Predicting the Difficulty of Multiple Choice Questions in a High-stakes Medical Exam

Le An Ha[1] **Victoria Yaneva**[2] **Peter Baldwin**[2] **Janet Mee**[2]

[1]Research Group in Computational Linguistics, University of Wolverhampton, UK
ha.l.a@wlv.ac.uk
[2]National Board of Medical Examiners, Philadelphia, USA
{vyaneva, pbaldwin, jmee}@nbme.org

Abstract

Predicting the construct-relevant difficulty of Multiple-Choice Questions (MCQs) has the potential to reduce cost while maintaining the quality of high-stakes exams. In this paper, we propose a method for estimating the difficulty of MCQs from a high-stakes medical exam, where all questions were deliberately written to a common reading level. To accomplish this, we extract a large number of linguistic features and embedding types, as well as features quantifying the difficulty of the items for an automatic question-answering system. The results show that the proposed approach outperforms various baselines with a statistically significant difference. Best results were achieved when using the full feature set, where embeddings had the highest predictive power, followed by linguistic features. An ablation study of the various types of linguistic features suggested that information from all levels of linguistic processing contributes to predicting item difficulty, with features related to semantic ambiguity and the psycholinguistic properties of words having a slightly higher importance. Owing to its generic nature, the presented approach has the potential to generalize over other exams containing MCQs.

1 Introduction

For many years, approaches from Natural Language Processing (NLP) have been applied to estimating reading difficulty, but relatively fewer attempts have been made to measure conceptual difficulty or question difficulty beyond linguistic complexity. In addition to expanding the horizons of NLP research, estimating the construct-relevant difficulty of test questions has a high practical value because ensuring that exam questions are appropriately difficult is both one of the most important and one of the most costly tasks within the testing industry. For example, test questions that are too easy or too difficult are less able to distinguish between different levels of examinee ability (or between examinee ability and a defined cut-score of some kind – e.g., pass/fail). This is especially important when scores are used to make consequential decisions such as those for licensure, certification, college admission, and other high-stakes applications[1]. To address these issues, we propose a method for predicting the difficulty of multiple choice questions (MCQs) from a high-stakes medical licensure exam, where questions of varying difficulty may not necessarily vary in terms of reading levels.

Owing to the criticality of obtaining difficulty estimates for items (exam questions) prior to their use for scoring, current best practices require newly-developed items to be pretested. Pretesting typically involves administering new items to a representative sample of examinees (usually between a few hundred and a few thousand), and then using their responses to estimate various statistical characteristics. Ideally, pretest data are collected by embedding new items within a standard live exam, although sometimes special data collection efforts may also be needed. Based on the responses, items that are answered correctly by a proportion of examinees below or above certain thresholds (i.e. items that are too easy or too difficult for almost all examinees) are discarded. While necessary, this procedure has a high financial and administrative cost, in addition to the time required to obtain the data from a sufficiently large sample of examinees.

Here, we propose an approach for estimating the difficulty of expert-level MCQs, where the

[1]Examples of well-known high-stakes exams include the TOEFL (Test of English as a Foreign Language) (https://www.ets.org/toefl), the SAT (Scholastic Assessment Test) (https://collegereadiness.collegeboard.org/sat), and the USMLE (United States Medical Licensing Examination) (https://www.usmle.org/).

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 38C (100.4F). 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

gold standard of item difficulty is defined through large-scale pretesting and is based on the responses of hundreds of highly-motivated examinees. Being able to automatically predict item difficulty from item text has the potential to save significant resources by eliminating or reducing the need to pretest the items. These savings are of even greater importance in the context of some automatic item generation strategies, which can produce tens of thousands of items with no feasible way to pretest them or identify which items are most likely to succeed. Furthermore, understanding what makes an item difficult other than manipulating its reading difficulty has the potential to aid the item-writing process and improve the quality of the exam. Last but not least, automatic difficulty prediction is relevant to automatic item generation as an evaluation measure of the quality of the produced output.

