Automated Capturing of Psycho-linguistic Features in Reading Assessment Text

Makoto Sano

Prometric

Correspondence may be sent to:

Makoto Sano

Prometric

makoto.sano@prometric.com

Paper presented at the annual meeting of

National Council on Measurement in Education (NCME)

Chicago, Illinois

April, 2015

Abstract

This study used PLIMAC, a natural language processing tool, developed by the author to automatically capture psycho-linguistic features of passage based multiple choice items. Items from the 2011 NAEP Grade 8 Reading assessment were evaluated and regression analyses were conducted to identify the psycho-linguistic features that best predicted overall item difficulty (*p*-value). Results indicated that three psycho-linguistic features captured by PLIMAC accounted for 30% and 56% of the variance of item difficulty using multiple linear and tree-based regression techniques, respectively. This is considerably smaller when compared to the 89% of the variance result found by Kirsch and Mosenthal (1990) for document literacy tasks. PLIMAC still has some room of improvement in capturing the psycho-linguistic features in applying further intensive natural language processing.

Automated Capturing of Psycho-linguistic Features in Reading Assessment Text

Passage based multiple-choice (MC) reading comprehension test items are commonly used on high stakes assessments as this type of item has the ability to measure higher order language functioning, can be easily machine or hand score across large groups of examinees, and if well-constructed, has more robust statistical qualities in comparison to open ended constructed response (CR) item types. Unfortunately, it is not always easy for test developers to formulate MC passage based reading comprehension items in a consistent manner as it is sometimes difficult to subjectively classify items into categories that reflect the appropriately outlined cognitive complexity of the item.

The primary goal of the research is to explore and identify, through the uses of natural language processing (NLP) techniques, the psycho-linguistic features associated with reading passage MC item types that can be used to predict item difficulty levels of these item types. It is presumed that once certain psycho-linguistic features are identified, these features can be used to alter both item content and format to help automate the item development process.

Research has extensively explored the cognitive processes, skills, and knowledge that examinees are required to have in order to successfully answer these types of questions (see Leighton & Gierl, 2007) and several cognitive models predicting item difficulty have been developed (Drum, Calfee & Cook, 1981; Embretson & Wetzel, 1987; Freedle & Kostin, 1991/1992; Gorin, 2005; Gorin & Embretson, 2006; Kirsch & Mosenthal, 1990; Rupp, Garcia, & Jamieson, 2001).

Embretson and Wetzel (1987) proposed a cognitive model on reading comprehension items using propositional analyses of the item texts by raters incorporating linguistic surface structure variables and an index of word frequency (see Drum, et al,1981; Kucera-Francis,1967).

Results from this study indicated that the difficulty of passage based MC items may depend more on the complexity of individuals' response decisions than the complexity of the propositions in the passages.

Research has also indicated the importance of structural relations (correspondence) between questions and document literacy tasks, indicating that three principle stages of cognitive processing occur during this type of task completion: 1) Distinguishing between *Given* and *Requested* information in a question or directive, 2) Matching information *Given* to corresponding information in the document, and 3) Looking for the *Requested* information that is likely to contain the answer to the question or directive (see Kirsch & Mosenthal, 1988/1990; Mosenthal & Kirsch,1991).

In addition to research focusing the relationship between cognition and item difficulty, there has also been significant advancements in the use of NLP tools to try to capture these relationships. (see Cherry & Vesterman, 1980; Embretson & Wetzel, 1987; Rupp et al., 2001). Coh-Metrix, (Graesser et al., 2004), an advanced NLP tool, has the capacity to analyze text on over 200 measures of cohesion, language and readability using lexicons, part-of-speech classifiers, syntactic parsers, templates, corpora, latent semantic analysis, and other components that are widely used in computational linguistics.

While the Coh-Metrix NLP tool attempts to provide generalized measures of passage based text cohesion and text difficulty, it does not focus on the cognitive complexity of MC items. This study with PLIMAC attempts to develop further optimized functions to capture the cognitive complexity of exam items structured by a passage, an item stem, and item options.

Methods

The approach by Kirsch and Mosenthal (1990) above is used in this study to capture the variables which represent the five degrees (difficulty levels) of correspondence between questions and documents: literally correspondence, synonymous correspondence, correspondence arrived at via a low text-based inference, correspondence arrived at via a high text-based inference, and correspondence determined based upon special prior knowledge. The five degrees of correspondence construct a significant valuable for item difficulty prediction in their model. Although most of the variables identified in their study were captured by human raters, their approach is relatively easy computationally since the variables captured are based on structures of item content texts rather than test takers' judgment processes (Embretson & Wetzel, 1987) and PLIMAC may just focus on capturing the structures of the texts.

