Linguistic Feature Representation with Statistical Relational Learning for Readability Assessment

Xinying Qiu[1] Dawei Lu[2] Yuming Shen[1*] Yi Cai[3]

[1] School of Information Science and Technology, Guangdong University of Foreign Studies
[2] School of Liberal Arts, Renmin University of China
[3] School of Software Engineering, South China University of Technology
ymshen2002@163.com

Abstract. Traditional NLP model for readability assessment represents document as vector of words or vector of linguistic features that may be sparse, discrete, and ignoring the latent relations among features. We observe from data and linguistics theory that a document's linguistic features are not necessarily conditionally independent. To capture the latent relations among linguistic features, we propose to build feature graphs and learn distributed representation with Statistical Relational Learning. We then project the document vectors onto the linguistic feature embedding space to produce linguistic feature knowledge-enriched document representation. We showcase this idea with Chinese L1 readability classification experiments and achieve positive results. Our proposed model performs better than traditional vector space models and other embedding based models for current data set and deserves further exploration.

Keywords: Linguistic Feature Embedding, Statistical Relational Learning, Readability Assessment

1 Introduction

Document-level readability assessment is an important research aspect in linguistic complexity for many different languages. It could be defined as measuring the comprehension difficulty perceived by humans when processing linguistic input at document level. The majority of machine-learning assessment methods are based on the framework of supervised learning with human-designed linguistic features [1]. Although such feature-driven classification models achieved some of the top performances that are hard to be transcended, their sparse and discrete characteristics did not take into consideration the latent relations among linguistic features. Recent development in readability assessment model learns word representation by encoding knowledge on word-level difficulty into word-embedding [2]. However, readability level does not reflect only at word-level complexity, but also at syntactic, structural, and discourse sophistication.

* Corresponding author: Yuming Shen (E-mail: ymshen2002@163.com)

We propose, therefore, to learn linguistic feature embedding models that cover four categories (i.e. shallow features, syntactic features, POS features, and discourse features) from their relation graphs to construct an enriched, dense, and low dimensional document representation for automatic readability assessment.

In particular, based on our observation on the linguistic feature data, we hypothesize that the linguistic features that demonstrate high impact on readability differences contain, among themselves, multiple types of correlation connected with latent linguistic factors. We illustrate our observation with the following examples:

Table 1. Examples of linguistic feature relations and their latent factors.

	Linguistic Feature 1	**Linguistic Feature 2**	**Correlation**	**Latent Factors**
1	Percentage of conjunctions	Average height of parse tree	Positive	Complex parse tree contains more conjunctions.
2	Average number of characters per word	Percentage of unique functional words	Negative	Length of Chinese functional word is short.
3	Number of punctuation clauses per sentence	Average number of unique idioms per sentence	Neutral	Unrelated

In the above table, the two linguistic features and their relation form a triplet that could be explained with the latent factors of linguistic implications. For example, documents of low readability may have more complex discourse structures such that the percentage of conjunctions and the average height of parse tree are both large because complex parse trees may contain more conjunctions. In the second example, since Chinese functional word is mostly composed of one or two characters, the higher the percentage of unique functional words within a document, probably the lower the average number of characters per word for that document. Both features may affect document readability but in different directions. We propose to automatically infer these latent factors and the existence of relationships among linguistic features by applying the latent feature model of Statistical Relational Learning (SRL) [3,4,5].

We showcase this linguistic feature embedding (LFE) model in the area of Chinese L1 readability assessment. By projecting the document representation vectors onto the space of linguistic feature embedding representation, we provide a linguistic knowledge-enriched and low-dimensional model that achieves better performance in readability prediction.

