ORIGINAL RESEARCH

Information-based methods for evaluating the semantics of automatically generated test items

Syed Latifi*[1], Mark Gierl[1], Ren Wang[2], Hollis Lai[3], Andong Wang[2]

[1] Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, Alberta, Canada
[2] Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada
[3] Office of Undergraduate Medical Education, University of Alberta, Edmonton, Alberta, Canada

Received: August 23, 2016 **Accepted:** October 30, 2016 **Online Published:** November 22, 2016
DOI: 10.5430/air.v6n1p69 **URL:** http://dx.doi.org/10.5430/air.v6n1p69

ABSTRACT

Multiple-choice questions are the popular type of test items that are used for testing the knowledge of health-science students in north America and elsewhere. The motivation of this article is to present the recent advances in the automatic item generation (AIG) and to propose a novel unsupervised approach that extends the information-based Compositional Distributional Semantic Model (CDSM) to measure the semantic relatedness among the pool of automatically generated items. We have used operational item bank from the medical science domain for developing the CDSM and demonstrated our approach using the concepts from AIG research. We illustrated our approach using eleven item models from the medical education domain, and discussed the possible applications to advance the AIG research.

Key Words: Automatic item generation, Technology-enhanced assessments, Question generation, Large-scale assessment, Semantics similarity, Computer based testing

1. INTRODUCTION

The growing popularity of automatic item generation (AIG) can be attributed to the increasing demand for large pools of operational test items that measure learning outcomes.[1,2] AIG is an algorithmic way of generating assessment tasks by combining cognitive theories, psychometric practices, and computer technologies. The outcome of this algorithmic transcription is often referred to as item modeling.[3–6] The item modeling process requires the identification of elements within the assessment task so that these elements can be used to create large set of items.[1,7] These generated items may or may not be similar to one another and thus the lexical similarity among the generated items is often unknown.

Consequently, it is imperative to develop a measure to quantify the semantic similarity among the generated items so that the relatedness of words within the items across the item pool can be evaluated. Information on semantic relatedness will enhance the quality and usability of generated item pools thereby leading to the selection of higher quality distractors[8] using semantic descriptions of the item content that, in turn, could help predict the item difficulty level,[9] the quantification of language diversity in the generated item pool, the use of similarity-based theory to control the difficulty of the multiple-choice questions – MCQs[10] and a better understanding what makes an item difficult for a group of students. However, visual evaluation of similarity and semantic relatedness is subjective and therefore ineffective and could be impractical for large set of automatically generated

*Correspondence: Syed Latifi; Email: syed.latifi@ualberta.ca; Address: Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, Alberta, Canada.

items, Unfortunately, quantifying item similarity is challenging because prior semantic knowledge about the assessment tasks is often not available.

Hence, the purpose of the present study is to introduce a novel method for assessing the semantic relatedness among the generated items. Our approach extends the compositional distributional semantic model (CDSM) for assessing the semantic relatedness among a set of generated items. We will also present a popular measure from the natural language processing (NLP) domain for the quantification of visual similarity of automatically generated items. We illustrate and discuss our approach using eleven item models from the medical education domain.

2. THEORETICAL FRAMEWORK

2.1 Approaches for automatic item generation

Continuous testing is one important benefit of computer-based testing (CBT). But it also requires large pools of operational test items. Traditional item development practices are resource intensive and are less viable for creating the content required for CBT. Another important aspect of CBT is the administration of parallel test forms. In which case, the examinee(s) may require a different version of test form, due to assessment needs or due to exceptional circumstances, that must assess the identical knowledge level across multiple test administrations. This is almost impossible to accomplish if the test forms are created manually and thus innovative methods, such as parallel test form tabu[11] and particle swarm optimization[12] were proposed to combine the test items using large item banks. However, examination agencies which adopt computer-assisted testing systems are faced with the daunting tasking of creating thousands of new and expensive items for developing the large item banks, that must meets multiple assessment and pedagogical objectives. This is a challenging issue in educational assessment,[13] and thus had promoted the AIG research and practice.

AIG employs three general steps.[6, 13] First, content experts create an item model by identifying the elements in the assessment task that can be manipulated. Second, the item model is programmed for algorithmic variation of identified elements using item generation software. Third, statistical models are used to estimate the psychometric properties of the generated items based on the combination of constrained-elements used in item assembly. The purpose of this study is to refine and extend step 2, the item generation process. AIG is used to produce the stem as well as the options as part of the item text. The stem is the part of the item which contains the context, content, item, and/or the question the examinee is required to answer. The options include the alternative answers with one correct option and one or more incorrect options or distracters. For the purpose of this study multiple-choice item models were used which generated both stem and four options.

The focus of the AIG process is to create items that share the same cognitive attribute and yet look different to the test takers. However, two broad item classification are discussed in the AIG literature; namely variants and clones (or isomorphs).[4, 6, 13] The generated items are classified as isomorphs if the elements of the assessment task is simplistically manipulated, where the resulting set of items contain only a slight variation in the values of the elements in an assessment task. Conversely, generated items are termed variants if the elements of the assessment task are sufficiently constrained while its instances are being generated to produce items that look different from one another. Regardless of whether the generated items are classified as isomorphs or variants, the process of building item models is iterative, which provides the opportunity for data collection and analysis,[13] and as a result, could be used to improve the quality of the generated items.

