4

Assessment of Clinical Skills

A Case Study in Constructing an NLP-Based Scoring System for Patient Notes

Polina Harik, Janet Mee, Christopher Runyon, and Brian E. Clauser

We begin this chapter with an overview of the evolution of automated scoring of text since Page's early work more than half a century ago (Page, 1966). As that review makes clear, NLP-based scoring systems have made it possible to augment or replace human ratings in scoring written responses (e.g., Burstein et al., 2001; Landauer et al., 2000; Monaghan & Bridgeman, 2005). In many cases this research focuses on the usefulness of these computer-based systems for approximating human scores. Details of the scoring algorithms are often viewed as secondary or, in some cases, they have been intentionally withheld as proprietary intellectual property. Even when details are provided, it is rare that researchers describe how each component – or module – within the system contributes to accuracy. In this chapter, in addition to describing our system, we provide a detailed evaluation focusing on the incremental improvement in accuracy associated with each component of the system. This type of evaluation is commonly referred to as ablation study. Results of this evaluation may have important practical implications for researchers developing similar applications for use in other assessments.

1. Automated Scoring of Written Responses

The history of automated scoring for written responses dates back at least to Page's work in the 1960s (Page, 1966, 1967). A complete historical overview is well beyond the scope of this chapter; instead, we will focus on three trends that have been evident since Page's work was published:

1. Early work on automated scoring of written responses focused on scoring form rather than content. Over time, it has been recognized that in many contexts, it is critical to score content.
2. Typically, the early systems used measures that might be described as surrogates or proxies rather than direct measures of the quality of the written response. Again, in many assessment contexts, there are advantages to replacing these surrogates with more direct measures of the quality of the response.

DOI: 10.4324/9781003278658-5

3. These early systems also typically were designed to model or approximate human ratings. It has become apparent that in some contexts, it is advantageous to model what content experts say raters *should* do rather than approximating what they *actually* do (Bennett & Bejar, 1998; Margolis & Clauser, 2020).

From a historical perspective, these changes in approach for automated scoring appear to be trends, but they might better be seen as variations in how automated scoring is implemented. In some contexts, scoring should focus on form, in others on content. In some contexts, surrogates will be both efficient and sufficient; in others, more direct measures will be needed. In some contexts, modeling what human raters do will be appropriate; in others, modeling what raters should do will be preferred. In what follows, we consider each of these trends or variations. Additionally, we place the scoring procedure we developed in the context of other recent state-of-the-art efforts.

1.1 Form vs. Content

Page's work focused on the form of the essay rather than the content (Page, 1966, 1967). A test-taker may well have received an excellent score even if the topic of the scored essay had little discernible relationship to the assigned topic. This was typical of many early efforts. Over time, researchers have tackled the problem of scoring that includes – or is entirely motivated by – the appropriateness of the content.

The historical development of automated scoring systems does not reflect an unequivocal trend away from scoring form to scoring content; in many instances (e.g., K-12 essays), form is central to the proficiency the assessment is intended to measure. There has, however, been a recognition that scoring written responses based solely on the form of the writing may be unacceptably limited. The testing context that we focus on in this chapter is an extreme example: the test-takers are asked to document the important information that they collected in interviewing a patient and completing a physical examination. Scoring is based almost exclusively on the presence (or absence) of critical information (referred to as *key essentials* or *key features*). Provided the writing is understandable, issues of form – such as complete sentences and standard punctuation – are relatively unimportant. In another assessment setting, for example, a middle school student might be asked to write an essay describing the causes of the First World War. That assessment might be scored based on both the presence of relevant and accurate historical information and on the use of appropriate grammar and structure. Finally, in some instances, an essay may be scored (almost) exclusively based on form, but an evaluation of content may be necessary to verify that the essay is 'on topic'.[1]

Two major approaches to scoring content have been developed. The first of these is commonly referred to as latent semantic analysis (Landauer et al., 1998). It provides a general measure of the extent to which the words in the response are similar to those in previously scored responses. The second approach explicitly attempts to match strings of words in the response to concepts delineated in the key.

With latent semantic analysis (and variations on this theme), a large corpus of relevant text is identified. In the case of scoring for a specialized content area – such as medicine – the corpus must be matched to the content area because it is essential that the vocabulary in the responses is included in the corpus. The corpus is then analyzed to produce a multidimensional semantic space in which every word (and document) in the corpus can be represented by a vector of numbers. The corpus will typically include hundreds of thousands of paragraphs of text, and the semantic space will have hundreds of dimensions. A set of previously scored essays can then be located in this semantic space. The associated vectors for new essays can then be

compared to those of previously scored essays by using a similarity metric such as the cosine of the angle between the new essay and each previously scored essay. The score for the most similar essay, or the average score for some set of similar essays, can be assigned to the new essay.

Latent semantic analysis has been demonstrated to be useful in a variety of settings. This approach has been incorporated into the topic analysis module in e-rater, the essay scoring system developed by the Educational Testing Service (Burstein et al., 2001). E-rater is designed for use in contexts where content contributes to an overall score but is not the driving focus of the assessment. Swygert et al. (2003) also demonstrated the usefulness of this procedure for scoring written summaries produced by medical students after interacting with standardized patients. Latent semantic analysis has also been incorporated into a system intended to provide writing instruction (Streeter et al., 2011) and has been used in both high-stakes (Shermis, 2014) and low-stakes (Landauer et al., 2000) writing assessments. Latent semantic analysis has generally been less successful in scoring shorter written responses where human raters identify highly specific concepts that can be matched to an answer key (LaVoie et al., 2020; Willis, 2015).

