Predicting Item Survival for Multiple Choice Questions in a High-stakes Medical Exam

Victoria Yaneva, Le An Ha, Peter Baldwin, Janet Mee

1 - National Board of Medical Examiners, Philadelphia, USA
2 - Research Institute in Information and Language Retrieval, University of Wolverhampton, UK
{vyaneva, pbaldwin, jmee}@nbme.org; l.a.ha@wlv.ac.uk

Abstract

One of the most resource-intensive problems in the educational testing industry relates to ensuring that newly-developed exam questions can adequately distinguish between students of high and low ability. The current practice for obtaining this information is the costly procedure of pretesting: new items are administered to test-takers and then the items that are too easy or too difficult are discarded. This paper presents the first study towards automatic prediction of an item’s probability to “survive” pretesting (item survival), focusing on human-produced MCQs for a medical exam. Survival is modeled through a number of linguistic features and embedding types, as well as features inspired by information retrieval. The approach shows promising first results for this challenging new application and for modeling the difficulty of expert-knowledge questions.

Keywords: Multiple Choice Questions, Difficulty Prediction, Educational Applications

1. Introduction

Large-scale testing relies on a pool of test questions, which must be replenished, updated, and expanded over time. Writing high-quality test questions is challenging as they must satisfy certain quality standards before they can be used to score examinees. These standards are based on statistical criteria and ensure that: i) items are not too easy or too difficult for the intended examinee population, and ii) the probability of success on each item is positively related to overall examinee performance (Section 3). While the exact thresholds vary, most exam programs have such a requirement. Even when item writers are well-trained and adhere to industry best practices, it has generally not been possible to identify which items will satisfy the various statistical criteria without first obtaining examinee responses through pretesting. Pretesting involves embedding new items within a standard live exam and, based on the collected responses, a determination is made about whether or not a given item satisfies conditions i) and ii). Items that meet the criteria are considered to have “survived” pretesting and can later be used to score examinees. The proportion of surviving items varies across programs; however, Brennan (2006) recommends pretesting at least twice the number of items needed.

While necessary, the enterprise of pretesting is costly. Scored items compete with pretest items for exam space, and the scarcity of which can create a bottleneck. As a result, it is sometimes not possible to pretest as many new items as needed and some exam programs may not be able to afford pretesting at all. This problem is expected to grow with advances in automatic question generation (Gierl et al., 2018), where a large amount of new questions are generated but there is no criteria on how to evaluate their suitability for live use. Conceivably, having advance knowledge of an item’s probability to survive can allow using the available pretesting slots for items that are more likely to pass the thresholds. To address these issues, we present a method for modeling item survival within a large-scale real-world data set of multiple choice questions (MCQs) for a high-stakes medical exam.

Contributions: i) The paper introduces a new practical application area of NLP related to predicting item survival for improving high-stakes exams. ii) The developed models outperform three baselines with a statistically significant difference, including a strong baseline of 113 linguistic features. iii) Owing to the generic nature of the features, the presented approach is generalizable to other MCQ-based exams. iv) We make our code available at: https://bit.ly/2EaTFNN

2. Related Work

Predicting item survival from item text is a new application area for NLP and, to the best of our knowledge, there is no prior work investigating this specific issue. The problem is, however, related to the limited available research on predicting question difficulty with the important difference that predicting survival involves predicting an additional item parameter that captures the relation between the probability of success for the individual item and overall examinee performance (Section 3).

With regards to estimating question difficulty for humans, the majority of studies focus on applying readability metrics to language comprehension tests, where the comprehension questions refer to a given piece of text and, therefore, there is a relationship between the difficulty of the two (Huang et al., 2017; Loukina et al., 2016). For example, Loukina et al. (2016) investigate the extent to which the difficulty of listening items in an English language proficiency test can be predicted by the textual properties of the prompt by using text complexity features (e.g. syntactic complexity, cohesion, academic vocabulary, etc). In another study, Beinborn et al. (2015) rank the suitability

---

1This constant need for new test questions arises as the population of test-takers grows, new topics for exam content are identified, item exposure threatens exam security, etc.

