A Language-Independent Approach to Automatic Text Difficulty Assessment for Second-Language Learners

Wade Shen[1], Jennifer Williams[1], Tamas Marius[2], and Elizabeth Salesky [†1]

[1]MIT Lincoln Laboratory Human Language Technology Group,
244 Wood Street Lexington, MA 02420, USA
{swade,jennifer.williams,elizabeth.salesky}@ll.mit.edu

[2]DLI Foreign Language Center, Bldg. 420, Room 119 Monterey, CA 93944, USA
tamas.g.marius.civ@mail.mil

Abstract

In this paper, we introduce a new baseline for language-independent text difficulty assessment applied to the Interagency Language Roundtable (ILR) proficiency scale. We demonstrate that reading level assessment is a discriminative problem that is best-suited for regression. Our baseline uses z-normalized shallow length features and TF-LOG weighted vectors on bag-of-words for Arabic, Dari, English, and Pashto. We compare Support Vector Machines and the Margin-Infused Relaxed Algorithm measured by mean squared error. We provide an analysis of which features are most predictive of a given level.

1 Introduction

The ability to obtain new materials of an appropriate language proficiency level is an obstacle for second-language learners and educators alike. With the growth of publicly available Internet and news sources, learners and instructors of foreign languages should have ever-increasing access to large volumes of foreign language text. However, sifting through this pool of foreign language data poses a significant challenge. In this paper we demonstrate two machine learning regression methods which can be used to help both learners and course developers by automatically rating documents based on the text difficulty. These methods can be used to automatically identify documents at specific levels in order to speed course or test development, providing learners with custom-tailored materials that match their learning needs.

ILR (Interagency Language Roundtable) levels reflect differences in text difficulty for second-language learners at different stages of their education. A description of each level is shown in Table 1 (Interagency Language Roundtable, 2013). Some levels differ in terms of sentence structure, length of document, type of communication, etc., while others, especially the higher levels, differ in terms of the domain and style of writing. Given these differences, we expect that both semantic content and grammar-related features will be necessary to distinguish between documents at different levels.

Level	Description
0	No proficiency
0+	Memorized proficiency
1	Elementary proficiency
1+	Elementary proficiency, plus
2	Limited working proficiency
2+	Limited working proficiency, plus
3	General professional proficiency
3+	General professional proficiency, plus
4	Advanced professional proficienty
4+	Advanced professional proficiency, plus
5	Functionally native proficiency

Table 1: Description of ILR levels.

Automatically determining ILR levels from documents is a research problem without known solutions. We have developed and adapted a series of rating algorithms and a set of experiments gauging the feasibility of automatic ILR level assignment for text documents. Using data provided by the Defense Language Institute Foreign Language Center (DLIFLC), we show that while the problem is tractable, the performance of automatic

[†] This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations, pages 30–38,
Sofia, Bulgaria, August 4-9 2013.

methods is not perfect.

Our general approach treats the ILR rating problem as one of text classification; given the contents and structure of a document, which of the ILR levels should this document be assigned to? This differs from traditional topic classification tasks where word-usage often uniquely defines topics, since we are also interested in features of text complexity that describe structure. Leveling text is a problem better fit to regression because reading level is a continuous scale. We want to know how close a document is to a given level (or between levels), so we measured performance using mean squared error (MSE). We show that language-independent features can be used for regression with Support Vector Machines (SVMs) and the Margin-Infused Relaxed Algorithm (MIRA), and we present our results for this new baseline for Arabic, Dari, English, and Pashto. To the best of our knowledge, this is the first study to systematically examine a language-independent approach to readability using the ILR rating scale for second-language learners.

This paper is structured as follows: Section 2 describes previous work on reading level assessment as a text classification problem, Section 3 describes the two algorithms that we used in our present work, Section 4 describes our data and experiments, Section 5 reports our results, Section 6 provides an analysis of our results, and Section 7 proposes different kinds of future work that can be done to improve this baseline.

2 Related Work

In this section we describe some work on the readability problem that is most closely related to our own.

