Text Complexity Analysis Task Report

Abstract

The task was to analyse how the complexity of language changes across levels of the story, using language processing techniques. The task of reading level classification is taken as a proxy task to analyse the text complexity. A classifier is made to classify the stories into different reading levels with various text related features which would probably help to decide the reading complexity level. The absence and presence of such features make in the classification accuracy is taken as the base for text complexity analysis. From various experiments it is found that

1. For Hindi and Telugu stories, both word level features like average number of words per sentence, total number of words in the text along with discourse features like number of pronouns and conjuncts were giving better classification accuracy. One reason for the effectiveness of the word level features could be the morphological richness of Indian languages.
2. For English, predominantly, discourse level features gave better classification accuracy. More than word level features, syntactic features gave better prediction.

To summarise, more the presence of discourse level features, more the complex texts are. Similarly the number of noun phrases.

Text Complexity Analysis

Text complexity analysis is the task of analysing the various factors which induce the challenges in understanding the text. This can be at the level of way in which the text is written or at the conceptual level. This analysis can be used for various purposes, however most commonly used in educational sector for deciding the levels of books based on the

Task

Using language processing techniques, analyse how the complexity of language changes across levels of the story.

Data

The given dataset contains a total of 3697 children stories details from the languages English, Hindi and Telugu. These stories are updated for their reading level complexity. L1 indicates the least complexity and L4 indicates the high complexity.

Domain	Languages	#Total stories	#reading level	reading level	#English	#Hindi	#Telugu

			labels	**labels**			
Children stories	English, Hindi & Telugu	3697	4	L1, L2, L3, L4	2433	920	344

Methodology

The task of reading level classification is taken as a proxy task to analyse the text complexity. A classifier is made to classify the stories into different reading levels with various text related features which would probably help to decide the reading complexity level. The absence and presence of such features make in the classification accuracy is taken as the base for text complexity analysis.

Data Analysis

Content column of the given dataset is the text which is given as input. From the data analysis it is found that there are missing values in the content column of few of the stories. Few of them only had white spaces. Apart from that, languages other than English, Hindi, and Telugu were encountered in content text. For few of the stories, the language of the text is given as 'English' and the content was given in Devanagari script. And vice versa is also true. Statistics of such cases have been given in the table below

#NaN	**#white-space**	**#Content in other languages**	**#total**
57	7	28	92

My decision

One of the option to handle such cases were to replace such cells with the text in the Synopsys column. However, in certain cases the language of such texts were different. Even if we are taking the content from synopsis field, it is meaningless with respect to the task of text Complexity Analysis as synopsis is comparatively very short to content. Moreover, in few of such cases, the labels were L3 and L4. If we take synopsis in such cases, this would be meaningless. Hence my decision was to drop such stories from train and test sets.

Language	**#L1**	**#L2**	**#L3**	**#L4**
Hindi	386	300	134	81
Telugu	122	107	75	38

English	992	828	395	147

Data cleaning

Lots of noise have been encountered while cleaning.

1. Html style texts were present along with most of the content.
    a. Solved to certain extent by regular expressions.
2. A lot white spaces within content which includes “\n”, “\t”, “\s”, “\r” etc.
    a. Solved by regular expressions.
3. Other language characters/words.
    a. Text was taken based on the unicode character range of the particular language in which the story is written.

Features employed in learning

Features for the reading level classifications are taken based on the paper “https://aclweb.org/anthology/W16-3620”.

1. **Surface Features**
    a. Average number of characters per word in the story.
    b. Average number of words per sentence in the story.
    c. Total number of words in the story

2. **Syntactic Features**
    a. Number of noun phrases in the story
    b. Number of verb phrases in the story

3. **Discourse and coherence features.**
    a. Number of pronouns in the story.
    b. Number of coordinate and subordinate conjuncts in the story.

Classification Parameters and other details.

1. Classification was done using Random Forest Classifier as it has the capability to find the important features which affected the decision making.
2. A 5 fold cross-validation is done on the cleaned story dataset.
3. n_estimators=300, max_depth=4
4. Since there exists data imbalance(visible from data statistics), class weights were balanced.
5. Pandas library for csv file handling and processing.
6. Scikit-learn for the classification task.
7. Numpy for vector operations.
8. Spacy parser for English

Experiments & Results

1. **Hindi & Telugu**
    a. Surface level features and discourse level features were used.
    b. It is found that when both the features are combined the F1 score is high.
    c. However, discourse feature alone can give good performance
2. **English**
    a. All the features, namely surface, syntactic and discourse features were used.
    b. It is found that performance is better when only discourse features were employed.

Hindi Experiment Results

FEATURES	precision	recall	f-measure
SURFACE FEATURES	0.5906	0.5929	0.5741
DISCOURSE FEATURES	0.6455	0.596	0.5858
SURFACE+ DISCOURSE FEATURES	0.6656	0.6167	**0.6098**

Telugu Experiment Results

FEATURES	precision	recall	f-measure
SURFACE FEATURES	0.6817	0.6569	0.6575
DISCOURSE FEATURES	0.7486	0.679	0.6863
SURFACE+ DISCOURSE FEATURES	0.7541	0.6878	**0.692**

English Experiment Results

ENGLISH	Precision	Recall	F-score
SURFACE FEATURES	0.5296	0.6790	0.4871
DISCOURSE FEATURES	0.5788	0.5418	**0.5445**

SYNTACTIC FEATURES	0.5862	0.5307	0.5413
SURFACE FEATURES + DISCOURSE FEATURES	0.5654	0.5361	0.5364
SYNTACTIC FEATURES+DISCOURSE FEATURES	0.5758	0.5374	0.5426
SURFACE FEATURES+SYNTACTIC FEATURES	0.5616	0.5250	0.5294
SURFACE FEATURES+DISCOURSE FEATURES+SYNTACTIC FEATURES	0.5679	0.5292	0.5330

Various possible further experiments

Due to the time constraints, I couldn't conduct few experiments. Those are given below.

1. Incorporating syntax level features for Hindi and Telugu with the help of POS taggers. Rules can be written to get noun phrases and verb phrases.
2. Research papers mention that presence of rare words indicate the complexity. For English, the way is to find the frequency of a particular word from Google ngram frequency. However for Indian languages, anchoring non frequency of a word would be a bad idea since the data would be sparse because of the morphological richness. One option is to anchor on the mutual information scores of character level ngrams.
3.