A New Approach to Automated Text Readability Classification based on Concept Indexing with Integrated Part-of-Speech n-gram Features

Abigail R. Razon
University of Birmingham
ARR125@cs.bham.ac.uk

John A. Barnden
University of Birmingham
J.A.BARNDEN@cs.bham.ac.uk

Abstract

This study is about the development of a learner-focused text readability indexing tool for second language learners (L2) of English. Student essays are used to calibrate the system, making it capable of providing realistic approximation of L2s' actual reading ability spectrum. The system aims to promote self-directed (i.e. self-study) language learning and help even those L2s who can not afford formal education.

In this paper, we provide a comparative review of two vectorial semantics-based algorithms, namely, Latent Semantic Indexing (LSI) and Concept Indexing (CI) for text content analysis. Since these algorithms rely on the *bag-of-words* approach and inherently lack grammar-related analysis, we augment them by incorporating Part-of-Speech (POS) n-gram features to approximate syntactic complexity of the text documents.

Based on the results, CI-based features outperformed LSI-based features in most of the experiments. Without the integration of POS n-gram features, the difference between their mean exact agreement accuracies (MEAA) can reach as high as 23%, in favor of CI. It has also been proven that the performance of both algorithms can be further enhanced by combining POS bi-gram features, yielding as high as 95.1% and 91.9% MEAA values for CI and LSI, respectively.

1 Introduction

Text Readability is often defined as how easily documents can be read and understood. Moreover, **Text Readability Indexing** (TRI) is the process wherein texts are classified according to their difficulty level based on educational standards set by institutions.

We take it as one of our working hypotheses that language learning is something personal and that text interpretations are greatly influenced by the learner's personality, preferences, experiences, and beliefs which are not something that can be easily set to a particular standard. Thus, TRI systems should be modelled from the learners themselves and that these systems should have the ability to adapt to the learner's learning progression.

Past researches on this domain, such as (Si and Callan, 2001) and (Heilman et al., 2007), rely greatly on syntactic features as indicators of text readability. Such features include sentence length, syllable and character counts per word, part-of-speech (POS) tags, and word frequency. Although these features are important linguistic components, these have not been sufficient to model reading difficulty levels. As a result, recent studies are geared towards using content learning techniques from the Natural Language Processing (NLP) area. Such techniques include Latent Semantic Indexing (LSI) and Concept Indexing (CI) which have the ability to extract text content features that can be used to model learner profiles within each school grade level.

There have been several attempts to combine grammar- and content-based features in readability analysis. However, it still hasn't been fully investigated so far and, to the best of our knowledge, the combined analysis of CI, a vectorial semantics-based algorithm similar to LSI, and POS n-gram features hasn't been explored at all for text readability indexing.

In this paper, we present a comparative study on LSI and CI with the integration of POS n-gram features. Section 2 and 3 give the summaries of related work and existing LSI versus CI researches, respectively. Our study's working assumptions are

Proceedings of Recent Advances in Natural Language Processing, pages 521–528,
Hissar, Bulgaria, Sep 7–9 2015.

then presented in Section 4. Section 5 describes the datasets we used in the experiments and details of the sampling procedure done on these datasets are provided in Section 6. Section 7 contains the discussion on methodology, followed by the experimental details and results presented in Sections 8 and 9, respectively. Finally, the conclusion of the study is provided in Section 10.

2 Related Work

In this section, we will discuss three researches which focused on the combination of features for text readability indexing. Authors of these studies were able to conclude that combining several text feature sets could yield improved classification metrics.

The study in (Si and Callan, 2001) combined content-based and surface linguistic features into a single text readability level classifier. The *Expectation Maximization* (EM) algorithm was then utilized to automatically calculate the weight values for their proposed models, namely, the *unigram language model* (i.e. using words in text) and the *sentence length distribution model*. The authors hypothesized that 1.) readability measures should be sensitive to content as well as to surface linguistic features and 2.) statistical language models could capture the content information related to reading difficulty. Experiments showed that 1.) Sentence length is a useful feature for readability analysis on their dataset since its mean value increases as the readability level of texts increase, while 2.) Syllable count is not a useful feature as it did not exhibit the same behavior. Si and Callan also achieved higher accuracy value of 70.5% for the unigram language model than the sentence length distribution model which only yielded 42.6%. Moreover, by combining these two models together, they were able to achieve their highest accuracy of 75.4%.