This limitation of latent semantic analysis was one of the motivations for developing alternative procedures intended to match concepts more directly in the key to text in a response. A range of related approaches exists. At the conceptually simple end of this range is an approach described by Yamamoto et al. (2017), which was designed for multilingual assessments. Their procedure is based on exact matches to a key or dictionary and works on the assumption that within a sample of examinee responses, the number of unique responses will be substantially smaller than the total number of responses. With this approach, any time content experts agree on the score to a specific response, that response will be added to the key (or dictionary), and the associated score will be assigned to all other responses that exactly match the scored response. Yamamoto et al. showed that this procedure is suitable for international assessments that include many languages, but relatively small sample sizes per language, and can substantially improve the efficiency of scoring for PISA administrations.

More complex approaches that make greater use of NLP technology include c-rater – also developed by Educational Testing Service (Leacock & Chodorow, 2003; Liu et al., 2014; Sukkarieh & Blackmore, 2009) – and INCITE, the procedure that is the focus of this chapter (Sarker et al., 2019). Again, the approaches used in these systems are designed to identify specific scorable concepts in the response. To implement these approaches, content experts create a key that contains the scorable concepts. Typically, content experts also review and annotate a sample of test-taker responses. The annotation identifies the words or phrases in the response that the annotator believes reflect the scorable concept. Then, using a variety of NLP techniques, the system is constructed to identify the concepts in future responses that may or may not exactly match previously scored responses.

Another example of the general approach of matching text to a key was presented by Willis (2015). With that system, content experts score a sample of responses; NLP technology is then used to develop rules that match the pattern of correct/incorrect judgments produced by the judges. The human judges can then edit the rules. The rules reflect the presence of specific terms in the response and the relative position of those specific terms. The system works iteratively so that human judgments are used to develop scoring rules; the rules can then be used to score additional responses, and humans can be included in the process to score response patterns not seen previously. These new judgments are then used to create additional rules. Numerous variations on the general approach of matching to a key exist (e.g., Cook et al., 2012; Jani et al., 2020).

1.2 Use of Surrogates in Scoring

Longer essays that include longer sentences, more sophisticated vocabulary, and more complex punctuation (e.g., semicolons) are not inherently better essays, but the presence of these

characteristics tends to correlate positively with scores. Early automated scoring procedures took advantage of these relationships. The demonstration that it was possible to use an automated system to produce scores that correlated well with human ratings represented an important breakthrough. And if all that is needed is an efficient means of providing an appropriate rank ordering of a set of essays, Page's early approach was useful as well (Page, 1966). There are, however, contexts in which this approach might be viewed as inadequate.

First, for formative assessment of writing, it is important to provide actionable feedback. Scores from this type of scoring procedure are likely to fall short. Telling a student to write longer essays using more sophisticated vocabulary is not likely to be particularly helpful.

Second, for tests used to make high-stakes decisions (graduation, admission, certification), scoring based on surrogates may create an opportunity for test-takers to game the system. If test-takers know that longer responses with more sophisticated vocabulary produce higher scores, they can use that knowledge to artificially inflate their scores. A short essay may be copied and pasted into the response interface multiple times; lists of sophisticated, but irrelevant, vocabulary can be memorized and inserted into the essay (Bejar, 2013; Bridgeman et al., 2012; Higgins & Heilman, 2014).

Third, scoring systems that make substantial use of indirect measures of the quality of an essay lack transparency (sometimes referred to as *traceability*). With systems that use surrogates – or other indirect measures – it is unlikely that stakeholders of the testing process will be able to understand how a specific performance resulted in an associated score.[2] Content experts are likely to be skeptical if they cannot see a relationship between what is taught and what is scored. In the case in which scores are used to make high-stakes decisions, test-takers are likely to believe that knowing how the test is scored is a prerequisite for fair testing.

Transparency was an important consideration in developing the INCITE system. The approach used by the system to match content from the response to a scorable key ensures transparency. The appropriateness of the key can be questioned, and the reliability of the matching process must be empirically evaluated, but the process itself is open to transparent evaluation.

1.3 Modeling Human Scores

Related to the use of correlation as the basis for scoring is the explicit intention to model – or predict – human scores. In some sense human scores are an obvious criterion, given that the automated system is often developed to replace human scoring. At the same time, it must be recognized that human scores are often unreliable. Without constant monitoring and feedback, humans tend to either systematically diverge from identified criteria or to apply criteria inconsistently. Cianciolo et al. (2021) reported on the development of an automated system to score diagnostic justification essays written by medical students. They note, 'Faculty ratings were insufficiently reliable for training machine scoring algorithms, so trained research assistants were employed to re-rate the essays using a more rigorous process' (p. 1027). In this case, the original faculty ratings had inter-rater reliabilities ranging from .13 to .33.

The example provided in the previous paragraph makes it clear that human ratings have practical limitations as a criterion either for modeling automated scores or for evaluating the quality of such scores. Although some researchers (e.g., Cianciolo et al., 2021) have continued to attempt to improve the quality of the ratings, the trend in educational measurement has been to recognize that these ratings may be both practically and theoretically limited. If, for example, content experts agree that an optimal justification for a specific diagnosis would cite four specific patient characteristics (identified through the history and physical examination) and include no additional inaccurate or irrelevant information, the criterion for evaluating the automated scoring system might be based on the accuracy of identifying these scorable features

within a set of essays. This approach requires a more detailed description and justification of the scoring criteria and explicitly excludes vaguely defined expert judgment, but it brings the automated scoring process into line with more contemporary principled approaches to test construction such as evidence-centered design (Mislevy et al., 2006).