Items models can be developed using either a weak-theory or a strong-theory approach.[14, 15] In weak theory approach, design guidelines are used to create item models that generate isomorphic item instances which imitates a given parent item (or family of parent items). The weak theory approach is well suited to broad content domains where few theoretical descriptions exist on the knowledge and skills required to solve test items.[6] Conversely, the strong-theory approach employs a cognitive model to specify and manipulate the elements that yield generated items with predictable psychometric characteristics (*e.g.*, item difficulty). To date, few comparable cognitive theories exist to guide the item development practices[16] thereby limiting the practical application of strong theory. Both approaches require the content experts to engage in the complex task of item modeling using a combination of design guidelines and principles acquired from experiences, theory, and research[14] For the purpose of this study, eleven multiple-choice item models were created using a strong-theory approach in the content area of medical education.

2.2 Measure of semantic relatedness

Comprehending the meanings of electronic text is a fundamental application of natural language processing (NLP), which involve the segmentation of large text (*e.g.*, text passage) into smaller units of text (*i.e.*, sentences or phrases) and employing NLP methods for representing the meaning of these units. These methods can be broadly categorized in three families,[17, 18] namely semantic networks, feature-based models, and semantic spaces. The semantic networks

representation can be visualized as graphs of words, in which each node is a word and the edges between the words represents the semantics relationships between the words. Here, the semantic relatedness between words is expressed by the path length (*i.e.*, number of edges between words). The shorter path represents semantically related words and the longer path represents less related words. Semantic networks are not practical for a large language corpus because it requires linguistic modelers to hand code the nodes and the edges. In feature-based modeling, native speaker (*i.e.*, human modeler) are asked to develop list-of-feature that they consider important for representing the meaning of a word. This approach is time consuming and often substantively unaligned because of the way the representations are coded and analyzed using human judgements,[17, 19] and could also be impractical for larger text corpus.

Semantic space is an attempt to model the characteristics of a human semantic memory which is driven by the principle (from brain research) that the words with similar meaning co occur in similar linguistic environment. For example, gene and molecules tend to occur in contexts of similar words, such as proteins, DNA, and hereditary. A semantic space is a vector-space that captures the meaning quantitatively in terms of co-occurrence statistics, where words or concepts are represented as vectors in a high-dimensional space.[17, 20] As a result, similarity of word meanings can be quantified by measuring their distance in high dimensional vector space. Semantic representation using semantic space is advantageous mainly because no explicit human judgments (*i.e.*, human modeler) are required and as a result larger lexicon (corpora) can be coded and analyzed.

The linguistic structures are compositional in nature because simpler language-elements are combined to form more complex ones. For example, morphemes (a minimal meaningful language unit) are combined into words, words into phrases, and phrases into sentences. Therefore it is reasonable to assume that the meaning of sentences is composed of the meanings of individual words or phrases.[17, 18, 21] Most importantly, the temporal order of words in a sentence represents the differences in meaning as a result of differences in word order or syntactic structure.[18] Early semantic space modeling approaches (*e.g.*, Foltz's[22]) were insensitive to word ordering and thus could not capture meaning differences that are modulated by differences in syntactic structure. Consider, for example, the following two sentences from Landauer and Dumais[23] that happen to use the same vocabulary but are still semantically unrelated:

(1) It was not the sales manager who hit the bottle that day, but the office worker with the serious drinking problem.
(2) That day the office manager, who was drinking, hit the problem sales worker with the bottle, but it was not serious.

The mechanism of compositionality, for quantifying the relatedness between text, depends on how high-dimensional representation are combined for constructing the sematic space.

Table 1. A hypothetical semantic space for "gene" and "molecules"

	Protein (w_1)	**DNA** (w_2)	**Hereditary** (w_3)	**Life** (w_4)	**Cell** (w_5)	**Acid** (w_6)
Gene	2	9	7	5	2	3
Molecules	7	5	0	7	3	1

Semantic spaces could be constructed either using an additive or a multiplicative model. Both additive and multiplicative approaches for constructing semantic space assumes that composition is a symmetric function of the constituents (words) and they apply the mathematical operation (addition or multiplication) for combining the co-occurrence vectors. But neither approach take into account the order of words. To illustrate basic composition functions, consider the simplified semantic space in Table 1 that represents the co-occurrence vector for words gene and molecules using a six-dimensional semantic space from a hypothetical corpora. A two word phrase p could be represented using its two constituents (words) q and r with a compositional-function acting on those constituents, that is, $p = f(q, r)$. Using Table 1, q = gene, and r = molecules, the additive model would yield the vector [9_{w1}, 14_{w2}, 7_{w3}, 12_{w4}, 5_{w5}, 4_{w6}] and multiplicative model would yield [14_{w1}, 45_{w2}, 0_{w3}, 35_{w4}, 6_{w5}, 3_{w6}] for representing the phrase p. These representations form the basis of distributional information for quantifying the semantic relatedness among constituents (*i.e.*, words or phrases). Such modeling, which approximates the meaning of words with vectors summarizing their patterns of co-occurrence in corpora, are called distributional semantic models (DSMs).

Researchers have extended the DSMs to incorporate the compositional structure of language and called these models

compositional-DSMs (CDSMs). CDSMs assume that the meaning of a word can be interpreted by its context and the meaning of a sentence can be derived from its compositions.[17,24] Central to CDSM is the notion of compositionality, *i.e.*, the meaning of complex expressions is determined by the meanings of their constituent expressions and the rules used to combine them. However, access to the annotated text and rules or corpora of symbolic-logic representation is challenging to evaluate in operational settings such as items generated from an item model.