In (Schwarm and Ostendorf, 2005), binary *Support Vector Machines* (SVM) were utilized to approximate the syntactic and semantic complexities of texts. Several text features including sentence length, syllable count, word instances- and uniqueness-based features, part-of-speech (POS) features (e.g. tags, parse tree height, average number of noun phrases, average number of verb phrases), and word uni-, bi-, and tri-gram features were used to train the classifiers. In the experiments, Schwarm and Ostendorf observed the contribution of individual features to the overall performance of the SVM classifiers and found out that 1.) no feature stood out as the most important one, and 2.) system performance was degraded when any particular feature was removed. They also realized that trigram models were noticeably more accurate than bigrams and unigrams. Results showed that their system could sometimes achieve *precision* value of 75%, *recall* of 87% and adjacent accuracy classification error (percentage of articles which are misclassified by more than one grade level) of 3.3%.

In (Heilman et al., 2007), the authors had concluded from their interactions with instructors of second language learners of English that combining grammatical and lexical features as predictors of text readability could outperform those measures based solely on one of the two. Heilman et al. combined vocabulary-based approach using *Multinomial Naive Bayes* classifier on unigrams, and grammar-based approach using *k-Nearest Neighbor* algorithm on parse trees, sentence length, verb forms, and part-of-speech tags features to evaluate text readability. Results of their study showed that vocabulary-based approach alone is better than grammar-based approach. However, the combined approach was proven to further enhance the performance of their system, reducing the mean squared error value by as much as 21% from 0.51 to 0.4.

3 LSI versus CI

LSI has been a well-known information retrieval algorithm. It was patented in 1988 by Scott Deerwester, Susan Dumais, George Furnas, Richard Harshman, Thomas Landauer, Karen Lochbaum and Lynn Streeter (Deerwester et al., 1989). CI, on the other hand, was proposed more recently by George Karypis and Eui-Hong (Sam) Han in 2000 (Karypis and Han, 2000) as a faster alternative for LSI. In this section, we present existing researches comparing the performances of LSI and CI on text content and readability analyses.

3.1 English Essay Content Analysis

The study presented in (Razon, 2010) focused on comparing LSI and CI as applied on English essay scoring. Both algorithms are based on vectorial semantics using dimensionality reduction.

Through several experiments, the study was able to prove that CI can outperform LSI in grad-

ing essays using content features alone. Below is the result of one of the experiments the authors conducted, where *accuracy* was calculated based on the exact agreement between the system's predicted scores and actual essay scores given by human checkers. As indicated on Table 1, CI outperformed LSI on all datasets reaching as high as 84.21% accuracy. It is also important to note that, as shown in the Grade8 dataset results, the difference between the accuracies of the two algorithms can reach as high as 18.75% in favor of CI.

Dataset	LSI Accuracy	CI Accuracy
Grade7	78.95	84.21
Grade8	62.50	81.25
Grade9 Set1	50.00	58.82
Grade9 Set2	64.10	69.23

Table 1: LSI vs. CI Accuracies (%) using Normalized Raw Term Frequency in (Razon, 2010)

3.2 Filipino Essay Content Analysis

The study in (Ong, 2011) was an attempt to implement a CI-based Filipino essay grader. Filipino language experts were consulted to validate the outputs. Experiments comparing CI and LSI showed that CI may perform better than LSI for some experts. The experimental results have proven that the system has a 95% probability of achieving accuracy from 75.5% to 79.9% in predicting the actual essay score given by human raters using the CI-based system. This range of accuracy values is comparable to those achieved among human raters which is between 70.6% and 70.9%.

As also stated in (Ong, 2011), CI, with small number of vectors representing each pre-defined class or group in the dataset, can run faster than LSI. The time complexity for CI is $O(iekn)$ while LSI is $O(en^2)$, where i is the number of iterations until convergence is achieved, k is the number of vectors representing each pre-defined class, e is the number of vocabulary entries, and n is the number of essays.