In the next section we describe the context in which the system was to be deployed operationally. We then provide a description and evaluation of the system.

1.4 Context

Between 2004 and 2020, the USMLE Step 2 Clinical Skills Examination (Step 2 CS) was part of the sequence of assessments required for allopathic medical practice in the United States.[3] This live simulation was administered to approximately 30,000 candidates each year at five locations across the United States. The examination was designed to measure patient-centered clinical skills. For each administration, examinees rotated through a sequence of twelve 25-minute encounters with actors trained to play patients with specific medical problems. Examinees had up to 15 minutes to interact with the standardized patient: taking a focused history, performing a physical examination, and discussing their findings with the patient. During the remaining 10 minutes, the examinee documented the encounter in a *patient note*. These patient notes consisted of two sections: (1) the *data gathering* section required test-takers to document the pertinent findings from the patient history and the physical examination; (2) in the *data interpretation* (DI) section, test-takers were instructed to produce an ordered list of up to three potential diagnoses, provide pertinent evidence from the data gathering section to support each diagnosis, and identify initial diagnostic studies that would be warranted. Examinee patient notes were assigned to physicians trained to rate the notes using case-specific algorithms that mapped patterns of performance onto a rating scale for both the data gathering and data interpretation subcomponents.

In order to pass the examination, it was necessary to receive a passing score on each of three separate components: *the Integrated Clinical Encounter*, *Spoken English Proficiency*, and *Communication and Interpersonal Skills*. The latter two scores were provided by the standardized patients. The *Integrated Clinical Encounter* score consisted primarily of the physician ratings of the data gathering and data interpretation subcomponents.[4]

The motivation for developing an automated scoring system for this examination was much the same as that for other assessments – to reduce the cost of scoring and improve reliability. The specifics of the plan for implementing this system were, however, somewhat different from those for most other large-scale tests. Because the examination was scored pass/fail and no numeric score was reported, minor differences between the scores produced by trained physicians and those produced by the automated system could be ignored for test-takers whose proficiency level was far above the cut score. This allowed for a scoring approach in which all test-takers could be scored by the automated system. Test-takers receiving scores well above the cut score would have a *pass* decision reported based on the computer-generated score. All other test-takers would then be re-scored by physician raters. Preliminary results indicated that this would cut the number of required human ratings by nearly 50% without impacting classification accuracy, which translated into eliminating the need for approximately 200,000 human ratings per year. This would result in substantial savings in time and money.

As we noted, the examination was in place from 2004 to 2020. USMLE Step 2 CS was discontinued during the COVID-19 pandemic because it was unsafe for test-takers to travel to the test sites, and because close interaction between the standardized patients and the test-takers represented a risk to both groups. The rollout of the automated scoring system was scheduled to begin within days of when the examination was terminated, so the system was never used

operationally. Nonetheless, the development of the system was completed, and as part of the development process, we evaluated how different components of the system contributed to the usefulness of the system for correctly identifying targeted concepts in the written responses. The remainder of this chapter describes the system and reports on our evaluation of the accuracy of the system for the data gathering section of the patient note.

2. The INCITE System

INCITE is a system designed to identify scorable concepts described as key essentials. The individual concepts can be expressed in many ways depending on the test-taker's choice of words. The number of variations can also grow substantially because identifiable misspellings are also considered correct responses. The system gives priority to high precision over high recall – that is, it gives priority to correctly identifying true matches (minimizing false-positive decisions) over maximizing the number of concepts identified. The system was designed in this way because of the operational context – as noted previously, it was developed for use in high-stakes testing and our intention was to use computer-based scores only for those test-takers with a proficiency level well above the cut score. Test-takers with INCITE-based scores at or below the cut score would be re-scored by human raters. This meant that giving credit for a concept that was absent from the response was a more serious error than failing to give credit for a concept that was present in the response.

The INCITE system attempts to match each key essential to the text of an individual note sequentially until a match is made or until the sequence is completed without a match. The processing sequence includes the following conceptual steps.

2.1 Preprocessing

The preprocessing step removes unnecessary characters and converts all text to lowercase. The common steps of stemming and stop-word removal[5] are not performed at this stage because the system relies on exact text matches.

2.2 Annotation

The annotation process was intended to identify specific strings of words in the patient notes that could be mapped to the key essentials. Two annotators annotated each case. They began by annotating three 'training' notes. For each of these notes they discussed what did and did not count as a match. For example, they considered whether 'abdominal pain for some time' would be considered a sufficient representation of 'LLQ abdominal pain x 2 weeks'.[6] Once the common annotation rules were established, the annotators were each given 22 notes; 12 of these were unique to each annotator. Five of the notes annotated in common by both annotators were used to cross-validate the results produced by INCITE and will be discussed in the evaluation section.

The annotations consisted of strings of text found in patient notes that conceptually matched case-specific key essentials. They included a wide range of lexical representations of each key essential: synonyms, misspellings, medical abbreviations, and alternative expressions. The example in Figure 4.1 shows a section of a patient note in which the concepts 'Relief with Pain meds' is documented as 'pain which improves with ice and pain medication' and 'No drug allergies' is recorded as 'NKDA'.