Contributions i) We develop and test the predictive power of a large number of different types of features (e.g. embeddings and linguistic features), including innovative metrics that measure the difficulty of MCQs for an automatic question-answering system. The latter produced empirical evidence on whether parallels exist between question difficulty for humans and machines. ii) The results outperform a number of baselines, showing that the proposed approach measures a notion of difficulty that goes beyond linguistic complexity. iii) We analyze the most common errors produced by the models, as well as the most important features, providing insight into the effects that various item characteristics have on the success of predicting item difficulty. iv) Owing to the generic nature of the features, the presented approach is potentially generalizable to other MCQ-based exams. We make our code available[2] at: `https://github.com/anomymous1/Survival-Prediction`.

2 Related Work

The vast majority of previous work on difficulty prediction has been concerned with estimating readability (Flesch, 1948; Dubay, 2004; Kintsch and Vipond, 2014; François and Miltsakaki, 2012; McNamara et al., 2014). Various complexity-related features have been developed in readability research (see Dubay (2004) and Kintsch and Vipond (2014) for a review), starting from ones utilising surface lexical features (e.g. Flesch (1948)) to NLP-enhanced models (François and Miltsakaki, 2012) and features aimed at capturing cohesion (McNamara et al., 2014).

There have also been attempts to estimate the difficulty of questions for humans. This has been mostly done within the realm of language learning, where the difficulty of reading comprehension questions is strongly related to their associated text passages (Huang et al., 2017; Beinborn et al., 2015; Loukina et al., 2016). Another area where question-difficulty prediction is discussed is the area of automatic question generation, as a form of evaluation of the output (Alsubait et al., 2013; Ha and Yaneva, 2018). In many cases such evaluation is conducted through some form of automatic measure of difficulty (e.g., the semantic similarity between the question and answer options as in (Ha

[2]The questions cannot be made available because of test security.

and Yaneva, 2018)) rather than through extensive evaluation with humans. Past research has also focused on estimating the difficulty of open-ended questions in community question-answering platforms (Wang et al., 2014; Liu et al., 2013); however, these questions were generic in nature and did not require expert knowledge. Other studies use taxonomies representing knowledge dimensions and cognitive processes involved in the completion of a test task to predict the difficulty of short-answer questions (Padó, 2017) and identify skills required to answer school science questions (Nadeem and Ostendorf, 2017). We build upon previous work by implementing a large number of complexity-related features, as well as testing various prediction models (Section 4).

While relevant in a broad sense, the above works are not directly comparable to the current task. Unlike community question answering, the questions used in this study were developed by experts and require the application of highly specialized knowledge. Reading exams, where comprehension difficulty is highly associated with text complexity, are also different from our medical MCQs, which are deliberately written to a common reading level (see Section 3). Therefore, the models needed to capture difficulty in this context that goes beyond linguistic complexity.

3 Data

Data comprises 12,038 MCQs from the Clinical Knowledge component of the United States Medical Licensing Examination®. An example of a test item is shown in Table 1. The part describing the case is referred to as the *stem*, the correct answer option is called the *key* and the incorrect answer options are known as *distractors*. The majority of the items in the data set used here had five or six answer options.

Item writing All items tested medical knowledge and were designed to emulate real-life scenarios wherein examinees must first identify the relevant findings and then, based on these findings, make a diagnosis or take a clinical action. Items were written by experienced item-writers following a set of guidelines. These guidelines stipulated that the writers adhere to a standard structure and avoid excessive verbosity, "window dressing" (extraneous material not needed to answer the item), "red herrings" (information designed to mislead the test-taker), overly long or complicated stems or options, and grammatical cues (e.g., correct answers that are longer, more specific, or more complete than the other options; or the inclusion of the same word or phrase in both the stem and the correct answer). Item writers had to ensure that the produced items did not have flaws related to various aspects of validity. For example, flaws related to irrelevant difficulty include: *Stems or options are overly long or complicated, Numeric data not stated consistently* and *Language or structure of the options is not homogeneous*. Flaws related to "testwiseness" are: *Grammatical cues; The correct answer is longer, more specific, or more complete than the other options;* and *A word or phrase is included both in the stem and in the correct answer*. Finally, stylistic rules concerning preferred usage of terms, formatting, abbreviations, conventions, drug names, and alphabetization of option sets were also enforced. The goal of standardizing items in this manner is to produce items that vary in difficulty and discriminating power due only to differences in the medical content they assess. This practice, while sensible, makes modeling difficulty very challenging.