3.3 Tagalog Text Readability Indexing

A comparative study between LSI- and CI-based algorithms, as applied on readability analysis for **Tagalog** text documents, was conducted in (Razon et al., 2011). In the experiments, the authors applied *Spearman's rho* onto the training and test cosine similarity matrices, such that, each test document vector of cosine similarity scores with respect to the semantic space created using the training documents, is correlated against the training set's vectors of cosine similarity scores. Grade levels were then assigned to each test document based on the grade level of the training document with the highest correlation to it.

Grade Level	RTF		TF-IDF	
	LSI	CI	LSI	CI
2	61.67	80.00	76.67	66.67
3	40.00	52.00	62.00	52.00
4	16.67	36.67	23.33	33.33
6	65.00	47.50	32.50	20.00

Table 2: Exact Agreement Accuracy (%) using Raw Term Frequency (RTF) and Term Frequency-Inverse Document Frequency (TF-IDF) Weighting Schemes in (Razon et al., 2011)

As shown on Table 2, CI using raw term frequency (RTF) weighting scheme outperformed LSI on all the datasets except Grade 6. On the other hand, for the term frequency-inverse document frequency (TF-IDF) case, LSI performed better than CI except on Grade 4 dataset.

4 Our Assumptions

Our main assumption in this study is that ***written essays by students can be used to approximate their lowest possible reading level***. This assumes that whatever the students can write, they can also read. In (Metametrics, 2009), it was empirically proven that people's reading ability is consistently higher than their writing ability, hence providing a justification to this claim. Aside from this main assumption, we have also drawn out the following working assumptions from the researches discussed in Section 2 and other references cited in this paper:

1. *Statistical and n-gram analysis of POS tags can yield useful information to approximate text readability levels.* (Schwarm and Ostendorf, 2005; Heilman et al., 2007)

2. *Combined grammar-related and content-based analyses can yield better results for text readability analysis.* (Heilman et al., 2007; Landauer and Way, 2012)

5 Our Datasets

One of the challenges in this study is creating a suitable dataset to model and test readability lev-

els of reading materials. There are two categories of data in this project. The first one is composed of English essays written by high school students. Under this category, we have the *2010 Gr 7-9* and *2014 Gr 7-9* datasets. These are used to model student reading abilities per school grade level. Each of these datasets is divided into two, $\frac{2}{3}$ for **training** and $\frac{1}{3}$ for **test**. The second data category is the teacher-prepared instructional materials which we call the *Reading Mats* dataset. These materials are selected by the schools' instructional materials experts and are classified from grade 7 to grade 9. In the experiments, these are used to create the **reference** set for both the training and testing processes which will be discussed in the later sections of this paper.

Dataset	Grade7	Grade8	Grade9	Total
2010 Gr 7-9	47	54	112	213
2014 Gr 7-9	67	62	64	193
Reading Mats	12	6	10	28

Table 3: Summary of Datasets Used

6 Our Sampling Procedure

Sampling is another very important factor to consider in the implementation of the system. For both the 2010 Grade7-9 and 2014 Grade7-9 datasets, a stratified 3-fold cross-validation is implemented, such that, essays in each grade level (i.e. Grade7, Grade8, Grade9) are roughly divided into three equal static partitions. One partition is always set aside for testing and the other two for training. Note that since there 3 grade levels with 3 partitions each, 27 Test-Training combinations are created to exhaust all possible partition combinations with 1:2 test-to-training partition ratio for each grade level.

7 Our Methodology

7.1 Content-based Analysis

7.1.1 Matrix Representation

After creating the vocabulary list from text samples, the three sets (i.e. training, test and reference) are converted to their term-by-document matrix representation. In this representation, each column is equivalent to one text sample vector, each row represents one word or term in the vocabulary, and each entry in the matrix is the number of occurrences of each term in each text sample.