3. Our Approach

3.1 Extension of compositional distributional semantic model

To address this challenge, we developed a unsupervised CDSM, which extends the work by Mitchell and Lapata.[17] We expanded the CDSM framework in three ways: 1) constructing word vector of co-occurrence weighted by distance, 2) using the bigram multiplicative model to compose sentence vector, and 3) computing term frequency–inverse document frequency (TF-IDF) weights for different words. Taken together, an enhanced CDSM method which explore the utility of making better use of the structural information for quantifying the semantic similarity.

One drawback for Mitchell and Lapata,[17,25] and Baroni and Zamparelli[26] approaches is that they do not take word distances into consideration and instead, only focus on neighboring words next to the target words. Semantic relatedness becomes stronger when words are closer, but long distance words can also convey semantic information about target words.[27] So, in our method, we not only focused on occurrences of neighboring words, but also occurrences of other words in the context window of a certain size, multiplied by some weights between 0 and 1 according to the distance. Implementing such weights serve as a reciprocal of the distance measure. For example, if context window = 3, and the sentence = A dog jump onto a tree, the vector of weights could be define for the target word jump relative to the distance of each ward in the context window, as shown in Figure 1.

A dog jump onto a tree.

	a	dog	onto	tree
jump	1/2+1/2	1	1	1/3

Figure 1. An example for weighting based on context window

The second modification we implemented is to use the bigram multiplicative model. In the original approach, Mitchell and Lapata[25] built the representation of sentences by composing its individual word representation. Normally, the composition method is only an additive operation, where all vectors are added to get the sentence vector. Although Mitchell and Lapata[17,25] proposed a multiplicative operation, they used it only to compare phrases, but not full sentences. We propose a bigram multiplicative operation that incorporates the syntactic sequence thereby taking into account the temporal order of words. Thus, their additive model for sentence vector $\sum_{i-1}^{n} v_i$ (where v_i is the vector representation) was extended by appending the bigram multiplicative model, as shown in equation 1:

$$\sum_{i=1}^{n} v_i + \alpha \sum_{j=1}^{n-1} v_j * v_{j+1} \quad (1)$$

Here the additive model is appended with the weighted multiplication as well as the component-wise multiplication of adjacent two words (where $v_j * v_{j+1}$ is the bigram representation; and α is the weight, farther the distance lower the weight). In this way we can exploit the multiplicative operation to compare sentences.

The last idea is to add weights to important words that affect the similarity. The important words were flagged using the TF-IDF score. TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a corpora collection. The words with higher TF-IDF scores are often the words that best characterize the topic of the document.[28] Intuitively, if a word is less frequent in the whole training set but appears often in one single sentence, then it means this word is of high probability to be significant to the theme of this sentence. Thus this word should be given more weight. Finally, the model representation vector of the sentence was updated by incorporating this modification, *i.e.*, words were weighted according to TF-IDF score (*i.e.*, $s_{w,i}$, $s_{w,j}$, and $s_{w,j+1}$ represents weighted TF-IDF score of a word at position i, j, and $j+1$). Mathematically, the final expression for sentence-vector relatedness, as shown in equation 2:

$$\sum_{i=1}^{n} s_{w,i} * v_i + \alpha \sum_{j=1}^{n-1} (s_{w,j} * v_j)(s_{w,j+1} * v_{j+1}) \quad (2)$$

We named our unsupervised measure the index of semantic heterogeneity or ISH. It ranges between 0.0 and 1.0. The minimum ISH value of 0.0 means high semantic diversity among the sentence pair and the maximum ISH value of 1.0 means high semantic relatedness among the sentence pairs. Simply put, the lower ISH indicates semantic diversity and

higher ISH indicates the semantic relatedness (*i.e.*, variant and isomorphs, respectively, in the AIG context). When more than two sentences (*i.e.*, items) are compared, the pair-wise ISH indices need to be computed and then averaged for evaluation of semantic relatedness. The standard deviation among the pair-wise ISH value could be computed for evaluating the amount of semantic variation among the set of sentence pairs.

3.2 Cosine similarity index

It is also important to quantify the visual similarly of the generated items. One numeric measure that serves as an alternate for visual similarity is the Cosine Similarity Index (CSI) . CSI is one among many type of NLP word similarity measures shown to produce high quality results across several text and document similarity domains[29–31] and could be used for the quantification of visual similarity of automatically generated items. CSI is based on a text-vector indexing technique[29] which measures the similarity between two vectors of co-occurring texts in terms of distance score. Mathematically, CSI is expressed in equation 3:

$$\cos(\overrightarrow{A}, \overrightarrow{B}) = \frac{\overrightarrow{A} \cdot \overrightarrow{B}}{\left\|\overrightarrow{A}\right\| \left\|\overrightarrow{B}\right\|} \tag{3}$$

