ performs even better on both datasets, although G^f performs poorest among all the graphs. In summary, the graph merging strategy is effective, and by combining the feature-based graph, the performance of GRAW can be improved. This demonstrates the potential of GRAW.

5 Conclusion

In this paper, we propose a graph-based readability assessment method using word coupling. The coupled bag-of-words model is designed, which exploits the correlation of readability among the words, and by applying TF-IDF, models the relationship among documents on reading levels. The model is employed in the graph-based classification framework for readability assessment, which involves graph building, merging, and label propagation. Experiments are conducted on both Chinese and English datasets. The results show that our method can outperform the commonly used

methods for readability assessment. In addition, the evaluation demonstrates the potential of the coupled bag-of-words model and the graph combination/merging strategies.

In our future work, we plan to verify the soundness of the results by applying our method on large volume corpus of both English and Chinese. In addition, we will investigate other ways of computing the word coupling matrices, such as incorporating word coherency or semantics, and develop efficient merging strategies which can be used for training classification models, as well as for building graphs.

Acknowledgments

This work was supported by the National NSFC projects under Grant Nos. 61373012, 61321491, and 91218302.

References

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

Xin Cheng, Duoqian Miao, Can Wang, and Longbing Cao. 2013. Coupled term-term relation analysis for document clustering. In *Proceedings of the 2013 International Joint Conference on Neural Networks*, pages 1–8. IEEE.

Kevyn Collins-Thompson and James P Callan. 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*, pages 193–200. Association for Computational Linguistics.

Samuel I Daitch, Jonathan A Kelner, and Daniel A Spielman. 2009. Fitting a graph to vector data. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 201–208. ACM.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics.

Thomas François and Cédrick Fairon. 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*, pages 466–477. Association for Computational Linguistics.

Itziar Gonzalez-Dios, Marıa Jesús Aranzabe, Arantza Dıaz de Ilarraza, and Haritz Salaberri. 2014. Simple or complex? assessing the readability of basque texts. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 334–344. Association for Computational Linguistics.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for german using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 1063–1080. Association for Computational Linguistics.

Tony Jebara, Jun Wang, and Shih-Fu Chang. 2009. Graph construction and b-matching for semi-supervised learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 441–448. ACM.

Zhiwei Jiang, Gang Sun, Qing Gu, and Daoxu Chen. 2014. An ordinal multi-class classification method for readability assessment of chinese documents. In *Knowledge Science, Engineering and Management*, pages 61–72. Springer.

Paul Kidwell, Guy Lebanon, and Kevyn Collins-Thompson. 2009. Statistical estimation of word acquisition with application to readability prediction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 900–909. Association for Computational Linguistics.

Doo Soon Kim, Kunal Verma, and Peter Z Yeh. 2013. Joint extraction and labeling via graph propagation for dictionary construction. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, pages 510–517.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Air Station, Memphis, TN.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

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

Natalia Ponomareva and Mike Thelwall. 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, pages 655–665. Association for Computational Linguistics.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing and management*, 24(5):513–523.

Sarah E Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics.

Amarnag Subramanya, Slav Petrov, and Fernando Pereira. 2010. Efficient graph-based semi-supervised learning of structured tagging models. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 167–176. Association for Computational Linguistics.

S. K. Michael Wong, Wojciech Ziarko, and P. C. N. Wong. 1985. Generalized vector space model in information retrieval. In *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 18–25. ACM.

Beverley L. Zakaluk and S. Jay Samuels. 1988. *Readability: Its Past, Present, and Future*. ERIC.

Xiaodong Zeng, Derek F Wong, Lidia S Chao, and Isabel Trancoso. 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*, pages 770–779. Association for Computational Linguistics.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University.