Distributed Vector Representations for Unsupervised Automatic Short Answer Grading

Oliver Adams[†‡] **Shourya Roy**[‡] **Raghuram Krishnapuram**[§]
[†]The University of Melbourne, Australia
[‡]Xerox Research Center India, Bangalore, India
[§]M. S. Ramaiah Institute of Technology, Bangalore, India
oliver.adams@gmail.com, shourya.roy@xerox.com, raghuk@msrit.edu

Abstract

We address the problem of automatic short answer grading, evaluating a collection of approaches inspired by recent advances in distributional text representations. In addition, we propose an unsupervised approach for determining text similarity using one-to-many alignment of word vectors. We evaluate the proposed technique across two datasets from different domains, namely, computer science and English reading comprehension, that additionally vary between high-school level and undergraduate students. Experiments demonstrate that the proposed technique often outperforms other compositional distributional semantics approaches as well as vector space methods such as latent semantic analysis. When combined with a *scoring scheme*, the proposed technique provides a powerful tool for tackling the complex problem of short answer grading. We also discuss a number of other key points worthy of consideration in preparing viable, easy-to-deploy automatic short-answer grading systems for the real-world.

1 Introduction

Grading is an important task in schools and colleges in order to assess students' understanding and guide teachers in providing instructive feedback. However, answer grading is tedious work and the prevalence of Computer Assisted Assessment has been limited to *recognition* questions with constrained answers such as multiple choice questions. In this paper, we delve into the topic of automatic assessment of students' constructed responses. In particular, we consider *short answers* which are a few words or a few sentences long, including everything in between fill-in-the-gap and essay-type answers (Burrows et al., 2015). Automatic short answer grading (ASAG) involves scoring a student answer given an instructor-provided model (reference) answer. Scoring schemes may optionally be provided to indicate the relative importance of different parts of the model answer. This is a complex natural language understanding task owing to linguistic variations (the same answer could be articulated in different ways), the subjective nature of assessment (multiple possible correct answers or no correct answer) and lack of consistency in human rating. For example, in Table 1, both student answers are correct, but this may not be apparent to a computer system.

In this paper, we employ distributed vector representation of words (Bengio et al., 2003; Mikolov et al., 2013b) for *unsupervised* ASAG, a task where graded student answers are not provided as training data (although a reference answer is still available). This has not yet been systematically explored even though such word embeddings have proven useful in natural language processing. (However, there has been work using embeddings for *supervised* ASAG Sakaguchi et al. (2015), and neural networks for essay grading Alikaniotis et al. (2016).) We conduct an empirical study to compare the proposed method against various other vector aggregation including naive vector addition, Word Mover's Distance (WMD) (Kusner et al., 2015) and paragraph vectors (Le and Mikolov, 2014) using two datasets that come from two different domains. The first is the undergraduate computer science dataset used in Mohler et al. (2011), while the second is a high-school English reading comprehension task which we present and intend to share with the community for future research. An important feature of the latter dataset is the

Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications,
pages 20–29, Osaka, Japan, December 12 2016.

Question	What is the unexpected fact stated by the writer? (2)
Model answer	The unexpected fact stated by the writer is that although the modern air-conditioned office is an unlikely place for work related injury, more and more people working in such places complain of disorders involving hands, wrists and other body parts.
Scoring scheme	Working in modern air-conditioned offices leads to more people getting work related injury (1.5) People complain of disorders involving hands, wrists and other body parts. (0.5)
Student 1 answer	More no. of white collar workers sitting in high tech offices are complaining of disorders involving hands, wrists, arms, shoulders, neck and back
Student 2 answer	The author states that many white collar workers in hi-tech offices complain of disorders in their hands, wrists etc. But this is very unexpected because an air conditioned office seem an unlike place for an injury to occur.

Table 1: Example of question, model answer, scoring scheme and two student answers from English reading comprehension dataset.

presence of a weighted scoring scheme for each question, which demonstrates promise in improving unsupervised ASAG performance when used.