Item administration The questions in our data set were pretested by embedding them within live exams. In practice, the response data collected during pretesting is used to filter out items that are misleading, too easy, or too difficult based on various criteria. Only those items satisfying these criteria are eligible for use during scoring on subsequent test forms. The current set of items contains pretest data administered for four standard annual cycles between 2012 and 2015. The questions were embedded within a standard nine-hour exam and test-takers had no way of knowing which items were used for scoring and which were being pretested. Examinees were medical students from accredited[3] US and Canadian medical schools taking the exam for the first time as part of a multistep examination sequence required for medical licensure in the US.

Determining item difficulty On average, each item was answered by 328 examinees ($SD = 67.17$). The difficulty of an item is defined by the proportion of its responses that are correct, which is commonly referred to in the educational-testing literature as its *P-value*[4]. The P-value is calculated

[3] Accredited by the Liaison Committee on Medical Education (LCME).

[4] We adopt this convention here but care should be taken not to confuse this usage with a p-value indicating statistical

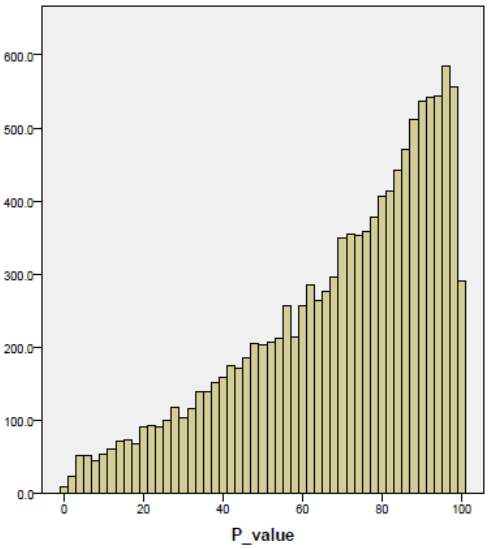

Figure 1: Distribution of the P-value variable

in the following way:

$$P_i = \frac{\sum_{n=1}^{N} U_n}{N},$$

where *Pi* 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. As an example, a P-value of .3 means that the item was answered correctly by 30% of the examinees. The distribution of P-values for the data set is presented in Figure 1.

4 Features

A number of features were modeled for P-value prediction and can be roughly divided into three classes. First, we extract embeddings, which have been found to have predictive power in many different applications. The second class of features included more than one hundred linguistic characteristics, which account for differences in the way the items are written. Finally, a third class of features were based on the difficulty an item posed to an automated question-answering system under the working hypothesis that this system difficulty had a positive relationship with the difficulty an item poses to human respondents. Information about each type of feature is presented below. Additional details can be found in the available code.

4.1 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 embeddings were generated using approximately 22,000,000 MEDLINE abstracts,[5], which were found to outperform other versions of the embeddings extracted from generic corpora (Google News Corpus[6] for Word2Vec and 1B Word (Chelba et al., 2013) for ELMo). Embeddings were aggregated at item level using mean pooling, where an average item embedding is generated from the embeddings of all words.

4.2 Linguistic features

This class of features includes the following subcategories.

Lexical Features Previous research has found surface lexical features to be very informative in predicting text readability (Dubay, 2004). Lexical features in our experiments include counts, incidence scores and ratios for *ContentWord, Noun, Verb, Adjective*, and *Adverb*; *Numeral 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) and include: *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;* and *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) Meta-thesaurus (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; and Average No Of Competing Concepts Per Term* (average number of UMLS concepts that each medical term can refer to).

significance.