One of the earliest formulas for reading level assessment, called the Flesch Reading Ease Formula, measured readability based on shallow length features (Flesch, 1948). This metric included two measurements: the average number of words per sentence and the average number of syllables per word. Although these features appear to be shallow at the offset, the number of syllables per word could be taken as an abstraction of word complexity. Those formulas, as well as their various revisions, have become popular because they are easy to compute for a variety of applications, including structuring highly technical text that is comprehensible at lower reading levels (Kincaid et al., 1975). Some of the revisions to the Flesch Reading Ease Formula have included weighting these shallow features in order to linearly regress across different difficulty levels.

Much effort has been placed into automating the scoring process, and recent work on this issue has examined machine learning methods to treat reading level as a text classification problem. Schwarm and Ostendorf (2005) worked on automatically classifying text by grade level for first-language learners. Their machine learning approach was a one vs. all method using a set of SVM binary classifiers that were constructed for each grade level category: 2, 3, 4, and 5. The following features were used for classfication: average sentence length, average number of syllables per word, Flesch-Kincaid score, 6 out-of-vocabulary (OOV) rate scores, syntactic parse features, and 12 language model perplexity scores. Their data was taken from the Weekly Reader newspaper, already separated by grade level. They found that the error rate for misclassification by more than one grade level was significantly lower for the SVM classifier than for both Lexile and Flesch-Kincaid. Petersen and Ostendorf (2009) later replicated and expanded Schwarm and Ostendorf (2005), reaffirming that both classification and regression with SVMs provided a better approximation of readabilty by grade level when compared with more traditional methods such as the Flesch-Kincaid score. In the current work, we also use SVM for regression, but have decided to report mean squared error as a more meaningful metric.

In an effort to uncover which features are the most salient for discriminating among reading levels, Feng et al., (2010) studied classification performance using combinations of different kinds of readability features using data from the Weekly Reader newspaper. Their work examined the following types of features: discourse, language modeling, parsed syntactic features, POS features, shallow length features, as well as some features replicated from Schwarm and Ostendorf (2005). They reported classifier accuracy and mean squared error from two classifiers, SVM and Logistic Regression, which were used to predict grade level for grades 2 through 5. While they found that POS features were the most predictive overall, they also found that the average number of words per sentence was the most predictive length

feature. This length feature alone achieved 52% accuracy with the Logistic Regression classifier. In the present work, we use the average number of words per sentence as a length feature and show that this metric has some correspondence with the different ILR levels.

Another way to examine readability is to treat it as a sorting problem; that is, given some collection of texts, to sort them from easiest to most difficult. Tanaka-Ishii et al., (2010) presented a novel method for determining readibility based on sorting texts using text from two groups: low difficulty and high difficulty. They reported their results in terms of the Spearman correlation coefficient to compare performance of Flesch-Kincaid, Dale-Chall, SVM regression, and their sorting method. They showed that their sorting method was superior to the other methods, followed by SVM regression. However, they call for a more modern and efficient approach to the problem, such as online learning, that would estimate weights for regression. We answer their call with an online learning approach in this work.

3 Algorithms

In this section, we describe two maximum margin approaches that we used in our experiments. Both are based on the principle of structural risk minimization. We selected the SVM algorithm because of its proven usefulness for automatic readability assessment. In addition, the Margin-Infused Relaxed Algorithm is advantageous because it is an online algorithm and therefore allows for incremental training while still taking advantange of structural risk minimization.

3.1 Structural Risk Minimization

For many classification and regression problems, maximum margin approaches are shown to perform well with minimal amounts of training data. In general, these approaches involve linear discriminative classifiers that attempt to learn hyperplane decision boundaries which separate one class from another. Since multiple hyperplanes that separate classes can exist, these methods add an additional constraint: they attempt to learn hyperplanes while maximizing a region around the boundary called the margin. We show an example of this kind of margin in Figure 1, where the margin represents the maximum distance between the decision boundary and support vectors. The maximum margin approach helps prevent overfitting issues that can occur during training, a principle called structural risk minimization. Therefore we experiment with two such margin-maximizing algorithms, described below.