In ASAG, student answers often contain information beyond the key concepts instructors are looking for, though those extra pieces of text typically do not affect their scores unless they are contradictory or wrong. In order to address the shortcomings of symmetric similarity techniques, we propose an intuitive technique, *Vecalign*, which uses word vector based representations for unsupervised ASAG. Our technique computes an aggregate of word-level distances based on one-to-many word vector alignment using the cosine similarity of the aligned words vectors. Importantly, this allows us to assess similarity of the student answer against the model answer asymmetrically as a textual entailment problem (whether the student answer implies the model answer), which vector space methods such as *paragraph vectors* cannot do. In summary, our contributions include:

1. A comparison of the applicability of popular distributed vector representations of text for unsupervised ASAG across domains.

2. Proposal of an asymmetric word vector alignment method, which exploits weighted scoring schemes. Results indicate that such schemes show promise for improving ASAG reliability by allowing improved expressiveness in specifying answer requirements.

3. An analysis focusing on qualitative assessment of the methods, in light of the shortcomings of correlation metrics such as Pearson's r. To this end, we discuss points of consideration relating to the methods and design of questions for unsupervised ASAG.

2 Background

Our work draws on two foundational bodies of research: that of unsupervised ASAG as well as distributional semantics.

2.1 Unsupervised ASAG

Two recent surveys (Roy et al., 2015; Burrows et al., 2015) provide comprehensive views of research in ASAG, where similarity-based ASAG techniques can be broken into categories including lexical, knowledge-based and vector space. Among the lexical measures, one of the earliest approaches is Evaluating Responses with BLEU (Perez et al., 2004). It adapted the most popular evaluation measure for machine translation, i.e., BLEU, for ASAG with a set of natural language processing techniques such as stemming, closed-class word removal, etc. Mohler and Mihalcea (2009) conducted a comparative study of different semantic similarity measures for ASAG including knowledge-based measures using Wordnet as well as vector space-based measures such as Latent Semantic Analysis (LSA) (Landauer et al., 1998) and Explicit semantic analysis (Gabrilovich and Markovitch, 2006). LSA has remained a popular approach for ASAG and been applied in many variations (Graesser et al., 2000; Wiemer-Hastings and

Zipitria, 2001; Kanejiya et al., 2003; Klein et al., 2011). Lexical and semantic measures have been combined to validate natural complementarity of syntax and semantics for ASAG tasks (Perez et al., 2005). Wael H Gomaa (2012) compared several lexical and corpus-based similarity algorithms (13 string-based and 4 corpus) and their combinations for grading answers on a 0-5 scale. Dzikovska et al. (2013) conducted a 5-way (non-ordinal scale) Student Response Analysis challenge as a part of SemEval-2013. However, the task had more emphasis on giving feedback on student answers, possibly using textual entailment techniques.

2.2 Compositional distributional semantics

In recent years there has been an abundance of distributional text representation techniques based on the distributional hypothesis that words that appear in similar contexts have similar meanings (Harris, 1968). Popular techniques include *word2vec* (Mikolov et al., 2013b) and *Glove* (Pennington et al., 2014) in recent times, but also concepts such as latent semantic analysis (Deerwester et al., 1990) and its variants, as well as measures of distributional similarity (Lee, 1999; Lin, 1998). Compositional techniques building on these word vectors derive vector representations of longer pieces of text, i.e., phrases, sentences, paragraphs and documents. An approach of averaging of bag of word representations of text snippets was employed by early researchers such as (Landauer and Dumais, 1997; Foltz et al., 1998). While they were the first ones to introduce the notion, these approaches do not incorporate word order and have the adaptive capacity to represent the variety of possible syntactic relations in a phrase. Additionally, Erk and Pad (2008) highlighted that a fixed dimensionality vector may suffer from *information scalability* and not able to represent text snippets of arbitrary length. Some related models include holographic reduced representations (Plate, 1995), quantum logic (Widdows, 2008), discrete-continuous models (Clark and Pulman, 2007) and compositional matrix space model (Rudolph and Giesbrecht, 2010). Grefenstette and Sadrzadeh (2011) analyze subject–verb–object triplets and find a matrix-based categorical model to correlate well with human judgments. In recent times there has been a slew of work towards vector composition using neural network models. Notable of those are the paragraph vector of Le and Mikolov (2014) and the recursive deep models of Socher et al. (2013).