2 Related Research

In applying NLP technology for readability assessment, Sung (2015) evaluated 30 linguistic features and classification model using primary school text books in traditional Chinese used in Taiwan [6]. Jiang et al. (2014) proposed classification model and feature sets for readability prediction using L1 primary school text books in simplified

Chinese. However, the features developed for Chinese are far from enough [7]. In their following work, Jiang et al. (2015) model word representation with their difficulty distribution in sentences and proposed a graph-based classification framework with coupled bag-of-word model [8]. Recently, Jiang et al. (2018) incorporated word-level difficulty from three knowledge source into a knowledge graph and trained an enriched word embedding representation [2]. However, the word-level difficulty knowledge did not distinguish source knowledge for L1 and L2 instruction. Besides, differences in document-level readability is not reflected solely on word-level complexity, but also contain discrepancy in syntactic, discourse, and structural sophistication.

3 Methodology

In this work, we propose a new linguistic feature embedding model to construct document representation for readability assessment. Our overall research structure as illustrated in Figure 1 consists of four stages: feature design, feature relation graph and embedding learning, document representation, and readability classification. We discuss technology detail in this section.

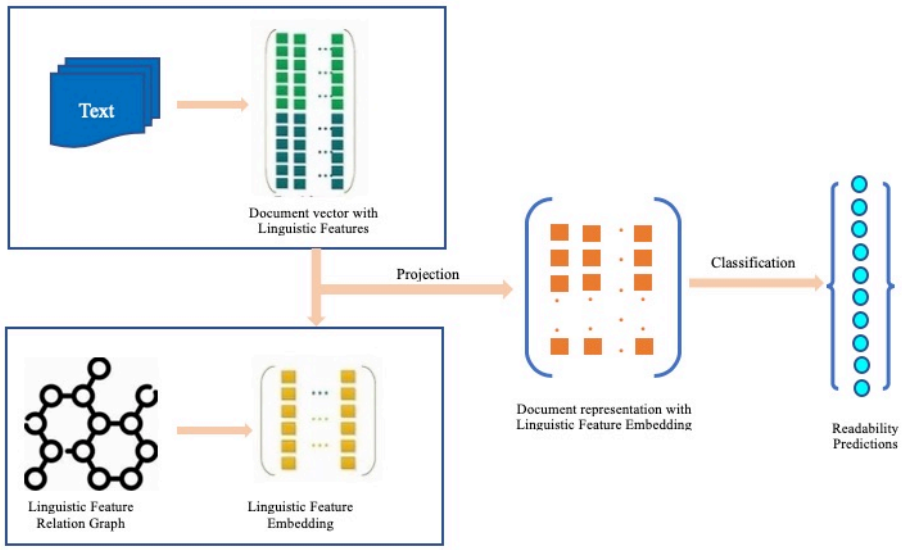

Fig. 1. Research structure

3.1 Linguistic Features

We designed 102 linguistic features of 4 categories: Shallow features, POS features, Syntactic features, and Discourse features. We cover the features used in Flesch index [9], Feng (2010) [10], Vajjala & Meurers (2012) [11], Todirascu (2016) [12], Qiu et al. (2017) [13] among many others as discussed in the Related Research section and adapted them for Chinese language. Please refer to Table 2 for feature descriptions.

For pre-processing, we use NLPIR[†] for word segmentation, LTP[‡] platform for POS tagging, and named entity recognition, and NiuParser[§] for syntactic parsing, grammatical labeling, and clause annotation.

Table 2: Summary of Linguistic Metrics

Feature category	Sub-category	Features used in metrics
Shallow Features	Character	common characters, stroke-counts
	Words	words of different character length,
	Sentence	sentence length by word count, n-gram count, and character count
	Document	document length by character count and symbol count
POS Features		adjective, functional words, verbs, nouns, content words, idioms, adverbs
Syntactic Features	Phrases	noun phrases, verbal phrases, prepositional phrases
	Clauses	independent clause, punctuation clause, dependency distance, word-count by punctuation clause
	Sentences	sentence count, parse tree height, sentence dependency distance
Discourse Features	Entity density	entities, named entities, entity nouns, named entity nouns
	Coherence	conjunctions, pronouns

3.2 Features Graph and Translation-based Method

In Statistical Relational Learning (SRL), the representation of an object can contain its relationships to other objects. Thus, the data is in the form of a graph, consisting of nodes (entities) and labelled edges (relationships between entities).