The NLP tool created by the author and used for this study is the Psycho-Linguistic Measures of Assessment Content (PLIMAC). The PLIMAC was developed in Python and its Natural Language Tool Kit (NLTK, Bird, Klein, and Loper, 2009) to capture psychological and linguistic features automatically from reading passage and MC item text. PLIMAC includes the following 10 functionalities (see Figure 1):

1. **Test content text parser:** Test content text data is usually semi-structured by reading passages, item stems, and options. Most input text data is not completely structured (i.e., XML format) and does not have clear boundaries/delimiters between the reading passages, item stems, and options. The test content text parser enables to detect the boundaries in plane texts and re-structure the test content text data into a CSV-formatted matrix.

2. **Tokenizer and part-of-speech tagger:** A tokenizer divides text into a sequence of words. A part-of-speech tagger classifies the tokenized words into a lexical category such as noun, verb, or adjective.
3. **Lemmatizer:** A lemmatizer provides a process of grouping together the different inflected forms of a word (e.g. run, runs, ran and running) into a lemma (e.g. run) so they can be analyzed as a single word. As the first step identifying the degree of correspondence (Kirsch & Mosenthal, 1990), the number of lemmas in the reading passages, item stems, and options were captured after the lemmatization.
4. **The ability to retrieve synonyms from Wordnet (Miller, 1995):** Wordnet is used to retrieve synonyms in the item stems and options and the number of synonyms were captured to identify the degree of *synonymous correspondence*.
5. **The ability to count of overlapping lemmas and synonyms in reading passage, item stem and options:** This function provides information to quantify the first and second degrees of correspondence (*literally correspondence* and *synonymous correspondence*, Kirsch and Mosenthal, 1990) between reading passage, item stem and options.
6. **The ability to retrieve lemma frequencies from lexical resources:** In PLIMAC, an open source lexical resource of The Open American National Corpus (Reppen, Ide, and Suderman, 2005) is used for the index of word frequency.
7. **The calculation of lemmas and synonyms location/distance in reading passage:** Freedle and Kostin (1991) hypothesized that relevant main idea information (as *Requested* information) that is located early in the reading passage will facilitate main idea item correctness. PLIMAC quantifies the location of lemmas in reading passage which are overlapping with lemmas/synonyms in item stems or options. Also PLIMAC

quantifies the distance (number of words away) across lemmas overlapping with lemmas/synonyms in item stems or options.

8. **The ability to reformat psycho-linguistic features and given item stats/information:** PLIMAC reformats captured psycho-linguistic features into CSV-formatted matrices together with item stats or item information initially given by the user of PLIMAC.
9. **The ability to produce summary statistical reports:** PLIMAC reports descriptive statistics, zero-order correlations regarding item stats, item information, and psycho-linguistic features with CSV-formatted charts.
10. **The ability to perform multiple liner regression and tree-based regression:** PLIMAC performs multiple linier regression analysis and tree-base regression analysis using psycho-linguistic features as predictors of item stats (i.e., item difficulty).