Most of the papers mentioned here emphasize obtaining a good generalized vector representation of text snippets. In the case of ASAG, our primary interest is to obtain a measure of similarity between them. We observe that for ASAG not all words in student and model answers are equally important. Rather, pairs of related words which appear in student and model answers are more important than some other words. Hence a measure which identifies and aggregates over such pairs would be meaningful to apply, such as Word Movers Distance (WMD) (Kusner et al., 2015).

3 Techniques

There are a wide range of approaches for generating scores of similarity between documents (Choi et al., 2010). We evaluate a variety of representative distributional semantics based approaches in the task of unsupervised grading and propose an asymmetric method based on aligning word vectors that exploits properties of grading tasks.

3.1 Document vector approaches

We consider two approaches that create document vector representations without composing individual word representations. We use implementations available in the *gensim* Python package (Rehurek, 2010).

Latent semantic analysis: Latent semantic analysis uses matrix factorization to create vector representation of words and documents in the same space. Since it has had a long history of use in ASAG, we consider it as a point of comparison in evaluating the other methods presented here.

Paragraph vectors: A number of more recent approaches have been proposed for creating vector representations of larger units of text (Mitchell and Lapata, 2010; Mikolov et al., 2013a; Grefenstette et al., 2013). The Paragraph vector method of (Le and Mikolov, 2014) provides a way to train vector representations of such larger units by using an approach similar to that of (Mikolov et al., 2013a). Importantly, paragraph vector representations are not compositions of word vectors and they also implicitly consider word order to some extent, which word vector compositions generally do not. We evaluated a variety of configurations, with the *distributed bag of words* model performing the best.

3.2 Word vector based approaches

We also evaluate a number of methods based directly on word vectors of the continuous bag-of-words (CBOW) model (Mikolov et al., 2013b). This method of word vector learning has allowed for word vectors to be trained on larger quantities of data than before, permitting state-of-the-art results in various tasks. Two key practical advantages of this is that the 100 billion word Google News Corpus can be harnessed and that hyperparameters do not need to be tweaked.[1]

Averaging word vectors: A naive approach to creating a document representation is by collapsing word vectors into a single vector through addition or multiplication. In ASAG, this approach has been used in the context of LSA word vectors (Perez et al., 2005). Though composition of few word vectors has demonstrated interesting results, one problem with this method of addition is that the best dimensionality for vectors of a set of paragraphs may not be the same as those for words.

Word mover's distance: This is a measure of similarity between groups of words that is equal to the minimum total distance the word vectors of one document must move in the vector space in order to become the word vectors of another document (Kusner et al., 2015). This method allows words to move to multiple other words, when a mismatch in document sizes occurs.

Targeting scoring schemes with *Vecalign*: We present a vector alignment method, *Vecalign-asym*, designed to capitalize on a feature of assessments: the asymmetry of scoring scheme items and student answer sizes, comparable to the word2vec alignment feature used in the supervised system of Sakaguchi et al. (2015).

Given two texts A and B (model and student answers, respectively), non-open-class words (words that are not nouns, verbs, adjectives or adverbs) are first removed from both. Each word is then replaced by its word vector representation such that we now have A consisting of vectors $(\mathbf{a_1}, \mathbf{a_2}, \dots, \mathbf{a_m})$ and B consisting of vectors $(\mathbf{b_1}, \mathbf{b_2}, \dots, \mathbf{b_n})$. We define the one-to-many similarity using the cosine similarity of the component word vectors:

$$asym(A,B) = \frac{\sum_{\mathbf{a}_i \in A} \max_{\mathbf{b}_j \in B}(cos(\mathbf{a}_i, \mathbf{b}_j))}{m} \tag{1}$$

This similarity measure is asymmetric, and is motivated by the observation that model answers are often more concise than student answers, since model answers typically contain only the salient points, while student answers are frequently less to the point, without necessarily being less correct.

