ISBN 978-981-11-0008-6
Proceedings of 2016 6th International Workshop on Computer Science and Engineering (WCSE 2016)
Tokyo, Japan, 17-19 June, 2016, pp. 580-584
doi: 10.18178/wcse.2016.06.099

Automated Essay Scoring by Combining Syntactically Enhanced Latent Semantic Analysis and Coreference Resolution

Gilbert Wonowidjojo[1], Michael S. Hartono[1], Frendy[1], Derwin Suhartono[1 +], Almodad B. Asmani[2]
[1]Bina Nusantara University, Computer Science Department, Jakarta, Indonesia
[2]Bina Nusantara University, English Department, Jakarta, Indonesia

Abstract. The objective of this research is to measure how much syntactic information (in the form of word order) and Coreference Resolution affect the result of Automated Essay Scoring (AES) using Latent Semantic Analysis (LSA). To incorporate the syntactic information, Syntactically Enhanced LSA (SELSA) is used, whilst Stanford CoreNLP Natural Language Processing Toolkit is used for the Coreference Resolution. To evaluate the results, we calculate the average absolute difference between the system score and human score for each essay. Based on the results, we can conclude that syntactic information, when combined with Coreference Resolution, do not have higher correlation to human score than LSA (an average absolute difference of 0.15748 as opposed to LSA's 0.12597). But interestingly, the two techniques work better when they are used together, rather than when they are used separately. We also develop a new algorithm to calculate the scores, with a better average absolute difference, which is as high as 0.08969.

Keywords: automated essay scoring, syntactically enhanced latent semantic analysis, word order, coreference resolution.

1. Introduction

There are several ways to evaluate students' understanding of a particular topic, most notably is to assign multiple choice questions or essays. In multiple choice questions, students are given several questions, each with a set of possible answers (usually 3 to 5 choices per question), in which one is correct. Students are given points (or scores) for a correct answer, but no points or sometimes point deductions are given for an incorrect answer. This type of assessment is easy to score both manually and automatically. Nowadays, optical answer sheets are used to assess students in multiple choice questions, to save time and work needed to score them. Even without optical answer sheets, multiple choice questions scoring is relatively fast, but they have a drawback. The drawback is, in fact, students can answer a multiple choice question correctly just by doing an educated guess or even a random guess. This may lead the instrument (assessment) to incorrectly evaluate their understanding of the topic. To overcome this issue, essays are used instead of multiple choice questions. An essay can show exactly how deep the students understand the topic. But like multiple choice questions, essays do have a drawback. The major drawback is the time and work needed to score the essays. An expert said that to score a single essay, 5 to 10 minutes are needed. It can reflect that to score a whole class' essays, several hours or even days are needed. The scorer's consistency is also in question, whether he or she can maintain the consistency in checking all the essays with the same standard. These conditions make automated essay scoring systems important to have.

Latent Semantic Analysis (LSA) is a most used technique to score essays automatically. It is a statistical-algebraic technique to represent usage of words in documents (Landauer et al., 1998). The representation is in form of a word-document co-occurrence matrix, which is then scaled, normalized, and approximated using Singular Value Decomposition (SVD) in a finite number of dimensions. The result can be used to project any

+ Corresponding author. Tel.: +62215345830; fax: +62215300244
E-mail address: dsuhartono@binus.edu

text document into a latent semantic space, and then compare two documents to find their cosine similarity measure, using their projection vectors. LSA is a 'bag-of-words' approach and so lacks the word-order or syntactic information in a text document. But for correct automatic evaluation of students' answers, a model should consider both syntax and semantics (Kanejiya et al., 2003). Syntactic information can affect the automated scoring results, since in LSA, a student can write all the keywords, ignoring word-orders, and get high scores. The other major issue of LSA is nowadays, a step known as text preprocessing is used, before LSA is applied to the texts. In text preprocessing, stopwords removal is included. Stopwords are words that are considered to have no meaning at all, but some stopwords may refer to entities that may be important in the texts, which is called pronouns. Because of that, it will be beneficial to resolve all of the pronouns first to their referred entities before applying LSA to optimize the results. Moreover, Miller (2003) said that adding Anaphora Resolution (a big part of Coreference Resolution) to LSA would be an interesting study.

2. Related Work

Many research about automated essay scoring have been undergone, and some of them even result in commercial applications such as Intelligent Essay Assessor or IEA (Foltz et al., 1999) which uses LSA as its basis on content scoring, and e-rater® (Attali & Burstein, 2006) which measures grammar, usage, mechanics, style, organization, development, lexical complexity, and prompt-specific vocabulary usage.

