Lexical and Syntactic features selection for an adaptive reading recommendation system based on text complexity

Mohamed Zakaria KURDI
Lynchburg College
Department of Computer Science
kurdi_m@lynchburg.edu

ABSTRACT
The goal of this work is to build a classifier that can identify text complexity in service of English as a Second Language (ESL) learners. In order to present language learners with texts that are suitable to their level of English, a set of features that can best describe the lexical and syntactic complexity of a given text were identified. Using a corpus of 355 texts which had already been classified into three different levels of difficulty, I built two different models with different sets of features. These models were tested using four popular machine learning algorithms. The recall of the SVM classifier, who achieved the best results, is 0.87 and its precision is 0.88.

Artificial intelligence→Natural language processing Information extraction • education →Interactive learning environments

Keywords
Intelligent Tutoring Systems, Text Mining, text reading, syntactic complexity, lexical complexity.

1. INTRODUCTION
The goal of this paper is to present an Intelligent Tutoring System ITS designed to develop the reading comprehension skill for English as a Second Language (ESL) learners. Written language understanding is an important skill where lexicon, syntax, semantics, and discourse are involved.

A typical usage scenario of the system presented here would be a learner having passed a pretest to determine his English level. Based on these results, the learner is presented with appropriate texts retrieved from the web (not necessarily written for educational purposes) corresponding to his level. To automatically determine the level of a text, we need to identify and extract those features that are most relevant to classification. As a first step, I experiment with existing educational texts since these have the advantage of being already classified by level.

After reviewing the literature, I will present and discuss the corpus I used in this paper. The proposed or adopted lexical and syntactic features will also be presented and discussed. Finally, I'll present the experiments and discuss their results.

2. RELATED WORK
Researchers from different disciplines proposed different frameworks to classify or score written texts according to their complexity.

In the area of language acquisition and psycholinguistics, Scarborough's Index of Productive Syntax (IPSyn) [17] was the one that attracted much of the attention from the Natural Language Processing community. IPSyn is a grammatical measure designed to show the individual differences in the acquisition of syntax. It includes sixty syntactic structures organized into four groups: noun phrases (N), verb phrases (V), questions and negations (Q), and sentence structures (S). IPSyn was implemented to automatically score children production using Charniak's statistical parser [16], [8]. In addition, within the framework of automatic scoring of children production, Gabriella Ramirez and her collaborators [15] proposed some statistical measures based on word class n-grams.

Within the framework of foreign language teaching, [19], [10] adopted Latent Semantic Analysis (LSA) to recommend appropriate reading material based on the relationships between the lexicon of the texts read by the learner and existing text candidates. In addition, several researchers have focused on the automatic assessment of ESL learners' written essays. Developed in the mid-nineties by Educational Testing Service (ETS), [2], e-rater combines syntactic criteria with discourse to detect and score abrupt shifts in topicality. Another system that is worth to mention is the *Intelligent Essay Assessor*. It evaluates the content of essays based on LSA. It also assesses syntactic structures and style based on statistical measures [6]. Based on previous studies on syntactic complexity measures and their relation to foreign language proficiency [14], [18], Xiofei Lu implemented a system for automatic analysis of syntactic complexity in second language writing [12]. To automatically score non-native speech, [3] collected and implemented a set of 17 key features, based on human-rated learners' transcriptions. They tested the correlation of five different models that cover each a different set of features and got encouraging results.
As we can see in the literature review, most of the previous work focused on using some features to solve slightly different problems: essay scoring, or measuring the development of children's language. In addition, I am experimenting with some new features such as the frequency of the word in the language, verb tenses, or percentage of discourse connectors that have not been used, to the extent of my knowledge, as features in a previous work about text classification.

3. Corpus
I collected a corpus of 355 texts of English from free professional websites[1]. The texts provided in these websites are organized by level. This made it easy to make three distinct groups for each complexity level. The distribution of the texts over the levels is as follows: 123 texts categorized as level one, 118 level two texts, and 114 level three. Texts of higher levels are usually of larger size. This makes the collections of inferior level smaller in size despite their larger number. The overall size of the corpus is 526 KB.

[1] I collected the texts from the following websites: http://www.newsinlevels.com/home/productscbm_550428/50/ http://learnenglishteens.britishcouncil.org/study-break/easy-reading and http://linguapress.com/inter.htm

ICISDM 2017, April 1–3, 2017, Charleston, SC, USA.