If a scoring scheme is available, the awarded score can be defined as the weighted average of the asymmetric Vecalign similarity of each element in the scoring scheme with the student answer. While other methods can similarly be used with scoring scheme, the asymmetry between the size of the student answer and the scoring scheme suits Vecalign-asym well.

We believe that scoring schemes represent a very promising approach for both human grading and ASAG for a variety of reasons: (a) they elicit clearer wording of specifically what the creator of the question is looking for; (b) they allow for explicit weighting of the importance of these components; (c) By introducing smaller scoring scheme items, each of which should be covered in a student answer, they decompose the problem into sub-problems that have a textual entailment flavor, and more readily permit the use of effective asymmetric metrics; (d) We additionally conjecture that scoring schemes improve agreement between graders by making the creator think about the question more carefully by providing clearer guidance to graders.

In addition to the asymmetric Vecalign-asym, a symmetric version can be defined by the average of the asymmetric similarity in both directions. This is similar to the approach of the lexical similarity methods of (Mohler and Mihalcea, 2009). We refer to this measure simply as *Vecalign*.

[1] While these vectors were trained in a supervised manner, our proposed method still remains unsupervised analogous to how LSA has been treated as an unsupervised textual similarity measure. An interesting future study would be to train domain specific vectors based on student answer corpora but our datasets were very small for doing the same.

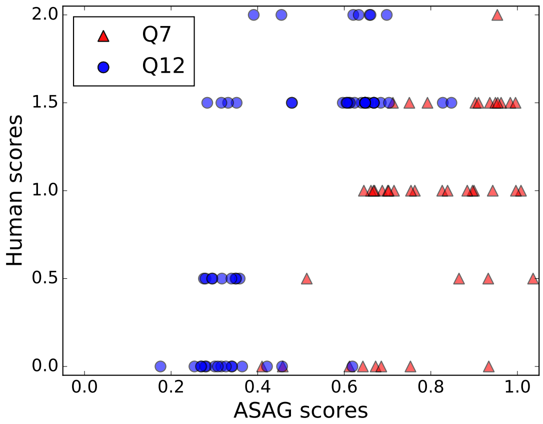

Figure 1: Questions 7 and 12 from the reading comprehension dataset. The ASAG scores come from the *vecalign* approach using the scoring scheme as the gold standard. Overall correlation is less than the average of its parts. Darker circles and triangles represent duplicate datapoints.

4 Data

We evaluate these methods on two datsets. The first is a set of questions (*CSDataset*), model answers and student answers taken from an undergraduate computer science course (Mohler et al., 2011). The dataset comprises of 87 questions from a number of assignments and examinations, with responses from 31 students. The second dataset comes from a high-school reading comprehension task for Class (Standard) XII (12) students as a part of a course on English in the Central Board of Secondary Education (CBSE) in India (*RCDataset*). The dataset comprises 14 questions, each with the recorded responses from 58 students. Along with model answers, questions in this dataset also have fine grained scoring schemes. We will share the dataset with people who are interested in pursuing research in this field and will be benefited by the dataset. Both datasets were tokenized with the *Punkt* tokenizer implementation of NLTK (Bird, 2006) before case normalization and lower-casing.

5 Quantitative evaluation

5.1 Metrics

Pearson's r has been used quite extensively in prior ASAG research. However, the suitability of Pearson's r has been questioned in the context of ASAG (Mohler and Mihalcea, 2009) as well as in general (Willmott, 1982). We too find issue with the use of Pearson's r. Firstly, very different scatter plots can yield similar correlations. Different questions result in different lines of best fit, which sullies the overall correlation. Consider Figure 1, which presents scatter plots of one ASAG system against the human scores. The individual correlations of the questions are both 0.72, but the overall correlation drops to 0.55. For this reason we favour question-wise correlation. Although this means that each number is determined by fewer datapoints and the calculations are statistically less meaningful, we still find that this evaluation is more informative. Furthermore, Pearson's r also penalizes non-linearities. For unsupervised ASAG this is not ideal, as in this case a nonlinear function cannot be learned to optimize for linear correlation. Spearman's ρ, while still subject to the first two problems, avoids this problem and thus we primarily report Spearman's ρ.