The annotated notes were used to develop the model used for identifying key essentials in the text. As this process proceeded, it became clear that strings of words identified by one

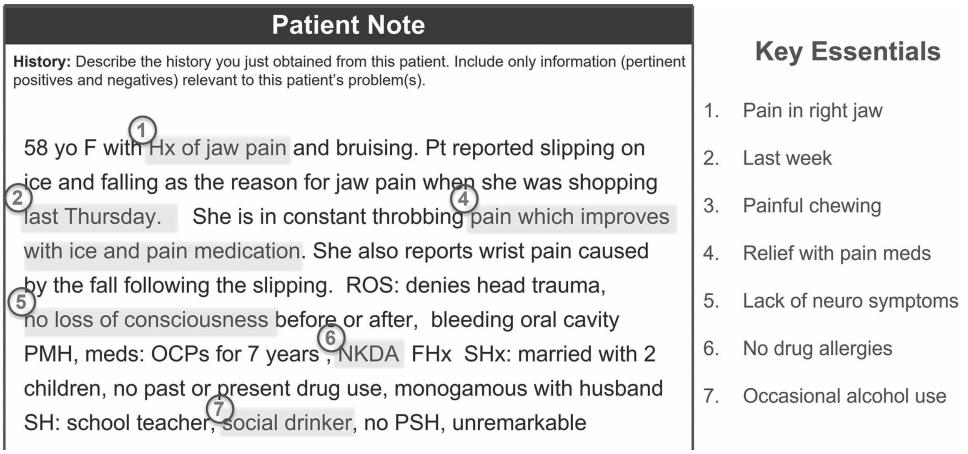

Figure 4.1 An example of mapping key essentials concepts to patient note text.

annotator did not consistently agree with the decisions made by the other annotator. To minimize the impact of these inconsistencies, the annotators reconciled all instances in which two annotators matched the same string of words to different key essential concepts. This was typically the result of a clerical error on the part of one of the annotators.

2.3 Exact Matching

The initial matching step includes a search for exact matches to the key essentials as well as matches to variations on the key essentials (different wording and misspellings) represented in several dictionaries. The first dictionary, referred to as the *global dictionary*, was developed without the use of case-specific annotations. The second dictionary – *dictionary-A* – was compiled from the annotations for notes annotated by the two annotators. The third dictionary – *dictionary-B* – included augmentation of the annotations provided by staff involved in fine-tuning the system in addition to the information in dictionary-A. These augmentations were created by combining information from individual annotations. For example, if the annotators had identified the following phrases as representing a specific key essential – 'acetaminophen helps reduce pain' and 'pain is controlled with Tylenol', the augmentation would be: 'Tylenol helps reduce pain' and 'pain is controlled with acetaminophen'.

2.4 Fuzzy Similarity and Dynamic Thresholding

The number of potential variations on each key essential far exceeded the number of variants contained in the dictionaries, so a fuzzy similarity matching module was included in the system. The fuzzy matching used a sliding window to evaluate strings of words. The size of the window depended on the length of the key essential. For each key essential, four windows were used: (1) a window one word shorter than the length of the key essential; (2) a window equal to the length of the key essential; (3) a window one word longer than the key essential; and (4) a window two words longer than the key essential (e.g., a key essential with four words would be evaluated with windows of three to six words).

To evaluate the similarity between the key essentials and the string in the window, we used the Levenshtein Ratio Method. With this approach, the distance between two strings is represented by the number of deletions, insertions, or substitutions that are required to transform

one string to the other. The Levenshtein Ratio equals the sum of the length of the two strings minus the distance between the strings, divided by the sum of the length of the strings:

$$Levenshtein\ Ratio = \frac{Length - Distance}{Length}.$$

If the two strings match exactly, the ratio equals one. As the distance between the strings increases, the ratio approaches zero.

Once the Levenshtein Ratio is calculated, a string of words can be classified as a match if the ratio exceeds an identified threshold. Evaluation of the classification accuracy for varying key essentials at different thresholds suggested that a fixed threshold would be suboptimal because the key essentials significantly varied in length. For shorter key essentials, a very high threshold is needed to ensure that, for example, the term 'contusion' is not matched to the very different concept 'concussion'. Decreasing the threshold for shorter key essentials would result in a large number of false positive matches. On the other hand, a high threshold would result in a large number of false negatives for a longer key essential, such as 'traveled abroad two weeks ago', where 'traveled to Kenya a few weeks back' and 'international travel 2–3 weeks ago' are matches that can be detected only with a threshold around 0.6. Experimentation on a pilot set of cases led us to select an approach that uses a dynamic threshold where longer key essential entries have a proportionally lower threshold, compared to shorter entries. The dynamic threshold was defined as

$$DT = T_i - \frac{k \times Length}{100}$$

where T_i is an initially set static threshold, Length is the length of the sliding window, and k is an index that determines the magnitude of the threshold change (for a detailed description, see Sarker et al., 2019).

2.5 Set Overlap and Intersection

Fuzzy matching is effective for identifying many of the variants of the key essentials that are not already included in the dictionaries. However, the approach is ineffective if the text in the note and the key essential use a substantially different word order. Consider the phrase 'Antibiotics taken in recent times for his symptoms – negative'. This clearly captures the same concept as the key essential 'Negative for recent antibiotics', but it would not be identified with the INCITE fuzzy-matching algorithm. To allow for matching under this scenario, a wider search window is employed with a bag-of-words approach.[7] The system searches for strings of words that overlap or intersect with the words in the key essential or alternative versions of the key essential that appear in the dictionaries. To account for misspellings, fuzzy matching with a high threshold is also incorporated in the matching used for the bag-of-words approach.