[5]https://www.nlm.nih.gov/bsd/medline.html

[6]https://news.google.com

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*, which indicates the ease with which a mental image of a word is constructed; *Familiarity* of the word for an adult; *Concreteness*; *Age Of Acquisition*; and finally *Meaningfulness Ratio Colorado* and *Meaningfulness Ratio Paivio*. The meaningfulness rating assigned to a word indicates the extent to which the word is associated with 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*.

4.3 Information Retrieval (IR) features

The working hypothesis behind this group of features is that there is a positive correlation between the difficulty of questions for humans and for machines. To quantify machine-difficulty, we develop features based on information retrieval that capture how difficult it is for an automatic question-answering (QA) system to answer the items correctly. This was accomplished following the approaches to QA presented in Clark and Etzioni (2016).

First, we use Lucene[7] with its default options to index the abstracts of medical articles contained in the MEDLINE[8] database. Then for each test item we build several queries, corresponding to 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 the QA system is more confident in its answer choice and the item is deemed relatively easy. This information can then be used to predict item difficulty.

[7]https://lucene.apache.org/

[8]https://www.nlm.nih.gov/bsd/medline.html

	r	RMSE
Random Forests	**0.24**	**23.15**
Linear Regression	0.17	25.65
SVM	0.17	25.41
Gaussian Processes	0.2	23.87
Dense NN (3 layers)	0.16	25.85

Table 2: Results for algorithm selection (validation set)

5 Experiments

In this section we present our experiments on predicting the P-value.

First, we randomly divide the full data set into training (60%), validation (20%) and test (20%) sets for the purpose of evaluating a number of different algorithms[9] on the validation set. This was done using all features. The most notable results on algorithm selection are presented in Table 2. As can be seen from the table, the best results are obtained using the Random Forests (RF) algorithm (Breiman, 2001), which was selected for use in subsequent experiments.

5.1 Baselines

Five baselines were computed to evaluate model performance. The first baseline is the output of the ZeroR algorithm, which simply assigns the mean of the P-value variable in the training set as a prediction for every instance. Each of the four remaining baselines was based on a common feature known to be a strong predictor of reading difficulty: *Word Count, Average Sentence Length, Average Word Length in Syllables*, and the *Flesch Reading Ease*[10] formula (Flesch, 1948). These

[9]Parameters for the Neural Network algorithm: 3 dense layers of size 100, activation function: RELU, loss function: MSE, weight initialization Xavier and learning rate = 0.001. Trained for 500 epochs with early stopping after 10 epochs with no improvement.

[10]While readability formulae are used as features in the models and their predictive power is assessed, it is acknowl-

simple baselines allow us to assess whether the difficulty of the items in our data set can be reliably predicted using heuristics such as "longer items are more difficult" or "items using longer words and sentences are more difficult". The performances of the baselines as single features in an RF model (except ZeroR, which is an algorithm of its own) are presented in Table 3. In terms of Root Mean Squared Error (RMSE), the strongest baseline was ZeroR, with *Average Word Length in Syllables* producing somewhat similar results. All other baselines performed worse than ZeroR, showing that item length (*Word Count*), as well as *Average Sentence Length* and especially *Flesch readability*, are rather weak predictors of item difficulty for our data. These results provide an empirical evidence in support of the claim that easy and difficult items do not differ in terms of surface readability, commonly measured through word and sentence length.

5.2 P-value Prediction

We use various combinations of the features presented in Section 4 as input to an RF model to predict P-value. The results are presented in Table 4. As can be seen from the table, using the full feature set performs best and is a statistically significant improvement over the strongest baseline (ZeroR) with an RMSE reduction of approximately one point (Training set (10-fold CV): $p = 7.684^{e-10}$ with 95% Confidence Intervals (CI) from 10,000 bootstrap replicates: -0.9170, -0.4749. Test set: $p = 2.20^{e-16}$ with 95% CI from 10,000 bootstrap replicates: -1.423, -0.952).