Here, A and B are two binary vectors which represent the word occurrence in sentence A, and sentence B, from the universe of unique words. To illustrate, suppose the words in Table 1 forms the universe of words using six words (*i.e.*, proteins, DNA, hereditary, life, cell, acid), and we want to compute the visual similarity between two sentences, sentence A: "Life DNA hereditary protein DNA", and sentence B: "Acid hereditary cell". In total, we have six unique words and each word represents a dimension from the universe of words, which in this case is represented as a six dimensional vector. Next a word frequency vector needs to be constructed for each sentence which corresponds to the value in the word dimension. For example, the word DNA occurred twice in A, thus the value of 2 in dimension w_2. The word frequency vector for $\overrightarrow{A} = \langle 1\ w_1, 2\ w_2, 1\ w_3, 1\ w_4, 0\ w_5, 0\ w_6 \rangle$, and for $\overrightarrow{B} = \langle 0\ w_1, 0\ w_2, 1\ w_3, 0\ w_4, 1\ w_5, 1\ w_6 \rangle$. Here, the frequency of each word corresponds to the components of vector A and B. The vector values are then used in the CSI equation to produce a similarity value between sentence A and B, as demonstrated in equation 4:

$$\begin{aligned} &(1.0 + 2.0 + 1.1 + 1.0 + 0.1 + 0.1)/ \\ &(\sqrt{1^2 + 2^2 + 1^2 + 1^2 + 0^2 + 0^2} \times \\ &\sqrt{0^2 + 0^2 + 1^2 + 0^2 + 1^2 + 1^2}) \approx 0.22 \end{aligned} \tag{4}$$

The CSI ranges from 0 to 1. The minimum CSI value of 0 means that no words overlapped between the two vectors. The maximum CSI value of 1 means that the text represented by the two vectors are identical. That is, lower CSI indicates lower similarity and higher CSI indicates higher similarity among the words in the text. When more than two items are compared, the pair-wise CSI indices need to be computed and then averaged for the evaluation of similarity. The standard deviation among the pair-wise CSI value could be computed for evaluating the amount of variation among the set of generated items.

Table 2. Random sample of generated items from an item model of Post Operative Fever

531. The man is on post operative day 3 and has received a left hemicholectomy. The patient has a temperature of 38.5℃. There are no other potential complications; his age is recorded as 40 years. Which one of the following is the most likely diagnosis? a. Jugular Vein Thrombosis b. Pneumonia* c. Staphylococcal infection d. Urinary tract infection
832. A 70-year-old man has had a appendectomy. On post operative recovery day 4, after returning to a general ward settings, the man has a temperature of 38.5 ℃. Physical examination reveal a red and tender wound. Which one of the following is the most likely diagnosis? a. Wound infection* b. Jugular Vein Thrombosis c. Atypical pneumonia d. Urinary tract infection
158. The man has a temperature of 38.5 ℃ and is on post operative day 3. Earlier this year, the patient was hospitalized for urinary catheterization. The patient's age is recorded as 70-years. He has had a left hemicholectomy. The best prognosis for this patient would be: a. Urinary tract infection* b. Jugular Vein Thrombosis c. Atypical pneumonia d. Nonobstructive atelectasis

Note. *- correct option.

4. METHODS

Eleven item models were used to generate a large pool of test items. Each item model belongs to a specific topic within the medical education domain. These item models were developed by the content experts during an item modeling session at an international medical licensing authority. To generate the items, the computer program IGOR[32] was used. IGOR, an acronym for Item GeneratOR, is a software program writ-

ten in JAVA that instantiates all possible combinations of elements into items based on the definitions and constrains within the item model. The item models from content experts were first transcribed into an XML format that IGOR can interpret, after which IGOR computes the necessary information and outputs items in either HTML or Word format. A sample item model is shown in Figure 2, and a small random-sample of generated items is presented in Table 2. Next we will present the methods for constructing the semantic space and operationalization of our measures.

<table>
<tr><th colspan="5">Parent Items</th></tr>
<tr><td colspan="5">A 40-year-old woman has a appendectomy. On post operative day 4 he has a temperature of 38.5 C. Physical examination reveal a red and tender wound at the opening. Which one of the following is the most likely diagnosis?
(A) Pneumonia
(B) Wound infection *
(C) Urinary tract infection
(D) Nonobstructive atelectasis</td></tr>
<tr><th colspan="5">Item Model</th></tr>
<tr><td>Stem</td><td colspan="4">[Patient Profile]
1: A [[AGE]]-year-old [[GENDER]] has a [[SURGERY]].
2: The [[GENDER]] has a temperature of [[TEMPERATURE]] C and is on post operative day [[DAY]].
3: The [[GENDER]] is on post operative day [[DAY]] and has gone through [[SURGERY]].
[Further Examination]
1: On post operative recovery day [[DAY]] and after returning to general ward settings, the [[GENDER]] has a temperature of [[TEMPERATURE]] C.[[CUE]]
2: [[CUE]]. Patient`s age is recorded as [[AGE]]-years and had a [[SURGERY]].
3: Patient has temperature of [[TEMPERATURE]]C. [[CUE]]. [[GENDER]] age is recorded as [[AGE]] years.
[Diagnosis]
1: Which one of the following is the most likely diagnosis?
2: The best prognosis for patient recovery would be:</td></tr>
<tr><td rowspan="4">Elements</td><td rowspan="2">[SURGERY]
1: gastrectomy
2: left hemicholectomy
3: appendectomy</td><td rowspan="2">[SCENARIO]
1: Urinary tract infection
2: Actelectasis
3: Wound infection
4: Pneumonia
5: Deep vein thrombosis</td><td>[AGE]
Range: 40 to 70 by 10</td><td>[DAY]
Range 1 to 6</td></tr>
<tr><td>[GENDER]
1: man, his
2: woman, her</td><td>[TEMPERATURE]
Range: 38.0 to 38.6 by 0.2.</td></tr>
<tr><td colspan="4">[CUE]
1: Earlier this year, patient was hospitalized for urinary catheterization.
2: Physical examination reveal a red and tender wound at the opening.
3: There are no other potential complications.</td></tr>
<tr><td>[OPTION1]
1: Urinary tract infection
2: Nonobstructive atelectasis
3: Staphylococcal infection
4: Atypical pneumonia
5: Jugular Vein Thrombosis
6: Wound infection
7: Pneumonia</td><td>[OPTION2]
1: Urinary tract infection
2: Nonobstructive atelectasis
3: Staphylococcal infection
4: Atypical pneumonia
5: Jugular Vein Thrombosis
6: Wound infection
7: Pneumonia</td><td colspan="2">[OPTION3]
1: Urinary tract infection
2: Nonobstructive atelectasis
3: Staphylococcal infection
4: Atypical pneumonia
5: Jugular Vein Thrombosis
6: Wound infection
7: Pneumonia</td></tr>
<tr><td>Options</td><td colspan="4">[[OPTION1]], [[OPTION2]], [[OPTION3]]</td></tr>
<tr><td>Key</td><td colspan="4">Conditional on [[Scenario]]</td></tr>
</table>

