Grammatical Templates: Improving Text Difficulty Evaluation for Language Learners

Shuhan Wang
Department of Computer Science
Cornell University
forsona@cs.cornell.edu

Erik Andersen
Department of Computer Science
Cornell University
eland@cs.cornell.edu

Abstract

Language students are most engaged while reading texts at an appropriate difficulty level. However, existing methods of evaluating text difficulty focus mainly on vocabulary and do not prioritize grammatical features, hence they do not work well for language learners with limited knowledge of grammar. In this paper, we introduce *grammatical templates*, the expert-identified units of grammar that students learn from class, as an important feature of text difficulty evaluation. Experimental classification results show that grammatical template features significantly improve text difficulty prediction accuracy over baseline readability features by 7.4%. Moreover, we build a simple and human-understandable text difficulty evaluation approach with 87.7% accuracy, using only 5 grammatical template features.

Keywords text difficulty evaluation, education, grammatical templates, language learners.

1 Introduction

Evaluating *text difficulty*, or *text readability*, is an important topic in natural language processing and applied linguistics (Zamanian and Heydari, 2012; Pitler and Nenkova, 2008; Fulcher, 1997). A key challenge of text difficulty evaluation is that linguistic difficulty arises from both vocabulary and grammar (Richards and Schmidt, 2013). However, most existing tools either do not sufficiently take the impact of grammatical difficulty into account (Smith III et al., 2014; Sheehan et al., 2014), or use traditional syntactic features, which differ from what language students actually learn, to estimate grammatical complexity (Schwarm and Ostendorf, 2005; Heilman et al., 2008; François and Fairon, 2012). In fact, language courses introduce grammar constructs together with vocabulary, and grammar constructs vary in frequency and difficulty just like vocabulary (Blyth, 1997; Manzanares and López, 2008; Waara, 2004). Ideally, we would like to have better ways of estimating the grammatical complexity of a sentence.

To make progress in this direction, we introduce *grammatical templates* as an important feature in text difficulty evaluation. These templates are what language teachers and linguists have identified as the most important units of grammatical understanding at different levels, and what students actually learn in language lessons. We also demonstrate that grammatical templates can be automatically extracted from the dependency-based parse tree of a sentence.

To evaluate, we compare the difficulty prediction accuracy of grammatical templates with existing readability features in Japanese language placement tests and textbooks. Our results show that grammatical template features slightly outperform existing readability features. Moreover, adding grammatical template features into existing readability features significantly improves the accuracy by 7.4%. We also propose a multilevel linear classification algorithm using only 5 grammatical features. We demonstrate that this simple and human-understandable algorithm effectively predicts the difficulty level of Japanese texts with 87.7% accuracy.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1692–1702, Osaka, Japan, December 11-17 2016.

2 Related Work

Text difficulty evaluation has been widely studied over the past few decades (Nelson et al., 2012; Sinha et al., 2012; Hancke et al., 2012; Jameel et al., 2012; Gonzalez-Dios et al., 2014; Sinha et al., 2014). Researchers have developed over 200 metrics of text difficulty (Collins-Thompson and Callan, 2004). For example, *Lexile* measures text complexity and readability with word frequency and sentence length (Smith III et al., 2014). *ATOS*[1] includes two formulas for texts and books, both of which take into account three variables to predict text difficulty: word length, word grade level and sentence length. *TextEvaluator* is a comprehensive text analysis system designed to help teachers and test developers evaluate the complexity characteristics of reading materials (Sheehan et al., 2014). It incorporates more vocabulary features, such as meaning and word type, as well as some sentence and paragraph-level features.

