Multi-Lingual Text Leveling

Salim Roukos, Jerome Quin, and Todd Ward

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598
{roukos,jlquinn,tward}@us.ibm.com

Abstract. Determining the language proficiency level required to understand a given text is a key requirement in vetting documents for use in second language learning. In this work, we describe our approach for developing an automatic text analytic to estimate the text difficulty level using the Interagency Language Roundtable (ILR) proficiency scale. The approach we take is to use machine translation to translate a non-English document into English and then use an English language trained ILR level detector. We achieve good results in predicting ILR levels with both human and machine translation of Farsi documents. We also report results on text leveling prediction on human translations into English of documents from 54 languages.

Keywords: Text Leveling, ILR Proficiency, Second Language Acquisition

1 Introduction

As computerized tools for second language teaching become more widely available, the selection of content that is appropriate to a learner's proficiency in a language may benefit from the use of automatic text leveling tools. These tools can enable more automated content selection for personalized adaptive self-service language teaching systems. They can also support educators select more efficiently content that is contemporary and with the appropriate text difficulty level for their classes.

The Interagency Language Roundtable proficiency scale [1] of 11 levels ranges from level 0 to level 5 using half steps (0, 0+, 1, 1+, 2, 2+, 3, 3+, 4, 4+, 5). Level 0 indicates *no proficiency*, level 5 indicates *functional native proficiency* and level 3 indicates *general professional proficiency*.

In some contexts, there has been a significant investment in determining the ILR level of texts covering a variety of topics for multiple levels. However, updating the collection of documents to cover recent news, events, and topics can be a daunting task. The ILR text leveling guidelines are quite complicated, and the authors are not aware of any inter-annotator agreement studies for ILR level assignment. We report in this paper on some initial human ILR level annotation and the inter-annotator agreement reached in a preliminary annotation exercise. While we do not have access to expertly trained linguists, we attempted to assess the IAA that can be achieved with some training on ILR text difficulty level assessment. We report on these results in Section 2.

Recently, Shen et al [3] introduced their work to develop an automatic text leveling analytic for each of 4 languages. Their regression model is trained on about 2k documents from each language that have been annotated for ILR text difficulty level. The effort required to develop such a model for each new language is not scalable since

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 19–26, 2014.

it requires extensive training in ILR level labeling and applying it correctly for a new language. We propose an approach that relies on using machine translation from the source language to English and using an English-trained automatic text leveling analytic. This approach does not require text leveling annotation for the new language, though it requires a machine translation system. We report our initial results for Farsi documents.

Earlier work on text difficulty addressed the readability of a document based on the Flesch Reading Ease Formula, which uses two simple length features [2]: the average number of words per sentence and the average number of syllables per word. There has been various attempts at exploring weighing these features (linear regresion models) to improve the accuracy of predicting different readability levels. More recent work [4,5] used a richer feature set such as:

- average sentence length,
- average number of syllables per word,
- Flesch-Kincaid score,
- six out-of-vocabulary (OOV) rate scores,
- syntactic parse features, and
- 12 language model perplexity scores.

They also explored both classification and regression with SVMs to estimate the grade level (grades 2, 3, 4, and 5) of documents from the Weekly Reader newspaper. The richer models outperfomed the simpler Flesch-Kincaid score.

A similar feature set was used in a recent readability experiment conducted under the DARPA Machine Reading Program where *readability* was re-defined as the ease of reading of various styles of text as opposed to text level difficulty as addressed in earlier work. The range of documents cover various genres such newswire, transcriptions of conversational content, machine translation into English from other languages [6]. The methods used similar features ranging from parser-based features to n-gram language models.

Shen et al [3] used a corpus of 200 documents for each of seven levels (1, 1+, 2, 2+, 3, 3+, 4) for a given language. In their data, each of the texts was labeled by two independent linguists expertly trained in ILR level scoring. The ratings from these two linguists were then adjudicated by a third linguist. They did not provide inter-annotator agreement measures but took the adjudicated decision as the reference truth for both training and testing their system.

Using the fine grained ILR level training data, Shen et al developed regression models of text difficulty and proposed the use of mean square error (mse) metric where the plus-levels were mapped to the mid-point (e.g. 2+ is 2.5). They used a 80/20 split for training and test and built a separate regression model for each of the 4 languages. The best results were an mse of 0.2 for Arabic, 0.3 for Dari, 0.15 for English, and 0.36 for Pashto. These would correspond to a root mean square error (rmse) of 0.45, 0.55, 0.39, and 0.60 for each of the languages, respectively. Note a change of one level is an interval of 0.5 in their study.

They used two types of features: Length features and Word-usage features. The length features were three z-normalized length features:

1. average sentence length (in words) per document,

2. number of words per document, and
3. average word length (in characters) per document.