3. Evaluation of the INCITE System

In what follows, we provide analyses that report how different parts of the system impact the accuracy of identifying key essential concepts in the patient note text. We first examine the usefulness of exact matching of text to the key essentials without and then with the various dictionaries. We then present related results that show the change in the performance of the system when the fuzzy-matching and bag-of-words approaches are incorporated into the system. The primary metric used for reporting these results is the F1 score. This metric is a commonly used index of accuracy in machine learning (Han et al., 2012). It represents the harmonic mean of the precision and recall.

$$\mathrm{F1} = \frac{2}{\frac{1}{Recall} \times \frac{1}{Precision}}.$$

The value can also be expressed as

$$\mathrm{F1} = \frac{True\,Positive}{True\,Positive + .5\left(False\,Positive + False\,Negative\right)}.$$

The F1 score takes on a value of one when the proportion of non-identified (false negatives) and wrongly identified (false positives) concepts are both zero, and a value of zero when no concepts are identified correctly.

3.1 Data

The dataset used for both developing the system and for the subsequent evaluation included patient notes written by examinees who took the Step 2 CS Examination between September 2017 and August 2019. For this study, a curated set of 18 cases representing the range of the exam's content was selected. For each case, a committee of physicians defined the essential concepts (key essentials) expected in a patient note produced by a competent physician. These key essentials reflected information about the patient that should be collected as part of a focused history and physical examination. The number of key essentials per case ranged from 10 to 20. The results presented in the next section are based on the five notes that were annotated by two annotators. These five notes were intended for independent cross-validation and so were not used in developing the system.

3.2 Results

Table 4.1 presents the F1 scores for identifying key essential concepts using exact matching for each of the 18 cases. For the results in this table, a concept was considered present if *either* of the annotators identified it in the note. The second column shows the F1 scores when the matching was implemented using only the key essential definition. The third column shows the same score when the global dictionary is added. The fourth and fifth columns present results associated with adding dictionary-A and dictionary-B, respectively.

Not surprisingly, the F1 scores were relatively low when only the key essential definition was used. Adding each of the three dictionaries improved the scores to a varying degree. Adding the global dictionary increases the mean F1 score by approximately .12. Dictionary-A, which reflects the results of annotation, additionally increases the mean F1 score by .31. Including the augmented annotations in dictionary-B additionally increases the F1 scores by approximately .09. Table 4.2 provides analogous information to that in Table 4.1, but for these results, a key essential was considered present in the note only if it was identified by *both* annotators. The results are similar to those in Table 4.1, reflecting similar incremental improvements as the dictionaries are added (.13, .30, and .08).

Table 4.3 presents F1 scores for the full INCITE system using dictionaries A or B (unlike the results reported in Tables 4.1 and 4.2, these results include exact matches as well as those produced by the fuzzy-matching and bag-of-words procedures). The inclusion of the fuzzy-matching and bag-of-words procedures substantially improve the performance of the system. For dictionary-A, the improvement results in an increase in the mean F1 score of between .19 and .20, depending on whether the criterion was identification of the key essential by annotator 1 or 2 or both annotators 1 and 2. With dictionary-B (which includes the augmented annotations), adding the additional matching procedures increases the mean F1 scores by between

Table 4.1 System Performance (F1 Scores) Against Combined Annotations, Exact Matching Only

	F1 Scores Exact Matching Only			
Case	KEs Only	KEs + Global Dictionaries	KEs, Global Dictionaries, Dictionary-A	KEs, Global Dictionaries, Dictionary-B
1	0.44	0.59	0.75	0.80
2	0.27	0.36	0.71	0.75
3	0.32	0.34	0.68	0.78
4	0.34	0.51	0.65	0.68
5	0.20	0.39	0.74	0.74
6	0.20	0.36	0.69	0.83
7	0.00	0.18	0.55	0.55
8	0.11	0.11	0.48	0.73
9	0.18	0.28	0.53	0.73
10	0.36	0.56	0.72	0.79
11	0.23	0.36	0.77	0.86
12	0.32	0.55	0.68	0.74
13	0.13	0.13	0.67	0.85
14	0.19	0.35	0.80	0.87
15	0.42	0.45	0.82	0.91
16	0.24	0.45	0.68	0.68
17	0.33	0.36	0.68	0.78
18	0.23	0.34	0.70	0.77
Mean	0.25	0.37	0.68	0.77
SD	0.11	0.14	0.09	0.08

Table 4.2 System Performance (F1 Scores) Against Matching Annotations, Exact Matching Only

	F1 Scores Exact Matching Only			
Case	KEs Only	KEs + Global Dictionaries	KEs, Global Dictionaries, Dictionary-A	KEs, Global Dictionaries, Dictionary-B
1	0.44	0.59	0.75	0.80
2	0.27	0.37	0.73	0.77
3	0.32	0.35	0.68	0.78
4	0.36	0.53	0.68	0.71
5	0.21	0.41	0.75	0.75
6	0.21	0.38	0.71	0.81
7	0.00	0.19	0.57	0.57
8	0.12	0.12	0.52	0.75
9	0.20	0.31	0.52	0.70
10	0.39	0.61	0.76	0.79
11	0.25	0.35	0.79	0.84
12	0.33	0.57	0.67	0.71
13	0.13	0.13	0.67	0.84
14	0.20	0.37	0.79	0.84