Nevertheless, most of these methods provide limited consideration of grammatical difficulty, which is a major challenge for foreign language learners (Callan and Eskenazi, 2007). In fact, text readability not only depends on sentence lengths or word counts, but on 'the grammatical complexity of the language used' as well (Richards and Schmidt, 2013). Based on this fact, recent readability evaluation systems improved performance by incorporating syntactic features like parse tree depth (Schwarm and Ostendorf, 2005) and subtree patterns (Heilman et al., 2008) to measure grammatical complexity. Moreover, researchers have developed an unified framework of text readability evaluation, which combines lexical, syntactic and discourse features, and predicts readability with outstanding accuracy (Pitler and Nenkova, 2008). The relationship between text readability and reading devices was also studied in the past two years (Kim et al., 2014). However, most of these approaches are intended for native speakers and use texts from daily news, economic journals or scientific publications, which are too hard to read for beginning and intermediate language learners. Ideally, we would have specific features and approaches for text difficulty evaluation for language learners.

Recently, language educational researchers conducted a bunch of studies on text readability evaluation for language learners in different languages, such as English, German, Portuguese and French (Blyth, 1997; Waara, 2004; Manzanares and López, 2008; Vajjala and Meurers, 2012; François and Fairon, 2012; Xia et al., 2016). However, they use traditional syntactic features such as sentence length, part of speech ratios, number of clauses and average parse tree height, which differ from the grammatical knowledge that students actually learn in language lessons. For example, Curto et al. measured text difficulty using traditional vocabulary and syntactic features, to predict text difficulty levels for Portuguese language learners (Curto et al., 2015). Unfortunately, 75% accuracy in 5-level classification with 52 features is not satisfactory. Instead, we extract grammatical features from *grammatical templates*, the knowledge units that language students actually learn in classes and that expert language instructors have identified and highlighted in textbooks. We also propose a novel technique that has a simpler and human-interpretable structure, uses only 5 grammatical template features, and predicts text difficulty with 87.7% accuracy in 5-level classification.

3 Grammatical Template Analysis

A key challenge in modeling text difficulty is to specify all prerequisite knowledge required for understanding a certain sentence. Traditional methods measure text difficulty mostly by evaluating the complexity of vocabulary (word count, word frequency, word type, etc.). This is effective for native speakers, who typically understand the grammar of their language but vary in mastery of vocabulary. However, these vocabulary-based methods underperform for language learners who have limited knowledge of grammar (Callan and Eskenazi, 2007; Curto et al., 2015).

To resolve this, we focus our research on grammatical difficulty. We introduce the idea of *grammatical templates*, units of grammar that expert language instructors and linguists have identified as the most important grammatical knowledge, and are typically emphasized as key points in every textbook lesson (Banno et al., 2011; People's Education Press, 2013). Since these grammatical templates are

[1] http://www.renaissance.com/Products/Accelerated-Reader/ATOS/ATOS-Analyzer-for-Text

taught explicitly in language lessons and learned directly by language students, we believe they reflect the conceptual units of grammar more closely than parse trees.

Grammatical templates play an important role in language understanding because:

- Many grammatical templates suggest sentence structure. For example, "hardly ... when ..." in English, "nicht nur ..., sondern auch ..." (not only ... but also ...) in German, and "必ずしも ... とはいえない" (it is not necessarily the case that ...) in Japanese;
- For languages like Chinese and Japanese, lacking knowledge of some grammatical templates will cause difficulties in segmentation. For example, consider the Japanese template "...つ...つ" (two opposite behaviors occuring alternately) in the phrase "行きつ戻りつ" (to walk back and forth), and the Chinese template "越...越好" (the more ... the better) in "越早越好"(the earlier the better);
- Some grammatical templates may refer to special meanings that cannot be understood as the combination of individual words. For example, "in terms of", "such that" in English, "mit etwas zu tun haben" (have something to do with ...) in German, and "... ことはない" (no need to ...) in Japanese.

We show some simple examples of grammatical templates for Japanese in Table 1[2]. Line 2 shows the pronunciation of the templates, line 3 shows the translations, and the uppercase letters in line 4 are provided for notation. We also provide examples of how the grammar of a sentence can be described as combinations of these grammatical templates in Table 2.

3.1 Difficulty Evaluation Standard

To evaluate the difficulty of texts and grammatical templates, we follow the standard of the Japanese-Language Proficiency Test (JLPT). The JLPT is the most widely used test for measuring proficiency of non-native speakers, with approximately 610,000 examinees in 62 countries and areas worldwide in 2011[3]. It has five different levels, ranging from from N5 (beginner) to N1 (advanced). A summary of the levels can be found at JLPT website [4].