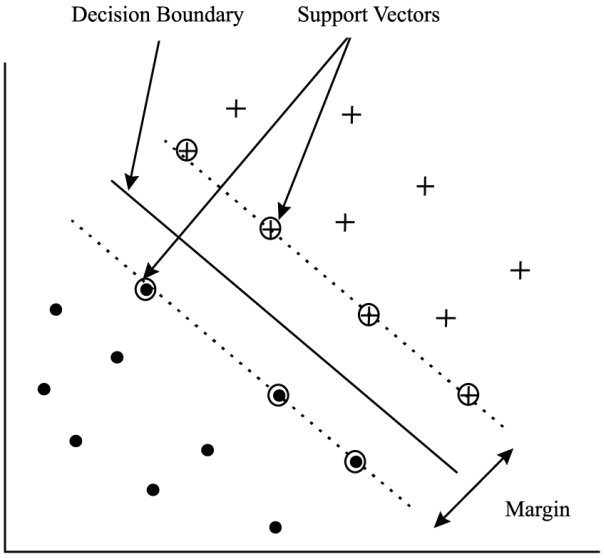

Figure 1: Graphical depiction of the maximum margin principle.

3.2 Support Vector Machines

For text classification problems, the most popular maximum margin approach is the SVM algorithm, introduced by Vapnik (1995). This approach uses a quadratic programming method to find the support vectors that define the margin. This is a batch training algorithm requiring all training data to be present in order to perform the optimization procedure (Joachims, 1998a). We used LIBSVM to implement our own SVM for regression (Chang and Lin, 2001).

Discriminative methods seek to best divide training examples in each class from out-of-class examples. SVM-based methods are examples of this approach and have been successfully applied to other text classification problems, including previous work on reading level assessment (Schwarm and Ostendorf, 2005; Petersen and Ostendorf, 2009; Feng et al., 2010). This approach attempts to explicitly model the decision boundary between classes. Discriminative methods build a model for each class c that is defined by the boundary between examples of class c and examples from all other classes in the training data.

3.3 Margin-Infused Relaxed Algorithm

Online approaches have the advantage of allowing incremental adaptation when new labeled examples are added during training. We implemented a version of MIRA from Crammer and Singer (2003), which we used for regression. Crammer and Singer (2003) proved MIRA as an online multiclass classifier that employs the principle of structural risk minimization, and is described as ultraconservative because it only updates weights for misclassified examples. For classification, MIRA is formulated as shown in equation (1):

$$c^* = \arg\max_{c \in \mathbf{C}} f_c(\mathbf{d}) \quad (1)$$

where

$$f_c(\mathbf{d}) = \mathbf{w} \cdot \mathbf{d} \quad (2)$$

and $\mathbf{w}$ is the weight vector which defines the model for class c. During training, examples are presented to the algorithm in an online fashion (i.e. one at a time) and the weight vector is updated accourding to the update shown in equation (2):

$$\mathbf{w}_t = \mathbf{w}_{t-1} + l(\mathbf{w}_{t-1}, \mathbf{d}_{t-1})\mathbf{v}_{t-1} \quad (3)$$

$$l(\mathbf{w}_{t-1}, \mathbf{d}_{t-1}) = ||\mathbf{d}_{t-1} - \mathbf{w}_{t-1}|| - \epsilon \quad (4)$$

$$\mathbf{v}_{t-1} = (sign(||\mathbf{d}_{t-1} - \mathbf{w}_{t-1}||) - \epsilon)\mathbf{d}_{t-1} \quad (5)$$

where $l(\cdot)$ is the loss function, ϵ corresponds to the margin slack, and $\mathbf{v}_{t-1}$ is the negative gradient of the loss vector for the previously seen example $||\mathbf{d}_{t-1} - \mathbf{w}_{t-1}||$. This update forces the weight vector towards erroneous examples during training. The magnitude of the change is proportional to the $l(\cdot)$. For correct training examples, no update is performed as $l(\cdot) = 0$. In a binary classification task, MIRA attempts to minimize the loss function in (4), such that the magnitude of the distance between a document vector and the weight vector is also minimized.