2The questions cannot be released because of test security.
A 55-year-old woman with small cell carcinoma of the lung is admitted to the hospital to undergo chemotherapy. Six days after treatment is started, she develops a temperature of 38°C (100.4°F). Physical examination shows no other abnormalities. Laboratory studies show a leukocyte count of 100/mm³ (5% segmented neutrophils and 95% lymphocytes). Which of the following is the most appropriate pharmacotherapy to increase this patient’s leukocyte count?

(A) Darbepoetin  (B) Dexamethasone
(C) Filgrastim  (D) Interferon alfa
(E) Interleukin-2 (IL-2)  (F) Leucovorin

Table 1: An example of a practice item

3 Accréditée par la Liaison Committee on Medical Education (LCME).
where $P_i$ is the p-value for item $i$, $Un$ is the 0-1 score (correct-incorrect) on item $i$ earned by examinee $n$, and $N$ is the total number of examinees in the sample.

- The correlation (referred to as $Rb$) between examinees’ responses on the given item and examinees’ total test score had to be greater than zero. In other words, examinees who perform better on the test overall, must be more likely to succeed on the item than those examinees who performed worse. This ensures that the item, if used for scoring, will improve the quality of the total test score.

In our sample, half of the items (2,959) are from the surviving class, while the other 2,959 are from the non-surviving class. This distribution is in line with the observation by Brennan (2006) that the pretested items should be at least twice the number of items needed. From the non-surviving items, only 22% violated the $Rb$ criterion, while 78% violated the P-value criterion.

### 4. Features

To model item survival (through $P$ and $Rb$), we are concerned with several types of item characteristics. First, following (Ha et al., 2019), we extract 113 linguistic features accounting for several levels of linguistic complexity in order to understand the extent to which item survival can be attributed simply to the way the items are written. Such an understanding is useful in two main ways. First, it represents a strong baseline that other models can be compared to and second, it allows examining whether the item writers successfully minimize text complexity variation (the latter has a high practical value for test development). As a step beyond linguistic complexity, we explore a deeper-level semantic characteristics of the items by generating two types of embeddings presented below.

While these two approaches (linguistic features and embeddings) are widely used in various NLP applications, they do not account for the type of difficulty that is related to MCQs in particular, e.g., the relationship between item components. For this reason we develop a set of features inspired by automatic question answering (Ha and Yaneva, 2019), the aim of which is to quantify the difficulty of solving an MCQ for an automatic system. The subsections below describe each type of feature group, together with its individual features. Additional details can be found in the available code.

#### 4.1. Linguistic features (baseline)

This class of features is inspired by readability research and its application to estimating question difficulty, forming a strong baseline for comparison to other approaches (Dubay, 2004; McNamara et al., 2014; Yaneva and Evans, 2015; Yaneva et al., 2019). It includes the following subcategories.

**Lexical Features**, such as counts, incidence scores and ratios for ContentWord, Noun, Verb, Adjective, and Adverb; Numerical Count; Type-Token Ratio; Average Word Length In Syllables; and Complex Word Count (> 3 syllables).

**Syntactic Features**: These were implemented using information from the Stanford NLP Parser (Manning et al., 2014): Average Sentence Length (words); Average Depth Of Tree; Negation Count; Negation In Stem; Negation In the Lead-In Question; NP Count; NP Count With Embedding (the number of noun phrases derived by counting all the noun phrases present in an item, including embedded NPs); Average NP Length; PP and VP Count; Proportion Passive VPs; Agentless Passive Count; Average Number of Words Before Main Verb; Relative Clauses and Conditional Clauses Count.