3.2 Grammatical Template Library

Due to their significance in Japanese education, grammatical templates are well-studied by Japanese teachers and researches. Grammatical templates are summarized and collected for both Japanese learners (common templates) and native speakers (templates used in very formal Japanese or old Japanese). We referenced 3 books about grammatical templates for Japanese learners (Sasaki and Kiko, 2010; Xu and Reika, 2015; Liu and Ebihara, 2012), all of which divide their templates into N1-N5 levels, for generating our template library at each corresponding level.

Although not common, books may have different opinions on the difficulty of the same template. For example, an N1 template in book A may be recognized as an N2 template in book B. In order to conduct our experiments on a reliable template library, we only pick the templates recognized as the same level by at least two of the three books. For example, if both book A and C recognized template t as an N3 template, we can incorporate template t into our N3 template library. Ultimately, we collected 147 N1 templates, 122 N2 templates, 74 N3 templates, 95 N4 templates and 128 N5 templates in our library. All selected grammatical templates are stored in the format of regular expressions for easy matching in parse trees.

3.3 Grammatical Template Extraction

The framework of grammatical template extraction is shown in Algorithm 1. The program requires the dependency-based parse tree of a sentence as input, runs from bottom to top and returns a set of

[2]A long list of Japanese grammatical templates with English translations can be accessed at the JGram website: http://www.jgram.org/pages/viewList.php. There is also a nice and comprehensive book of Japanese grammatical templates, written by Japanese linguists, with English, Korean and Chinese translations: (Tomomatsu Etsuko and Masako, 2010).

[3]http://www.jlpt.jp/e/about/message.html

[4]http://www.jlpt.jp/e/about/levelsummary.html

Template	– は	– の	– を	– では ない	–(名詞) に	–(動詞連用形) に
Pronunciation	– *wa*	– *no*	– *o*	– *dewa nai*	–(noun) *ni*	–(verb, i-form) *ni*
Translation	(topic)	(genitive)	(object)	is not	to (location)	for (purpose)
Notation	A	B	C	D	E	F

Table 1: Grammatical Templates in Japanese, with hyphens denoting words to be filled in. Note that some grammatical templates may impose requirements of some properties (e.g. part of speech or form) on the missing words.

Sentence	彼	**は**	すぐ	東京	**に**	到着する	
Pronunciation	kare	**wa**	sugu	*toukyou*	**ni**	touchakusuru	
Translation	he	**(topic)**	soon	*Tokyo*	**to (location)**	arrive	
Templates		A			E		
	“ he will soon arrive in Tokyo ”						
Sentence	僕	**の**	彼女	**を**	見	**に**	行く
Pronunciation	boku	**no**	kanojo	**o**	*mi*	**ni**	iku
Translation	I	**(genitive)**	girlfriend	**(object)**	*see*	**for (purpose)**	go
Templates		B		C		F	
	“ I go to see my girlfriend ”						
Sentence	これ	**は**	君	**の**	本	**では**	**ない**
Pronunciation	kore	**wa**	kimi	**no**	hon	**dewa**	**nai**
Translation	this	**(topic)**	you	**(genitive)**	book	**is**	**not**
Templates		A		B		D	
	“ this is not your book ”						

Table 2: Identified grammatical templates of Japanese sentences. In sentences, pronunciations and translations, grammatical templates are in bold. The word *toukyou* in the first sentence is a noun (Tokyo,東京), as characterized by template E. The word *mi* (to see, 見) in the second sentence is the i-form (動詞連用形) of a verb, as required by template F.