(*Continued*)

Table 4.2 System Performance (F1 Scores) Against Matching Annotations, Exact Matching Only *(Continued)*

	F1 Scores Exact Matching Only			
Case	KEs Only	KEs + Global Dictionaries	KEs, Global Dictionaries, Dictionary-A	KEs, Global Dictionaries, Dictionary-B
15	0.44	0.47	0.82	0.87
16	0.25	0.47	0.71	0.71
17	0.36	0.39	0.68	0.78
18	0.25	0.36	0.69	0.76
Mean	0.26	0.39	0.69	0.77
SD	0.12	0.14	0.09	0.07

Table 4.3 System Performance (F1 Scores) for Exact, Fuzzy and Bag-of-Words Matching

	F1-Score			
Case	A1OrA2 and INCITE-A	A1AndA2 and INCITE-A	A1OrA2 and INCITE-B	A1AndA2 and INCITE-B
1	0.89	0.89	0.91	0.91
2	0.86	0.85	0.86	0.85
3	0.87	0.88	0.90	0.91
4	0.94	0.97	0.94	0.97
5	0.88	0.88	0.88	0.88
6	0.88	0.91	0.95	0.94
7	0.89	0.87	0.89	0.87
8	0.89	0.88	0.96	0.93
9	0.86	0.85	0.91	0.87
10	0.85	0.88	0.91	0.90
11	0.88	0.86	0.94	0.89
12	0.86	0.86	0.91	0.89
13	0.83	0.82	0.96	0.93
14	0.93	0.93	0.95	0.93
15	0.89	0.89	0.95	0.92
16	0.84	0.85	0.87	0.87
17	0.86	0.83	0.91	0.88
18	0.90	0.88	0.93	0.91
Mean	0.88	0.88	0.92	0.90
SD	0.03	0.04	0.03	0.03

Table 4.4 Counts of Classifications for Cross-Validation Samples Using INCITE

	Classification			
Case	False Negative	False Positive	Ture Negative	True Positive
1	9	1	29	51
2	11	6	31	52
3	7	2	26	40

	Classification			
Case	False Negative	False Positive	Ture Negative	True Positive
4	3	1	15	31
5	11	1	31	42
6	5	1	21	58
7	6	5	19	45
8	4	0	20	46
9	7	1	15	42
10	8	2	35	50
11	5	1	14	50
12	6	2	26	41
13	2	3	17	58
14	4	2	31	53
15	2	3	20	50
16	7	7	26	45
17	6	2	34	43
18	4	3	28	50
Mean	5.9	2.4	24.3	47.1

.13 and .15. Again, the use of dictionary-B provides modestly better overall performance than dictionary-A.

To provide a more detailed evaluation of the classification accuracy for the system, Table 4.4 presents counts of false-negative, false-positive, true-negative, and true-positive classifications for the cross-validation sample for each case using the full INCITE system with classifications made by annotators *A and B* as the criterion. Consistent with our intentions in designing the system, the number of true-positive and true-negative classifications is high, and when classification errors occur, false-negative error rates (failing to identify a concept) were substantially higher than false-positive error rates (giving credit for a concept that was not present). The ratio of these errors is in excess of two to one.

4. Discussion

Numerous papers cited in this chapter have reported results showing that automated systems are capable of accurately scoring text. Depending on the context, these systems have been used in conjunction with human raters or independently. As we noted, previous (unpublished) results indicated that the INCITE system could reduce the number of human ratings required by half with no change in classification accuracy. Such information is important because the primary reason for introducing computerized scoring is to improve efficiency. Although these results provide encouragement about the usefulness of automated scoring of text-based responses, they provide little guidance for researchers hoping to develop new scoring systems. Often there is little detail about the specifics of the system; it is even rarer that information is provided about how different components of a system improve the accuracy of the scoring. The results presented in this chapter represent a step towards filling that gap.

Results of the type presented in this chapter can fill a number of needs. First, they provide information about the relative contribution of different scoring components. Introducing each component will have a cost. That cost might be in: (1) the human effort required to develop the

module (e.g., programing time); (2) the human effort required to implement the component (e.g., annotation time); or (3) the computer time required to implement the module for each response that must be scored. The results reported in this chapter reflect primarily on the second of these costs, although the third is important as well.

When we consider the human effort required to implement the INCITE system, there are two separate aspects of that effort to consider. The first of these is, to what extent is the system improved by customizing the scoring for each case? The primary effort required for this customization is the work done by the annotators. Setting aside the notes required for the cross-validation, customizing the system for each case required each of two annotators to annotate 20 notes. This is not a trivial amount of work, but it represents hours – not days – of effort for each annotator in each case. The return on this investment is represented by the increase in the F1 scores presented in the third and fourth columns of Tables 4.1 and 4.2. On average, that increase is approximately 0.30, which is substantial.

A second, separable effort involved in preparing the case-specific algorithms is represented by the augmentation step. The augmentation process requires careful review and comparison of the terms produced as part of annotation. Again, the effort per case is represented by hours of work, not days. The payoff of this effort is shown by comparing the fourth and fifth columns of Tables 4.1 and 4.2 or by comparing the INCITE-A and INCITE-B columns in Table 4.3. This improvement is meaningful, but more modest than that associated with the original annotation.