The Word-usage features were weighted word frequencies using TF-LOG weighted word frequencies on bag-of-words for each document. They compared length-based features which are not lexical to *Words-usage* features which are lexical items. The lexical features reduces the mse by 14% (Dari) to about 80% (Pashto). We are concerned about the word usage features. We surmise that the data used is more homogeneous than what is required for general second language acquisition (SLA) and may be influencing the significant performance improvement due to the *Words-usage* features since their leading examples of useful lexical features for English (which yielded a reduction of mse by 58%) appears to be topical. For example for level 3, the top ten lexical features, shown in Table 1, appear to be US politics centric.

Table 1. Top 10 Words-usage features and their weights for level 3

Word	Weight
obama	1.739
to	1.681
republicans	1.478
?	1.398
than	1.381
more	1.365
cells	1.355
american	1.338
americans	1.335
art	1.315

While it is hard to make solid claims about topicality without having access to the data, we are concerned about the robustness of the above results as we expect a sensitivity to topic change over time and geography for SLA content. For example, what would happen when the news is about French politics? Surely the names and parties will be different from the top indicators shown above.

In this work, we had access to data with single ILR level annotatio for 4 levels with coarser granularity spanning two consecutive levels (0+/1, 1+/2, 2+/3, 3+/4). The data had 5 broad topical areas and covered 54 languages. The texts were available in both the source language and its English human translation. We used these data to develop an ILR level detector based on English translations. While our work builds on Shen et al's results, we are different in 3 aspects: 1) we report initial measurements of ITA for human ILR text difficulty annotation, 2) our data set has coarser ILR annotation where a document was assigned a two level value (e.g 2+/3), and 3) our data have very broad variety of topics since it comes from 54 different languages.

The larger quantization interval of 1 versus an interval of 0.5 in the Shen et al study, implies that our mse error would be larger by 0.06 by definition, other factors being equal. Another aspect of our data is a skewed distribution of the levels with a severe

under-representation of the 0+/1 level at 2% of the documents with the other categories at 23%, 58%, and 17%, respectively. We present, in Section 2, our work on ILR level annotation, in Section 3, the data set, in Section 4, our text leveling results, and in Section 5, our conclusions.

2 Text Leveling Annotation

We attempted to train a small pool of 5 annotators to perform the ILR text leveling of English documents. We had access to ILR level annotation to a set of documents and a multimedia course on ILR annotation that requires five to ten hours to go through. We performed our own training of the annotators by explaining the principles as best as we could. We conducted 5 rounds of annotations followed by feedback sessions comparing the annotation of the five annotators on the same set of about 40 English documents per round. These were human translations of various languages and covered content for both reading and listening comprehension.

Table 2. Text leveling: human annotator performance

	kt	mn	mr	rx
AVG TIME	0:02:31	0:07:24	0:03:14	0:03:30
AVG ABS ERROR	0.64	0.57	0.55	0.67
NUM ERRORS	34	28	28	36
LARGE ERR	3	3	4	7

We report our results on the fifth round of ILR Text leveling annotation (we dropped one of the annotators due to consistently poorer scores). We compared each annotator to the reference truth as provided in our data set and to the other annotators. We used 60 documents covering source languages: Dari, Persian, Somali, and Spanish. They covered 3 levels nominally 2, 3 and 4 (strictly speaking these should be represented as 1.75, 2.75, and 3.75 as our data was annotated by an intervals such as 1+/2 meaning a midpoint of 1.75). We show in Table 2, the average time an annotator took to perform the task per document, the average absolute error between the human and the reference, the number of documents that had a different label, and the number of documents where the error was more than one level (interval of 1) for each of our four annotators. We show, in Table 3, the mse and rmse comparing each annotator to the reference. The mse in this work by definition is 0.06 higher than the results of Shen et al due to the coarser granularity of our reference truth (1 unit interval instead of 0.5). On average all 4 annotators have a rmse of 0.72.

We also computed the Pearson correlation between the annotators as shown in Table 4. We computed the average correlation of one annotator to the other three, and found mr and rx to have the highest average correlation of 0.74 and 0.73 respectively. Computing the mse and rmse between mr and rx, we get 0.24 and 0.49, respectively which indicates interestingly a better agreement between the two annotators than with the reference where the average rmse for the 2 annotators is 0.73.

Table 3. Text leveling: human annotator performance

	kt	mn	mr	rx
mse	0.48	0.53	0.51	0.56
rmse	0.69	0.73	0.72	0.75

Table 4. Interannotator correlation

	mn	mr	rx
cor	mn	mr	rx
kt	0.59	0.64	0.68
mn		0.79	0.72
mr			0.79

We were concerned that our annotators did not achieve a lower rmse than 0.72 relative to the reference and felt that the task is quite difficult for them. We decided not pursue our own annotation of text difficulty due to the larger investment required. The results with the human annotators' performance can be used as an indicator to assess how well our automatic text leveling analytic performs.