	N1 Texts	N2 Texts	N3 Texts	N4 Texts	N5 Texts
N1 Templates	0.902%	0.602%	0.077%	0.074%	0.056%
N2 Templates	2.077%	1.571%	1.072%	0.298%	0.056%
N3 Templates	4.070%	3.679%	1.531%	0.894%	0.222%
N4 Templates	16.635%	15.449%	13.323%	12.071%	1.832%
N5 Templates	76.316%	78.699%	83.997%	86.662%	97.834%

Table 3: Distribution of grammatical templates of level N1(hard)-N5(easy)

	N1 Texts	N2 Texts	N3 Texts	N4 Texts	N5 Texts
N1 Templates	3.536	2.342	0.295	0.230	0.146
N2 Templates	8.141	6.110	4.130	0.922	0.146
N3 Templates	15.954	14.308	5.900	2.765	0.582
N4 Templates	65.214	60.081	51.327	37.327	4.803
N5 Templates	299.178	306.059	323.599	267.972	256.477

Table 4: Number of templates of level N1(hard)-N5(easy) per 100 sentences

all identified grammatical templates $\mathbf{T}(node_0)$. Line 7 extracts the templates in the children of $node_0$ (and ignores the descendants of the children), by matching the phrase associated with the child nodes $[node_1, node_2, \cdots]$ to all templates stored in terms of regular expressions in our library. The matching is based on both the structure of the phrases and the properties of the words. Line 8 shows $\mathbf{T}(node_0)$ covers all templates identified in subtrees rooted at $node_0$'s children and the templates extracted in the phrase associated with the child nodes $[node_1, node_2, \cdots]$.

Algorithm 1 Grammatical Progression Extraction

Require: A dependency-based parse tree of the sentence

Ensure: $\mathbf{T}(node_0)$ = set of identified grammatical templates in (sub)parse tree rooted at $note_0$.

1: **if** $node_0$ is leaf node **then**
2: return $\mathbf{T}(node_0) = \{\}$
3: **end if**
4: $node_1, node_2, \cdots \leftarrow$ children of $node_0$
5: Calculate: $\mathbf{T}(node_1), \mathbf{T}(node_1), \cdots$ // templates identified in subtrees rooted at $node_0$'s children
6: $\mathbf{T_1}(node_0) \leftarrow \mathbf{T}(node_1) \cup \mathbf{T}(node_2) \cup \cdots$
7: $\mathbf{T_2}(node_0) \leftarrow$ identified templates in phrase $[node_1, node_2, \cdots]$
8: return $\mathbf{T}(node_0) = \mathbf{T_1}(node_0) \cup \mathbf{T_2}(node_0)$

We use Cabocha (Kudo and Matsumoto, 2002) for parsing Japanese sentences. This tool generates the hierarchical structure of the sentence as well as some properties (e.g. base form, pronunciation, part of speech, etc.) of each word. We execute Algorithm 1 on the parse tree to extract all identified templates of a Japanese sentence.

4 Statistics of Grammatical Templates

4.1 Corpus

We build our corpus from two sources: past JLPT exams and textbooks. The reading texts from JLPT exams are ideal for difficulty evaluation experiments since all of them are tagged authoritatively with difficulty levels, and JLPT problem sets before 2010 are publicly released[5]. We also collected reading texts from two popular series of Japanese textbooks: *Standard Japanese* (People's Education Press, 2013) and *Genki* (Banno et al., 2011). *Standard Japanese I* and *Genki I* are designed for the N5 level (the first semester) and *Standard Japanese II* and *Genki II* are designed for the N4 level (the second semester). Ultimately, our corpus consists of 220 texts (150 from past JLPT exams and 70 from textbooks), totaling 167,292 words after segmentation.

4.2 Results

For texts with different difficulties, we calculate the distribution of N1-N5 grammatical templates, which are shown in Table 3. We can see that N1 texts have higher portion of N1 and N2 templates than N2 texts, implying that the difficulty boosts from N2 to N1 are derived from increasing usage of advanced grammar. It is also clear that even in the texts of advanced levels, the majority of the sentences are organized by elementary grammatical templates, and the advanced ones are only used occasionally for formality or preciseness.