However, unlike topic classification or classification of words based on their semantic class where the classes are generally discrete, the ILR levels lie on a continuum (i.e. level 2 >> level 1 >> level 0). Therefore we are more interested in using MIRA for regression because we want to compare the predicted value with the true real-valued label, rather than a class label. For regression, we can redefine the MIRA loss function as follows:

$$l(\mathbf{w}_t, \mathbf{d}_t) = |l_t - \mathbf{d}_t \cdot \mathbf{w}_t| - \epsilon \quad (6)$$

In this case, l_t is the correct value (in our case, ILR level) for training document $\mathbf{d}_t$ and $\mathbf{d}_t \cdot \mathbf{w}_t$ is the predicted value given the current weight vector $\mathbf{w}_t$. We expect that minimizing this loss function cumulatively over the entire training set will yield a regression model that can predict ILR levels for unseen documents.

This revised loss function results in a modified update equation for each online update of the MIRA weight vector (generating a new set of weights $\mathbf{w}_t$ from the previously seen example):

$$\mathbf{w}_t = \mathbf{w}_{t-1} + l(\mathbf{w}_{t-1}, \mathbf{d}_{t-1})\mathbf{v}_{t-1} \quad (7)$$

$$\mathbf{v}_{t-1} = (sign(|l_{t-1} - \mathbf{d}_{t-1} \cdot \mathbf{w}_{t-1}|) - \epsilon)\mathbf{d}_{t-1} \quad (8)$$

$\mathbf{v}_{t-1}$ defines the direction of loss and the magnitude of the update relative to the current training example $\mathbf{d}_{t-1}$. Since this approach is online, MIRA does not guarantee minimal loss or maximum margin constraints for all of the training data. However, in practice, these methods perform as well as their SVM counterparts without the need for batch training (Crammer et al., 2006).

4 Experiments

4.1 Data

All of our experiments used data from four languages: Arabic (AR), Dari (DAR), English (EN), and Pashto (PS). In Table 2, we show the distribution of number of documents per ILR level for each language. All of our data was obtained from the Directorate of Language Science and Technology (LST) and the Language Technology Evaluation and Application Division (LTEA) at the Defense Language Institute Foreign Language Center (DLIFLC). The data was compiled using an online resource (Domino). Language experts (native speakers) used various texts from the Internet which they considered to be authentic material and they created the Global Language Online Support System (GLOSS) system. The texts were used to debug the GLOSS system and to see how well GLOSS worked for the respective languages. Each of the texts were labeled by two independent linguists expertly trained in ILR level scoring. The ratings from these two linguists were then adjudicated by a third linguist. We used the resulting adjudicated labels for our training and evaluation.

We preprocessed the data by doing the following tokenization: removed extra whitespace, normalized URIs, normalized currency, normalized

Level	AR	DAR	EN	PS
1	204	197	198	197
1+	200	197	197	199
2	199	201	204	200
2+	199	194	196	198
3	198	195	202	198
3+	194	194	198	200
4	198	195	190	195
Overall	1394	1375	1390	1394

Table 2: Total collection documents per language per ILR level.

numbers, normalized abbreviations, normalized punctuation, and folded to lowercase. We identified words by splitting text on whitespace and we identified sentences by splitting text on punctuation.

4.2 Features

It is necessary to define a set of features to help the regressors distinguish between the ILR levels. We conducted our experiments using two different types of features: word-usage features and shallow length features. Shallow length features are shown to be useful in reading level prediction tasks (Feng et al., 2010). Word-usage features, such as the ones used here, are meant to capture some low-level topical differences between ILR levels.

Word-usage features: Word frequencies (or weighted word frequencies) are commonly used as features for topic classification problems, as these features are highly correlated with topics (e.g. words like *player* and *touchdown* are very common in documents about topics like *football*, whereas they are much less common in documents about *opera*). We used TF-LOG weighted word frequencies on bag-of-words for each document.