3 Text Leveling Data Set

Through our cooperation with the Center of Advanced Study of Language (CASL) at the University of Maryland, we were able to obtain a document collection with a single ILR text leveling annotation. The documents covered 5 broad topical areas and were evenly split between written (4.5k texts) and human transcribed genres (5k texts). The data were also provided with human English translations in addition to the spource language from 54 non-English languages. Table 5 shows the division by topic.

Table 5. Text leveling data set

Culture/ Society	Defense/ Security	Ecology/ Geography	Economics/ Politics	Science/ Technology
3,635	1,046	823	2,904	1,007

We received the data in two batches. The first one with about 2k documents and the second about 9k. Most of our results are based on the initial set of 2k documents. For the smaller condition we created a test set of 125 documents. We refer to the full set, as the larger condition, with a corresponding test set of 881 documents.

4 Experimental Results

We experimented with the following features:

- number of words in document length,
- average sentence length,
- average word length in characters,
- ratio of count of unique words (types) to total words,
- pronoun histogram,
- POS bigrams, and
- log term frequency.

We measure the performance by the classification accuracy, the mean square error (mse), and its nding root mean square (rms) error. We used a maximum entropy regression model. When we use the first three features, which are similar to the basic length features of earlier work, the level assignment accuracy is 66%, the mse is 0.37 with an rms of 0.60. Adding the remaining features listed above improves the accuracy to 77% and reduces the rmse to 0.51. Table 6 shows the confusion matrix for the full feature set.

Table 6. Confusion matrix between the 4 levels using the full feature set classifier

Level R/S	0.75	1.75	2.75	3.75
0.75	-	-	2	-
1.75	-	**15**	5	1
2.75	-	4	**70**	1
3.75	-	1	15	**11**

To evaluate the effect of machine translation on text leveling performance, we identified the largest subset of text material by source language in the smaller set which turned out to be Farsi. We had a Farsi test set of of 60 documents. We used a phrase-based Farsi-English translation system produce the machine translation version of the documents. We used the basic three feature set with the addition of ten binned vocabulary rank histogram. Table 7 compares human to machine translation in terms of accuracy, mse, and rms error. We can see that MT is relatively close to human translation though the rms on Farsi at 0.64 is higher than on the original set of 125 documents at 0.51.

Table 7. Performance with human and machine translation

	Accuracy	mse	rmse
Human translation	65%	0.41	0.64
Machine translation	57%	0.47	0.69

4.1 Experiments with the Larger Data Set

For the full set of 9k documents, we show the distribution by level in Table 8 which indicates the paucity of data for the first level and the dominance of the third level.

Table 8. Count of documents for each of the 4 levels

Level	0.75	1.75	2.75	3.75
Count	148	2,214	5,531	1,569

We compared the small and large training and test conditions. As can be seen in Table 9, the small trained model's rms error increases to 0.69 on the large test set. The large training set reduces the rms error from 0.69 to 0.54 on the large test set.

Table 9. RMSE using both the large and small training and test sets

Train/Test	small	large
small	0.63	0.69
large	0.58	0.54

5 Conclusion

We have built a text leveling system using an English training set of about 9k documents. The rms error of 0.54 achieved is comparable to the earlier work of Shen et al which had an average rms error across the 4 languages of 0.50 in spite of the larger quantization error in our data. Our approach depends on using machine translation instead of annotating for each new source language. Our results outperfom what our human annotators were able to achieve over 5 rounds of training annotations.

Acknowledgments. We gratefully acknowledge the help of the Center for the Advanced Study of Language at the University of Maryland. In particular, we want to thank Amy Weinberg, Catherine Daughty, and Jared Linck from CASL and Carol Van Ess-Dykema from DOD for their support in getting us access to the ILR annotated data set and the ILR annotation multimedia course.

References

1. Interagency Language Roundtable: ILR Skill Scale. `http://www.govtilr.org/Skills/ILRscale4.htm`, accessed June 15, (2014)

2. R. Flesch: A new readability yardstick. Journal of Applied Psychology, 32(3) pp. 221–233. (1948)
3. Wade Shen, Jennifer Williams, Tamas Marius and Elizabeth Salesky: A Language-Independent Approach to Automatic Text Difficulty Assessment for Second-Language Learners. Proceedings of the Workshop on Predicting and Improving Text Readability for Target Reader Populations. Sofia, Bulgaria, pp. 30–38. (2013)
4. Sarah E. Schwarm and Mari Ostendorf: Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. (2005)
5. Sarah E. Petersen and Mari Ostendorf: A machine learning approach to reading level assessment. Computer Speech and Language, 23, pp. 89–106. (2009)
6. Rohit J. Kate, Luo Xiaoqiang, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J. Mooney, Salim Roukos and Chris Welty: Learning to predict readability using diverse linguistic features. Proceedings of COLING '10, the 23rd International Conference on Computational Linguistics, pp. 546–554. (2010)