In terms of individual feature groups, Linguistic, W2V, and ELMo achieved comparable performance (RMSE $\approx$ 22.6 for Test Set). The IR features performed notably worse, (RMSE = 23.4 for Test set), which is also the only result that does *not* outperform the ZeroR baseline ($p = 0.08$, 95% CI: -0.5336, 0.0401). For reference, the next "worst" result is obtained by combining the IR and Linguistic features (RMSE = 22.63); nevertheless, this is a significant improvement over ZeroR ($p = 5.517^{e-14}$ with 95% CI: -1.279, -0.756). Combining the Linguistic, W2V and ELMo features leads to a slight improvement in performance over their individual use, indicating that the signals captured by the different feature groups do not overlap entirely.

5.3 Error Analysis

Analysis of the 500 test-set items with largest error residuals between predicted and actual values (the bottom 20% of the test-set predictions) revealed that the largest errors occur for items with very low P-values (μ = 32, SD = 13.39, min = 0, max = 62). This was expected given the skewness of the P-value variable towards the high end of the scale. These items (P-value < 62) account for 34.5% of the full data. Therefore, one possible explanation for these large errors is the fact that these items are underrepresented as training examples.

As a follow-up study, we looked into items with P-values under .20, which account for only 4.5% of the full data. These items are considered to be either highly misleading and/or very difficult, as test-takers systematically chose incorrect answer options and performed worse than chance (the majority of items had five or six answer options). Excluding this small percentage of items from the training and test sets resulted in substantial improvements in RMSE (20.04 after excluding the items compared to 22.45 before excluding them), and outperformed ZeroR again a similar margin (20.98). This result shows that the success of the proposed approach would be higher for test samples with fewer extremely difficult or misleading items. It is acknowledged, however, that which items are too difficult or misleading can rarely be known a priori.

5.4 Feature Importance

Understanding the contributions of individual feature classes from the Linguistic set is useful for interpreting the models, as well as for informing future item-writing guidelines. To address this, we perform an ablation study where we remove one feature class at a time from the model using all Linguistic features.

As shown in Table 5, the removal of individual classes does *not* lead to dramatic changes in RMSE, suggesting that the predictive power of the Linguistic model is not dependent on a particular feature type (e.g. lexical, syntactic, semantic, etc). Nevertheless, removal of the Semantic Ambiguity and the Cognitively-motivated features led to a slightly lower performance for both cross-validation on the training set and for the test set. Indeed, a correlation analysis between individual

edged that the various formulae were validated on different types of texts than the MCQs in our data. Therefore, their performance is expected to be lower compared to applications which use the intended types of materials.

	Training set (10-fold CV)			Test set		
	r	MAE	RMSE	r	MAE	RMSE
ZeroR	-0.02	19.9	24.09	0	19.67	23.65
Word Count	0.01	20.13	24.5	0.05	19.81	23.87
Av. Sent. Length	-0.006	20.76	25.52	0.04	20.2	24.58
Av. Word Length	0.05	19.89	24.14	0.07	19.6	23.63
Flesch Reading Ease	0.02	22.05	27.53	-0.01	22.27	27.61

Table 3: Baseline results using 10-fold cross validation on the training set and evaluating the models on the test set (r = correlation coefficient, MAE = Mean Absolute Error, RMSE = Root Mean Squared Error).

	Training set (10-fold CV)			Test set		
	r	MAE	RMSE	r	MAE	RMSE
All	0.27	18.88	23.15	**0.32**	**18.53**	**22.45**
Linguistic	0.26	19	23.22	0.29	18.73	22.62
IR	0.17	19.58	23.91	0.18	19.28	23.4
W2V	0.27	18.94	23.18	0.3	18.61	22.58
ELMo	0.27	18.95	23.18	0.29	18.77	22.64
Ling + IR	0.26	19.04	23.25	0.29	18.75	22.63
Ling + ELMo	0.27	19.08	23.19	0.3	18.79	22.61
Ling + W2Vec	0.28	18.9	23.14	0.31	18.65	22.54
IR + W2V	0.27	18.94	23.18	0.3	18.67	22.56
IR + ELMo	0.26	18.95	23.26	0.31	18.53	22.55
W2V + ELMo	0.28	18.84	23.13	0.32	18.51	22.5
IR + W2V + ELMo	0.27	18.88	23.18	0.3	18.56	22.56
IR + Ling + W2V	**0.289**	**18.9**	**23.11**	0.31	18.6	22.52
IR + Ling + ELMO	0.27	19	23.2	0.327	18.64	22.48

Table 4: Results for the training (10-fold CV) and test sets for various feature combinations.