**Semantic Ambiguity Features**: This subcategory concerns the semantic ambiguity of word concepts according to WordNet (WN), as well as medical concepts according to the UMLS (Unified Medical Language System) Metathesaurus (Schuyler et al., 1993). The features include Polysemic Word Index; Average Number of Senses of: Content Words, Nouns, Verbs, Adjectives, Adverbs; Average Distance To WN Root for: Nouns, Verbs, Nouns and Verbs; Total No Of UMLS Concepts; Average No Of UMLS Concepts; Average No Of Competing Concepts Per Term (average number of UMLS concepts that each medical term can refer to).

**Readability Formulae**: Flesch Reading Ease (Flesch, 1948); Flesch Kincaid Grade Level (Kincaid et al., 1975); Automated Readability Index (ARI) (Senter and Smith, 1967); Gunning Fog index (Gunning, 1952); Coleman-Liau (Coleman, 1965); and SMOG index (McLaughlin, 1969).

**Cognitively-motivated features**: These are calculated based on information from the MRC Psycholinguistic Database (Coltheart, 1981), which contains cognitive measures based on human ratings for a total of 98,538 words. These features include Imagability, indicating the ease to construct a mental image of that word; Familiarity, or how familiar the word seems to an adult; Concreteness; Age Of Acquisition; and finally Meaningfulness Ratio Colorado and Meaningfulness Ratio Paivio. The meaningfulness rating assigned to a word indicates how associated the word is to other words.

**Word Frequency Features**: These include Average Word Frequency, as well as threshold frequencies such as words not included in the most frequent words on the BNC frequency list (Not In First 2000/3000/4000 or 5000 Count).

**Text Cohesion Features**: These include counts of All Connectives, as well as Additive; Temporal; and Causal Connectives, and Referential Pronoun Count.

### 4.2. Embeddings

We experiment with two types of embeddings: Word2Vec (300 dimensions) (Mikolov et al., 2013) and ELMo (1,024 dimensions) (Peters et al., 2018). The results presented in this paper refer to embeddings generated using approximately 22,000,000 MEDLINE abstracts, which were found to outperform other versions of the embeddings extracted from generic corpora (Google News Corpus for Word2Vec and 1B Word (Chelba et al., 2013) for ELMo). The embeddings were aggregated at item level using mean pooling, where an average item embedding is generated from the embeddings of all words.

---

4https://news.google.com
4.3. Information Retrieval (IR) features
This group of features measures how difficult it is for an automatic question-answering (QA) system to answer the items correctly. The hypothesis behind the design of these features is that MCQs which are more difficult to answer automatically would also be more difficult for humans (Ha and Yaneva, 2019). This idea stems from observations such as the fact that both humans and machines need to retrieve a given information in order to answer the questions, where humans use their subject knowledge, while machines query a database. Another parallel between the two processes is the need to reason over facts, where humans still have a great advantage. In spite of the limitations of current QA systems, it is conceivable that since humans and machines need to perform similar processes in order to answer a question, these processes might be challenged by the same types of questions. Even if such parallels are not found, testing this working hypothesis is a useful by-product of the current research, as it may inform better strategies for approaching automatic QA tasks. To test the hypothesis, we develop and train a full automatic question-answering system following approaches presented in Clark et al. (2018).

After that, we use the scores obtained by the system to extract the following features, as explained below.

First, we index the abstracts of medical articles contained in the MEDLINE\(^6\) database using Lucene\(^5\), with its default options. Then we query the database as follows. For each test item we build several queries, where each query contains the stem and one answer option. We use three different settings: i) All words, ii) Nouns only, and iii) Nouns, Verbs, and Adjectives (NVA). We then get the top 5 MEDLINE documents returned by Lucene as a result of each query and calculate the sum of the retrieval scores. These scores represent the content of the IR features (Stem Only, Stem + Correct Answer, and Stem + Options, where for each of these configurations we have a different feature for All words, Nouns only and NVA.). The scores reflect how difficult it is for a QA system to choose the correct answer. Specifically, if the IR scores of Stem + Correct Answer are much higher than those of Stem + Options, then it is easy for the QA system to answer that item correctly by picking the option that has the highest scores. This information can thus be used to predict item difficulty.