Length features: In addition to word-usage, we added 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. We used these as a basic measure of language level complexity. These features are easily computed by automatic means, and they capture some of the structural differences between the ILR levels.

Figures 2, 3, 4, and 5 show the z-normalized average word count per sentence for Arabic, Dari, English, and Pashto respectively. The overall data set for each language has a normalized mean of

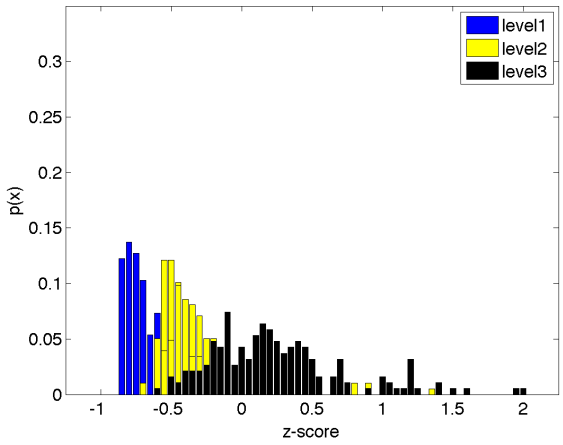

Figure 2: Arabic, z-normalized average word count per sentence for ILR levels 1, 2 and 3.

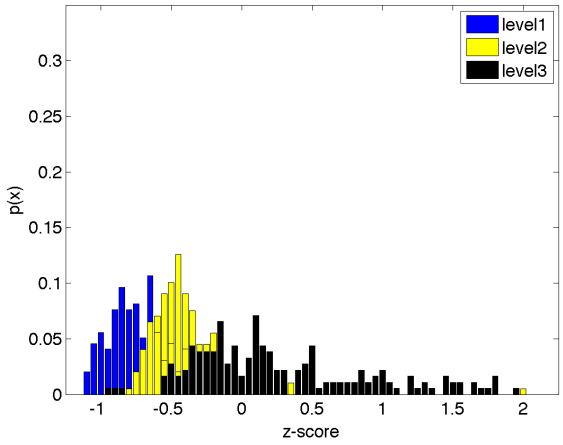

Figure 3: Dari, z-normalized average word count per sentence for ILR levels 1, 2 and 3.

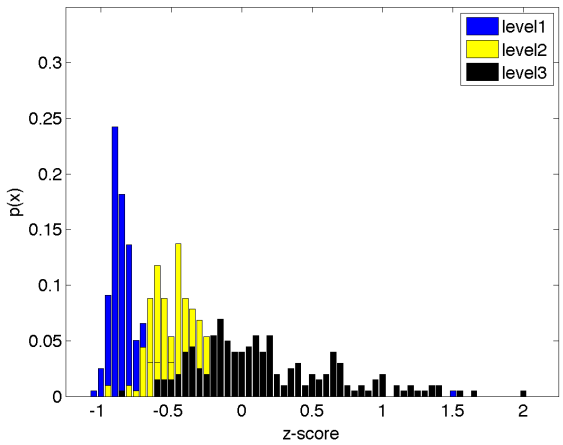

Figure 4: English, z-normalized average word count per sentence for ILR levels 1, 2 and 3.

	MIRA			SVM (linear)		
	LEN	**WORDS**	**COMBINED**	**LEN**	**WORDS**	**COMBINED**
AR	4.527	0.283	**0.222**	0.411	0.263	**0.198**
DAR	5.538	0.430	**0.330**	0.473	0.409	**0.301**
EN	5.155	0.181	**0.148**	0.430	0.181	**0.147**
PS	5.371	0.410	**0.360**	1.871	0.393	**0.391**

Table 3: Performance results (MSE) for SVM and MIRA on Arabic, Dari, English and Pashto for three different kinds of features/combinations.