Figure 2. A sample item model

4.1 Data corpus for constructing the semantic space

In total, 2,049 operational test items from the medial education domain were extracted from the item bank of an international medical licensing authority to build the corpus used in the current study. These multiple-choice items contain a stem and five options (*i.e.*, four distractor and one key). Each items belongs to one of six content area within the medical education domain, and was develop by the context expert using item development guideline. The content-wise count of these items is shown in Table 3. We used these item as a corpus for parameter tuning as well as for building the semantic space.

Table 3. Summary of Item counts as a function of content area

Content area	Item count
Medicine	669
Obstetrics and Gynecology	402
Pediatrics	274
Preventive Medicine & Community Health	54
Psychiatry	319
Surgery	331
Total	**2,049**

All items in the corpus were used to construct the semantic space of co-occurrence, which was then used to quantify the semantic relatedness for the AIG items. First, the text for each item (including the stem and the five options) was transcribed into sentences and was supplied as a TAB delimited text file to the semantic-space creation module. Several parameters were tuned so the optimal semantic space could be develop. We used 205 items as our development set (*i.e.*, 10% of 2,049 items) to tune the parameters. The final parameters were used to build the semantic space that, in turn, was used to quantify the semantic relatedness of the items from eleven AIG models.

4.2 Parameter tuning for constructing the semantic space

ISH is an extension of CDSM that requires the algorithm to learn the best set of parameters for representing distributional semantic space by means of parameter tuning. Parameter tuning is a process of learning an optimized set of parameters using independent dataset, that improves the overall performance of the system as evaluated by some measureable metric. For this step, we used independent dataset of 10,000 sentence pairs, that are rich in the lexical, syntactic and semantic phenomena and was annotated for relatedness in meaning by trained human annotator.[33] We used the gradient descent parameter learning method[34] for optimizing the performance metric of Pearson correlation between ISH and the human score of relatedness in meaning.

There are several parameters that must be tuned in order to achieve the best result for our improved version of CDSM. The first parameter is window size. When we compute the co-occurrence of words we only take into consideration the words whose distances are within a certain window size. By means of machine learning, we found that the best results are obtained when the window size is between 4 and 6. Smaller or larger range will not increase the accuracy of the results. Whether or not to add <s> and </s> at the start and end of sentences, respectively (which effects the word boundary selection for bigram multiplicative model) is our second parameters. We found that the correlation decreases to about 0.10 when we appended the start and end symbols, thus we omitted these symbols. For stopwords (most common words, *e.g.*, a, is, this), three options are available: 1) not deleting any, 2) delete using assigned list of common stopwords, or 3) deleting the n most frequent words. We tried all three methods as well as setting different thresholds for top n words and found that the best result is produced when we deleted the top 25 frequent words. The second and third modifications exhibited a satisfactory outcome. When we set a weight of 0.10 (α = 0.10) for bigram multiplicative vectors and took TF-IDF score of each word into consideration, the Pearson correlation result boosted significantly. The optimal parameters for our model are shown in Table 4.

Table 4. The best parameter values after model tuning

Parameters	Value
Window Size	6
Sentence Symbols	FALSE
Stopwords	25
Distance Weights	FALSE
Multiplicative Weight	0.1
TF-IDF	TRUE

4.3 Data analysis

For each item model, one hundred items were randomly selected from the set of generated items and then passed onto two software application modules that we develop to compute ISH-relatedness and CSI-similarity.

4.3.1 *For ISH*

Each AIG item (stem + options) were transcribed as sentence, and then paired with every other items. That is, for 100 randomly-selected items, 4,950 sentence-pair were created. Each sentence-pair was then quantified using the semantic space that we constructed using operational test item corpus. As a result, for each item model, 4,950 ISH values were computed and then summarized as a mean and standard deviation.

4.3.2 *For CSI*

To begin, the porter stemmer algorithm[35] was used to eliminate the post-fixes (*e.g.*, shapes into shape) and common words (*e.g.*, a, is, this). Next, the matrix of word occurrence was compiled for each sample of items, where each row represents an item and each column represents a unique word, and each row-by-column cell is enumerated dichotomously to determine whether a word occurred in a specific item. Finally, the CSI was calculated for each unique item-pair. That is, for a sample of hundred items, 4,950 CSI values were computed and summarized as a CSI mean and standard deviation.