4. LEXICAL FEATURES

4.1 Frequency of the words in the language

A word frequency score obtained from a large-scale corpus provides information about how common a word is, and here I used it as an indicator of vocabulary level. The more common the words in a text are, the easier is the text.

To avoid the bias introduced by grammatical words (determiners, copulas, adverbs of degree, etc.), I removed the stop words by using the NLTK list of stop words. Using NLTK's tagger, I also filtered proper nouns. Furthermore, I stem the words using the porter stemmer. This helps match the plural of the regular forms (e.g. book, books). However, this does not eliminate the irregular forms such as (make, made). This fits perfectly with the purpose of this score as a learner of English is supposed to know the simple plural forms of nouns but not necessarily the irregular forms of some nouns and verbs.

I then calculated the score based on the remaining words using the equation (1).

$$\frac{\sum_{i=1}^{n} freq(word\ i)}{n} \quad (1)$$

In the equation (1), *n* is the number of words in the text that remain after filtering. The frequency of every word is obtained from a freely available list[2] of the 5000 most frequent words in English. This list is based on a large corpus of contemporary English [4]. 77% of the words in the level one texts are included in this list. Only 66% of the words in the third levels are included. Furthermore, 71.50% of the texts of level 2 are covered. This shows that this feature is promising. The words that are not in the list are given the same frequency value. This value is lower than the lowest score in the list.

As shown in table 1, the correlation of this feature is -0.42. It is negative because the less common the vocabulary (smaller frequency coefficient) the higher is the level of the text.

4.2 Percentage of discourse connectors

Discourse connectors, such as *therefore*, *hence*, *or*, etc., are an indication of longer and elaborate sentences as well as an advanced structure of the text. I use equation (2) to calculate the percentage of discourse connectors based on the counts of the words and the connectors in the text.

$$\frac{\#\text{discourse connectors}}{\#\text{ words}} \quad (2)$$

As we can see in table 1, the correlation of this feature is both positive and statistically significant.

4.3 Diversity of lemmas

The basic assumption here is that an advanced text would use more diverse lemmas than an elementary one. A lemma being a simplified form of a word (see [11] chapter 3, for a detailed presentation and discussion of this concept). As indicated in table 1, the correlation of this lemma diversity feature is positive but significantly lower than the other two lexical features. A possible reason for this difference could be that the authors of educational texts try to focus the vocabulary regardless of the level.

Table 1. Correlation of the three lexical features

Phenomenon	Correlation
Frequency of the words in the language	-0.42
Percentage of discourse connectors	0.36
Diversity of the lemmas	0.14

5. SYNTACTIC FEATURES

To extract syntactic features, I used two sets of freely available tools. For parsing, I used the Stanford Parser [9]. I also used some tools from NLTK such as the sentence tokenizer and the POS tagger [1].

5.1 Verb tenses

Verb tenses are one of the most fundamental elements of the grammar of a language that the learner should master. Classic grammar books, classify these verbs according to complexity levels. For example, simple present and simple past are usually covered in level one while future perfect continuous is covered in level three or even later. In this study, I will examine how these verb forms can be used as features to describe the complexity of a text.

I implemented a module that uses regular expressions of POS tags sequences to recognize 13 different verb forms. The recall of this module is 0.92 while its precision is 0.90. Most of the errors are due to issues with the POS taggers. For example, when the tagger mistakenly tags *spirit* as a verb rather than a noun. it is inserted as simple past verb[3]. [… ('in', 'IN'), ('the', 'DT'), ('Christmas', 'NNP'), **('spirit', 'VBD')]**. Please note that not all tagging errors involving a verb lead systematically to verb tense categorization issues.

Table 2. Verb tenses with a correlation equal or superior to two

Tense	Correlation
Simple present	-0.61
Simple past	0.48
Present perfect	0.28
Past perfect	0.24
Infinitive	0.20

After measuring the correlations, I found that most of the verb forms have a zero or near-zero correlation. As presented in table 2, only five verb forms have a significant correlation. The simple present has a negative correlation because this tense is more common in lower level texts.

To understand the weak correlation of some verb forms, it is useful to observe the data in table 3. Given the sparseness of usage of some verb forms, their correlation was not statistically significant. For example, future perfect, future continuous, and, future perfect continuous were rarely present.