We also calculate the per-100-sentence number of templates at each level, which are shown in Table 4. When comparing any two adjacent levels (e.g. N2 and N3), the templates at those levels or above seem to be the most significant. For instance, N1/N2 texts differ in numbers of N1 and N2 templates while they have similar numbers of N3-N5 templates, and the numbers of N1, N2 and N3 templates differentiate

[5]For example, the second exam in 2009 is published in (Japan Educational Exchanges et al., 2010).

the N2/N3 texts while the numbers of N4 and N5 templates seem relatively similar. This phenomenon inspires us to build a simple and effective approach to differentiate the texts of two adjacent levels.

5 Difficulty Level Prediction

5.1 Multilevel Linear Classification

We differentiate two adjacent levels by looking at the knowledge 'on the boundary' and 'outside the boundary'. Concretely, when judging whether a text is harder than level N_i, we consider a grammatical template as:

- *within the boundary*, if the template is easier than N_i (N_{i+1} to N_5);
- *on the boundary*, if the template is exactly at N_i level;
- *outside the boundary*, if the template is harder than N_i (N_1 to N_{i-1}).

We found that texts of adjacent levels are nearly linear-separable with two features: templates 'on the boundary' and templates 'outside the boundary'. For example, Figure 1 shows how N1 and N2 texts are linearly separated based on the numbers of N1 and N2 templates: we can easily obtain a two-dimensional linear classifier separating N1 and N2 texts with 83.4% accuracy. This phenomenon is even more obvious at lower levels. Figure 2 shows N4 and N5 texts are almost perfectly linearly separated with two features: 'number of N5 templates per 100 sentences' (on the boundary) and 'number of N1-N4 templates per 100 sentences' (outside the boundary).

Taking advantage of this phenomenon, we build 4 linear classifiers for 4 pairs of adjacent levels. For example, the N4 classifier judges whether a text is harder than N4 (N1-N3). Our *Multilevel Linear Classification (MLC)* algorithm combines all 4 linear classifiers: A text is judged by the N5 classifier first. If it is no harder than N5, it will be labeled as an N5 text; otherwise, it will be passed to the N4 classifier in order to decide if it is harder than N4. The process continues similarly, until if it is judged to be harder than N2, it will be labeled as an N1 text. Figure 3 shows how the algorithm works.

5.2 Features

We conduct our experiments on the following 4 feature sets:

First, our *grammatical template feature set* has only 5 features:

- Average number of N1-N5 grammatical templates per sentence

We compare our work with recent readability evaluation studies (Kim et al., 2014; Pitler and Nenkova, 2008). In our experiments, the *baseline readability feature set* consists of the following 12 features:

- Number of words in a text
- Number of sentences in a text
- Average number of words per sentence
- Average parse tree depths per sentence
- Average number of noun phrases per sentence
- Average number of verb phrases per sentence
- Average number of pronouns per sentence
- Average number of clauses per sentence
- Average cosine similarity between adjacent sentences
- Average word overlap between adjacent sentences
- Average word overlap over noun and pronoun only
- Article likelihood estimated by language model