A feature graph is a multi-relational graph, composed of the linguistic features as nodes and three types of relations as edges: the positive, negative and irrelevant correlations. An instance of edge is a triplet of fact (*head feature*, *relation*, *tail feature*). For example, the triplet of fact (*percentage of conjunctions, positive correlation*, *average height of parse tree*), represents a relation type of positive correlation between two linguistic features as head and tail features respectively. While the triplet of fact (*average number of characters per word, negative correlation, percentage of unique functional words*) represents a negative relation with the two linguistic features as head and tail connected by relation edge. We speculate that we could infer such a multi-relational graph for the linguistic features that impact document-level readability differences.

[†] http://ictclas.nlpir.org/

[‡] http://www.ltp-cloud.com/

[§] http://www.niuparser.com/

The translation-based approach has been proposed to model multi-relational data, which attempts to embed a multi-relational graph into a continuous vector space while preserving certain properties of the original graph. Generally, each entity *h* (or *t*) is represented as a *k*-dimensional vector **h** (or **t**) and relation *r* is characterized by the translating vector **r**. For example, in **Trans E** [3]**,** given two entity vectors **h,t** and a translation vector **r** between them, the model requires **h**+**r**≈**t** for the observed triple (*h*,*r*,*t*). Hence, **Trans E** assumes the score function

$$f_r(h,t)=\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_2$$

is low if (*h*,*r*,*t*) holds, and high otherwise. To differentiate between correct and incorrect triples, **Trans E** score difference is minimized using margin based pairwise ranking loss. Formally, we optimize the following function:

$$\sum_{x\in S^+}\sum_{y\in S^-}\max(0, f(x)-f(y)+\gamma),$$

with respect to the entity and relation vectors. The γ is a margin separating correct and incorrect triples. S+ is the set of all positive triples, i.e., observed triples in the graph. The negative set S−are randomly corrupting the correct triples, that is, for a given correct triplet (h,r,t), a negative triplet (h',r,t')is obtained by randomly sampling a pair of entities (h',t') from S+. Then we get the set:

$$S^- = \{(h',r,t)\,|\,(h',r,t)\notin S^+\}\cup\{(h,r,t')\,|\,(h,r,t')\notin S^+\}$$

In this paper, we use **Trans E** to learn the linguistic feature embeddings on feature graphs.

3.3 Document Representation with Linguistic Embedding

With our hand-crafted linguistic features, we first design document representation model as vectors of discrete linguistic feature value. We also learn the embedding representation for each feature from their relation graph. Therefore, we project the document vectors onto the linguistic feature embedding space to obtain an enriched representation with linguistic feature embedding. Specifically, given a document vector $d = (x_1, x_2, \ldots x_n)$ where *n* is the number of linguistic features, x_i is the value of the i^{th} feature for document *d*, and a linguistic feature embedding matrix $L \in R^{n\times m}$ where *m* is the embedding dimension, we project the document vector *d* onto the linguistic feature embedding space by taking a vector-matrix multiplication to form a new representation as: $d_t = (l_1, l_2, \ldots l_m)$ where l_i is the projected value of linguistic features at dimension *i.*

4 Experiment

We evaluate our proposed Linguistic Feature Embedding (LFE) model from the following perspectives. **RQ1**: Whether LFE is effective for readability assessment, compared with using hand-crafted feature (HCF)? **RQ2**: Whether LFE can improve the performance of traditional readability assessment model with Bag of Word document representation? **RQ3**: Whether LFE is effective compared with other embedding-based

representation model? We showcase our evaluation of LFE in the area of Chinese L1 readability assessment.