[2] http://www.wordfrequency.info/free.asp

[3] For a presentation of the Penn tag set used in the tagged example, please refer to the following link: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Table 3. Verb tenses frequency of usage in the corpus

Tense	% of texts	Tense	% of texts
simple present	0.96	future continuous	0.01
simple past	0.74	past continuous	0.20
present perfect	0.23	present perfect continuous	0.04
present participle	0.18	future perfect continuous	0
present continuous	0.19	past perfect	0.025
simple future	0.20	infinitive	0.81
future perfect	0.01		

5.2 Features related to phrases and clauses

Despite the inevitable parsing errors, we can see in table 5 that the correlations of these features are generally high. This confirms the high correlation of these features as observed on spoken productions by [3].

Table 4. Correlation of features related to phrases and clauses

Phenomenon	Correlation
Mean number of verb phrases per sentence	0.62
Mean number of prepositional phrases per sentence	0.72
Mean lengths of dependent clauses	0.38
Mean lengths of phrases	0.60
Mean number of phrases per sentence	0.75
Mean number of dependent clauses per sentence	0.35

5.3 Other syntactic features

As we saw in the literature review, some previous studies tried to capture different aspects of syntactic complexity through n-gram based models [15]. Here, I propose a simple count of the number of bi-grams, tri-grams and four-grams of part of speech sequences within a sentence. The idea here is that the more diversified are the sequences of tags, and richer the syntactic structures the more advanced is the text.

Table 5. Correlation of general syntactic structure features

Phenomenon	Correlation
Mean number of bi-grams per sentence	-0.01
Mean number of tri-grams per sentence	0.11
Mean number of four-grams per sentence	0.20
Mean lengths of sentences	0.76

As we can see in table 5, only the means of four-grams have a statistically significant correlation to text complexity. A possible interpretation for this difference is that four-grams cover a bigger syntactic window. Therefore, they cover more syntactic phenomena.

The length of a sentence is a simple but good indicator of the complexity of its structure as well as the diversity of the words used in it. As shown in table 5, the correlation of this feature is high which confirms its importance as described by the previous studies who adopted it [3], [12].

6. EXPERIMENTS AND RESULTS

I built and tested two models with two different sets of features. Model one contains the 17 features with a correlation equal to or higher than two. Model 2, has nine features each with a correlation equal to or higher than 4.

I used the open source machine learning and data visualization tool Orange [5]. I experimented with four popular machine learning algorithms: Support Vector Machine (SVM), random forest, naïve Bayes, and decision tree. I used the following parameters in orange for these algorithms. For SVM, the parameters are as follows. SVM type: V-SVM, kernel: Radial Basis Function (RBF), Iteration limit: 50 and numerical tolerance: 0.001010. For random forest, the parameters are the number of trees: 15, and limit of splitting subsets: 5. The parameters for decision tree are the number of instance leaves: 2, the limit of splitting subsets: 10, stop when majority reaches 90%, and the maximum tree depth: 10.

Random sampling was used with three repetitions. Given the limited size of the used data, data was split into two sets: 80% for the training set and 20% for the test set. The results of the experiments are provided in table 6. To measure the performance of the different algorithms, I adopted the following measures: recall, precision and F-score. For a detailed introduction to these measures, please refer to [13] chapter 8.

Table 6. Text classification results

Model	Classifier	Recall	Precision	F-score
Model 1	SVM	0.845	0.850	0.847
	Random forest	0.831	0.837	0.833
	Naïve Bayes	0.803	0.806	0.804
	Decision tree	0.793	0.802	0.796
Model 2	SVM	0.873	0.880	0.875
	Random forest	0.845	0.847	0.846
	Naïve Bayes	0.812	0.817	0.814
	Decision tree	0.789	0.795	0.791

7. DISCUSSION

As shown in table 6, the difference in terms of performance between the two models is not big with a slight improvement with model 2 where we have fewer features with a higher correlation.

SVM outperforms the other learning algorithms for this task. The difference between the SVM and Naïve Bayes and decision tree is significant. However, the difference in terms of performance between SVM and the random forest is smaller. As an advanced learning algorithm, SVM is known for its good capacity to generalize and avoid local maxima/minima, this explains the superior performance of this algorithm. Compared to the decision trees, the random forests make it possible to average multiple deep decision trees to reduce the variance and consequently boost the performance. This explains the difference in terms of performance between the random forest and the decision tree. Being a probabilistic model, naïve Bayes is known to be sensitive to the size of the training data which explains its lower performance.

8. CONCLUSION AND PERSPECTIVES

In this paper, I proposed some new features for text classification by complexity level. I also explored the usage of some existing features that were proposed for other tasks. I experimented with two models based on two different sets of features using four popular learning algorithms. The results showed small differences between the two models but important differences in performance between the learning algorithms, with SVM being the best with both models.