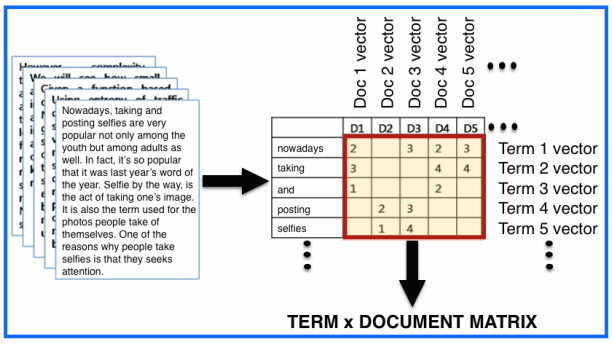

Figure 1: Term-by-Document Matrix

7.1.2 Dimensionality Reduction

As in the study of Razon in (Razon, 2010), both LSI and CI dimensionality reduction strategies are implemented separately on the training sets. These are *Singular Value Decomposition* (SVD) for LSI and *Concept Decomposition* (CD) for CI. SVD is defined as the decomposition of matrix X using $X = UDV^T$ where $U = XX^T$, $V = X^TX$ and D is a matrix whose diagonals are the singular values of matrix X. On the other hand, CD is defined as the decomposition of matrix X using $X_p = C_pZ^*$, where C_p is a matrix created using the normalized mean column vectors of each sub-cluster in the training set, and Z^* is the least-squares approximation with closed-form solution of $Z^* = (C_k^TC_k)^{-1}C_k^TX$ (Karypis and Han, 2000). A *sub-cluster* (sub) is derived from the stratified clustering of the vector representations of text documents by grade level. *K-means* clustering algorithm is utilized to accomplish this task.

7.1.3 Folding-In

Folding-in refers to the projection of the original training, test and reference document vectors onto the reduced semantic space derived in the previous step. For LSI, this process involves solving the equation $q_i = q_i^TU_kD_k^{-1}$ for all document vectors q_i of the training, reference and test sets. For CI, we solve the equation $q^* = (C_p^TC_p)^{-1}C_p^Tq$, where q^* is the reduced dimensionality matrix representation of the original training, reference or test matrix.

7.1.4 Similarity Measurement

After folding-in all column vectors of the training, test and reference sets onto the LSI- and CI-based reduced semantic spaces, cosine similarity values between the column vectors of both the training and test sets, against the column vectors of the ref-

erence set are calculated. Consequently, this step yields two sets of similarity vectors as shown in Figure 2. One set corresponds to the similarity values between all reference set vectors against a training document vector and the other corresponds to the similarity values between all reference set vectors against a test document vector. We will refer to these vectors as *training document-to-reference similarity vector* and *test document-to-reference similarity vector*, respectively. These vectors serve as training and test inputs of our SVM classifier.

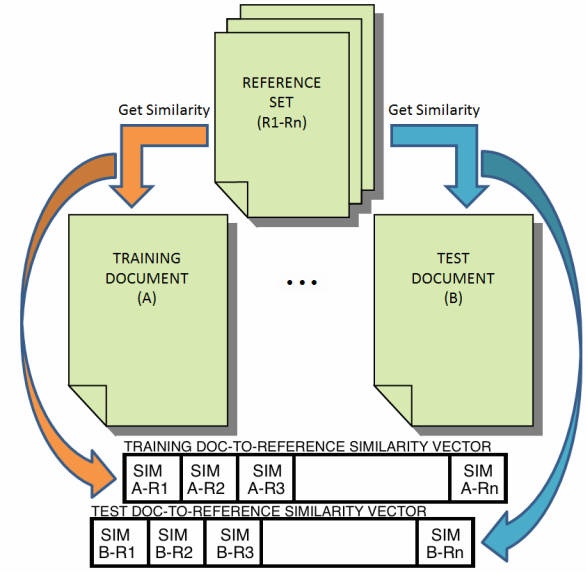

Figure 2: Similarity Vector Diagram

7.2 POS-based Grammar Analysis

Grammar features are necessary to model texts for each grade level. As part of our working assumptions discussed in Section 4, POS n-grams can be used to provide a rough approximation of the texts' syntactic information at the least. For example, POS unigrams can provide information regarding which of the POS tags are prevalent for each grade level and which are not. On the other hand, POS bi- and tri-grams can capture grammar-related information which can serve as basis for syntax complexity.

