Experimental Evaluation of Similarity Measures for Educational Items

Jaroslav Čechák
Masaryk University
Brno, Czech Republic
xcechak1@fi.muni.cz

Radek Pelánek
Masaryk University
Brno, Czech Republic
pelanek@fi.muni.cz

ABSTRACT

Measuring similarity of educational items has several applications in the development of adaptive learning systems, and previous research has already proposed a wide range of similarity measures. In this work, we provide an experimental evaluation of selected similarity measures using a large dataset. The used items are alternate-choice questions for the practice of English grammar for second language learners; the dataset contains thousands of items and over 10 million student answers. Our results provide warnings about the generalizability of results presented in EDM works: 1) the results vary significantly between knowledge components and 2) the size of available data is an important factor.

Keywords

item similarity, evaluation, generalizability

1. INTRODUCTION

Learning environments often contain thousands of educational items (questions, problems). A useful data mining contribution is to quantify the pairwise similarity of these items [9]. Such similarity measures have many applications. There are useful particularly for the management of the content, e.g., adding and deleting new items, preparing and revising explanations and hints, or deciding when to split knowledge components. Similarity measures can also be used in algorithms that guide the presentation of the content, e.g., in the presentation of error explanations, it may be useful to group similar items together; in sequencing items, we may want to avoid giving students two very similar questions in close succession. Item similarities may also be used for student modeling [6, 12].

Item similarity can be computed in many ways [9]; the basic two approaches are to use the item content data (e.g., the text of the question) and student performance data (e.g., the correctness of answers and response times). The content-based measures are, to a large degree, dependent on the specific type of data. The techniques based on student performance data are content-agnostic and widely applicable; the disadvantage is that they require (potentially large) student data. Previous research has proposed several specific measures [11, 7, 10].

In this work, we focus on the evaluation of previously proposed measures on a large and interesting dataset. The used items are alternate-choice questions for the practice of English grammar for second language learners (see examples in Table 1). The dataset contains thousands of items, which are categorized into knowledge components and difficulty levels. The items are alternate-choice questions, i.e., they consist of a stem, correct answer, and a single distractor. Items also have explanations, which are written in the Czech language. The dataset contains approximately 10 million student answers.

For this dataset, we evaluate various similarity measures and explore their relations. We focus particularly on the relation between performance-based measures and measures based on the text of explanations. We explore the issue of the sufficient size of data on student performance. In EDM research, this issue is often neglected; the performance of techniques is often studied using a fixed dataset ("all available data"). Our experiment shows that the studied methods are quite data-hungry; they require thousands of answers per item and the amount of available data seems to be more important than differences caused by choice of a measure (which is a type of result common with other machine learning applications [2, 4]). Experiments also show large differences in results between different knowledge components, even though all of these knowledge components come from a single domain (English grammar) and all the used items are of the same, simple format (alternate-choice questions). This result provides a warning about the generalizability of research results in educational data mining.

2. EXPERIMENTAL SETTING

In this section, we describe the data we used for experiments and the specific similarity measures.

2.1 Data

For the evaluation, we use data from the adaptive learning system Umíme anglicky, `umimeanglicky.cz`. The system contains various exercises for English grammar and vocabulary learning for second language learners (for Czech native speakers). We use only one type of exercise—alternate

Table 1: Examples of items from the knowledge component *Present simple vs. present continuous.* For the sake of readability, explanations are given here in English; in the used data, they are in the Czech language.

item stem	correct	distractor	explanation
I _ to the gym once a week.	go	am going	When talking about periodical events, we use present simple tense.
I _ the film that we're watching.	hate	am hating	The verb to hate is not used in continuous form. We use present simple form instead.
I can't hear you! Everybody _ so loudly.	is talking	talks	When the activity is still in progress, we use present continuous tense.

choice question of the form fill-in-the-blank with two options (the correct answer and a distractor). The number of options is not crucial and our analyses could also be applied to questions with multiple distractors. The questions have explanations (in the Czech language).

The questions are divided into *item sets*. Each item set contains questions of similar difficulty from a single knowledge component. The system uses three difficulty levels. An example of an item set is *Present simple vs. present continuous, medium difficulty*, for which examples of questions are provided in Table 1.

Our dataset consists of 54 knowledge components divided into 68 item sets that in total contain 4 348 items. Some item sets share the same knowledge component, and they only differ in the difficulty of items. Concerning student performance, we use the answer (correct or incorrect) and response time (measured in milliseconds). We have 9 752 957 answers from 151 904 students.