	CV RMSE	Test RMSE
All Linguistic	23.22	22.62
Without Lexical	23.3	22.49
Without Syntactic	23.23	22.66
Without Sem. ambiguity	23.31	22.89
Without Readability	23.22	22.59
Without Word Frequency	23.27	22.63
Without Cognitive	23.3	22.74
Without Cohesion	23.29	22.51

Table 5: Changes in RMSE after removing individual feature classes

features and the P-value variable reveals that the top three features with highest correlations are Age of Acquisition (-.11), Familiarity (.1038) and Referential Pronoun Incidence (.1035). Since the texts are domain-specific and contain a great deal of medical terminology, it is likely that the Age of Acquisition and Familiarity indices reflect the use of terms, however, further analysis is needed to confirm this.

6 Discussion

The experiments presented in the previous section provided empirical evidence that the difficulty of expert-level[11] multiple-choice questions can be predicted with accuracy significantly higher than various baselines. It was shown that simple metrics of complexity such as item length or average word and sentence length performed poorer than the ZeroR baseline, indicating that the difficulty of the items could not be predicted using surface readability measures. Best results were achieved when combining all types of available features (Linguistic, IR, Word2Vec, and ELMo), which showed a statistically significant improvement over the baselines. In terms of individual feature classes, the IR features performed poorly and were outperformed by the Linguistic, Word2Vec, and ELMo features – with the latter two being the strongest classes of predictors. Nevertheless, the fact that the combination of all the feature classes performed best supports the idea that the signals from the different feature groups did not overlap entirely and instead complemented each other. To understand whether the way the items were written had an effect on difficulty prediction and to gain insight into how item-writing could be improved, we analyzed the performance of the different types of Linguistic features. It was shown that the strength of the predictions were *not* due to a single linguistic feature; however, the strongest predictors were features related to semantic ambiguity and cognitively-motivated features (espe-

[11] Requiring expert knowledge as opposed to general knowledge

cially Age of Acquisition and Familiarity). Errors were largest for items at the lower end of the P-value scale, potentially because these items were underrepresented as training examples. Further experiments are needed to corroborate this.

In terms of generalizability, the presented approach is not test-specific and can therefore be applied to other exams containing MCQs. The results are, however, highly dependent on the population of test-takers. In fact, predicting the P-value in our particular case was arguably more challenging than for other exams owing to the homogeneity of the test-taker population. The majority of items were answered correctly by the majority of examinees because the test-takers were highly-able and highly-motivated medical students, who had already passed many other competitive high-stakes exams, including those for medical school admission. All else being equal, the expectation is that the performance of these models would improve for exams administered to, for example, examinees from K-12, where the ability of the test-takers has a higher variance and the distribution of P-values is less-skewed. However, all else is not equal and K-12 exams have substantially different test questions, the effects of which is unknown. Further research is needed here.

The presented approach is a first step toward predicting item difficulty and, therefore, there are a number of avenues for future work that could lead to better results. One of these relates to having separate embeddings for the stem and answer options as opposed to item-level embeddings. Another interesting approach would be to divide the items by content category (e.g. psychiatric, cardiac, etc). Content categories are not used as features in the current approach because there was no practical value in learning that, say, cardiac items are more difficult than psychiatric ones. However, it might be worthwhile to build content-specific models that predict item difficulty within-category (e.g., what are the relative item difficulties within the pool of psychiatric items). Finally, the performance of the IR features could be improved if the information is extracted from corpora that are more relevant (such as textbooks and examinee study materials) as opposed to medical abstracts.