In this study, `Apache OpenNLP Maxent POS Tagger` is used to tag all documents. After getting uni-, bi- and tri-gram features from the text documents, we constructed the term-by-document matrices for the training, test and reference sets, where the POS n-grams are treated as the terms of the said matrices (i.e. POS-n-gram-by-document matrices). Next, we constructed the corresponding *training document-to-reference* and *test document-to-reference similarity vectors* for these matrices the same way as discussed in Section 7.1.4. Finally, *sparsification* of these matrices have been considered to further enhance the performance of the systems.

Sparsification is the removal of sparse term vectors (i.e. n-gram vectors) or the exclusion of those term vectors which have mostly zero values. This procedure aims to reduce the dimensionality of the POS n-gram-by-document matrix without sacrificing the loss of significant information inherent in the matrix. In this study, the term *sparsity* refers to the maximum sparse percentage, called the *sparse index* (SI), to consider in the experiment. For example, SI value of 0.7 means that all term vectors which are 70% sparse and below will be considered. Therefore, higher sparsity values allow more POS n-gram vectors to be included in the analysis.

8 Our Experiments

Five (5) feature sets are investigated in this study. These are: 1.) **POS**: POS n-gram features only, 2.) **LSI**: LSI-based features only, 3.) **CI**: CI-based features only, 4.) **LSI+POS**: Combined LSI-based and POS n-gram features, and 5.) **CI+POS**: Combined CI-based and POS n-gram features.

The following experimental phases are implemented using the training and test document-to-reference similarity vectors discussed in Sections 7.1.4 and 7.2

1. **Phase 1**: Baseline Experiments using Feature Sets 1, 2 and 3
2. **Phase 2**: Combined Grammar and Content Features Experiments using Feature Sets 4 and 5 with NO Sparsification
3. **Phase 3**: POS n-gram Sparsification Experiments using Feature Sets 1, 4 and 5 with SI from 0.1 to 0.9

Radial basis function (RBF) is used as the kernel function for all SVM classifiers in all the experiments. For each phase discussed above, the following SVM parameters are held constant: 1.) γ: kernel parameter which controls the shape of the RBF, 2.) C: misclassification cost or penalty constant, and 3.) k: number of folds in training cross-validation (i.e. k-fold cross-validation constant). Having constant values for these parameters allows us to focus our investigation on the

POS n-gram sparsification and the primary parameter of each baseline feature set, namely, the *n* in POS n-grams (i.e. uni-, bi-, tri-), LSI's dimensionality constant, *dim*, with values from 0.5 to 0.9, and CI's number of sub-cluster representations per grade level, *sub*, with values from 1 to 5.

Optimal values for dim and sub are derived in Phase 1. Then, LSI- and CI-based features corresponding to these values are combined with the full POS uni-, bi- and tri-gram feature sets in Phase 2, where we aim to find out 1.) if the combination of content- and grammar-based feature sets could yield higher mean exact agreement accuracy (MEAA), and 2.) which of the combined feature sets would perform best among the others. Finally, an investigation on the effect of POS n-grams' sparsity index is performed in Phase 3 to optimize the LSI+POS and CI+POS combination processes.

9 Our Experimental Results

9.1 Phase 1: Baseline Experiments

Baseline experiments are those experiments done using isolated feature sets (i.e. POS only, LSI only and CI only). For the 2010 Grade 7-9 dataset, the highest MEAA of 89.72% is achieved by CI using 2-sub-cluster vector representation per grade level. This is followed by POS bigrams with a value of 85.39%, making LSI the last with a value of 68.28% at reduced dimensionality of 50%. Furthermore, baseline CI-based features also outperformed LSI- and POS-based features yielding as high as 93.40% MEAA for the 2014 Grade 7-9 dataset. This is also followed by POS bigrams with a value of 87.38%, making LSI the last again with a value of 79.80% at reduced dimensionality of 70%.

9.2 Phase 2: Combined Features Experiments

In this phase, we directly combined the LSI- and CI-based features with POS n-gram features (i.e. LSI+POS and CI+POS) **without sparsification**. In general, LSI's performance has improved while the reverse is true for CI.