Table 5. Summary statistics of Semantic Index of Heterogeneity (ISH) and Cosine Similarity Index (CSI) across eleven content area of health sciences

Content Area of Item Models	Semantic Heterogeneity (ISH)				Cosine Similarity (CSI)			
	Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
Abdominal Aneurysm	1.00	1.00	0.99	0.00	0.00	1.00	0.47	0.38
Adverse reactions to drugs	0.91	1.00	0.96	0.04	0.52	0.97	0.71	0.08
Diabetes	0.40	1.00	0.71	0.20	0.00	1.00	0.49	0.31
Diverticulitis	0.41	1.00	0.76	0.17	0.00	1.00	0.34	0.25
Gallstone	0.97	1.00	0.99	0.01	0.00	1.00	0.70	0.21
Hernia	0.27	1.00	0.70	0.17	0.00	1.00	0.53	0.16
Hypertension	0.44	1.00	0.79	0.17	0.00	1.00	0.48	0.24
Infection & Pregnancy	0.98	1.00	0.99	0.01	0.00	1.00	0.53	0.25
Menopause	0.45	1.00	0.77	0.15	0.00	1.00	0.36	0.23
Post Operative Fever	0.98	1.00	0.99	0.00	0.00	1.00	0.39	0.31
Post Operative Management	0.19	1.00	0.66	0.23	0.00	1.00	0.43	0.18

5. RESULTS

Table 5 summarizes the results using the semantic-heterogeneity and visual-similarity indices using the eleven medical education item models. Across the eleven item models, the mean SHI ranged from 0.66 to 0.99, with corresponding SD of 0.24 and 0.00. Item models that generated similar items (*i.e.*, item clones) had a higher mean ISH (> 0.80) and lower SD (< 0.10) values, suggesting that on average these item models generated items that are highly related in their meaning (*i.e.*, less lexical diversity in the generated items). Conversely, item models that generated semantically heterogeneous items (*i.e.*, item variants) had a lower mean SHI ($\leq$ 0.80) with higher SD ($\geq$ 0.10) values.

For our data, the mean CSI ranged from 0.34 to 0.71, with corresponding SD of 0.25 and 0.08. Item models that generated similar items (*i.e.*, item clones) had a higher mean CSI (> 0.70) and lower SD (< 0.10) values, suggesting that on average these item models generated the set of isomorphs. Conversely, item models that generated distinct items (*i.e.*, item variants) had a lower mean CSI ($\leq$ 0.70) with higher variability (SD $\geq$ 0.10) among the set of generated items.

These results suggest that six of the eleven item models generated variants. Consider for example, the Hypertension item model. This model had generated the most semantically heterogeneous items (ISH = 0.79) and yet visually the items are quite distinct (CSI = 0.48). The corresponding SD of 0.17 and 0.24, respectively, also suggest a high degree of heterogeneity in the generated item pool of this model. The Diabetes, Diverticulitis, Hernia, Hypertension, Menopause, and PostOperative Management item models produced the same pattern of results.

6. DISCUSSION AND CONCLUSIONS

Unfortunately, there are few empirical methods available for quantifying the similarity of generated test items,[36] and hence to date, similarity is often established more subjectively using judgments from test development specialists. To address this limitation in literature, we describe two measure of item similarity that can be used to evaluate the comparability of the generated items. Despite the feasibility and potential usefulness of using the AIG methodology to generate test items, the semantic cohesiveness of the generated items must also be evaluated. We illustrated how item quality can be evaluated using information-based approaches so that the semantic relatedness among the generated items can be determined.

AIG requires that the generated items are equivalent on the cognitive requirements but at the same time should appear to be different items to the test takers. The purpose of this study was to present an improved version of CDSM to represent test-items by composing different words vectors and then use these word representation vectors to compose sentence representation vectors and thus, compute the semantic similarity between item pairs using the semantic space. Semantics relatedness in the context of this study has a rather broad meaning. By "semantics relatedness" we refer to the whole test-item (stem + options) meaning that a word can be interpreted by its context and the meaning of a sentence can be derived from its compositions.

Although the first improvement we proposed in the CDSM (*i.e.*, adding distance weight) turned out to be unsatisfactory, the other two improvements we introduced weighted bigram multiplicative model and TF-IDF score produced optimum

result for the learned semantic space. Our unsupervised approach could be used to quantify the semantic heterogeneity among the automatically generated items thereby providing a more intuitive representation about the quality of generated item pools. The item models used in this study were found to be visually distinct and moderately related on meaning with high semantic variability, which suggests near ideal generative capacity of the medical item models. These findings embodied quantitative and cognitive indicator of distinctiveness for the generated items. Both measures, ISH and CSI, provided methods for concretely measuring distinctiveness and could be informative and helpful for several item generation and item bank management tasks.

For example, these NLP measures could be used as a feedback mechanism to the test developers for enhancing the elements in item model in order to produce more diverse items from an item model. These measures could also be used as a method for choosing items at the time of test assembly process to ensure that the final test form contains a diverse sample of items that measure different outcomes. Administering test on computers also facilitates delivery of individualized items by providing different difficulty levels, content emphases, and possibility of immediate feedback to the test takers.[37] Excluding the items with overlapping content and answer curing (enemies[38]) is another area in which these measures could be used to enhance the item delivery and development practices because as the item bank grows the human ability is restricted to compare and detected the enemies. The other promising application is statistical precalibration of the generated items,[39] that uses the concept of item families to develop a statistical model for calibrating siblings (*i.e.*, generated items) based on the commonality within generated items. In the absence of empirical measures of commonality, the sibling membership must be established more subjectively using judgements and ratings from test development specialists. Finally, comparing the items in active item bank against those on the shared public domain (*e.g.*, learning websites, materials from coaching school, *etc.*) can identify compromised items, and the investigation can proceed accordingly for the stolen items.[30]