5. Experiments
We explore and compare two approaches to predicting item survival, namely: i) modeling P-value and Rb individually and applying a function to select which items to retain, and ii) modeling survival as a binary classification between surviving and non-surviving items. Both these approaches are compared to three baseline models.

5.1. Baselines
The results of various models are compared to the following baselines.

- **Random**: Since the surviving and non-surviving classes are of the same size, the random baseline is

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>P-value</th>
<th>Rb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.31</td>
<td>0.13</td>
</tr>
<tr>
<td>SVM (regr.)</td>
<td>0.22</td>
<td>33.13</td>
</tr>
<tr>
<td>Gaussian Processes</td>
<td>0.22</td>
<td>29.71</td>
</tr>
<tr>
<td>Linear regression</td>
<td>0.18</td>
<td>36.86</td>
</tr>
<tr>
<td>Random Forests</td>
<td>0.31</td>
<td>26.82</td>
</tr>
</tbody>
</table>

F1 = 0.5. For predicting P-value and Rb individually, the model performance is first compared to the output of the ZeroR rule (assigning the mean of the distribution as a predicted value to each instance), the cutoff function is applied to classify the items based on the predicted values and then classification accuracy is measured and compared to the random baseline.

- **Item length in words**: This baseline is applied to rule out the possibility that item survival is simply a function of item length (e.g., that longer items may be more difficult). Using item length in words as a single variable in an Random Forests model (algorithm chosen for comparability, see below) in a 10-fold cross validation set up results in F1 = 0.51.

- **Linguistic features**: This rich baseline consists of the full set of linguistic features as input to a Random Forests classifier (see below) (10-fold CV) and the result is an F1 score of 0.54. This result significantly outperforms both the random baseline (p < 0.0001) and the item length one (p < 0.001), making the Linguistic feature model the strongest of our baselines.

5.2. Algorithm selection
First, we assign the items to a training (60%), validation (20%), and test (20%) sets. We then evaluate various neural and non-neural models on the validation set. As shown in Table 2 the neural approaches consistently performed better than most non-neural ones in predicting P and Rb, but were convincingly outperformed by the Random Forests (RF) algorithm (Breiman, 2001). For example, the RF algorithm outperformed a neural network with three dense layers by reducing RMSE for P-value with approximately one point (Table 2). For the classification task, non-neural approaches proved consistently more suitable than neural ones, with RF again performing best (F1 = 55.8). The rest of the results on algorithm selection for the classification task are presented in Table 2. Subsequent experiments involved testing various feature combinations.

5.3. Modeling P an Rb
This subsection presents results for item survival obtained by modeling P and Rb individually. Of all feature combinations, best performance was achieved using all features. P is predicted with r = 0.31 for the validation and 0.26 for the

---

Table 3: Evaluation of five algorithms (among others) for predicting item survival in a classification task (NN parameters include: 3 dense or LSTM layers of size 100, activation function: RELU, weight initialization Xavier and learning rate = 0.001. Trained for 500 epochs with early stopping after 10 epochs with no improvement.)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td></td>
</tr>
<tr>
<td>NN with 3 dense layers</td>
<td>47.2</td>
<td></td>
</tr>
<tr>
<td>LSTM (3 layers)</td>
<td>47.5</td>
<td></td>
</tr>
<tr>
<td>BayesNet</td>
<td>55.1</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>54.3</td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td>55.8</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Root Mean Squared Error (RMSE) for P-value and Rb (full feature set) with bootstrapped significance

