4 Readability Measures

This section describes the range of linguistic features explored and the machine learning framework applied to the WeeBit data that constitute a general readability assessment system. The set of features used in our experiments is an extension to those used in previous work (Feng et al., 2010; Pitler and Nenkova, 2008; Vajjala and Meurers, 2012; Vajjala and Meurers, 2014), and their predictive power for reading difficulty assessment is investigated in our experiments. We have extended the feature set with the EVP-based features, GR-based complexity measures and the combination of language modeling features that have not been applied to readability assessment before.

4.1 Features

Traditional Features The traditional features are easy-to-compute representations of superficial aspects of text. The metrics that are considered include: the number of sentences per text, average and maximum number of words per sentence, average number of characters per word, and average number of syllables per word. Two popular readability formulas are also included: the Flesch-Kincaid score (Kincaid et al., 1975) and the Coleman-Liau readability formula (Coleman and Liau, 1975).

Lexico-semantic Features Vocabulary knowledge is one of the most important aspects of reading comprehension (Collins-Thompson, 2014). Lexico-semantic features provide information about the difficulty or familiarity of vocabulary in the text.

A widely used lexical measure is the *type-token ratio* (TTR), which is the ratio of the number of unique word tokens (referred to as types) to the total number of word tokens in a text. However, the conventional TTR is influenced by the length of the text. *Root TTR* and *Corrected TTR*, which take the logarithm and square root of the text length instead of the direct word count as denominator, can produce a more unbiased representation and are included in the experiment.

Part of speech (POS) based lexical variation and lexical density measures (Lu, 2011) are also examined. *Lexical variation* is defined as the type-token ratio of lexical items such as nouns, adjectives, verbs, adverbs and prepositions. *Lexical density* is defined as the proportion of the five classes of lexical items in all word tokens. The percentage of content words (nouns, verbs, adjectives and adverbs) and function words (all the remaining POS types) are two other indicators of lexical density.

Vajjala and Meurers (2012; 2014) reported in their readability classification experiment that the proportion of words in the text that are found in the Academic Word List is one of the most predictive measures among all the lexical features they considered. The Academic Word List (Coxhead, 2000) is comprised of words that frequently occur across all topic ranges in an academic text corpus. The *proportion of academic vocabulary words* in the text can be viewed as another measure of lexical complexity.

A similar but more refined approach to estimate lexical complexity is based on the use of the *English Vocabulary Profile* (EVP).[4] The EVP is an online vocabulary resource that contains information about which words and phrases are acquired by learners at each CEFR level. It is collected from the Cambridge Learner Corpus (CLC), a collection of examination scripts written by learners from all over the world (Capel, 2012). It provides a more fine-grained lexical complexity measure that captures the relative difficulty of each word by assigning the word difficulty to one of the six CEFR levels. Additionally, the EVP indicates the word difficulty for L2 learners rather than native speakers, which makes it more informative in non-native readability analysis. In our experiments, the proportion of words at each CEFR level is calculated and added to the feature set.

Parse Tree Syntactic Features A number of *syntactic measures* based on the RASP parser output (Briscoe et al., 2006) are used to describe the grammatical complexity of text, including average parse tree depth, and average number of noun, verb, adjective, adverb, prepositional phrases and clauses per sentence.

Grammatical relations (GR) between constituents in a sentence may also affect the judgement of syntactic difficulty. Yannakoudakis (2013) applied 24 GR-based complexity measures in essay scoring and showed good results. These complexity measures capture the grammatical sophistication of the text through the representation of the distance be-

[4] http://www.englishprofile.org/

tween the sentence constituents. For instance, these measures calculate the longest/average distance in the GR sets generated by the parser and the average/maximum number of GRs per sentence. A set of 24 GR-based measures used by Yannakoudakis (2013) are generated by RASP for each sentence. We take the average of these measures across the text to incorporate the GR-related aspect of its syntactic difficulty.

Other types of complexity measures that are derived from the parser output include: *cost metric*, which is the total number of parsing actions performed for generating the parse tree; *ambiguity of the parse*, and so on. A total number of 114 non-GR based complexity measures are extracted. These complexity measures are averaged across the text and used to model finer details of the syntactic difficulty of the text.

Language Modeling Features Statistical language modeling (LM) provides information about distribution of word usage in the text and is in fact another way to describe the lexical dimension of readability. To avoid over-fitting to the WeeBit data, two types of language modeling based features are extracted using the SRILM toolkit (Stolcke, 2002): (1) *word token n-gram models*, with n ranging from 1 to 5, trained on the British National Corpus (BNC), and (2) *POS n-grams*, with n ranging from 1 to 5, trained on the five levels in the WeeBit corpus itself. The LMs are used to score the text with log-likelihood and perplexity.

Discourse-based Features Discourse features measure the cohesion and coherence of the text. Three types of discourse-based features are used.

(1) Entity density features

Previous work by Feng et al. (2009; 2010) has shown that entity density is strongly associated with text comprehension. An entity set is a union of named entities and general nouns (including nouns and proper nouns) contained in a text, with overlapping general nouns removed. Based on this, *9 entity density features*, including the total number of all/unique entities per document, the average number of all/unique entities per sentence, percentage of named entities per sentence/document, percentage of named entities in all entities, percentage of overlapping nouns removed, and percentage of unique named entities in all unique entities, are calculated.