To incorporate the syntactic information there are numerous research: *Tagged LSA* (Wiemer-Hastings & Zipitria, 2001) which pairs every word with its part-of-speech (POS) tag, *Syntactically Enhanced LSA* (Kanejiya et al., 2003) which pairs every word with the POS tag of the preceding word which indicates some kind of syntactic neighbourhood around the focus word, and *Generalized LSA* (Islam & Hoque, 2012), which uses n-grams (in form of unigrams, bigrams, and trigrams) to pair successive words in a document to be a single term in the term-document matrix.

Research about Coreference Resolution has created many algorithms, such as *Hobbs Algorithm* (Hobbs, 1978) which uses syntactic parse-tree traversal, *Lappin and Leass Algorithm* (Lappin & Leass, 1994) which calculates salience values of the potential referents for every pronoun, and *Centering Theory* (Grosz et al., 1995) which determines the antecedent from document segments. Because of its nature in determining the antecendent of every pronoun, *Centering Theory* can also be used to measure cohesion and coherence of essays (Putri, Fadilah, Ivan, Suhartono & Wiannastiti, 2016). Most recently, the state-of-the-art of Coreference Resolution is a usable toolkit, named *The Stanford CoreNLP Natural Language Processing Toolkit* (Manning et al., 2014), which uses a multi-pass sieve for the Coreference Resolution (Lee et al., 2013).

In this research, *Syntactically Enhanced LSA* (SELSA) and *The Stanford CoreNLP Natural Language Processing Toolkit* is used to incorporate the syntactic information and Coreference Resolution, respectively.

3. Methodology

To implement the research idea, we have built a system that consists of two phases, which are the training phase and the evaluation phase. The flow of the training phase can be seen on figure 1.

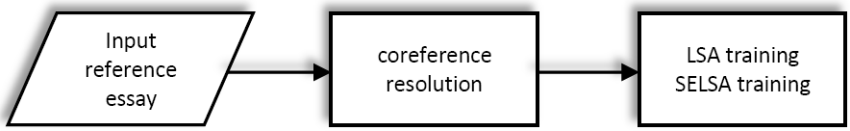

Fig. 1: System's training phase.

In the training phase, the first step is to input the reference essays. These are pre-graded essays (by human scorer) that are used to create the LSA-based model used in evaluation. After these essays are inputted, all pronouns in the essays are resolved using Coreference Resolution. After that, the essays will undergo the LSA-based training (LSA and SELSA) which is shown on Fig. 2.

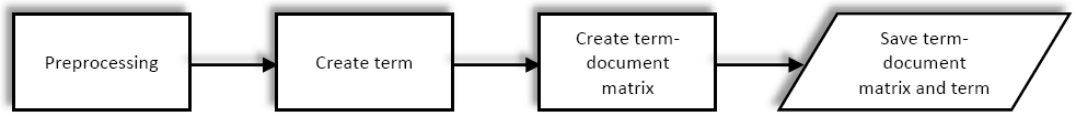

Fig. 2: LSA and SELSA in training phase.

The first step in LSA-based training is to pre-process the essays. This step includes lowercasing, token-checking (to break down the essays into sentences, and to eliminate non-alphabetic words), lemmatization (reverting each word into its root word), and stopwords removal (remove stopwords that do not refer to anything e.g. "is"). We use NLTK library (Bird et al., 2009) for the two latter processes. To implement the syntactic information for SELSA, each word is *prevtagged* (Kanejiya et al., 2003). After the essays are successfully preprocessed, terms for the term-document matrix are created by removing duplicates present in the essays. Mistyped words are also corrected and weighted according to the *term similarity* algorithm (Sendra et al., 2015). After that, the whole term-document matrix is created, where the rows denote terms, and the columns denote essays, so that each cell will contain the number of occurrences of the term denoted by the row, in the essay denoted by the column. After this matrix is created, it along with the term list is stored for use in the evaluation phase. The flow of the evaluation phase can be seen on Fig. 3.

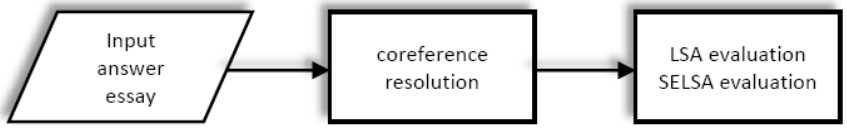

Fig. 3: System's evaluation phase.