Most importantly, all of these possible applications will be based on semantic vectors of numbers which would not requires access to the actual item bank there by providing an additional security that is seldom possible in costly conventional methods, which requires the content specialist to initially create and then transform each individual item as it moves through the creation, editing, and revision process. Given the high cost of item development, the proposed empirical methods for reviewing and identifying commonality within large item banks will help focus resources on unique items rather than on item editing and revision.[36]

Nevertheless, the semantic space constructed in this study uses a corpora from a relatively small item bank. We expect that if the semantic space had been constructed using a large sample of operational test items, we would have access to more co-occurrence information thereby leading to the construction of word vectors with close to true semantic meanings. This study had proposed two methods that could be used as an evidence to describe the similarity among the generated items. However, more research is needed to establish the gold standard to classify automatically generated items based on their measure of distinctiveness. Hence, one possible future direction of this research could be to develop an effect size measure with interpretative criteria[40] for classifying the generated items as isomorphic or variants and to evaluate the effectiveness of proposed distinctiveness measures in other content domains. Finally, AIG is not free from human subjectivity because content experts must still make decisions about which aspect of assessment task will contribute in the item difficulty, discrimination and semantic relevance. It is expected that human subjectivity will be reduced by incorporating more systematic approach to item development such that large number of items can be generated effectively to populate item bank for satisfying the demands of varying assessment criteria.

REFERENCES

[1] Lai H, Gierl MJ. Generating Items Under the Assessment Engineering Framework. In: Gierl MJ, Haladyna TM, editors. Automatic item generation Theory and Practice. New York, NY: Taylor & Francis; 2013. p. 77-101.

[2] Bennett RE. The Changing Nature of Educational Assessment. Rev Educ Res. 2015; 39(1): 370-407. http://dx.doi.org/10.3102/0091732X14554179

[3] Bejar II. Generative response modeling: Leveraging the computer as a test delivery medium (ETS Research Report). Princeton, NJ: Educational Testing Service; 1996. Report No.: 96-13.

[4] Bejar II. Generative testing: From conception to implementation. In: Irvine SH, Kyllonen PC, editors. Item generation for test development. Mahwah, NJ: Erlbaum; 2002. p. 199-217.

[5] Bejar II, Lawless R, Morley ME, et al. A feasibility study of on-the-fly item generation in adaptive testing. Journal of Technology Learning and Assessment. 2003; 2(3).

[6] Gierl MJ, Lai H, Breithaupt K. Methods for Creating and Evaluating the Item Model Structure Used In Automatic Item Generation.

Paper presented at: the annual meeting of the National Council on Measurement in Education; 2012 April; Vancouver, BC, Canada.

[7] Dennis I, Handley S, Bradon P, et al. Approaches to modeling item-generative tests. In: Irvine SH, Kyllonen PC, editors. Item generation for test development. Mahwah, NJ: Erlbaum; 2002. p. 53-72.

[8] Mitkov R, Ha LA, Varga A, et al. Semantic similarity of distractors in multiple-choice tests: extrinsic evaluation. Proceedings of the Workshop on Geometrical Models of Natural Language Semantic, Association for Computational Linguistics; 2009 March; 49-56. Available from: http://www.aclweb.org/anthology/W09-0207

[9] Bejar II. Item Generation: Implications for a Validity Argument. In: Gierl MJ, Haladyna TM, editors. Automatic item generation Theory and Practice. New York, NY: Taylor & Francis; 2013. p. 40-55.

[10] Alsubait T, Parsia B, Sattler U. A similarity-based theory of controlling mcq difficulty. Proceedings of IEEE Second International Conference on e-Learning and e-Technologies in Education (ICEEE). IEEE. 2013 September: 283-8. http://dx.doi.org/10.1109/icelete.2013.6644389

[11] Hwang GJ, Chu HC, Yin PY, et al. An innovative parallel test-sheet composition approach to meet multiple assessment criteria for national tests. Comput Educ. 2008; 51(1): 1058-72. http://dx.doi.org/10.1016/j.compedu.2007.10.006

[12] Yin PY, Chang KC, Hwang GJ, et al. A particle swarm optimization approach to composing serial test sheets for multiple assessment criteria. J Educ Techno Soc. 2006; 9(3): 3-15.

[13] Haladyna TM. Automatic Item Generation: A Historical Perspective. In: Gierl MJ, Haladyna TM, editors. Automatic item generation Theory and Practice. New York, NY: Taylor & Francis; 2013. p. 13-25.

[14] Drasgow F, Luecht RM, Bennett R. Technology and testing. In: Brennan RL, editor. Educational Measurement. 4th ed. Washington, DC: American Council on Education, 2006. p. 471-516.

[15] Gierl MJ, Lai H. Using Weak and Strong Theory to Create Item Models for Automatic Item Generation: Some Practical Guidelines with Examples. In: Gierl MJ, Haladyna TM, editors. Automatic item generation Theory and Practice. New York, NY: Taylor & Francis; 2013. p. 26-39.