Finally, we note that using measures of correlation sidesteps the real-world issue of allocating an actual score to the students. If the ASAG system is used for merely ranking the students, this is not a problem. However, if we need to scale ASAG scores to appropriately grade students, it is a nontrivial problem that is task dependent and likely requires some degree of supervision. The limitations of these metrics motivated us to perform the qualitative evaluation of Section 6, which was influenced largely by manual inspection of the ASAG system outputs and the relationship between scores and the gold-standard average of human scores.

Method	Spearman's ρ of question number													
	All	1	3	12	27	41	45	50	60	69	73	82	84	87
IAA	*.68*	*.76*	*.62*	*.96*	*.99*	*.54*	*.06*	*.47*	*1*	*.52*	*.58*	*und*	*und*	*.85*
VAA	**.50**	.83	**.60**	**.64**	.01	.17	.24	.59	**.72**	**.88**	**.54**	-.04	-.04	.43
VA	.49	**.85**	.55	.58	.25	.19	.08	.62	.34	.55	.51	-.01	.01	.46
W2V-Add	.32	.74	.03	.29	.12	**.21**	**.41**	**.64**	.26	.55	.21	-.04	.01	.18
WMD	.42	**.85**	.02	.57	.35	.02	.17	**.64**	.34	.55	.44	.06	.01	**.65**
Para-Vec	.49	.75	-.14	-.33	.17	.07	.19	.44	.30	.53	.47	**.32**	**.04**	.62
LSA	.39	.65	.07	-.45	**.47**	-.21	.22	.52	.13	.53	-.06	-.08	**.04**	.54

Table 2: Spearman's ρ of the ASAG systems on the CSDataset of Mohler et. al. (2011). There were 87 questions in total, and so only a representative sampling is presented, along with the overall correlation across all 87 questions. Undefined numbers are indicated by *und.* (the correlation is undefined when the variance along any dimension is zero).

Method	Gold	Spearman's ρ of question number														
		All	1	2	3	4	5	6	7	8	9	10	11	12	13	14
IAA	scheme	*.75*	*.73*	*.77*	*.80*	*.27*	*.86*	*und.*	*.71*	*.70*	*.60*	*.61*	*.93*	*.65*	*.55*	*.75*
VAA	scheme	.16	.66	**.75**	**.73**	**.59**	*und.*	*und.*	**.73**	**.69**	.56	**.48**	**.74**	**.69**	**.71**	**.76**
VA	scheme	.09	.62	.51	.70	.55	*und.*	*und.*	.68	.65	.53	.39	.58	.63	.66	**.76**
W2V-Add	scheme	-.16	.66	.37	.72	.46	.21	*und.*	.64	.55	**.57**	.43	.48	.43	.63	.68
WMD	scheme	.08	.56	.16	.58	.52	-.01	*und.*	.51	.37	.45	.36	.60	.59	.58	.68
Para-Vec	scheme	-.04	.54	.09	.61	.04	.10	*und.*	.49	.35	.38	.47	.29	.52	.55	.61
LSA	scheme	.09	.14	.17	.60	.48	**.31**	*und.*	.35	.19	.27	.33	*und.*	.57	.43	.50
VAA	ref.	.29	**.68**	**.75**	.69	.36	.05	*und.*	.71	.49	.48	**.48**	**.74**	.44	**.71**	.66
VA	ref.	.32	.61	.47	.65	.26	-.11	*und.*	.67	.54	.46	.46	.58	.56	.66	.72
W2V-Add	ref.	.13	.65	.74	.62	.38	.03	*und.*	.61	.36	.39	.46	.48	.36	.60	.52
WMD	ref.	.27	.62	.69	.67	.35	-.07	*und.*	.70	.45	.46	.46	.60	.54	.61	.68
Para-Vec	ref.	**.42**	**.68**	.42	.61	.35	-.02	*und.*	.60	.43	.38	**.48**	.32	.45	.46	.45
LSA	ref.	.20	.10	.47	.65	.22	.26	*und.*	.48	.19	.16	.25	.22	.45	.35	.14

Table 3: Overall and question-wise performance of methods on the reading comprehension dataset. *Gold* indicates whether the gold standard is the marking scheme or model answer.