4.1 Data

We provide a corpus for L1 readability assessment using textbooks from most widely used for primary school (Grades 1 through 6), secondary school (Grade 7 and 8), and high-school (Grade 10) education from three publishers, (i.e., People's Press, Jiangsu Education Press, and Beijing Normal University Press). We excluded playwrights, poetry, and classical literature to keep the genre of the text more simplistic and monotonous.

Table 3: Data Statistics

Grade	1	2	3	4	5	6	7	8	9	10	**Total**
# of Docs	93	147	164	157	148	163	96	138	94	32	1232
Percentage	7.6%	12.01%	13.4%	12.83%	12.09%	13.32%	7.84%	11.27%	7.68%	2.61%	100%

4.2 Learning Linguistic Feature Embedding

We use two types of feature graphs in the paper. The first type of feature graph is obtained by learning the positive, negative and irrelevant correlations among linguistic features of 4 categories. We set the positive correlation between two linguistic features if the Pearson correlation coefficient is above 0.7, the negative correlation if the coefficient is below -0.7 and the irrelevant correlation if the coefficient is between 0.7 and -0.7. The second type of feature graph is obtained by using human annotation for 4 categories linguistic features. In our experiments, we use 25 and 102 features of L1 for constructing the above two type of feature graphs. In training Trans E, the optimal parameters are determined by the validation set. After parameter tuning, we use the learning rate α for stochastic gradient descent at 0.01, the margin γ of 1, the embedding dimension k of 300, and batch size of 50.

4.3 Models and Experiment Setting

We have a total of 102 linguistic metrics for L1. We identify the features that are correlated with readability levels at 90%, 95%, and 99% confidence interval with linear regression. The 90% confidence interval gives us 25 out of 102. Our model comparisons with 95% and 99% confidence interval metrics produces similar results. We only present experimentation with the complete set of 102 features and the 90% interval set of 25 features. We use SVM and Logistic Regression as our multi-class classifiers to build predictive models for document-level readability.

To represent documents, we use the following approaches:

25HCF and 102HCF: We use only the scores of 25 linguistic features at 90% confidence interval to construct document vector representation. For 102HCF, we use the complete set of features.

25LFE and 102LFE: This is to represent documents by projecting 25-feature vectors onto the 25 linguistic feature embedding space learned with Trans-E. For 102LFE, we use the complete set of features.

25LFE-Anno: We have one of our coauthors, a linguistics Ph.D and professor, to manually annotate the pair-wise relations among the 25 linguistic features. We then use the annotated feature graph to learn feature embedding and then infer document representation by taking a vector-matrix multiplication.

BOW: This is the default baseline representation where each document is a vector of terms weighted with *ltc* variant of TF*IDF.

BOW+25HCF, BOW+25LFE, BOW+25LFE_Anno, BOW+25HCF+25LFE, BOW+25HCF+25LFE_Anno, BOW+102HCF, BOW+102LFE, BOW+102HCF+102LFE: We append different sets of feature representation to the BOW vector for each document.

W2V_Emb: We use the word embedding [14] of 300-dimension trained with Wikipedia to represent each word. We use the word vector average to represent each document.

CNNF: We trained a CNN model [15] for predicting readability with epoch of 100, batch size of 50, and learning rate of 0.1. We use the hidden layer output as features to represent document. For each document we have a 400-dimension vector representation.

For evaluating multi-class classification, we use Accuracy and Distance-1 Adjacent Accuracy. Adjacent-level Accuracy is often used in computational linguistics where predicting a text to be within one level of the true level label is still considered accurate [9]. According to our data distribution as shown in Table 4, the Majority Vote accuracies is 13.4%, and adjacent accuracy is 38.24%. With Uniform Random evaluation, we have a baseline of 10% accuracy and 30% adjacent accuracy. We perform 10-fold training-test cross-validation and paired two-tailed T-test for significance test.