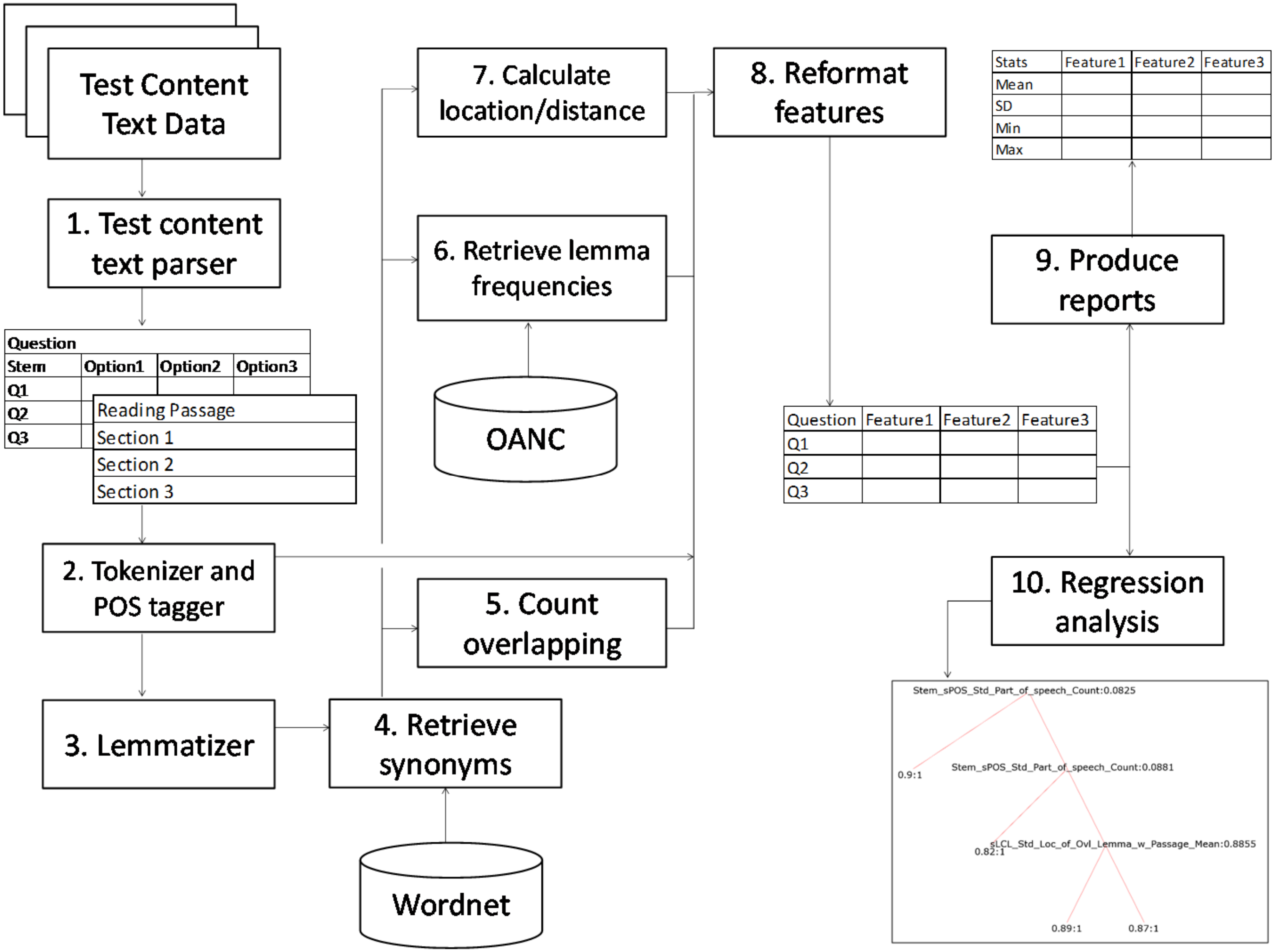

Figure 1. PLIMAC Function Diagram

For this study, the following psycho-linguistic features were automatically captured by PLIMAC:

1) The number of lemmas/synonyms in a keyed option, overlapping with the lemmas in a passage/item stem/distractor options [OLP/OSP/OLS/OSS/OLO/OSO],
2) The number of lemmas/synonyms in an item stem, overlapping with the lemmas in a passage/all options [OLP/OSP/ OLO/OSO],
3) The total Part-Of-Speech[1] count in a passage/item stem/keyed option [POS],
4) The Mean/SD of locations of lemmas in a passage, overlapping with the lemmas/synonyms in an item stem/keyed option [LCL/LCS],
5) The distance between lemmas in a passage, overlapping with the lemmas/synonyms in an item stem, and overlapping with the lemmas/synonyms in a keyed option [DTL/DTS],
6) The Mean/SD/Min/Max of the frequency of lemmas in a passage/item stem/keyed option [FRQ],
7) The standardized (divided by the total Part-Of-Speech count in a passage) number of lemmas/synonyms in an item stem/keyed option, overlapping with the lemmas in a passage [sOLP/sOSP],
8) The standardized total Part-Of-Speech count in an item stem [sPOS],
9) The standardized Mean of locations of lemmas in a passage, overlapping with the lemmas/synonyms in an item stem/keyed option [sLCL/sLCS], and

[1] Part-of-speech is a word class or a lexical category. In this study, a Part-Of-Speech also means an individual word tokenized and classified into a lexical category

10) The standardized distance between lemmas in a passage, overlapping with the lemmas/synonyms in an item stem, and overlapping with the lemmas/synonyms in a keyed option [sDTL/sDTS].

Real Data Evaluation

For the initial evaluation of the PLIMAC functionalities, content from the NAEP 2011 Grade 8 Reading assessment was used. The assessment content included four reading passages with 21 MC questions.

Captured Psycho-linguistic Features

Table 1 shows the descriptive statistics of captured psycho-linguistic features in a keyed option, an item stem and a reading passage. The acronyms [POS], [FRQ], [LCL], [LCS], [DTL], [DTS], [OLP], [OSP], [OLS], [OSS], [OLO], [OSO], [sOLP], [sOSP], [sLCL], [sLCS], [sDTL], [sDTS], and [sPOS] shown at the beginning of the psycho-linguistic features correspond to the PLIMAC psycho-linguistic features listed in the previous section. Note that the Mean/SD/Min/Max of the frequency of lemmas [FRQ] in a passage/item stem/keyed option is calculated based on the word (lemma) frequency information retrieve from The Open American National Corpus (Reppen, Ide, & Suderman, 2005). If a particular word included in a passage/item stem/keyed option is not retrieved from the corpus, the frequency of the word is recognized as zero. Also each NAEP 2011 Grade 8 Reading assessment passage has multiple sections divided by the page breaks. If an item stem explicitly specifies a particular section (page) of the reading passage, the structural relationships between the question and passage are captured just from the specified section(s). Some of the questions share the same section of the reading passage. However, PLIMAC repetitively calculates the psycho-linguistic features from the corresponded section(s) for each question. So some of the reading passage sections were double-counted across questions in the last portion of Table 1 (Psycho-linguistic Features in A Passage).