5.2 Observations

Tables 2 and 3 present the overall and question-wise correlations of the ASAG systems against the average of the human scores. Only the reading comprehension dataset had scoring scheme associated, hence Table 3 shows results with respect to both model answer and scoring scheme. We additionally present the annotators' correlation with one another as a point of reference (**IAA**), though these aren't ASAG scores. Due to space constraints, we opted to present a representative sample of the questions from the CSdataset.[2] In each table, we compare vecalign-asym (**VAA**), vecalign (**VA**), WMD (**WMD**), LSA (**LSA**), Paragraph Vectors (**Para-Vec**), and word vector addition (**W2V-Add**). We observe that the relative performance of models stays approximately the same across both datasets. In most cases, Paragraph Vectors and LSA underperform the word2vec based approaches, which is an indication that the large dataset afforded by these methods is a key advantage. Furthermore, note that results of paragraph vector and LSA were from among the best performing hyperparameter configurations, as a number of them were trained.

5.2.1 Comparison with human correlation

Note that the correlation of the ASAG grades cannot be fairly compared with the correlation of the human-assigned grades, since the ASAG grades are evaluated against the *average* of the human scores. However, it is nevertheless meaningful to consider when ASAG correlation is comparable to or exceeds that of humans, as it highlights questions where automated marking might be as effective as manual marking. It is worth noting that in the the RCDataset, the asymmetric approach harnessing the scheme

[2] Note, however, that the reported figure for *all* is across all 87 questions.

has a higher correlation than the inter-annotator agreement on 5 out of the 14 questions. In other questions the ASAG system is not so far behind the human agreement. The overall correlation is very low compared to that of humans, since human marks are scored on a consistent scale between questions, whereas ASAG grades are not. As mentioned in Section 5.1, this makes establishing correlation over a number of questions not very informative.

5.2.2 Performance using scoring schemes

As can be seen in Table 3, in 10 out of 14 questions of RCDataset, the asymmetric approach using the scoring scheme is the best performing approach: it has 6 out of 14 questions with a correlation over 0.7 and another 3 are over 0.6. The asymmetric vecalign method has an advantage over the other methods that measured student answers against scheme items, since scoring scheme items are typically far smaller than student answers. The asymmetric vecalign method is also frequently the best performer even when evaluating against the model answer. This is also a reflection on the asymmetry of the model answer. Student answers tend to be longer, while the provided model answer is shorter, capturing only the salient points.

5.2.3 Ensemble performance

We experimented with a few ensemble methods where the score for answers assigned by different methods were averaged to produce a final score. These combinations involved averaging the results of *vecalign* against the model answer along with asymmetric *vecalign* against the scoring scheme (thus intending to harness information both from the scoring scheme and model answer), as well as combinations considering LSA and Paragraph Vectors. However, in all cases, the ensemble approach underperformed the best constituent approach, since different methods implicitly score on different scales.

6 Discussion

We present a discussion involving qualitative assessment of the ASAG systems with respect to selected questions from the CSDataset ("CSQ").

6.1 Questions yielding low or undefined correlations

A number of questions have a low or undefined correlation because of a clear deficiency in the system. For example, the model answer for CSQ27 is "run-time error". Since "run-time" doesn't exist in the vocabulary of the CBOW model trained on the Google News Corpus, only "error" is considered as relevant, and as such responses such as "compilation error" receive high scores, while "run-time" receives none. Note that LSA did not completely fall down on this question.

In other cases, low correlation is not representative of poor system performance. Consider CSQ84. The model answer is "push and pop". This response was present in every student answer, and every student answer was awarded 5/5, except for one who was awarded 4.5. Vecalign-asym awarded almost every student's answer a perfect score, except for one that included "pop-" as a token. Infrequent deviation from perfect scores yields unreliable correlations.