4.4 Results and Analysis

To address RQ1, we compare hand-crafted feature representation and linguistic feature embedding as presented in Tables 4 and 5. We can see that compared with 25HCF, the LFE model performs significantly better with Logistic Regression (LR) in both accuracy and adjacent accuracy, and with SVM in adjacent accuracy. LFE performs similarly as 25HCF in accuracy with SVM classifier. Embedding learned with human annotation (25LFE-anno) performs significantly better than both hand-crafted model of 25HCF and 25LFE which is inferred with machine-learning. When experimenting with the complete set of 102 features, the LFE model performs significantly better HCF with both classifiers and for both accuracy and adjacent accuracy.

To address RQ2, we present results in Tables 6 and 7. Significant results are bolded. We can see that appending HCF features to the BOW vector achieves better results than BOW model alone. But, BOW+25LFE and BOW+102LFE performs even better than

augmenting with HCF. Furthermore, BOW+25LFE_Anno, which is embedding learned with human annotated relation provides us with the best performance in adjacent accuracy, and best accuracy when combined with 25HCF and BOW. We observe similar performance of LFE with 102-feature experiments.

Overall, LFE model alone achieves significantly better performance for readability assessment. When combined with other HCF and BOW representation, it also contributes to the improvement in predictive performance.

Table 4: Comparing Linguistic Feature Embedding (25LFE, 25LFE-Anno) with Hand-Crafted Feature (25HCF)

Representation Model	25 HCF		25 LFE		25 LFE-Anno	
Classifiers	**Accuracy**	**Adj. Accu.**	**Accuracy**	**Adj. Accu.**	**Accuracy**	**Adj. Accu.**
SVM	0.2637	0.6106	0.2638	**0.6127**	0.2627	0.6032
LR	0.2394	0.5815	**0.3259**	**0.7003**	**0.3381**	**0.71**

Table 5: Comparing Linguistic Feature Embedding (102LFE) with Hand-Crafted Feature (102HCF)

Representation Model	102 HCF		102 LFE	
Classifiers	**Accuracy**	**Adj. Accu.**	**Accuracy**	**Adj. Accu.**
SVM	0.2529	0.5398	**0.2538**	**0.6466**
LR	0.2818	0.6223	**0.3221**	**0.7063**

Table 6. Compare 25-feature LFE model with traditional models with BOW representation

Representation Model	**Classifiers**	**Accuracy**	**Adjacent Accuracy**
BOW (baseline)	SVM	0.377	0.7812
	LR	0.377	0.7546
BOW+25HCF	SVM	**0.385**	0.7764
	LR	**0.378**	**0.7569**
BOW+25LFE	SVM	**0.4055**	**0.8067**
	LR	**0.3983**	**0.7788**
BOW+25LFE_Anno	SVM	**0.409**	**0.8091**
	LR	**0.4004**	**0.7821**
BOW+25HCF+25LFE	SVM	**0.4021**	**0.8058**
	LR	**0.3998**	**0.7775**
BOW+25HCF+25LFE_Anno	SVM	**0.4104**	**0.8084**
	LR	**0.399**	**0.7774**

Table 7. Compare 102-feature LFE model with traditional models with BOW representation

Representation Model	BOW (baseline)		BOW+102HCF		BOW+102LFE		BOW+102HCF +102LFE	
Classifiers	SVM	LR	SVM	LR	SVM	LR	SVM	LR
Accuracy	0.377	0.377	**0.389**	**0.3817**	**0.4006**	**0.3908**	**0.3931**	**0.396**
Adjacent Accuracy	0.7812	0.7546	**0.7836**	0.7501	**0.8012**	**0.7768**	**0.7917**	**0.7659**

Figure 4 presents experiment results for RQ3, where we compare LFE with two popular word-embedding based representation with LR as the classifier. We can see that 25-LFE, 25LFE-Anno, and 102LFE all perform significantly better than other models based on CNN feature or word embedding feature.