The flow of the evaluation phase is almost the same as the training phase, they only have two differences. The first difference is that the essay inputted is the one to be evaluated (answer essay). It is pre-graded only for research purposes. The other difference is in the LSA-based evaluation which quite differs from the LSA-based training explained earlier. It can be seen on Fig. 4.

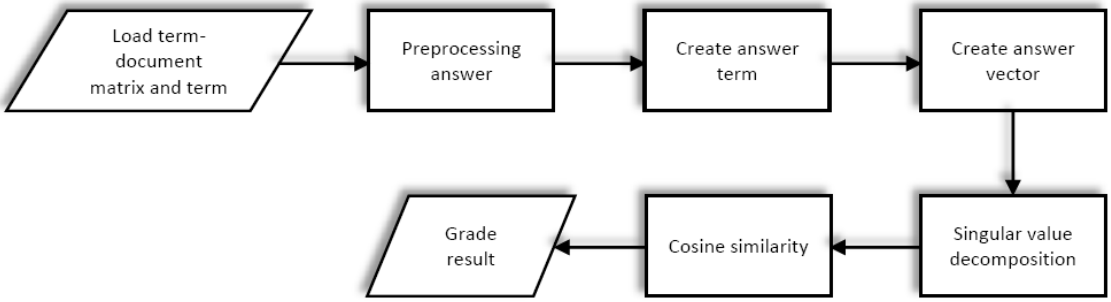

Fig. 4: LSA and SELSA in evaluation phase.

The first step is, of course, to load the term-document matrix and the term list that is previously saved. After that, the answer essay is preprocessed (similar to the one at the training phase). Thereafter, the answer term is created, this is done by removing duplicate words in the essay and removing words which are not in the term list. After that, a term-document matrix with one column (which corresponds to the answer essay) is created. This is called as the answer vector. The original matrix then undergoes Singular Value Decomposition, to enable the calculation of cosine similarity between the answer vector and the LSA-based model in the LSA space. The cosine similarity is then used to determine the score of the answer essay.

4. Experiment

The essays used in the experiment are obtained by collecting students' assignment. All 142 essays are written in English based on a predetermined topic "What are the qualities of a best friend?", and have been pre-graded with a score between 0 and 4. They are separated into two sets. The first set consists of 15 essays used in the training phase, and the remaining 127 essays in the testing set are used for the evaluation phase. We used 15 essays for the training phase because based on the experiment in a smaller domain, 15 essays give the best LSA-based model and a relatively fast training speed. Increasing the number any further will only make the training phase slower without any significant improvement of the performance.

The experiment consists of two scenarios. The first scenario is a document classification based on the cosine similarity measure between the answer essay and the essays in the training set. Every human score of the essays in the training set is rounded down to the nearest 0.25, this is to create clusters (which is needed for supervised document classification). There are 3 variables to be compared, which are LSA vs. SELSA, Using vs. Not Using Coreference Resolution, and Maximum vs. Average Similarity. In Maximum Similarity, the score of the essay is obtained from the pre-graded score of the essay with the maximum cosine similarity, while in Average Similarity the score is obtained from the pre-graded score of the cluster with the maximum average cosine similarity. The result of the experiment is shown in Fig. 5.

Technique Used	Without Coreference Resolution	With Coreference Resolution
LSA Maximum Similarity	0.16732	0.16929
LSA Average Similarity	0.12597	0.14173
SELSA Maximum Similarity	0.22047	0.18898
SELSA Average Similarity	0.17913	0.15748

Fig. 5: The result of the first scenario experiment.

The numbers shown above are the average absolute difference between human score and system score of the 127 essays. This means that the smaller the value, the higher the correlation the system has compared to the human score. The table above shows that LSA without Coreference Resolution with average similarity measure has the highest correlation to human score (with a difference of only 0.12597).

We can see that LSA's performance dropped when combined with Coreference Resolution, but the interesting point is that SELSA yields the opposite result when compared to LSA. From the table above, SELSA's performance increases significantly when combined with Coreference Resolution. This is a good sign that SELSA and Coreference Resolution work well with each other.

From the document classification method used, Average Similarity gives better results than Maximum Similarity. This is true because of the nature of text classification, where a text is best classified to a cluster when it has the highest average similarity to the whole cluster, as opposed to the highest similarity to a single text in a cluster.