Zero-order correlations between each feature variable and item difficulty (*p*-value) were calculated in Table 2. The features which have the absolute value of correlations equal to or greater than 0.30 are shown as bold styled. The strongest negative correlation shown over the all 21 items was [OSS] Number Overlapping Synonym w/ Stem in a keyed option (-0.42) followed by [POS] Part-of-speech Count in a keyed option (-0.39) and [OLP] Number Overlapping Lemma w/ Passage in a keyed option (-0.38). That means much longer keyed options which have more overlapping with the synonyms in item stems, and with the lemmas in reading passages make questions more difficult. The strongest positive correlation shown over the all 21 questions was [sPOS] Standardized Part-of-speech Count in a stem (0.31) followed by [FRQ] Lemma Frequency Max in a stem (0.30). That means if the relative length of an item stem compared to the corresponded reading passage is longer and the words in the item stem are more common, the question is easier.

Multiple Linear Regression Analysis

Multiple linear regression analyses were performed using the feature variables as predictors of item difficulty. Results from regression analyses indicated that the best predictors of item difficulty were (1) [POS] Part-of-speech Count in a keyed option, (2) [OSS] Number Overlapping Synonym w/ Stem in a keyed option, and (3) [sPOS] Standardized Part-of-speech Count in a stem. The amount of variance captured by the three predictors was 0.30 (as the adjusted R^2). This suggests that just 30% of the variance of item difficulty can be explained by the set of independent variables which is considerably smaller when compared to the 89% of the variance result found by Kirsch and Mosenthal (1990). Since the PLIMAC result also shows the zero-order correlations relatively smaller than Kirsch and Mosenthal's (1990; the maximum of the correlations was 0.85), the linear regression by PLIMAC does not perform as well as expected.

Tree-based Regression Analysis

Tree-based regression analyses were performed using the feature variables as predictors of item difficulty in Figure 2. In this study, the Classification and Regression Trees (CART) algorithm (Breiman, Friedman, Olshen, & Stone, 1984) was used as the same manner as Gao and Rogers (2011), Rupp et al. (2001), and Sheehan (1997). The CART algorithm forms clusters of questions which have similar psycho-linguistic feature values, by successively splitting the questions into increasingly homogeneous subsets called nodes. As the first step of the recursive partitioning, all possible splits of question groups are evaluated by deviance. In this study, the deviance *D* is the sum of squared differences between an item *p*-value and the mean *p*-value of all questions belonging to a single node.

$$D(y, \hat{y}) = \sum (y_i - \hat{y})^2 \qquad (1)$$

where $\hat{y}$ is the mean *p*-value and y_i is the *p*-value of item *i*.

Then the best split (by a particular psycho-linguistic feature value) is found as maximizing the difference in deviance ΔD between the parent node and the sum of the two child nodes.

$$\Delta D = D(y, \hat{y}) - D_{split}(y, \hat{y}_L, \hat{y}_R) \qquad (2)$$

$$D_{split}(y, \hat{y}_L, \hat{y}_R) = \sum (y_i - \hat{y}_L)^2 + \sum (y_i - \hat{y}_R)^2 \qquad (3)$$

where $\hat{y}_L$ is the mean *p*-value in the left child node and $\hat{y}_R$ is the mean *p*-value in the right child node.