The third largely separable component of the scoring system is represented by the fuzzy-matching and bag-of-words modules. Again, these procedures substantially improve the matching accuracy: mean increases of .19 to .20 were observed when they were applied to dictionary-A and .13 to .15 when they were applied to dictionary-B. These results suggest that some of the benefit associated with the augmentation process could be achieved simply by introducing the fuzzy-matching and bag-of-words modules, without augmentation. Nonetheless, the full system including augmentation continues to outperform the system without augmentation.

Adding these NLP-based matching procedures (fuzzy matching and bag of words) clearly enhances the system. It provides this enhancement without additional human review and intervention. That said, it is certainly not without cost. In addition to the programming time, experimentation was necessary to identify optimal search windows and thresholds for both the fuzzy-matching and bag-of-words modules. Introducing these procedures also makes the system more computationally intensive.[8]

The results make it clear that each component of the system adds to the accuracy of the identification of key essentials. These same results also suggest that the benefits are not strictly additive.[9] We have already commented that a proportion of the incremental matches resulting from the augmented annotations would have been produced by instituting the fuzzy-matching and bag-of-words procedures without including the augmented annotations. In this context, it is worth examining results for individual cases. As represented in Tables 4.1 and 4.2, case 7 stands out. For this case, exact matching based on the key essentials is essentially worthless. Using the variants represented in the dictionaries similarly results in the lowest F1 scores for any of the 18 cases. However, after including the fuzzy-matching and bag-of-words procedures, the case is no longer an outlier. Cases 1 and 15 represent the opposite pattern. These cases have the highest F1 values for exact matches both without and with the various dictionaries; the F1 scores for these cases are high after including the fuzzy-matching and bag-of-words procedures in the processing, but they are no longer the highest scores. This general pattern is confirmed by examining the standard deviations (across cases) for the scores reported in Tables 4.1 through 4.3. The standard deviations for the scores in Tables 4.1 and 4.2, reflecting variability

in F1 scores for exact matching, range from 0.07 to 0.14. This indicates a moderate level of variability across cases. In Table 4.3, which reports F1 scores for the full INCITE system including the fuzzy-matching and bag-of-words procedures, the variability across cases is reduced to between 0.03 and 0.04. This suggests that these more computationally intensive procedures are, relatively speaking, more useful when the exact-matching procedures are less useful.

The results reported in this chapter reflect a reasonably high level of accuracy for the INCITE system, but the scores are not perfect. As we have already mentioned, although the INCITE system was targeted for operational use at the time the Step 2 CS examination was discontinued, the system has continued to evolve. We are currently making two enhancements to the system that we expect will result in incremental improvements in performance. The first of these will allow us to introduce unique/customized thresholds for applying the Levenshtein Ratio for each key essential. The current form of INCITE uses thresholds that are a function of the length of the key essential. This approach proved to work better than using a single fixed threshold, but it ignores the fact that some key essentials are inherently less likely to produce false-positive matches and so can be associated with a lower threshold – presumably leading to more true-positive matches. The second enhancement to the system will record the specific position in the text where the match was made. This will allow for evaluation of the specific text that resulted in each false-positive match. This type of evaluation will both support the identification of an optimal threshold for individual key essentials and will provide a basis for identifying other aspects of the system that could be modified.

With regard to identifying optimal thresholds, it is worth returning to the results reported in Table 4.4. As we noted, in the context of the intended application, priority was given to precision over recall, and the results in the table reflect this choice; the false-negative rates are more than double the false-positive rates. This suggests that it might be possible to increase the overall accuracy – as reflected in the F1 scores – by using a lower threshold.

One final issue is worth mentioning in interpreting the results presented in this chapter. Although the decisions made by the annotators have been treated as truth, those decisions are not error free. The mean F1 scores reported for the full INCITE system in Table 4.3 (labeled INCITE-B) varies from .90 to .92, depending on whether the criterion is defined by concepts identified independently by both annotators *or* by concepts identified by at least one annotator. This difference reflects the less-than-perfect agreement between the annotators. This is admittedly a small difference, but it is, nonetheless, a meaningful consideration as we attempt to improve the accuracy of the system beyond the current level.

5. Conclusion

In this chapter we have described the INCITE system, an NLP-based system for computerized scoring of patient notes. The emphasis has been on the specifics of the system and how each component contributes to overall accuracy. In interpreting these results or in adopting aspects of the system for use in another context, it is important to remember the specifics of the context in which the system was developed. First, because it is used to score patient notes, it uses a specialized vocabulary. Scoring responses that use a different vocabulary may be more or less challenging, depending on the specifics. A second consideration is that we constructed the system to support transparency. This resulted in excluding some widely used approaches to evaluating text. Additionally, our focus was limited to scoring content. This decision will certainly impact the applicability of a system like the one we described for use in other contexts. Finally, we decided to prioritize precision over recall. This decision may have impacted the overall accuracy of the system and may be inappropriate in some other settings.