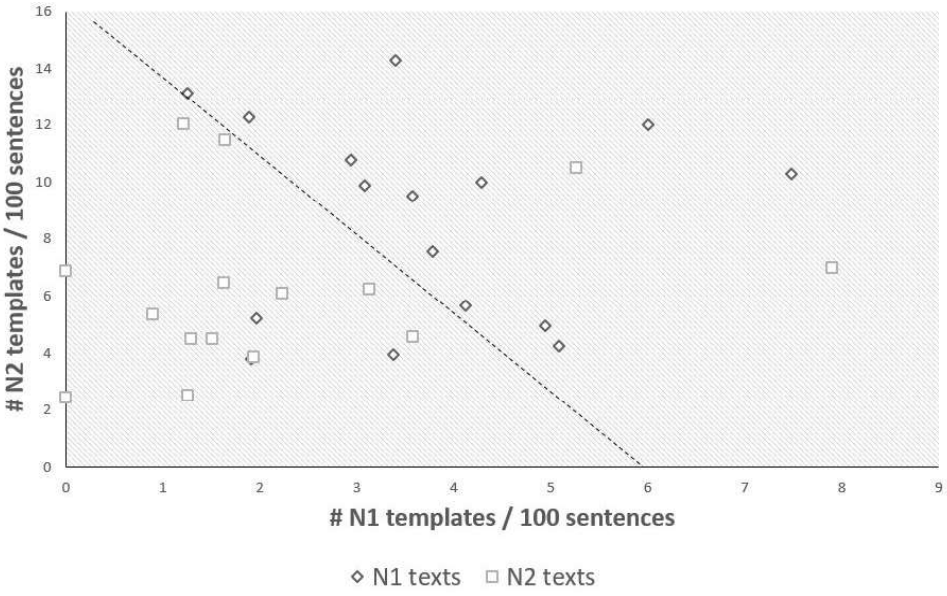

Figure 1: Grammatical difficulty in the N1/N2 texts

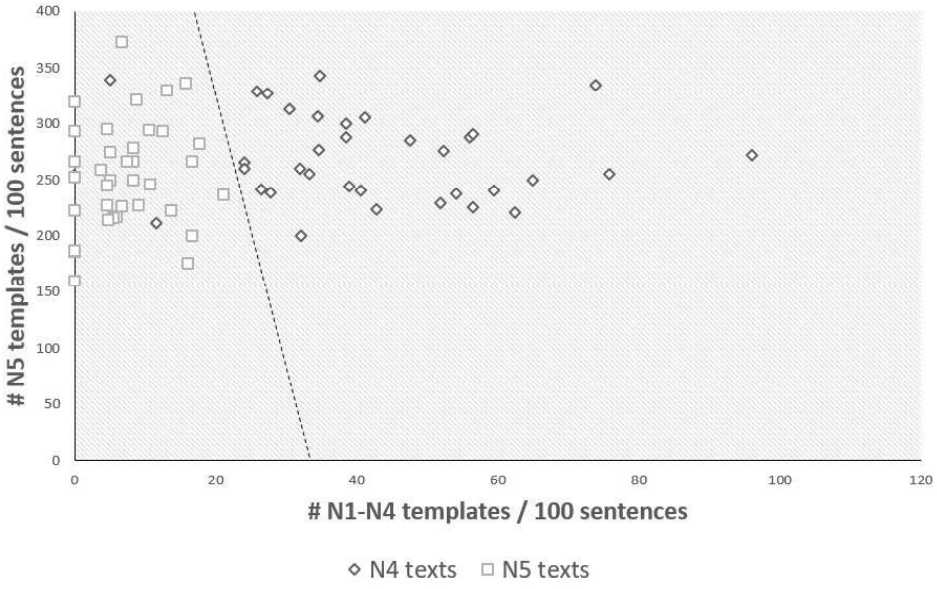

Figure 2: Grammatical difficulty in the N4/N5 texts

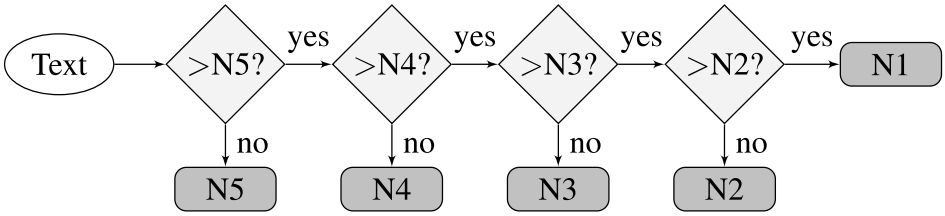

Figure 3: Multilevel Linear Classification (MLC). '>N5?' represents the linear classifier judging whether a text is harder than N5. The classifiers are similar for the other levels.

Feature Set (number of features)	Algorithm	Accuracy
TF-IDF Features (5100)	kNN	69.1%
	SVM	80.5%
Baseline Readability Features (12)	kNN	72.3%
	SVM	80.9%
Grammatical Template Features (**5**)	kNN	78.0%
	SVM	81.1%
	MLC	**87.7%**
Hybrid Features (17)	kNN	85.7%
	SVM	**88.5%**

Table 5: Accuracies of classifying N1-N5 texts

Moreover, we combine these 12 traditional readability features with our 5 grammatical template features, forming a *'hybrid' feature set*, since we would like to see if grammatical template features are really able to improve text difficulty evaluation.