<table>
<thead>
<tr>
<th></th>
<th>P-value RMSE</th>
<th>Rb RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Valid.</td>
<td>Test</td>
</tr>
<tr>
<td>ZeroR</td>
<td>28.18</td>
<td>29.1</td>
</tr>
<tr>
<td>RF</td>
<td>26.82</td>
<td>28.08</td>
</tr>
<tr>
<td>P (sign.)</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>CI</td>
<td>-1.7/-68</td>
<td>-1/-0.03</td>
</tr>
</tbody>
</table>

Table 4: Root Mean Squared Error (RMSE) for P-value and Rb (full feature set) with bootstrapped significance

5.4. Classification

Approaching survival as a classification task, we use the RF algorithm with various feature combinations as input. The evaluation is based on 10-fold cross-validation. The results for various combinations are presented in Table 3.

As can be seen from the table, all feature combinations led to a statistically significant improvement over both the Random baseline and the Item length baseline, with the exception of the IR features, which outperform only the Random baseline ($p = 0.0047$, 95% CI: 0.79, 4.39). The strongest baseline, the Linguistic feature model (F1 = 54), is outperformed by the full feature set with an F1 of 55.8 ($p = 0.04$, 95% CI: 0.0072, 3.5912). The Area Under ROC Curve for this result is 57.8. Notably, ELMo achieved a result that was very close to the best result (55.6) but it did not outperform the Linguistic feature baseline. Nevertheless, ELMo emerged as the strongest predictor among all features. The fact that the best result is achieved using all feature types indicates that they complement rather than overlap each other.

Table 6: Baselines

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR</td>
<td>52.6</td>
</tr>
<tr>
<td>Ling + W2V + ELMo</td>
<td>34.5*</td>
</tr>
<tr>
<td>W2V</td>
<td>34.6</td>
</tr>
<tr>
<td>W2V + ELMo</td>
<td>34.8*</td>
</tr>
<tr>
<td>IR + W2V</td>
<td>35.2*</td>
</tr>
<tr>
<td>Ling + ELMo</td>
<td>35.2*</td>
</tr>
<tr>
<td>IR + Ling + W2V</td>
<td>35.3*</td>
</tr>
<tr>
<td>IR + ELMo</td>
<td>35.4*</td>
</tr>
<tr>
<td>IR + W2V + ELMo</td>
<td>35.4*</td>
</tr>
<tr>
<td>IR + Ling + ELMo</td>
<td>35.5*</td>
</tr>
<tr>
<td>ELMo</td>
<td>35.6*</td>
</tr>
<tr>
<td>All</td>
<td>55.8**</td>
</tr>
</tbody>
</table>

5.5. Error analysis

Analysis of the errors for the full feature set reveals that the model results in more false positives than false negatives, where 40% of the false positives occur due to failure to recognize negative Rb values as non-surviving. The set of false negatives indicates that items with higher P-values tend to be mistakenly classified as non-surviving ($\mu = .73$, min = .31, max = .95). Roughly 45% of the items with negative Rb values are mistakenly classified as surviving and these are mostly instances with P-values within the surviving range.

To understand the behaviour of the different feature classes better, we analyse the model output of each feature class (Linguistic, IR, Word2Vec and ELMo). The linguistic features perform more or less comparable to the other types in terms of identifying true positives but have the lowest rate of true negatives, especially in predicting the Rb component. Conversely, ELMo do best among all feature types in recognizing that negative Rb values should be labeled as non-surviving but has the highest rate of false positives above the .95 threshold. No specific pattern emerged for the IR features and Word2Vec.

5.6. Feature importance

As can be seen from Table 5, ELMo performs best among all feature types, followed by Word2Vec, the linguistic features and the IR features at the last place. A more detailed analysis of the linguistic and IR features is presented in Table 7 which shows the top five highest correlated features from each set. Individual features are very weakly correlated with the class labels, indicating that no single feature has the predictive power to dominate the models. This result was corroborated with a backward feature elimination study for the IR set, where the removal of individual IR features did not lead to notable improvement in the performance of the set but instead reduced the accuracy.
6. Discussion

The classification models achieved modest, but statistically significant better results compared to three baselines: random assignment, item length and a strong baseline of 113 linguistic features. The fact that adding the IR features and embeddings achieved significant improvement over the linguistic features is an encouraging evidence that the models captured difficulty beyond the linguistic complexity represented by this baseline. In terms of task definition, the results from the classification approach were better than those from modeling P-value and Rb individually, suggesting that there may be value in learning the two parameters simultaneously.