Notes

1 There are a number of reasons a test-taker might choose to diverge from an assigned topic. At the extreme, this might include memorizing a well-written essay that the individual test-taker would have been unable to write. Identifying this sort of effort to game the system may require a fairly minimal evaluation of the content, but that evaluation could be critical for appropriate scoring.
2 Although systems that use surrogates are likely to lack transparency, other methods that use indirect measures of the quality of the response, such as latent semantic analysis, also have this limitation.
3 Practice for physicians with an MD degree.
4 These ratings were then combined with scores from the standardized patient that indicated whether the examinee correctly completed important components of the physical examination, referred to as the *physical exam* score.
5 Stemming and stop-word removal are common preprocessing steps for NLP systems. Stemming is a process in which words are reduced to their root or stem by eliminating suffixes. Stop words are common words in English (e.g., articles, prepositions, pronouns, conjunctions). They are typically removed because they tend to carry relatively little information that can be used in NLP.
6 This, in fact, would not be considered a match.
7 Bag of words refers to matching based on the number of words in one sample that are also found in a second sample, without regard to word order.
8 One aspect of the INCITE system was not included in our evaluation, but nonetheless warrants comment. In our description, we noted in passing that the process is sequential. The efficiency of the system – in terms of the time required for processing a note – was maximized by sequentially moving to more and more computationally intensive steps. Each note must be searched for each key essential associated with the case. The sequence for each search begins with exact matching to the key essential, followed by exact matching to the various dictionaries. This is followed by the more computationally intensive fuzzy-matching procedure and the bag-of-words procedure. Whenever a match occurs, the search is terminated. As Tables 4.1 and 4.2 suggest, although exact matching is in itself insufficient, these less computationally intensive procedures identify a substantial proportion of the variants.
9 We also note that the F1 scores do not represent an additive (or equal interval) scale. It is reasonable to interpret higher F1 scores as representing more accurate matching than lower F1 scores. It is not appropriate to interpret an increase from .50 to .55 as being equivalent to a change from .95 to 1.00.

References

Bejar, I. I. (2013, April). *Gaming a scoring engine: Lexical and discourse-level construct irrelevant response strategies in the assessment of writing*. Paper presented at the annual meeting of the National Council on Measurement in Education.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, *17*(4), 9–16.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*, 27–40.

Burstein, J., & Leacock, C., & Swartz, R. (2001). *Automated evaluation of essays and short answers*. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=4d7766c71bc1d03dc5b53511cac4ca947b017034

Cianciolo, A. T., LaVoie, N., & Parker, J. (2021). Machine scoring of medical students' written clinical reasoning: Initial validity evidence. *Academic Medicine*, *96*, 1026–1035.

Cook, R., Baldwin, S., & Clauser, B. (2012, April). *An NLP-based approach to automated scoring of the USMLE® step 2 CS patient note*. Paper presented at the annual meeting of the National Council on Measurement in Education.

Han, J. W., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques* (3rd ed.). Morgan Kaufmann Publishers.

Higgins, D., & Heilman, M. (2014). Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior. *Educational Measurement: Issues and Practice*, *33*(3), 36–46.

Jani, K. H., Jones, K. A., Jones, G. W., Amiel, J. B., & Elhadad, N. (2020). Machine learning to extract communication and history-taking skills in OSCE transcripts. *Medical Education*, 1–12.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Process*, *25*, 259–284.

Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The intelligent essay assessor. *IEEE Intelligent Systems*, *15*(5), 27–31.

LaVoie, N., Parker, J., Legree, P. J., Ardison, S., & Kilcullen, R. N. (2020). Using latent semantic analysis to score short answer constructed responses: Automated scoring of the consequences test. *Educational and Psychological Measurement*, *80*, 399–414.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*(4).

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

Margolis, M. J., & Clauser, B. E. (2020). Automated scoring in medical licensing. In D. Yan, A. A. Rupp, & P. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 445–467). Taylor and Francis.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., & Lucas, J. F. (2006). Concepts, terminology, and basic models of evidence centered design. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring for complex tasks in computer-based testing* (pp. 123–167). Lawrence Erlbaum Associates.

Monaghan, W., & Bridgeman, B. (2005, April). *E-Rater as a quality control on human scores*. ETS R&D Connections, ETS.

Page, E. B. (1966). *Grading essays by computer: Progress report*. Notes from the 1966 Invitational Conference on Testing Problems, Educational Testing Service.

Page, E. B. (1967). The imminence of grading essays by computer. *Phi Delta Kappan*, *47*, 238–243.

Sarker, A. D., Klein, A. Z., Mee, J., Harik, P., & Gonzalez-Hernandez, G. (2019). An interpretable natural language processing system for written medical examination assessment. *Journal of Biomedical Informatics*, *98*.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53–76.

Streeter, L., Bernstein, J., Foltz, P., & DeLand, D. (2011). *Pearson's automated scoring of writing, speaking, and mathematics* (White Paper). http://kt.pearsonassessments.com/download/PearsonAutomatedScoring-WritingSpeaking-Math-051911.pdf

Sukkarieh, J. Z., & Blackmore, J. (2009). *C-rater: Automatic content scoring for short constructed responses* (pp. 290–295). Proceedings of the Twenty-Second International FLAIRS Conference, Association for the Advancement of Artificial Intelligence.

Swygert, K., Margolis, M., King, A., Siftar, T., Clyman, S., Hawkins, R., & Clauser, B. (2003). Evaluation of an automated procedure for scoring patient notes as part of a clinical skills examination. *Academic Medicine (RIME Supplement)*, *78*(10), S75–S77.

Willis, A. (2015). *Using NLP to support scalable assessment of short free text responses* (pp. 243–253). Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics.

Yamamoto, K., He, Q., Shin, H. J., & von Davier, M. (2017). *Developing a machine-supported coding system for constructed-response items in PISA (ETS RR-17-47)*. Educational Testing Service.