Since the text difficulty level prediction can be regarded as a special text classification problem, we also extract *TF-IDF features* (Sparck Jones, 1972) (Nelson et al., 2012) as an extra baseline, in order to see how general text classification techniques work on text difficulty evaluation.

5.3 Result

We test k-Nearest Neighbor and Support Vector Machines for each feature set. The implementations of these two popular classification algorithms are provided by the WEKA toolkit (Hall et al., 2009) and LibSVM (Chang and Lin, 2011). The SVMs use RBF kernels (Chang et al., 2010). We also test our Multilevel Linear Classification (MLC) algorithm on the grammatical template feature set. We use 5-fold cross validation to avoid overfitting. Table 5 shows the results.

Comparing the results of kNN and SVM across the four different feature sets in Table 5, it is clear that TF-IDF features have the largest feature set yet lowest accuracy, indicating the general word-based text classification techniques do not work well on text difficulty level prediction. Compared with baseline readability features, our grammatical template features have smaller number of features but higher accuracy (slightly higher with SVM but significantly higher with kNN). Moreover, the hybrid features, which combine baseline readability features with grammatical template features, decisively outperform baseline readability features, confirming our expectation that adding grammatical template features to existing readability techniques improves text difficulty evaluation for language learners.

Additionally, our Multilevel Linear Classification algorithm achieves excellent accuracy with only 5 grammatical template features. An accuracy of 87.7% , although slightly lower than hybrid features + SVM (more features, more complexity), still significantly outperforms baseline readability techniques. In conclusion, the Multilevel Linear Classification algorithm has high accuracy, a small number of features, and a simple, human-understandable structure.

6 Conclusions and Future Work

We proposed a new approach for evaluating text difficulty that focuses on grammar and utilizes expert-identified grammatical templates, the grammar knowledge that students actually learn in language lessons. This approach significantly improved the accuracy of text difficulty evaluation for Japanese language learning. We also introduced a simple, human-understandable, and effective text difficulty evaluation approach using only five grammatical template features.

In future work, we are interested in extending our work to other languages like English, and adapting grammatical templates for various languages. To achieve this, we need to itemize the grammar knowledge that students learn from language lessons. We can also develop a machine learning system that can automatically discover discriminative grammatical templates from texts. Moreover, we would like

to study if the topic of a text has considerable impact on text difficulty for language learners, just like vocabulary and grammar.

We also hope to use our approach to recommend reading texts to individual learners at appropriate difficulty levels. For instance, Japanese news articles could be good learning materials for advanced Japanese language learners. We want to build an online tool to collect reading texts from current news reports in specific target languages, and select appropriate ones for language learners, especially intermediate and advanced learners.

Finally, we plan to leverage some novel ideas from Human-Computer Interaction and educational technology (Andersen et al., 2013) to build an *adaptive* Computer-Assisted Language Learning (CALL) system. Using our new approach introduced in this paper, we can decompose the difficulty of a text into several basic skills (grammatical templates), and model the internal hierarchical structure of a sequence of texts with a partial ordering graph. Using this structure, we can comprehensively assess a student's knowledge and tailor optimal learning progressions for individual students.

Acknowledgements

Special thanks to Xiang Long for his help during the writing of this paper.

References

Erik Andersen, Sumit Gulwani, and Zoran Popovic. 2013. A trace-based framework for analyzing and synthesizing educational progressions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 773–782. ACM.

Eri Banno, Yoko Ikeda, and Yutaka Ohno. 2011. *GENKI: An Integrated Course in Elementary Japanese*. Japan Times and Tsai Fong Books.

Carl Blyth. 1997. A constructivist approach to grammar: Teaching teachers to teach aspect. *The Modern Language Journal*, 81(1):50–66.

Jamie Callan and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of NAACL HLT*, pages 460–467.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Yin-Wen Chang, Cho-Jui Hsieh, Kai-Wei Chang, Michael Ringgaard, and Chih-Jen Lin. 2010. Training and testing low-degree polynomial data mappings via linear svm. *The Journal of Machine Learning Research*, 11:1471–1490.