The second scenario is not using document classification, but instead calculating the average score based on the cosine similarities across all essays in the training set. The score of the essays in the training set are not rounded down. The score of each essay will be computed with the following formula:

$$\sum_{i=1}^{n} \frac{C_i X_i}{\sum_{j=1}^{n} C_j}$$

where **n** is the number of essays in the training set,

$\mathbf{C_i}$ is the cosine similarity between the answer essay and the i^{th} essay in the training set, and

$\mathbf{X_i}$ is the pre-graded human score of the i^{th} essay in the training set.

The result of the second scenario experiment is shown on Fig. 6.

Technique Used	Average Absolute Difference to Human Score
LSA Without Coreference Resolution	0.08969
LSA With Coreference Resolution	0.09472
SELSA Without Coreference Resolution	0.09266
SELSA With Coreference Resolution	0.09340

Fig. 6: The result of the second scenario experiment.

From the result shown above, the average absolute difference to human score is significantly smaller than the results from the first scenario. Even the highest difference in the second scenario (0.09472) is smaller than the smallest difference in the first scenario (0.12597), so this approach is better in terms of correlation. In the second scenario result, we can also see that Coreference Resolution dropped LSA's performance, but in SELSA it is quite stable.

5. Conclusion

From the experiments explained earlier, we can conclude that:

- SELSA has lower correlation to human scores than LSA in both experiments, whether using document classification or calculating the average score based on the cosine similarities across all essays in the training set. This is similar to the conclusion made by Kanejiya et al. (2003). But SELSA can evaluate some essays far better than LSA.
- LSA's performance dropped when combined with Coreference Resolution but oppositely, SELSA's performance increased significantly then combined with Coreference Resolution. This signify that SELSA and Coreference Resolution works well together.
- The result scores of the proposed algorithm for calculating essay scores automatically (second scenario) have a very high correlation to human score. The average absolute difference between the two scores is as small as 0.08969.

Some suggestions that are workable for future research:

- Combining SELSA and Coreference Resolution for other applications, such as document classification based on themes, or automatic summarization.
- Implementing the research in various languages other than English.
- Developing new algorithms for automated essay scoring, to achieve higher degree of correlation to human score.

6. References

[1] T. K. Landauer, P. W. Foltz, and D. Laham. Introduction to Latent Semantic Analysis. *Discourse Processes*. 1998, 25: 259-284.

[2] D. Kanejiya, A. Kumar, and S. Prasad. Automatic Evaluation of Students' Answers Using Syntactically Enhanced LSA. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*. 2003, 2: 53-60.

[3] T. Miller. Essay Assessment with Latent Semantic Analysis. *Journal of Educational Computing Research*. 2003, 29(4): 495-512.

[4] P. W. Foltz, D. Laham, and T. K. Landauer. Automated Essay Scoring: Applications to Educational Technology. In *Proc. Of the ED-MEDIA'99 conference*, Charlottesville. AACE. 1999.

[5] Y. Attali and J. Burstein. Automated Essay Scoring with E-Rater® v.2. *Technology, Learning, and Assessment*. 2006, 4(3).

[6] P. Wiemer-Hastings and I. Zipitria. Rules for Syntax, Vectors for Semantics. In *Proc. 23rd Annual Conf. of the Cognitive Science Society*. Mahwah, New Jersey: Erlbaum. 2001.

[7] M. M. Islam and A. S. M. L. Hoque. Automated Essay Scoring Using Generalized Latent Semantic Analysis. *Journal of Computers*. Academy Publisher, 2012, 7(3): 616-626.

[8] J. R. Hobbs. Resolving Pronoun References. *Lingua*, 44: 311-338. 1978. Also in *Readings in Natural Language Processing*. B. Grosz, K. Sparck-Jones, and B. Webber, editors, Morgan Kaufmann Publishers, Los Altos, California. pp. 339-352.

[9] S. Lappin and H. J. Leass. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*. 1994, 20(4): 535-561.

[10] B. J. Grosz, A. Joshi, and S. Weinstein. Centering : A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*. 1995, 21(2): 203-225.

[11] E. H. Putri, D. R. Fadilah, Ivan, D. Suhartono, and M. Wiannastiti. Thematic Development for Measuring Cohesion and Coherence Between Sentences in English Paragraph. *4th International Conference on Information and Communication Technology (ICOICT), Bandung, Indonesia*. 2016.

[12] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* 2014.

[13] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky. Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. *Computational Linguistics.* 2013, 39(4).

[14] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[15] M. Sendra, R. Sutrisno, J. Harianata. Pengembangan Algoritma Latent Semantic Analysis pada Penilai Esai Otomatis dalam Bahasa Inggris. Bachelor Thesis of Bina Nusantara University, Jakarta, Indonesia, 2015.