Referring to Tables 4 and 5, we can see that both, LSI+POS uni-grams and CI+POS uni-grams, have achieved higher MEAA values than POS uni-grams alone. It is also important to note that the MEAA for the combination of POS bi- and tri-grams with CI and LSI have resulted to values

Feature Set	Primary Param.	2010 Gr7-9		2014 Gr7-9	
		MEAA	SD	MEAA	SD
POS n-gram	n=1, uni	0.749	0.064	0.786	0.096
	n=2, bi	**0.854**	0.027	**0.874**	0.041
	n=3, tri	0.853	0.035	0.845	0.044
CI	sub=1	0.891	0.052	0.933	0.039
	sub=2	**0.897**	0.051	**0.934**	0.041
	sub=3	0.884	0.071	0.931	0.042
	sub=4	0.873	0.045	0.927	0.042
	sub=5	0.882	0.053	0.929	0.042
LSI	dim=0.5	**0.683**	0.054	0.781	0.056
	dim=0.6	0.660	0.055	0.783	0.050
	dim=0.7	0.666	0.040	**0.798**	0.048
	dim=0.8	0.659	0.044	0.789	0.053
	dim=0.9	0.655	0.039	0.785	0.060

Table 4: Phase 1: Baseline Experiment Summary

equal to or very close to that of isolated POS bi- and tri-grams, respectively. Therefore, it can be inferred that POS bi- and tri-gram features dominate the content-based features from CI and LSI, clipping the MEAA to the values achieved in the POS n-grams baseline experiments and this consequently affected CI's MEAA values negatively. Hence, we can say that simply adding the features together, which has been a common practice in other existing researches, does not guarantee a much better feature set. This anomalous behaviour led us to go forward and conduct Phase 3 to investigate the effects of POS n-gram sparsification.

Base Feature	Modified Feature Set	2010 Gr7-9		2014 Gr7-9	
		MEAA	SD	MEAA	SD
CI (sub=2)	CI+POS uni	0.838	0.064	0.850	0.073
	CI+POS bi	0.854	0.027	0.874	0.041
	CI+POS tri	0.853	0.035	0.845	0.044
LSI (dim-0.7)	LSI+POS uni	0.768	0.060	0.816	0.058
	LSI+POS bi	0.854	0.027	0.874	0.041
	LSI+POS tri	0.855	0.033	0.852	0.041

Table 5: Phase 2: Combined Features Experiment Summary

9.3 Phase 3: POS n-gram Sparsification

As evident on Figures 3 and 4, the MEAA tends to increase as we increase the SI value, reaching its peak in the range of 0.6 to 0.9 for all n-grams (i.e. uni-grams, bi-grams, tri-grams). Results also show that CI with POS bi-grams has yielded the highest MEAA of 90.9% and 95.1%, with low standard deviations (SD) of 0.045 and 0.021, on both the 2010 and 2014 Grade 7-9 datasets, respectively.