[16] Leighton JP, Gierl MJ. The learning sciences in educational assessment: The role of cognitive models. Cambridge, UK: Cambridge University Press; 2011. http://dx.doi.org/10.1017/CBO9780511996276

[17] Mitchell J, Lapata M. Composition in distributional models of semantics. Cogn Sci. 2010; 34(8): 1388-429. PMid:21564253. http://dx.doi.org/10.1111/j.1551-6709.2010.01106.x

[18] Tai KS, Socher R, Manning CD. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. arXiv preprint arXiv:1503.00075v3, 2015. Available from: http://arxiv.org/pdf/1503.00075v3.pdf

[19] Sloman SA, Rips LJ. Similarity as an explanatory construct. Cognition. 1998; 65: 87-101. http://dx.doi.org/10.1016/S0010-0277(97)00048-6

[20] Lund K, Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence. Behav Res Methods Instrum Comput. 1996; 28(2): 203-8. http://dx.doi.org/10.3758/BF03204766

[21] Duffy SA, Henderson JM, Morris RK. Semantic facilitation of lexical access during sentence processing. J Exp Psychol Learn Mem Cogn. 1989; 15(5): 791–801. PMid:2528603. http://dx.doi.org/10.1037/0278-7393.15.5.791

[22] Foltz PW, Kintsch W, Landauer TK. The measurement of textual coherence with latent semantic analysis. Discourse Process. 1998; 25(2-3): 285-307. http://dx.doi.org/10.1080/01638539809545029

[23] Landauer TK, Dumais ST. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychol Rev. 1997; 104(2): 211-40. http://dx.doi.org/10.1037/0033-295X.104.2.211

[24] Marelli M, Bentivogli L, Baroni M, et al. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. Proceedings of the SemEval 2014: International Workshop on Semantic Evaluation. Dublin, Ireland: Association for Computational Linguistics. 2014: 1-8. PMid:24275290. http://clic.cimec.unitn.it/marco/publications/marelli-etal-semeval14-task1.pdf

[25] Mitchell J, Lapata M. Vector-based models of semantic composition. Proceedings of ACL-08: HLT. Columbus, Ohio: Association for Computational Linguistics. 2008: 236-44. Available from: http://www.aclweb.org/anthology/P08-1028

[26] Baroni M, Zamparelli R. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2010 October: 1183-93. Available from: https://www.aclweb.org/anthology/D/D10/D10-1115.pdf

[27] Van Urk C, Richards N. Two components of long-distance extraction: Successive cyclicity in Dinka. Linguist Inq. 2015; 40(1): 113-55. http://dx.doi.org/10.1162/LING_a_00177

[28] Leskovec J, Rajaraman A, Ullman JD. Mining of massive datasets. 2nd ed. Cambridge, UK: Cambridge University Press; 2014. PMid:25327001. http://dx.doi.org/10.1017/CBO9781139924801

[29] Bayardo JR, Ma Y, Srikant R. Scaling Up All Pairs Similarity Search. Proceedings of the 16th International Conference on World Wide Web, Banff-Alberta, Canada. 2007: 131-40. Available from: http://www2007.cpsc.ucalgary.ca/papers/paper342.pdf

[30] Becker K, Kao S. Finding stolen items and improving item banks. Paper presented at the annual meeting of the American Educational Research Council; 2009 April: 14-7; San Diego, CA.

[31] Spertus E, Sahami M, Buyukkokten O. Evaluating Similarity Measures: A Large Scale Study in the Orkut Social Network. Proceedings of the 11th ACM-SIGKDD International Conference on Knowledge Discovery in Data Mining. 2005: 678-84. http://dx.doi.org/10.1145/1081870.1081956

[32] Gierl MJ, Zhou J, Alves C. Developing a Taxonomy of Item Model Types to Promote Assessment Engineering. Journal of Technology, Learning, and Assessment. 2008; 7(2). Available from: http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1629

[33] Marelli M, Menini S, Baroni M, et al. A SICK cure for the evaluation of compositional distributional semantic models. Proceedings of International Conference on Language Resources and Evaluation (LREC-2014). Reykjavik, Iceland: European Language Resources Association. 2014: 216-23. Available from: http://clic.cimec.unitn.it/marco/publications/marelli-etal-sick-lrec2014.pdf

[34] Bach F. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. J Mach Learn Res. 2014; 15(1): 595-627. Available from: http://jmlr.org/papers/volume15/bach14a/bach14a.pdf

[35] Porter MF. An algorithm for suffix stripping. Program. 1980; 14(3): 130-7. http://dx.doi.org/10.1108/eb046814

[36] Gierl MJ, Latifi S, Lai H, et al. Using Automated Procedures to Generate Test Items That Measure Junior High Science Achievement. In: Rosen Y, Ferrara S, Mosharraf M, editors. Handbook of Research on Technology Tools for Real-World Skill Development. Hershey,

PA: IGI Global; 2016. p. 590-610. http://dx.doi.org/10.4018/978-1-4666-9441-5.ch022

[37] Chou C. Constructing a computer-assisted testing and evaluation system on the World Wide Web-the CATES experience. IEEE Trans Ed. 2000; 43(3): 266-72. http://dx.doi.org/10.1109/13.865199

[38] Van der Linden WJ. Linear models for optimal test design. New York, NY: Springer Science & Business Media; 2005. http://dx.doi.org/10.1007/0-387-29054-0

[39] Sinharay S, Johnson MS, Williamson DM. Calibrating item families and summarizing the results using family expected response functions. J Educ Behav Stat. 2003; 28(4): 295-313. http://dx.doi.org/10.3102/10769986028004295

[40] Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. New Jersey: Lawrence Erlbaum; 1988.