6.1.1 Answer open-endedness and pattern matching

We observe that the more open-ended the expresssion of a legitimate answer can be, the less useful the model answer is. However, open-endedness is an important motivation for short answers as opposed to multiple choice questions. Therefore, a balance must be struck. For CSQ41, the model answer describes the main advantage of linked lists as "the linked lists can be of variable length". The notion of "variable length" can be described in many ways that are not so easily capturable even by semantic vectors. An answer "its resizable" was given a low score by the ASAG system. Other answers described linked lists as being able to be "grown dynamically" and that "elements can be added to a linked list w/o defining any size.".

At the other end of the spectrum, there were questions where the model answer indicated that simple pattern matching approaches for grading would suffice, and perhaps be more effective than distributional semantics based approaches. This is particularly true for jargon such as the previously mentioned "enqueue and dequeue" and "push and pop". Some questions were not suitable for a short answer framework at all, such as yes/no questions. In designing assessments for automatic grading, identifying where simple pattern matching or multiple choice would suffice would improve system performance.

6.1.2 Fundamental limitations of ASAG and question quality

The example of CSQ12 also highlights some fundamental limitations. Student answers that received full marks by graders included answers such as "any number you want" and "as many as needed". However, the model answer was "unlimited number", making it difficult for ASAG.

CSQ45 asks "what is the main advantage of a doubly-linked list over a basic linked list?" and uses the model answer "all the deletion and insertion operations can be performed in constant time, including those operations performed before a given location in the list or at the end of the list". The correlation between human answers for this question was only 0.06, indicating problems with the question and model answer. Since there are multiple advantages and disadvantages of doubly-linked lists, such a question may be more suited to a scoring scheme reference comprised of shorter items that can be matched against the student answer. Another notable reason is that the mention of "speed" and "fast" cannot easily be related to "constant time".

6.1.3 Length of model answer

The model answers in the computer science course were often short and sweet, which is likely why Vecalign-asym outperformed Vecalign. In many cases, student answers elaborated beyond what was in the model and were thus punished by the symmetric method. The asymmetric method avoided this, which explains its significantly better performance on some questions (e.g. CSQ60), where the model answer is a single word.

6.1.4 Text normalization

One key step of normalization that we performed was case normalization, which demonstrated its importance for word vectors. CSQ12 has a model answer "Unlimited number". "Unlimited" has a cosine similarity with "infinite" of only 0.22. But when lowercased, the similarity jumps to 0.48.

6.1.5 Interpretability

Since Vecalign aligns word vectors, it is usually simple to interpret how a score was arrived at for a student answer. However aragraph vectors and LSA are notably more opaque.

6.1.6 Appropriateness of scoring schemes

Use of scoring schemes work well in conjunction with the Vecalign-asym method. However, performance is notably worse when the scoring schemes are used with the other symmetric approaches. Since the elements of the scoring scheme are small, it is a textual entailment problem, which is inherently asymmetric.

7 Conclusion

Although ASAG has been investigated for many years, adoption is not widespread. This is partly because ASAG systems often do not perform adequately and those that do perform well enough for real-world use typically require significant manual supervision (Liu et al., 2014). In order to make ASAG more practicable in the real-world, effort should be placed in situating available models so that it is feasible to create reliable ASAG systems without an overly large amount human of supervision. In this paper, we have made a contribution towards minimizing the impact of this compromise by presenting a simple, effective, and interpretable method of word vector alignment in conjunction with the use of weighted marking schemes, as well as evaluating a variety of alternative approaches. Another important part of making ASAG feasible is examination of which types of questions are amiable to automation. This is an important consideration since questions on which ASAG systems perform poorly are often also the ones on which humans disagree.

References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725, Berlin, Germany, August. Association for Computational Linguistics.

Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155.

S. Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

S. Choi, S. Cha, and C. Tappert. 2010. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48.

Stephen Clark and Stephen Pulman. 2007. Combining symbolic and distributional models of meaning. In *AAAI Spring Symposium: Quantum Interaction*, pages 52–55. AAAI.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Myroslava O Dzikovska, Rodney D Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa T Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. Technical report, DTIC Document.