Kevyn Collins-Thompson and James P Callan. 2004. A language modeling approach to predicting reading difficulty. In *HLT-NAACL*, pages 193–200.

Pedro Curto, Nuno Mamede, and Jorge Baptista. 2015. Assisting european portuguese teaching: Linguistic features extraction and automatic readability classifier. In *Computer Supported Education*, pages 81–96. Springer.

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.

Glenn Fulcher. 1997. Text difficulty and accessibility: Reading formulae and expert judgement. *System*, 25(4):497–513.

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Haritz Salaberri. 2014. Simple or complex? assessing the readability of basque texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 334–344, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080, Mumbai, India, December. The COLING 2012 Organizing Committee.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 71–79. Association for Computational Linguistics.

Shoaib Jameel, Xiaojun Qian, and Wai Lam. 2012. N-gram fragment sequence based unsupervised domain-specific document readability. In *Proceedings of COLING 2012*, pages 1309–1326, Mumbai, India, December. The COLING 2012 Organizing Committee.

Japan Educational Exchanges, Services, and Japan Foundation. 2010. *The 2009-2 Japanese Language Proficiency Test Level 1 and 2: Questions and Correct Answers*. Bonjinsha Inc.

A-Yeong Kim, Hyun-Je Song, Seong-Bae Park, and Sang-Jo Lee. 2014. Device-dependent readability for improved text understanding. In *EMNLP*, pages 1396–1404.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–7. Association for Computational Linguistics.

Wenzhao Liu and Hiroshi Ebihara. 2012. *New JLPT N1 Grammar Description*.

Javier Valenzuela Manzanares and Ana María Rojo López. 2008. What can language learners tell us about constructions? *APPLICATIONS OF COGNITIVE LINGUISTICS*, 9:197.

Jessica Nelson, Charles Perfetti, David Liben, and Meredith Liben. 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. *Council of Chief State School Officers, Washington, DC2012*.

People's Education Press. 2013. *Standard Japanese of China-Japan Exchanges for Beginners*. Mitsumura Tosho Publishing Co.Ltd.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing*, pages 186–195. Association for Computational Linguistics.

Jack C Richards and Richard W Schmidt. 2013. *Longman dictionary of language teaching and applied linguistics*. Routledge.

Hitoko Sasaki and Matsumoto Kiko. 2010. *Japanese Language Proficiency Test N1 GRAMMAR Summary*.

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.

Kathleen M Sheehan, Irene Kostin, Diane Napolitano, and Michael Flor. 2014. The textevaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal*, 115(2):184–209.

Manjira Sinha, Sakshi Sharma, Tirthankar Dasgupta, and Anupam Basu. 2012. New readability measures for Bangla and Hindi texts. In *Proceedings of COLING 2012: Posters*, pages 1141–1150, Mumbai, India, December. The COLING 2012 Organizing Committee.

Manjira Sinha, Tirthankar Dasgupta, and Anupam Basu. 2014. Influence of target reader background and text features on text readability in bangla: A computational approach. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 345–354, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Malbert Smith III, Anne Schiano, and Elizabeth Lattanzio. 2014. Beyond the classroom. *Knowledge Quest*, 42(3):20.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Miyamoto Atsushi Tomomatsu Etsuko and Waguri Masako. 2010. *Essential Japanese Expression Dictionary: A Guide to Correct Usage of Key Sentence Patterns (New Edition)*. ALC Press.

Sowmya Vajjala and Detmar Meurers. 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*, pages 163–173. Association for Computational Linguistics.

Renee Waara. 2004. Construal, convention, and constructions in l2 speech. *Cognitive linguistics, second language acquisition and foreign language pedagogy*, pages 51–75.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22. Association for Computational Linguistics.

Xiaoming Xu and Reika. 2015. *Blue Book All-in-one: JLPT N1-N5 Grammar*.

Mostafa Zamanian and Pooneh Heydari. 2012. Readability of texts:: State of the art. *Theory and Practice in Language Studies*, 2(1):43.