The results presented in this paper have both practical and theoretical importance. Being able to predict the P-value of an MCQ reduces the cost of pretesting while maintaining exam quality. From a theoretical perspective, assessing difficulty beyond readability is an exciting new frontier that has implications for language understanding and cognition. Last but not least, such an application could also potentially be useful for assessing the performance of question-answering systems by predicting the difficulty of the questions for humans.

7 Conclusion

The paper presented an approach for predicting the construct-relevant difficulty of multiple-choice questions for a high-stakes medical licensure exam. Three classes of feature were developed: linguistic features, embeddings (ELMo and Word2Vec), and features quantifying the difficulty of items for an automatic question-answering system (IR features). A model using the full feature set outperformed five different baselines (ZeroR, Word Count, Average Sentence Length, Average Word Length in Syllables, and the Flesch Reading Ease formula) with a statistically significant reduction of RMSE of approximately one point. Embeddings had the highest predictive power, followed by linguistic features, while the IR features were ranked least useful. Largest errors occurred for very difficult items, possibly due to the skewness of the data distribution towards items with a higher proportion of correct responses. Amongst the linguistic features, all classes contributed to predicting item difficulty, with the semantic-ambiguity and cognitively-motivated features having a slightly higher predictive power.

These results indicate the usefulness of NLP for predicting the difficulty of MCQs. While this study is an early attempt toward the goal of automatic difficulty prediction for MCQs, it has both theoretical and practical importance in that it goes beyond predicting linguistic complexity and in that it has the potential to reduce cost in the testing industry. Next steps include the application of the approach to other exam content administered to a different population of examinees, as well as various improvements in the methodology.

References

Tahani Alsubait, Bijan Parsia, and Ulrike Sattler. 2013. A similarity-based theory of controlling mcq difficulty. In *e-Learning and e-Technologies in Education (ICEEE), 2013 Second International Conference on*, pages 283–288. IEEE.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2015. Candidate evaluation strategies for improved difficulty prediction of language tests. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Peter Clark and Oren Etzioni. 2016. My computer is an honor student - but how intelligent is it? standardized tests as a measure of ai. *AI Magazine*, 37:5–12.

E. B. Coleman. 1965. *On understanding prose: some determiners of its complexity*. National Science Foundation, Washington, D.C.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

William H. Dubay. 2004. *The Principles of Readability*. Impact Information.

R. Flesch. 1948. A New Readability Yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Thomas François and Eleni Miltsakaki. 2012. Do nlp and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 49–57. Association for Computational Linguistics.

Robert Gunning. 1952. *The technique of clear writing*. McGraw-Hill, New York.

Le An Ha and Victoria Yaneva. 2018. Automatic distractor suggestion for multiple-choice tests using concept embeddings and information retrieval. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 389–398.

Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question difficulty prediction for reading problems in standard tests. In *AAAI*, pages 1352–1359.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. *Derivation of new readability formulas (Automatic Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*, 8-75 edition. CNTECHTRA.

Walter Kintsch and Douglas Vipond. 2014. Reading comprehension and readability in educational practice and psychological theory. *Perspectives on learning and memory*, pages 329–365.

Jing Liu, Quan Wang, Chin-Yew Lin, and Hsiao-Wuen Hon. 2013. Question difficulty estimation in community question answering services. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 85–90.

Anastassia Loukina, Su-Youn Yoon, Jennifer Sakano, Youhua Wei, and Kathy Sheehan. 2016. Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3245–3253.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Harry G. McLaughlin. 1969. SMOG grading - a new readability formula. *Journal of Reading*, pages 639–646.

Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

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

Farah Nadeem and Mari Ostendorf. 2017. Language based mapping of science assessment items to skills. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 319–326.

Ulrike Padó. 2017. Question difficulty–how to estimate without norming, how to use for automated grading. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Peri L Schuyler, William T Hole, Mark S Tuttle, and David D Sherertz. 1993. The umls metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2):217.

R. J. Senter and E. A. Smith. 1967. Automated Readability Index. Technical Report AMRL-TR-6620, Wright-Patterson Air Force Base.

Quan Wang, Jing Liu, Bin Wang, and Li Guo. 2014. A regularized competition model for question difficulty estimation in community question answering services. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1115–1126.