The overall performance is encouraging. However, some improvements are possible. Aside from the need of using a larger

corpus, a possible path for improvements could be to explore other syntactic features such as T-unit. T-unit is defined as the shortest grammatically allowable sentence into which writing can be split[4]. It has been utilized extensively as a mean to measure syntactic complexity [7], [3]. Another feature that is worth to explore is the lexical variation over the n-grams that was proposed by [15].

I adopted here a static approach based on text classification. In the future, I plan to use this work as a part of a dynamic system that would propose the best-suited text based on users' profiles. In this scenario, the features of a read text would be used to update the learner's profile which would be the cornerstone in selecting the next text to read.

9. REFERENCES

[1] Bird, Steven, Klein, Ewan, Loper, Edward. 2009. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, O'reilley, ISBN-10: 0596516495.

[2] Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. 1998. Automated essay scoring using a hybrid feature identification technique. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computation Linguistics, pp. 206-210.

[3] Chen, Miao, Zechner, Klaus. 2011. Computing and Evaluating Syntactic Complexity Features for Automated Scoring of Spontaneous Non-Native Speech. 49th annual meeting of the ACL pp 722-731. Portland, Oregon. June 9-24.

[4] Davies, Mark. 2009. The 385+ million word *Corpus of Contemporary American English* (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14.2: 159-190.

[5] Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina M, Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M., Zupan, B. 2013. Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research 14(Aug):2349−2353.

[6] Foltz, P. W., Laham, D., & Landauer, T. K. 1999. Automated essay scoring: Applications to educational technology. Proceedings of the ED-MEDIA 1999 World Conference on Educational Multimedia, Hypermedia, and Telecommunications.

[7] Gaies, S. J., T-Unit analysis in second language research: applications, problems, and limitations, TESOL quarterly 14, 53-60.

[8] Hassanali, Khairun-nisa, Liu, Yang, Iglesias, Aquiles, Solorio, Thamar, Dollaghan, Christine A. 2013. Automatic Generation of the Index of Productive Syntax For Child Language Transcripts, Behavior Research Methods, 46(1), pp. 254-262.

[9] Klein, Dan, Manning, Christopher D. 2003. Fast Exact Inference with a Factored Model for Natural Language Parsing. In Advances in Neural Information Processing Systems 15 (NIPS 2002), Cambridge, MA: MIT Press, pp. 3-10.

[10] Kurdi, Mohamed Zakaria, 2011. Personalized language learning through adaptive Computer software: application to French. International Conference for Computer Applications ICCA. 20-23 May.

[11] Kurdi, Mohamed Zakaria. 2016. Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax. ISTE-Wiley. ISBN-10: 1848218486.

[12] Lu, Xiofei. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474-496.

[13] Manning, Christopher D., Schütze, Hinrich, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999, ISBN 0-262-13360-l.

[14] Ortega, L. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. Applied Linguistics, 24 (4), 492-518.

[15] Ramirez de la Rosa, Gabriela, Solorio, Thamar, Montes-y-Gomez, Manuel, Tonantzintla, Sta. Maria, Liu, Yang, Iglesias, Aquiles, Bedore, Lisa, Pena, Elizabeth. 2013. Exploring word class n-grams to measure language development in children, Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013), pages 89–97, Sofia, Bulgaria, August 4-9.

[16] Sagae, Kenji, Lavie, Alon, MacWhinney, Brian. 2005. Automatic parsing of parental verbal input. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, Michigan — June 25 - 30, , pp 197-204.

[17] Scarborough, H. S. 1990. Index of Productive Syntax. Applied Psycholinguistics,11, 1-22.

[18] Wolfe-Quintero, K., Inagaki, S., Kim, H.-Y. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. Honolulu, HI: University of Hawaii Press.

[19] Zampa, V., Lemaire, B. 2002. Latent Semantic Analysis for Student Modeling, Journal of Intelligent Information Systems, Special Issue on Education Applications 18(1), 15-30.

[4] https://en.wikipedia.org/wiki/T-unit

Authors' background

Your Name	Title*	Research Field	Personal website
M. Zakaria Kurdi	Assistant Professor	NLP, Text Mining, Intelligent Tutoring Systems	http://www.lynchburg.edu/academics/majors-and-minors/computer-science/faculty-and-staff/m-zakaria-kurdi/