Since details of data collection can often have a nontrivial impact on the results of the evaluation [8], we provide a basic description of the core aspects of system behavior that influence the collected data:

- In the system, students answer a sequence of items from a single item set in random order.
- The system uses mastery learning on the level of item sets. Students are motivated to answer a sufficient number of items correctly to satisfy the mastery criterion.
- The choice of an item set that a student solves can be done in a variety of ways: student free choice, assignment by a teacher (homework, assignment within a class), or recommendation by the system (based on past activity).
- The item sets differ widely in their difficulty. The samples of solvers may differ significantly for individual item sets (e.g., *Second conditional, hard* is solved by more advanced students than *Present simple tense, easy*).
- Items may move between difficulty levels ("design level adaptivity" [1]). This aspect may be important for some measures.

2.2 Similarity Measures

In our experiments, we use similarity measures that are variations on previously studied measures [9].

2.2.1 *Measures Based on Item Content*

One type of measure utilizes that available data about items. One possibility is to utilize item statements, e.g., to measure the similarity of item texts or match on options (the correct answer and distractor). In the case of grammar learning, this approach is hard to use: two questions that practice the same grammar rule can have completely different texts, answers, and distractors. We have performed preliminary experiments with various measures based on item text; these experiments showed very weak results. Therefore, we do not discuss these measures in more detail.

A more applicable content data are explanations. In the used dataset, each item has an associated explanation shown as feedback to students (particularly when they make a mistake). To quantify similarity based on explanations, we compute the text similarity of the explanations. To do so, we considered two common methods: Levenshtein edit distance [5] and Jaccard index.

Both methods compute the pairwise similarity of two explanations. Levenshtein edit distance operates at the character level and computes the minimal number of edits (character addition, removal, and substitution) required to transform one explanation into another explanation. Jaccard index only compares sets of words appearing in the two explanations regardless of their position. It is defined as

$$\frac{|E_1 \cap E_2|}{|E_1 \cup E_2|}$$

where E_1 is a set of words in one explanation and E_2 is a set of words in another explanation.

2.2.2 *Measures Based on Student Performance*

For computing similarity based on student performance, we consider two basic aspects: the correctness of answers and response times. These aspects are easy to collect and relevant for a vast range of items. In our experiments, we use similarity measures based on either of the two types of data and their combination.

Answer Correctness. The correctness of a student's answer is a simple binary indication of whether the student has answered an item correctly (selected the correct option

Table 2: Agreement matrix for items i and j. Values a, b, c, and d are numbers of students that answered both items in a particular way. For example, c is number of students that answered item i correctly but answered item j incorrectly.

$n = a + b + c + d$		item i	
		correct	incorrect
item j	correct	a	b
	incorrect	c	d

$$S_p = \frac{(ad - bc)}{\sqrt{(a+c)(a+b)(b+d)(c+d)}}$$

$$S_c = \frac{(P_o - P_e)}{(1 - P_e)}$$

$$P_o = \frac{(a+d)}{n}$$

$$P_e = \frac{((a+b)(a+c) + (b+d)(c+d))}{n^2}$$

$$S_{kl} = \frac{(ad - bc)}{(a+c)(c+d)}$$

in our case). Similarity measures based on the answer correctness then measure "agreement" between answers given by the same students to different items. This is best illustrated on an agreement matrix for two items i and j. There are only four possible ways a student can make binary responses to two items, as illustrated in Table 2. Similarity measures then differ in how exactly they compute the agreement from the individual components of the matrix. In our experiments, we use Pearson correlation coefficient (S_p), Cohen's Kappa (S_c) [3], and Kappa Learning [7] (S_{kl}).

Answer correctness measures can be extended by including a "second step" [9], i.e., computing similarity of similarities. In the first step, binary vectors of student answers for two items are compared to obtain the two items' similarity. The result is a similarity matrix with real-valued elements $s_{i,j}$ equal to the similarity of items i and j. The second step compares real-valued vectors $s_{i,*}$ and $s_{j,*}$ to obtain similarities of items i and j. In our experiments, we use Pearson-Pearson which is a Pearson correlation coefficient used in both first and second step.

Response Time. Response time is measured as the time it takes a student to answer the item (read the item statement and click on one of the options in our case). Student response times can vary due to external distractions during answering or even technical reasons like unreliable internet connection. To make the measure more robust, we opted to bin each item's response times into percentiles. The similarity of two items i and j is then measured as Pearson correlation coefficient of student response time percentiles vectors for items i and j.