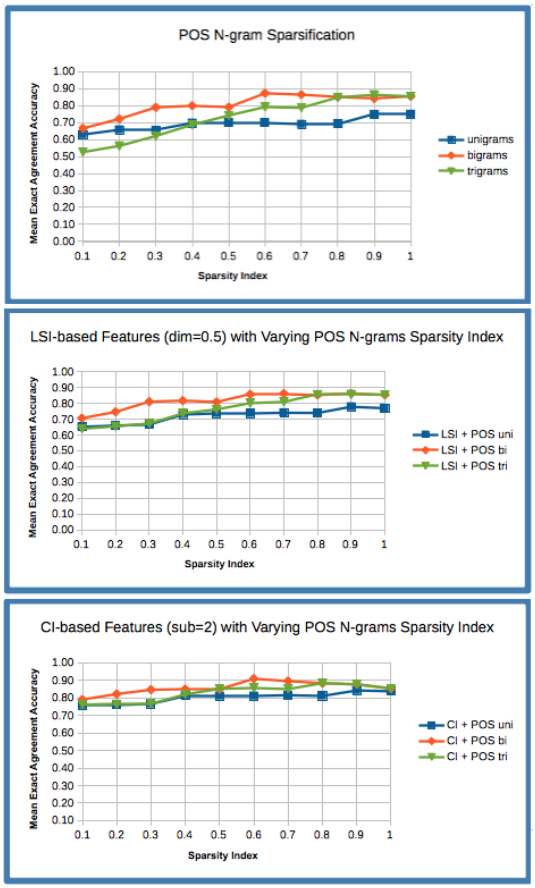

Figure 3: POS n-grams Sparsification Experimental Results on the 2010 Grade 7-9 Dataset using Feature Sets 1-POS only, 4-LSI+POS and 5-CI+POS

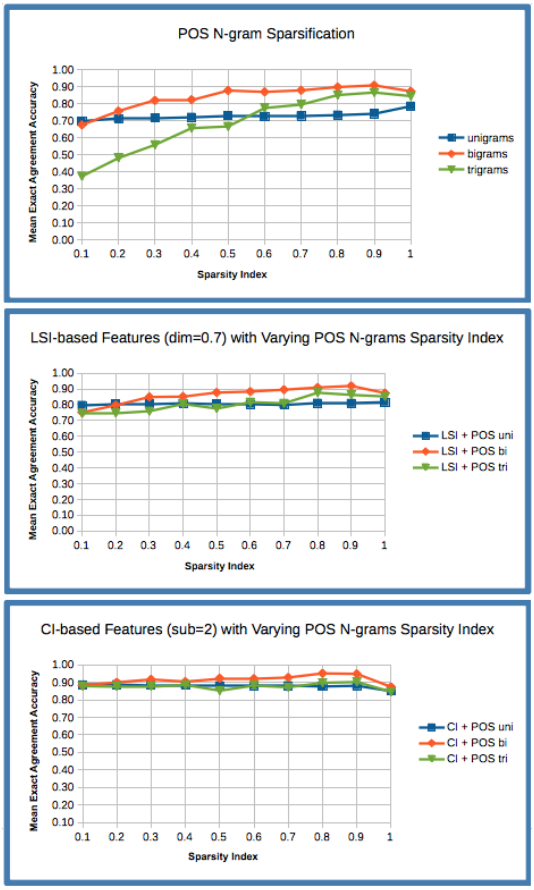

Figure 4: POS n-grams Sparsification Experimental Results on the 2014 Grade 7-9 Dataset using Feature Sets 1-POS only, 4-LSI+POS and 5-CI+POS

9.4 General Observations

The following statements summarize the overall results of the experiments:

1. For baseline experiments, CI-based similarity features alone can yield good results, outperforming the LSI- and POS-based similarity features.

2. LSI's performance can be greatly improved by combining it with the full set POS-based features (i.e. SI=1.0, no sparsification). However, the opposite is true for CI's.

3. The combined CI and POS Bi-grams feature sets (i.e. CI+POS bi-grams) consistently yield the highest MEAA in Phase 3, ranging from 80% to 95% for SI values between 0.2 to 0.8 as shown by the red lines on Figures 3 and 4.

4. POS N-gram features sparsification improves the MEAA of isolated POS-, combined LSI+POS-, and combined CI+POS-based feature sets (i.e. feature sets 1, 4 and 5 as discussed in Section 8). Optimal MEAA values can be achieved within 0.6 to 0.8 SI values, then slope downwards all the way to 1.0. Note that at SI=1.0, no term is removed from the feature sets' vocabulary (i.e. full combined vocabulary is being used without sparsification).

5. POS bi-gram feature set is superior among the other n-grams' (i.e. uni- and tri-grams). This is exhibited on Figures 3 and 4, where bi-grams almost always yield the highest MEAA all throughout the SI spectrum.

6. SI has greater influence on bi- and tri-grams than uni-grams in terms of MEAA. Unigrams tend to exhibit gradual changes in the MEAA graphs.

10 Conclusion

In this study, we have successfully implemented a learner-based text readability indexer using combined content and grammar features for the English language. Superiority of the combined CI and POS bi-grams feature set has been established in the experiments, yielding as high as 95.1% MEAA. Moreover, the results of the Phase2 and Phase3 experiments also prove that POS n-gram sparsification is important to optimize the feature combination process. This goes to show that careful analysis is necessary in combining feature sets and that merely adding the features together does not guarantee a better feature set.