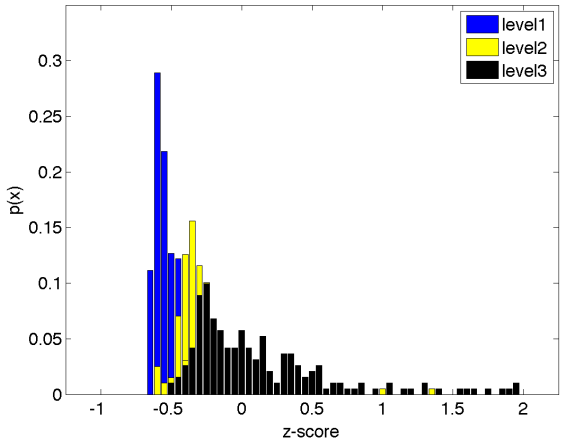

Figure 5: Pashto, z-normalized average word count per sentence for ILR levels 1, 2 and 3.

zero and unit variance, which were calculated separately for a given length feature. The x-axis shows the deviation of documents relative to the data set mean, in units of overall standard deviation. It is clear from the separability of the levels in these figures that sentence length could be an important indicator of ILR level, though no feature is a perfect discriminator. This is indicated by the significant overlap between the distributions of document lengths at different ILR levels.

4.3 Training

We split the data between training and testing using an 80/20 split of the total data for each language. To formulate the ILR scale as continuous-valued, we assumed that "+" levels are 0.5 higher than their basis (e.g. 2+ = 2.5). Though this may not be optimal if distances between levels are non-constant, the best systems in our experiments show good prediction performance using this assumption.

Both of the classifiers were trained to predict the ILR value as a continuous value using regression. We measured the performance of each method in terms of the mean squared error on the unseen test documents. We tested the following three conditions: length-based features only (LEN), word-usage features only (WORDS), and word and length features combined (COMBINED). Since each algorithm (SVM and MIRA) has a number of parameters that can be tuned to optimize performance, we report results for the best settings for each of the algorithms. These settings were determined by sweeping parameters to optimize performance on the training data for a range of values, for both MIRA and SVM. For both algorithms, we varied the number of training iterations from 500 to 3100 for each language, with stepsize of 100. We also varied the minimum word frequency count from 2 to 26, with stepsize 1. For MIRA only, we varied the slack parameter from 0.0005 to 0.0500, with stepsize 0.00025. For SVM (linear kernel only), we varied the C parameter and γ at a coarse setting of 2^n with values of n ranging from -15 to 6 with stepsize 1.

5 Results

We compared the performance of the online MIRA approach with the SVM-based approach. Table 3 shows the overall performance of MIRA regression and SVM regression, respectively, for the combinations of features for each language. Mean squared error was averaged over all of the levels in a given language. MIRA is an approximation to SVM, however one of the advantages of MIRA is that it is an online algorithm so it is adaptable after training and training can be enhanced later with more data with a small number of additional data points.

Figures 6 and 7 show the per-level performance for each classifier with the overall best features (COMBINED) for each language. The highest level (Level 4) and lowest levels (Level 1) tend to

exhibit the worst performance across all languages for each regression method. Poorer performance on the outlying levels could be due to overfitting for both SVM and MIRA on those levels. The ILR scale includes 4 major levels at half-step intervals between each one. We are not sure if using a different scale, such as grade levels ranging from 1 to 12, would also exhibit poorer performance on the outlying levels because the highest ILR level corresponds to native-like fluency. This U-shaped performance is seen across both classifiers for each of the languages.

6 Analysis

Our results show that SVM slightly outperformed MIRA for all of the languages. We believe that the reason why MIRA performed worse than SVM is because it was overfit during training whereas SVM was not. This could be due to the parameters that we set during our sweep in training. We selected C and γ as parameters to SVM linear-kernel for the best performance. The γ values for English and Arabic were set at more than 1000 times smaller than the values for Pashto and Dari (AR:γ=6.1035156 $\times$ $10^{-}5$, DAR:γ=0.0078125, EN:γ=3.0517578 $\times$ $10^{-}5$, PS:γ=0.03125). This means that the margins for Pashto and Dari were set to be larger respective to English and Arabic. One reason why these margins were larger is because the features that we used had more discriminative power for English and Arabic. In fact, both MIRA and SVM performed worse on Pashto and Dari.