(2) Lexical chain features

Lexical chains model the semantic relations among entities throughout the text. The lexical chaining algorithm developed by Galley and McKeown (2003) is implemented. The semantically related words for the nouns in the text, including synonyms, hypernyms, and hyponyms, are extracted from the WordNet (Miller, 1995). Then for each pair of the nouns in the text, we check whether they are semantically related. Finally, lexical chains are built by linking semantically related nouns in text. A set of *7 lexical chain-based* features are computed, including total number of lexical chains per document, total number of lexical chains normalized with text length, average/maximum lexical chain length, average/maximum lexical chain span, and the number of lexical chains that span more than half of the document.[5]

(3) Entity grid features

Another entity-based approach to measure text coherence is the entity grid model introduced by Barzilay and Lapata (2008). They represented each text by an entity grid, which is a two-dimensional array that captures the distribution of discourse entities across text sentences. Each grid cell contains the grammatical role of a particular entity in the specified sentence: whether it is a subject (S), object (O), neither a subject nor an object (X), or absent from the sentence (-). A local entity transition is defined as the transition of the grammatical role of an entity from one sentence to the following sentence. In our experiments, we used the Brown Coreference Toolkit v1.0 (Eisner and Charniak, 2011) to generate the entity grid for the documents. The *probabilities of the 16 types* of local entity transition patterns are calculated to represent the coherence of the text.

4.2 Implementation and Evaluation

In our experiments, we cast readability assessment as a supervised machine learning problem. In particular, a pairwise ranking approach is adopted and compared with a classification method. We believe that the reading difficulty of text is a continuous rather than discrete variable. Text difficulty within a level can also vary. Instead of assigning an abso-

[5]The *length* of a chain is the number of entities contained in the chain. The *span* of a chain is the distance between the indexes of the first and the last entities in the chain.

feature set	Classification		Ranking	
	ACC	PCC	pairwise ACC	PCC
traditional	0.586	0.770	0.862	0.704
lexical	0.578	0.726	0.863	0.743
syntactic	0.599	0.731	0.824	0.692
LM	0.714	0.848	0.872	0.769
discourse	0.563	0.688	0.848	0.659
all combined	**0.803**	**0.900**	**0.924**	**0.848**

Table 3: Classification and ranking results on the WeeBit corpus with feature sets grouped by their type

lute level to the text, treating readability assessment as a ranking problem allows prediction of the relative difficulty of pairs of documents, which captures the gradual nature of readability better. Because of this, we hypothesize that the ranking model can generalize better to unseen texts and texts with different level annotation.

Support vector machines (SVM) have been used in the past for readability assessment by many researchers and have consistently yielded better results when compared to other statistical models for the task (Kate et al., 2010). We use the LIBSVM toolkit (Chang and Lin, 2011) to implement both multiclass classification and pairwise ranking. Five-fold cross validation is used for evaluation. We report two popular performance metrics, accuracy (ACC) and Pearson correlation coefficient (PCC), and use pairwise accuracy to evaluate ranking models. Pairwise accuracy is defined as the percentage of instance pairs that the model ranked correctly. It should be noted that accuracy and pairwise accuracy are not directly comparable. Thus, PCC is introduced to compare the results of the classification and the ranking models.

4.3 Results

In predicting the text reading difficulty on the WeeBit data, the best result is achieved with a combination of all features and a classification model, with ACC=0.803 and PCC=0.900. We performed ablation tests and found that all feature sets have contributed to the overall model performance. Although there have been readability assessment studies on similar datasets, the results obtained in our experiments are not directly comparable to those. One of the major reasons is the modifications that we have made to the corpus (as discussed in Section 3.1). Vajjala and Meurers (2012) reported that a multilayer perceptron classifier using three traditional metrics alone yielded an accuracy of 70.3% on their version of the WeeBit corpus. Their final system achieved a classification accuracy of 93.3% on the five-class corpus. Nonetheless, the best system in our experiments yields results competitive to most existing studies. For reference, Feng et al. (2010) reported an accuracy of 74.01% using a combination of discourse, lexical and syntactic features for readability classification on their Weekly Reader Corpus and an accuracy of 63.18% when using all feature sets described in Schwarm et al. (2005).

Comparing the classification and the ranking models, we note that the results of the two models vary across feature sets and none of the two models is consistently better than the other. When all features are combined, the classification model outperforms the ranking one. It suggests that a ranking model is not necessarily the best model in predicting readability overall when trained and tested on the same dataset.

5 Readability Assessment on L2 Data

So far we have studied the effect of various readability measures on the task of readability assessment and built two different types of models to predict text difficulty. However, the WeeBit corpus consists of texts aimed at native speakers of different ages rather than at L2 readers. Although there are certain similarities concerning reading comprehension between these two groups, the perceived difficulty of texts can be very different due to the difference in the pace and stages of language acquisition. Since the goal of our research is to automatically detect readability levels for language learners, it would be more helpful to work with data that are directly annotated with reading difficulty for L2 learners.

Ideally, it would be good to train a model directly on text annotated with L2 levels and then use this model to estimate readability for the new texts. However, the Cambridge Exams data we have compiled is relatively small, and the model trained on it will likely not generalize well. Therefore, we examined several approaches to make use of the WeeBit corpus for readability assessment on the L2 data.