For future work, it would be interesting to find out what happens if we use the combined POS n-gram features such that we have: 1.) uni-grams and bi-grams, 2.) bi-grams and tri-grams, 3.) uni-grams and tri-grams, and 4.) uni-grams, bi-grams, and tri-grams, together with CI- or LSI-based features. Then, we can also attempt to optimize the combination process through sparsification as we did in this study. Adding more grade levels and text documents into the system can also be done to further validate the results. Furthermore, the flexibility of the system can be tested by applying it on languages other than English.

Acknowledgments

We would like to acknowledge the support provided by the Department of Science and Technology (DOST) of the Philippines through its Engineering Research and Development for Technology (ERDT) program. We would also like to thank the University of the Philippines Integrated School (UPIS) for giving us permission to use their essay samples and reading materials as data.

References

Scott C. Deerwester et al. (Bell Communications Research, Inc.). *Computer Information Retrieval using Latent Semantic Structure*. US Patent 4,839,853. June 13, 1989.

George Karypis and Euihong Han. 2000. *Concept Indexing: A Fast Dimensionality Reduction Algorithm with Applications to Document Retrieval and Categorization*. In *Proc. of the 9th International Conference on Information and Knowledge Management* McLean, Virginia.

Luo Si and Jamie Callan. 2001. *A Statistical Model for Scientific Readability*. In *Proc. of the 2001 ACM CIKM. 10th International Conference on Information and Knowledge Management*, Atlanta, GA, USA.

Kevyn Collins-Thompson and Jamie Callan. 2004. *A Language Modeling Approach to Predicting Reading Difficulty*. In *Proc. of HLT/NAACL 2004*, Boston, USA.

Constantinos Boulis and Mari Ostendorf. 2005. *Text Classification by Augmenting the Bag-of-Words Representation with Redundancy-Compensated Bigrams*. In *Proc. of the SIAM International Conference on Data Mining at the Workshop on Feature Selection in Data Mining*.

Sarah E. Schwarm and Mari Ostendorf. 2005. *Reading Level Assessment using Support Vector Machines and Statistical Language Models*. In *Proc. of the 43rd Annual Meeting on Association for Computational Linguistics*, Michigan, USA.

Patric Larsson. 2006 *Classification into Readability Levels - MS Thesis*. Uppsala University, Uppsala, Sweden.

Michael J. Heilman et al. 2007. *Combining Lexical and Grammatical Geatures to Improve Readability Measures for First and Second Language Texts*. In *Proc. of NAACL HLT 2007*, Rochester, New York.

Sarah E. Petersen and Mari Ostendorf. 2009. *A Machine Learning Approach to Reading Level Assessment*. In *Journal of Computer Speech and Language, Volume 23 Issue 1*, London, United Kingdom.

Malbert Smith III. 2009. *The Reading-Writing Connection*. *MetaMetrics Position Paper*, Durham, North Carolina.

Abigail R. Razon. 2010. *A New Approach to Automated Essay Content Analysis using Concept Indexing - MS Thesis*. University of the Philippines-Diliman, Manila, Philippines.

Abigail R. Razon et al. 2011. *Readability Analysis of Grade School Reading Books using Concept Learning with K-Means Clustering*. In *Proc. of 2011 International Symposium on Multimedia and Communication Technology*, Hokkaido, Japan.

Darrel Alvin N. Ong. 2011. *Automated Content Scoring of Filipino Essays using Concept Indexing - MS Thesis*. University of the Philippines-Diliman, Manila, Philippines.

Thomas Landauer and Denny Way. 2012 *Improving Text Complexity Measurement through the Reading Maturity Metric*. In *Annual meeting of the National Council on Measurement in Education*, Vancouver, British Columbia, Canada.

Jessica Nelson et al. 2012. *Measures of Text Difficulty: Testing their predictive value for grade levels and student performance*. *Council of Chief State School Officers*, Washington, DC.