Since the method described here makes use of linear classifiers that weigh word-usage and length features, it is possible to examine the weights that a classifier learns during training to see which features the algorithm deems most useful in discriminating between ILR levels. One way to do this is to use a multiclass classifier on our data for the categorical levels (e.g. 1, 1+, 2, etc.) and examine the weights that were generated for each class. MIRA is formulated to be a multiclass classifier so we examined its weights for the features. We chose MIRA instead of SVM, even though LIBSVM supports multiclass classification, because we wanted to capture differences between levels which we could not do with one vs. all. We examined classifier weights of greatest magnitude to see which features were the most indicative and most contra-indicative for that level. We report these two types of features for Level 3 and Level 4 in Tables 4 and 5, respectively. Level 3 documents can have some complex topics, such as *politics* and *art*, however it can be noted that some of the more abstract topics like *love* and *hate* are contra-indicative of Level 3. On the other hand, we see that abstract topics are highly indicative Level 4 documents where topics such as *philosophy*, *religion*, *virtue*, *hypothesis*, and *theory* are discussed. We also note that *moral* is highly contra-indicative of Level 3 but is highly indicative of Level 4.

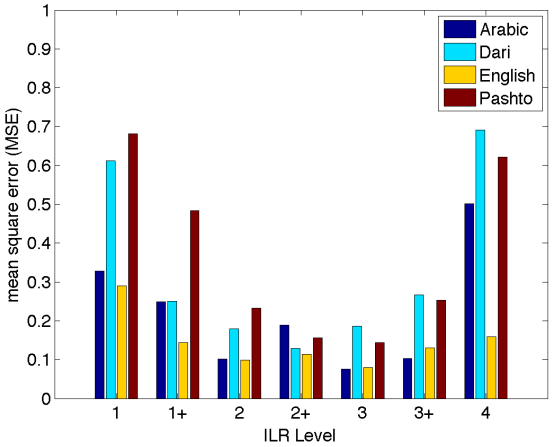

Figure 6: MIRA performance (MSE) per ILR level for each language.

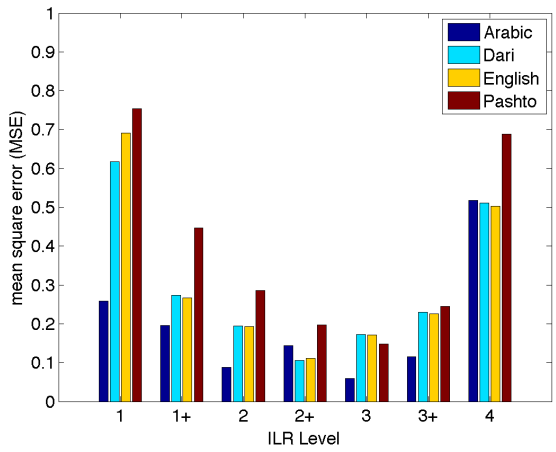

Figure 7: SVM performance (MSE) per ILR level for each language.

7 Discussion and Future Work

We have presented an approach to score documents based on their ILR level automatically using language-independent features. Measures of structural complexity like the length-based fea-

Most Indicative	+	Most Contra-Indicative	-
obama	1.739	said	-2.259
to	1.681	your	-1.480
republicans	1.478	is	-1.334
?	1.398	moral	-0.893
than	1.381	this	-0.835
more	1.365	were	-0.751
cells	1.355	area	-0.751
american	1.338	love	-0.730
americans	1.335	says	-0.716
art	1.315	hate	-0.702
it's	1.257	against	-0.682
could	1.180	people	-0.669
democrats	1.143	body	-0.669
as	1.139	you	-0.666
a	1.072	man	-0.652
but	1.041	all	-0.644
america	0.982	over	-0.591

Table 4: Dominant features for English at ILR Level 3.