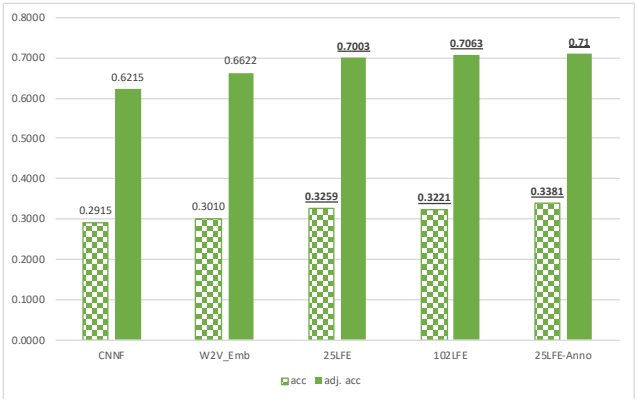

Fig. 4. Compare LFE model with other word-embedding models for readability assessment. Significant performance is bolded and underlined.

5 Conclusions

We present in this paper a model to learn distributed representation of linguistic features for readability assessment. Our assumptions include the following: 1) Distributed model could be extended to features beyond word-level differences for knowledge-enriched representation; 2) There may exist latent factors that connects linguistic features to form certain types of relationship; and 3) The similarities and inter-relations among the linguistic features and their membership to different feature categories demonstrate that the linguistic features possess the statistical properties of feature graphs of "homophily", "block structure" and "global and long-range statistical dependencies".

We propose to automatically infer the multi-relations among linguistic features and project the document representations onto the linguistic feature embedding space. We showcase the model implementation in the area of Chinese L1 readability assessment

with positive results. We hope to extend the current research on extra datasets and other types of latent factor models to refine and strengthen the linguistic knowledge informed representation models.

Acknowledgements. This work was supported by National Social Science Fund (Grant No. 17BGL068). We thank Prof. Jianyun Nie and anonymous reviewers for their valuable suggestions and thoughtful feedback. We thank undergraduate students Zhiwei Wu, Yuansheng Wang, Xu Zhang, Yuan Chen, Hanwu Chen, Licong Tan, and Hao Zhang for their helpful assistance and support.

References

[1] Collins-Thompson, K.; Callan, J. A language-modelling approach to predicting reading difficulty. In: Proceedings of HLT-NAACL. Boston. (2004)
[2] Jiang et al., “Enriching Word Embeddings with Domain Knowledge for Readability Assessment.” In: Proceedings of COLING 2018, pages 366–378, (2018)
[3]Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems. pages 2787–2795. (2013)
[4] Lise Getoor , Ben Taskar, Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning), The MIT Press, (2007)
[5] Judea Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan Kaufmann Publishers Inc., San Francisco, CA, (1988)
[6] Sung Y T, et al. Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. The Modern Language Journal 99(2): 371-391(2015).
[7] Jiang et al., “An Ordinal Multi-Class Classification Method for Readability Assessment of Chinese Documents.” R. Buchmann et al. (Eds.): KSEM 2014, LNAI 8793, pages. 61–72, (2014)
[8] Jiang et al. A graph-based readability assessment method using word coupling. In: Proceedings of EMNLP 2015, pages 411–420. (2015).
[9] Flesch R. A new readability yardstick. Journal of applied psychology, 32(3): 221 (1948).
[10] Feng L. Automatic readability assessment. Ph.D Thesis. The City University of New York, (2010).
[11] Vajjala and Meurers. On improving the accuracy of readability classification using insights from second language acquisition. In Proceedings of the ACL 2012 BEA 7th Workshop, pages 163–173. (2012)
[12] Todirascu A, et al. Are Cohesive Features Relevant for Text Readability Evaluation? In: Proceedings of COLING 2016, pages. 987-997. (2016).
[13] Qiu, X., et al. Exploring the Impact of Linguistic Features for Chinese Readability Assessment. In Proceedings of NLPCC, pages 771-783. (2017)
[14] Tomas Mikolov, et.al. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, 3111–3119. (2013)
[15] Kim, Yoon. "Convolutional Neural Networks for Sentence Classification." arXiv preprint arXiv:1408.5882 (2014)