Table 1. Descriptive Statistics of Psycho-Linguistic Features

	P-value	Psycho-linguistic Features in A Keyed Option				
		[POS] Part-of-speech Count	[FRQ] Lemma Frq. Mean	[FRQ] Lemma Frq. SD	[FRQ] Lemma Frq. Min	[FRQ] Lemma Frq. Max
Count	21	21	21	21	21	21
Mean	0.70	7.00	35002.53	44534.68	812.95	126374.57
SD	0.14	2.53	29470.56	40127.79	1123.54	118709.58
Min	0.39	2.00	2067.75	1740.58	0.00	4969.00
Max	0.90	13.00	85647.29	114876.51	4440.00	339971.00
	[LCL] Location of Overlapping Lemma w/ Passage Mean	[LCL] Location of Overlapping Lemma w/ Passage SD	[LCS] Location of Overlapping Synonym w/ Passage Mean	[LCS] Location of Overlapping Synonym w/ Passage SD	[DTL] Distance of Overlapping Lemmas in Passage	[DTS] Distance of Overlapping Synonyms in Passage
Count	21	21	18	18	20	17
Mean	333.11	116.84	337.00	151.50	39.53	131.44
SD	187.56	102.13	215.77	136.62	40.48	162.89
Min	98.25	0.00	53.00	0.00	0.42	3.63
Max	829.17	419.40	773.18	427.12	125.10	508.00
	[OLP] Number Overlapping Lemma w/ Passage	[OSP] Number Overlapping Synonym w/ Passage	[OLS] Number Overlapping Lemma w/ Stem	[OSS] Number Overlapping Synonym w/ Stem	[OLO] Number Overlapping Lemma w/ Distractor Options	[OSO] Number Overlapping Synonym w/ Distractor Options
Count	21	21	21	21	21	21
Mean	3.52	3.33	1.10	0.38	3.43	0.38
SD	1.57	2.76	1.18	0.67	3.11	0.67
Min	1	0	0	0	0	0
Max	8	9	3	2	9	2
	[sOLP] Std. Number Overlapping Lemma w/ Passage	[sOSP] Std. Number Overlapping Synonym w/ Passage	[sLCL] Std. Location of Overlapping Lemma w/ Passage Mean	[sLCS] Std. Location of Overlapping Synonym w/ Passage Mean	[sDTL] Std. Distance of Overlapping Lemmas in Passage	[sDTS] Std. Distance of Overlapping Synonyms in Passage
Count	21	21	21	18	20	17
Mean	0.0092	0.0076	0.7043	0.6463	0.1122	0.2332
SD	0.0051	0.0065	0.1897	0.2352	0.1182	0.2013
Min	0.0013	0.0000	0.3712	0.2335	0.0004	0.0048
Max	0.0198	0.0274	1.1123	1.1825	0.3504	0.6680

Table 1. Descriptive Statistics of Psycho-Linguistic Features (continued)

Psycho-linguistic Features in A Stem

	[POS] Part-of-speech Count	[FRQ] Lemma Frq. Mean	[FRQ] Lemma Frq. SD	[FRQ] Lemma Frq. Min	[FRQ] Lemma Frq. Max
Count	21	21	21	21	21
Mean	18.71	41078.38	70715.88	135.52	245689.29
SD	9.31	20842.29	37087.29	462.71	127914.45
Min	6.00	9543.90	13372.75	0.00	35829.00
Max	49.00	79668.23	116752.25	2150.00	339971.00

	[OLP] Number Overlapping Lemma w/ Passage	[OSP] Number Overlapping Synonym w/ Passage	[OLO] Number Overlapping Lemma w/ All Options	[OSO] Number Overlapping Synonym w/ All Options
Count	21	21	21	21
Mean	9.38	4.48	3.48	0.86
SD	7.62	4.11	3.54	1.01
Min	0	0	0	0
Max	37	14	9	4

	[sPOS] Std. Part-of-speech Count	[sOLP] Std. Number Overlapping Lemma w/ Passage	[sOSP] Std. Number Overlapping Synonym w/ Passage
Count	21	21	21
Mean	0.0537	0.0259	0.0102
SD	0.0316	0.0174	0.0081
Min	0.0066	0.0000	0.0000
Max	0.0913	0.0645	0.0278

Psycho-linguistic Features in A Passage

	[POS] Part-of-speech Count	[FRQ] Lemma Frq. Mean	[FRQ] Lemma Frq. SD	[FRQ] Lemma Frq. Min	[FRQ] Lemma Frq. Max
Count	21	21	21	21	21
Mean	513.62	33840.15	63836.28	0.00	332931.29
SD	326.05	5596.79	9785.41	0.00	31206.67
Min	227.00	26569.70	49647.07	0.00	258821.00
Max	1127.00	45247.23	81683.80	0.00	347939.00

Table 2. Correlations Between Psycho-Linguistic Features and *p*-value

Psycho-linguistic Features in A Keyed Option	Correlation to P-value
[POS] Part-of-speech Count	**-0.39**
[FRQ] Lemma Frq. Mean	0.03
[FRQ] Lemma Frq. SD	0.05
[FRQ] Lemma Frq. Min	0.20
[FRQ] Lemma Frq. Max	0.04
[LCL] Location of Overlapping Lemma w/ Passage Mean	-0.12
[LCL] Location of Overlapping Lemma w/ Passage SD	**-0.34**
[LCS] Location of Overlapping Synonym w/ Passage Mean	**-0.42**
[LCS] Location of Overlapping Synonym w/ Passage SD	**-0.30**
[DTL] Distance of Overlapping Lemmas in Passage	0.11
[DTS] Distance of Overlapping Synonyms in Passage	-0.23
[OLP] Number Overlapping Lemma w/ Passage	**-0.38**
[OSP] Number Overlapping Synonym w/ Passage	-0.18
[OLS] Number Overlapping Lemma w/ Stem	0.15
[OSS] Number Overlapping Synonym w/ Stem	**-0.42**
[OLO] Number Overlapping Lemma w/ Distractor Options	-0.03
[OSO] Number Overlapping Synonym w/ Distractor Options	-0.05
[sOLP] Std. Number Overlapping Lemma w/ Passage	0.07
[sOSP] Std. Number Overlapping Synonym w/ Passage	-0.05
[sLCL] Std. Location of Overlapping Lemma w/ Passage Mean	**0.33**
[sLCS] Std. Location of Overlapping Synonym w/ Passage Mean	**-0.33**
[sDTL] Std. Distance of Overlapping Lemmas in Passage	**0.37**
[sDTS] Std. Distance of Overlapping Synonyms in Passage	0.04