Most Indicative	+	Most Contra-Indicative	-
of	3.298	+number+	-2.524
this	2.215	.	-2.514
moral	1.880	government	-1.120
philosophy	1.541	have	-1.109
is	1.242	people	-1.007
theory	1.138	would	-0.909
in	1.131	could	-0.878
absolute	1.034	after	-0.875
religion	1.011	you	-0.874
hyperbole	0.938	,"	-0.870
mind	0.934	were	-0.827
as	0.919	was	-0.811
hypothesis	0.904	years	-0.795
schelling	0.883	your	-0.747
thought	0.854	americans	-0.746
virtue	0.835	at	-0.745
alchemy	0.828	they	-0.720

Table 5: Dominant features for English at ILR Level 4.

tures used in this work are important to achieving good ILR prediction performance. We intend to investigate further measures that could improve this baseline, including features from automatic parsers or unsupervised morphology to measure syntactic complexity. Here we have shown that higher reading levels in English correspond more with abstract topics. In future work, we also want to capture some of the stylistic features of text, such as the complexity of dialogue exchanges.

For both SVM and MIRA, the combination of length and word-usage features had the best impact on performance across languages. We found better performance on this task overall for SVM and we believe that MIRA was overfitting during training. For MIRA, this is likely due to an interaction between a small number of features and the stopping criterion (mean squared error = 0) that we used in training, which tends to overfit. We intend to investigate the stopping criterion in future work. Still, we have shown that MIRA can be useful in this task because it is an online algorithm, and it allows for incremental training and active learning.

Our current approach can be quickly adapted for a new subset of languages because the features that we used here were language-independent. We plan to build a flexible architecture that enables language-specific feature extraction to be combined with our method so that these techniques can be easily used for new languages. We will continuously improve this baseline using the approaches described in this paper. We found that these two algorithms along with these types of features performed pretty well on 4 different languages. It is surprising that these features would correlate across languages even though there are individual differences between each language. In future work, we are interested to look deeper into the nature of language-independence for this task.

With respect to content, we are interested to find out if more word features are needed for some languages but not others. There could be diversity of vocabulary at higher ILR levels, which we could measure with entropy. Additionally, since the MIRA classifier that we are using is an online classifier with weight vector representation for each feature, we could examine the weights and measure the mutual information by ILR level above a certain threshold to find which features are the most predictive of an ILR level, for each language. Lastly, we have assumed that the ILR rating metric is approximately linear, and although we have used linear classifiers in this task, we are interested to learn if other transformations would give us a better sense of ILR level discrimination.

References

Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Koby Crammer and Yoram Singer. 2003. Ultraconservative Online Algorithms for Multiclass Problems. *Journal of Machine Learning Research*, 3(2003):951-991.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online Passive-Agressive Algorithms. *Journal of Machine Learning Research*, 7(2006):551-585.

George R. Doddington, Mark A. Przybocki, Alvin F. Martin, and Douglas A. Reynolds. 2000. The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective. *Speech Communication*, 31(2-3):225-254.

Lijun Feng, Martin Jansche, Matt Huenerfauth, Noémie Elhadad. 2010. A Comparison of Features for Automatic Readability Assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters.* Association for Computational Linguistics, 2010.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221-233.

Interagency Language Roundtable. ILR Skill Scale. http://www.govtilr.org/Skills/ILRscale4.htm, 2013. Accessed February 27, 2013.

Thorsten Joachims. 1998a. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of the European Conference on Machine Learning,* pages 137-142, 1998a.

Peter J. Kincaid, Lieutenant Robert P. Fishburne, Jr., Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas for Navy enlisted personnel. Research Branch Report 8-75, U.S. Naval Air Station, Memphis, 1975.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23(2009):89-106.

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 of the Association for Computational Linguistics*.

Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada. 2010. Sorting texts by readability. *Computational Linguistics*, 36(2):203-227.

Vladimir Vapnik. 1995. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.