While the IR features alone performed worse than the linguistic ones, there are two interesting observations that can be made regarding their role. First, they significantly outperformed the random assignment baseline, which shows that they did provide some signal. Second, they were included in the best performing model, as well as the majority of the better models, which shows that the signal they provided did not overlap with that of the embeddings or linguistic features. This is an evidence that there is value in approaches to question difficulty estimation that rely on input from automatic question-answering systems and that there are some similarities between questions that are difficult for machines and those that are difficult for humans.

The exact extent to which this is the case is the subject of a separate investigation. The most challenging part for the models was the correct classification of items with a negative Rb component which had P-values within the range of surviving items. This is explained by the ability of the models to predict the P-value with higher accuracy than the Rb component, as shown by modeling the two parameters individually. As revealed by the error analysis, the great advantage of ELMo compared to other feature types is the comparatively higher accuracy at predicting Rb. Nevertheless, ELMo did worse at identifying non-surviving items at the high end of the P-value spectrum. It is these differences in the predictive power of the different feature sets that make them complement each other rather than overlap. In terms of feature importance, ELMo performed best, followed by Word2Vec, the linguistic features and the IR features. As shown in the feature analysis section, no individual features from the linguistic or IR sets had a significantly higher correlation with class labels than the rest.

One of the more serious limitations of this study is the relatively low overall accuracy achieved by the models compared to other classification tasks. This is largely explained by the difficulty of the problem, since predicting item survival for highly-specialised questions with no varying reading levels requires going beyond the measurement of linguistic complexity. While this limitation remains true, predicting item survival represents low-risk and high-benefit strategy, as even small improvements in accuracy can lead to saving substantial resources from pretesting. This can be achieved if items are first filtered automatically and pretesting slots are then assigned to items that are more likely to pass the thresholds. For example, the best result of 55.8 translates to having 3,304 correctly classified items, which is 345 more than the random assignment baseline. In practice, one standard nine-hour exam can only afford to have one block of pretesting items, which is 40 items (the bottleneck problem referred to in the introduction section) and each item is seen by an average of 328 examinees. This illustrates the suitability of NLP to help with this practical task by reducing its cost without introducing any major risks, especially in the context of evaluating automatically generated questions.

One of the strengths of the proposed approach is its generic nature, which means that it could be applicable to other exams. While such data sets are typically not freely available because of test security (exam questions become unusable if there is a chance that examinees might have seen them in advance), they exist in most high-stakes exams that test subject knowledge using the multiple-choice format. It is therefore conceivable that the proposed approach could be used in other exams. These results are a first step towards the challenging but highly beneficial application of predicting item survival. Future work includes the exploration of multi-task learning for predicting P and Rb simultaneously, as well as searching for better predictors for the Rb component.

7. Conclusion

This paper presented a new practical application for NLP, namely predicting item survival from item text. The task involves differentiating between good quality exam questions and ones that are too easy or too difficult, as well as ones that do not improve the quality of the overall test score. Four feature types were extracted, consisting of 113 linguistic features (baseline), Word2Vec and ELMo embeddings, as well as features related to the difficulty a QA system would have with answering the questions. Results indicated statistically significant improvement over the linguistic model, a random assignment baseline and an item length baseline, by using all feature types in a binary classification task. ELMo had the highest predictive power, followed by Word2Vec, the linguistic features and the IR features. The proposed approach is generic and has the potential to generalize over other high-stakes exams that test subject knowledge through multiple-choice questions.

8. Bibliographical References