Psycho-linguistic Features in A Stem	Correlation to P-value
[POS] Part-of-speech Count	0.01
[FRQ] Lemma Frq. Mean	0.17
[FRQ] Lemma Frq. SD	0.23
[FRQ] Lemma Frq. Min	-0.10
[FRQ] Lemma Frq. Max	**0.30**
[OLP] Number Overlapping Lemma w/ Passage	-0.12
[OSP] Number Overlapping Synonym w/ Passage	-0.28
[OLO] Number Overlapping Lemma w/ All Options	0.08
[OSO] Number Overlapping Synonym w/ All Options	-0.04
[sPOS] Std. Part-of-speech Count	**0.31**
[sOLP] Std. Number Overlapping Lemma w/ Passage	0.12
[sOSP] Std. Number Overlapping Synonym w/ Passage	-0.24

Psycho-linguistic Features in A Passage	Correlation to P-value
[POS] Part-of-speech Count	**-0.31**
[FRQ] Lemma Frq. Mean	0.14
[FRQ] Lemma Frq. SD	0.07
[FRQ] Lemma Frq. Min	-
[FRQ] Lemma Frq. Max	0.01

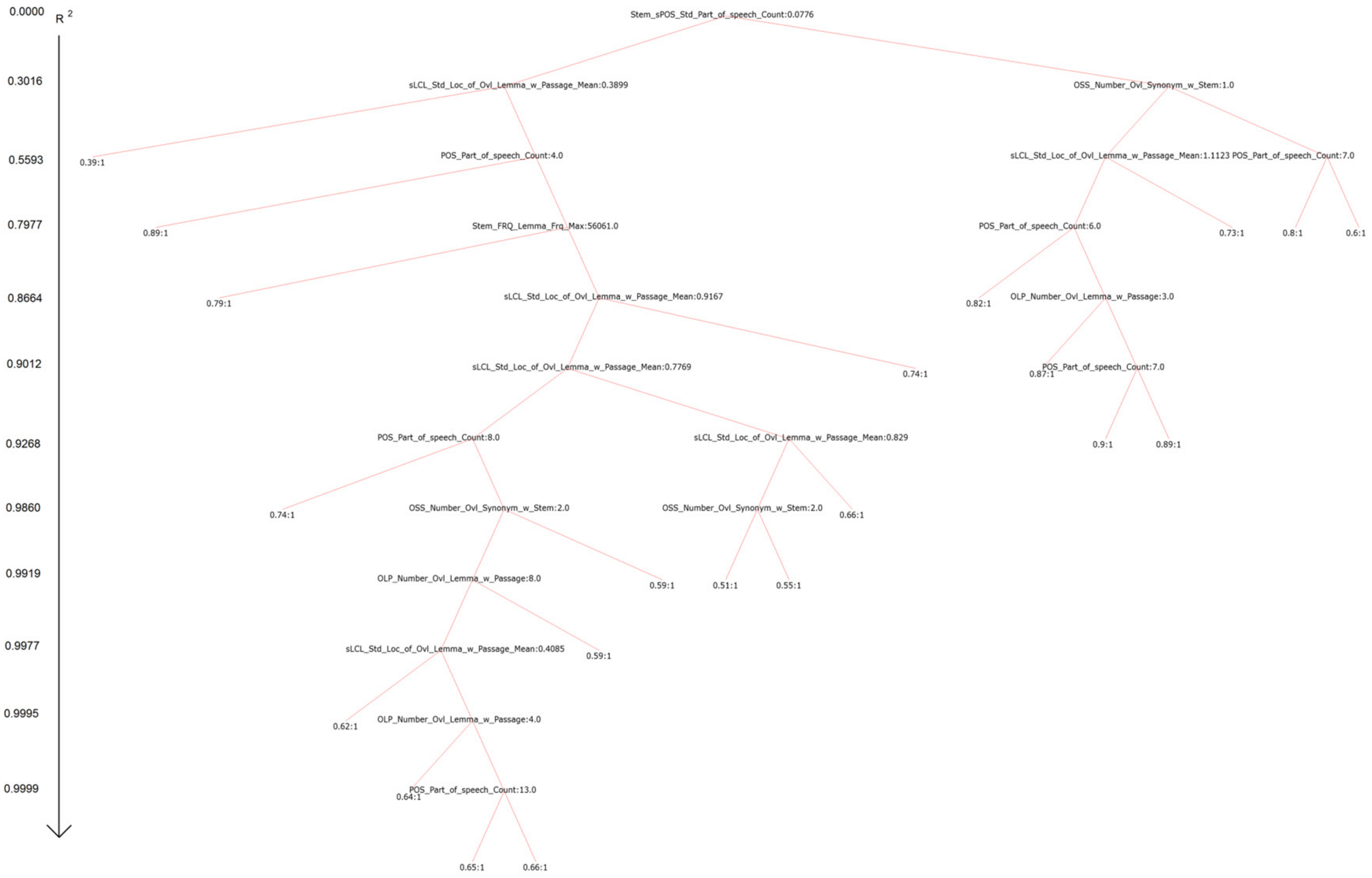
0.0000
R 2
0.3016
0.5593
0.7977
0.8664
0.9012
0.9268
0.9860
0.9919
0.9977
0.9995
0.9999
Stem_sPOS_Std_Part_of_speech_Count:0.0776
sLCL_Std_Loc_of_Ovl_Lemma_w_Passage_Mean:0.3899
OSS_Number_Ovl_Synonym_w_Stem:1.0
0.39:1
POS_Part_of_speech_Count:4.0
sLCL_Std_Loc_of_Ovl_Lemma_w_Passage_Mean:1.1123
POS_Part_of_speech_Count:7.0
0.89:1
Stem_FRQ_Lemma_Frq_Max:56061.0
POS_Part_of_speech_Count:6.0
0.73:1
0.8:1
0.6:1
0.79:1
sLCL_Std_Loc_of_Ovl_Lemma_w_Passage_Mean:0.9167
0.82:1
OLP_Number_Ovl_Lemma_w_Passage:3.0
sLCL_Std_Loc_of_Ovl_Lemma_w_Passage_Mean:0.7769
0.74:1
0.87:1
POS_Part_of_speech_Count:7.0
POS_Part_of_speech_Count:8.0
sLCL_Std_Loc_of_Ovl_Lemma_w_Passage_Mean:0.829
0.9:1
0.89:1
0.74:1
OSS_Number_Ovl_Synonym_w_Stem:2.0
OSS_Number_Ovl_Synonym_w_Stem:2.0
0.66:1
OLP_Number_Ovl_Lemma_w_Passage:8.0
0.59:1
0.51:1
0.55:1
sLCL_Std_Loc_of_Ovl_Lemma_w_Passage_Mean:0.4085
0.59:1
0.62:1
OLP_Number_Ovl_Lemma_w_Passage:4.0
0.64:1
POS_Part_of_speech_Count:13.0
0.65:1
0.66:1

Figure 2. A Tree-based Regression Analysis with Six Psycho-linguistic Features