Katrin Erk and Sebastian Pad. 2008. A structured vector space model for word meaning in context. In *EMNLP*, pages 897–906. ACL.

P. W. Foltz, W. Kintsch, and T. K. Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25:285–307.

Evgeniy Gabrilovich and Shaul Markovitch. 2006. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In *AAAI*, pages 1301–1306. AAAI Press.

Arthur C. Graesser, Peter M. Wiemer-Hastings, Katja Wiemer-Hastings, Derek Harter, and Natalie K. Person. 2000. Using latent semantic analysis to evaluate the contributions of students in autotutor. *Interactive Learning Environments*, 8(2):129–147.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404. Association for Computational Linguistics.

E. Grefenstette, G. Dinu, Y. Zhang, M. Sadrzadeh, and M. Baroni. 2013. Multi-step regression learning for compositional distributional semantics. *arXiv preprint arXiv:1301.6939*.

Zellig Harris. 1968. *Mathematical Structures of Language*. John Wiley and Son, New York.

Dharmendra Kanejiya, Arun Kumar, and Surendra Prasad. 2003. 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 - Volume 2*, pages 53–60.

Richard Klein, Angelo Kyrilov, and Mayya Tokman. 2011. Automated assessment of short free-text responses in computer science using latent semantic analysis. In *ITiCSE*, pages 158–162. ACM.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In Francis R. Bach and David M. Blei, editors, *ICML*, volume 37 of *JMLR Proceedings*, pages 957–966. JMLR.org.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

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

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Lillian Lee. 1999. Measures of distributional similarity. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL98*, Montreal, Canada.

O. Liu, C. Brew, J. Blackmore, L. Gerard, J. Madhok, and M. Linn. 2014. Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*, 33(2):19–28.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Michael Mohler and Rada Mihalcea. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 567–575.

Michael Mohler, Razvan C. Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *ACL*, pages 752–762.

J. Pennington, R. Socher, and C. Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP)*, 12:1532–1543.

Diana Perez, Enrique Alfonseca, and Pilar Rodrguez. 2004. Application of the bleu method for evaluating free-text answers in an e-learning environment. In *LREC*. European Language Resources Association.

Diana Perez, Alfio Gliozzo, Carlo Strapparava, Enrique Alfonseca, Pilar Rodriguez, and Bernardo Magnini. 2005. Automatic assessment of students' free-text answers underpinned by the combination of a BLEU-inspired algorithm and latent semantic analysis. In *Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference, FLAIRS*, Clearwater Beach, FL, United states.

T. A. Plate. 1995. Holographic reduced representations. *IEEE Transactions on Neural Networks*, 6:623–641.

R. Rehurek. 2010. Fast and faster: A comparison of two streamed matrix decomposition algorithms.

Shourya Roy, Y Narahari, and Om D Deshmukh. 2015. A perspective on computer assisted assessment techniques for short free-text answers. In *Computer Assisted Assessment. Research into E-Assessment*, pages 96–109. Springer.

Sebastian Rudolph and Eugenie Giesbrecht. 2010. Compositional matrix-space models of language. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 907–916. Association for Computational Linguistics.

Keisuke Sakaguchi, Michael Heilman, and Nitin Madnani. 2015. Effective feature integration for automated short answer scoring. *Proceedings of NAACL, Denver, Colorado, USA*.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer.

Aly A Fahmy Wael H Gomaa. 2012. Short Answer Grading Using String Similarity And Corpus-Based Similarity. *International Journal of Advanced Computer Science and Applications(IJACSA)*, 3(11).

Dominic Widdows. 2008. Semantic vector products: Some initial investigations. In *Second AAAI Symposium on Quantum Interaction*, volume 26, page 28th. Citeseer.

P. Wiemer-Hastings and I. Zipitria. 2001. Rules for syntax, vectors for semantics. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, Mahwah, NJ. Erlbaum.

C. J. Willmott. 1982. Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society*, 63:1309–1369, #nov#.