Figure 2 shows a result with applying six psycho-linguistic features. These six features were selected as the six highest absolute values of correlations to *p*-value over the all 21 questions. At the root node, the best split was found using the [sPOS] Standardized Part-of-speech Count in a stem variable at 0.0776. The left hand side of the child node has a group of questions which feature values of [sPOS] are below 0.0776, while the right hand side of the childe node has a group of questions which feature values of [sPOS] are equal to or greater than 0.0776. Even just applying this one psycho-linguistic feature variable at the root node, the clustering result explains 30% of the variance as shown by the estimated (i.e., mean) *p*-value of questions in the left-side and right-side groups (0.64 and 0.80 respectively). The second splits (at the second nodes) were identified using the [sLCL] Standard Location of Overlapping Lemma w/ Passage Mean in a keyed option variable at 0.3899, and the [OSS] Number Overlapping Synonym w/ Stem in a keyed option variable at 1.0. Applying these three values from the root to the second nodes, the clustering result explains 56% of the valiance of item difficulty. Therefore, at least for the 21 questions from NAEP 2011 Grade 8 Reading assessment, tree-based regression show much better performance on predicting item difficulty than the multiple linear regression.

Each end of the branch in Figure 2 shows the *p*-value of question(s) belonging to the branch followed by the number of question(s). For example, the end of the branch located on the far left side shows 0.39:1. This means that the branch has one item whose *p*-value is 0.39. Even though Figure 2 shows 100% of the variance is explained after all 12 recursive partitions, it does not mean the item difficulty of NAEP Grade 8 Reading assessment questions can be explained 100% by the six psycho-linguistic features. This prediction model is just applicable for the certain 21 questions and further research should study if the all or a part of this prediction model can be applied to other NAEP Grade 8 Reading assessment questions.

Discussion

Taking account of the item content to review the tree-based regression analysis result, the 30% of the variance explained by the first node can be interpreted as follows: Eight out of the 14 harder questions (in the left-side group, the mean *p*-value was 0.64) have more than one section (page) of the corresponded reading passages (specified/not specified by the item stems), while six out of the seven easier questions (in the right-side group, the mean *p*-value was 0.80) have just one section (the remaining question has two sections). Also five out of the seven questions from the right-side group have repetitive phrases in the item stem quoted from the reading passages (in order to accommodate the clarification of the *Required* information). They make item stems in the right-side group much longer than the left-side group and affect the feature values of [sPOS] as Standardized (divided by the total Part-Of-Speech count in a passage) Part-of-speech Count in a stem. These two things accommodate fewer sections to be referred and more information given by the item stems for the right-side group, then they make questions much easier. Finally, the questions which have much larger feature value of [sPOS] (in the right-side group) have resulted in much higher *p*-value.

It was initially expected that the psycho-linguistic features related to the correspondence across reading passages, item stems and options, especially the features of [OLP/OLS/OLO] Number Overlapping Lemma w/ Passage/Stem/Distractor Options, should take a major role in explaining item difficulty. But, at least for the 21 questions of NAEP Grade 8 Reading assessment, they did not contribute as much as expected. One of the reasons may be that the captured overlapping lemmas do not necessarily represent the *literally correspondence* suggested by Kirsch and Mosenthal (1990). Most of the overlapping lemmas captured were something like common words as “for”, “have, “her”, or “his”. They were distributed through the reading passages, not only at the position where the *Required* information located. This shows PLIMAC has some room of improvement in capturing the psycho-linguistic features which represent correspondence more directly in applying further intensive NLP.

As a part of future steps, the function of the tree-based regression analysis needs to be enhanced in some points. First, it should have a feature to predict item difficulty for the new questions not included in building the regression model. Second, a pruning the tree feature is necessary. This would allow PLIMAC users to stop recursive partitioning at any preferred node level so as to avoid over fitting to a particular set of questions and build more widely applicable prediction models. Third, the graphic representations need to be improved in order to observe the mean *p*-value of questions of the left/right groups in a figure. Sheehan's (1997) approach, which shows item difficulty on the horizontal axis, is one of the possible implementations to follow.

Acknowledgements

I would like to express my deepest appreciation to our corporate advisor, Ric Luecht, my colleagues, Tammy J. Trierweiler, and Ian Clifford, whose contribution in stimulating suggestions, helped me to coordinate my research especially in writing this paper.

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media, Inc.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression

trees. Monterey, CA: Wadsworth, Inc.

Cherry, L. L., & Vesterman, W. (1980). Writing tools-The STYLE and DICTION programs. *Computing Science Technical Report No. 91*, Murray Hill, NJ: Bell Laboratories.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly*, 16, 486-514.

Embretson, S. E. & Wetzel, C. D. (1987). Component latent models for paragraph comprehension. *Applied Psychological Measurement*, 11, 175-193.

Freedle, R., & Kostin, I. (1991). *The Prediction of SAT Reading Comprehension Item Difficulty for Expository Prose Passages.* ETS Research Report RR-91-29. Princeton, NJ: Educational Testing Service.

Freedle, R., & Kostin, I. (1992).*The prediction of GRE reading comprehension item difficulty for expository prose passages for each of three item types: main ideas, inferences, and explicitstatements*. ETS Research Report RR-91-59, Princeton, NJ: Educational Testing Service.

Gao, L., & Rogers, W. T. (2011). Use of tree-based regression in the analyses of L2 reading test items. *Language Testing*, 28, 77-104.

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, 42, 351-373.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, 3, 394-411.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers*, 36, 193-202.

Kirsch, I. S., & Mosenthal, P. B. (1988). *Understanding document literacy: Variables underlying the performance of young adults.* ETS Research Report RR-88-62, Princeton, NJ: Educational Testing Service.

Kirsch, I. S., & Mosenthal, P. B. (1990). Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly*, 25, 5-30.

Kucera, H., & Francis, W. (1967). *Computational analysis of presentday American English. Providence*, RI: Brown University Press.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, 26, 3-16.

Miller,G. A. (1995). WordNet: A Lexical database for English. *Communications of the ACM*, 38, 39-41.

Mosenthal, P.B., & Kirsch, I.S. (1991). Toward an explanatory model of document literacy. *Discourse Processes*, 14, 147-180.

Reppen, R., Ide, N., & Suderman, K. (2005). *American National Corpus (ANC) Second Release* LDC2005T35, Web Download, Philadelphia, PA: Linguistic Data Consortium.

Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing*, 1, 185-216.

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement*, 34, 333-352.