Combined. Both correctness and response time can be combined to extract more bits of information. There are multiple ways to combine correctness and response time into a single score [9]. In our experiments, we use linear time transformation for correct answers as a combined score defined as $r = c \cdot max(1 - t/2\tau), 0)$ where $c \in \{0, 1\}$ is correctness, $t \in \mathbb{R}^+$ is response time, and τ is the median time for a given item. Similarities of items i and j are then Pearson correlation coefficient of score vectors for items i and j.

Table 3: Overview of all item similarity measures used in this study.

name	measure type	data used
Levenshtein edit distance	content	explanations
Jaccard index	content	explanations
Pearson corr. coef.	performance	correctness
Cohen's Kappa	performance	correctness
Kappa Learning	performance	correctness
Pearson-Pearson	performance	correctness
Response time percentile	performance	response time
Response time score	performance	correctness + response time

3. RESULTS

In this section, we present our findings. We use the explanations as "ground truth" for item similarity. The reasoning is that explanation describes the aspect of knowledge component that the item is practicing, and similar aspects are described in a similar way (e.g., same tense or conditional). This approach has its limitations, and it is heavily dependent on the quality of explanations. Not all explanations are necessarily ideal (different granularity between knowledge components, human errors), but it is a reasonable proxy.

For intuition behind the performed evaluation, Figure 1 provides an illustration using two knowledge components. The figure shows a PCA projection of items into plain based on the Pearson similarity measure that uses only the correctness of answers. The color of points is based on the explanations provided in the system. As we can see, these two approaches to measuring item similarity to a large degree agree—the points with the same color (similar with respect to explanations) are close to each other (similar with respect to performance). We now explore these relations in a more qualitative manner.

3.1 Relations Among Measures

Table 3 provides an overview of measures introduced in Section 2.2. Other measures can be defined in a similar fashion. An obvious question is whether they differ in any significant way or measure the same thing. To explore relations among measures, we first look at how much they are correlated. The correlation of two measures is computed as the Pearson correlation coefficient of item similarity matrices, each produced by applying item similarity measure to all pairs of items. A high correlation of two measures means that they generally agree on which pairs of items are similar.

Figure 2 shows correlations among measures based on performance and explanation averaged across all item sets. Both explanation-based item similarity measures are strongly correlated, and they also have comparable correlations with all performance-based measures. Therefore, it is not important

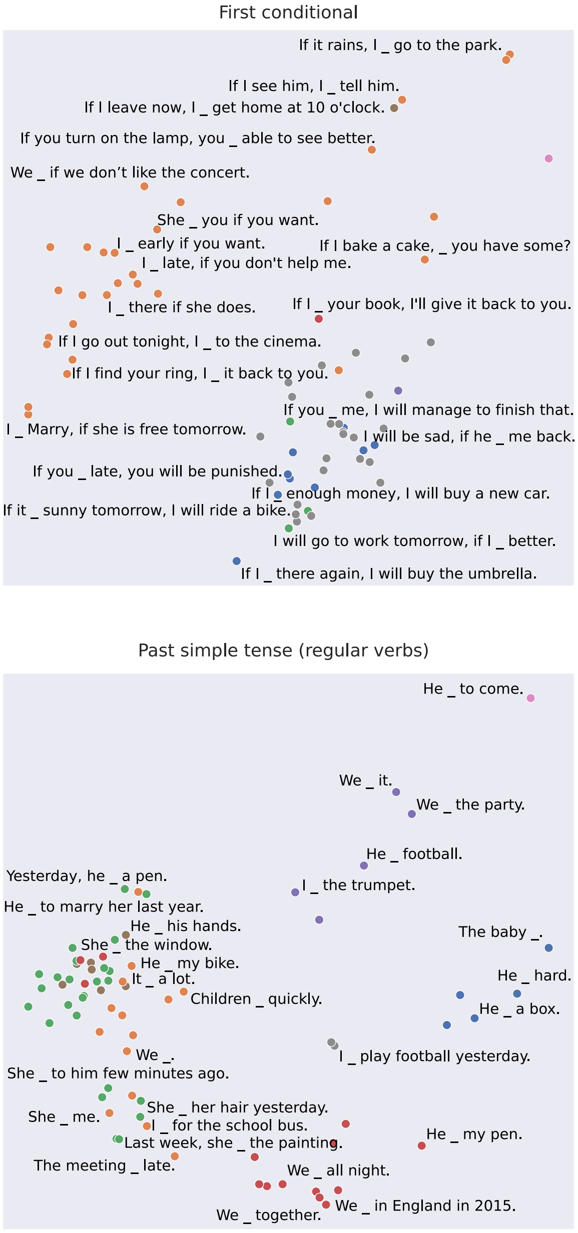

Figure 1: PCA projections based on measures using performance data (Pearson correlation). Points with the same color share the same explanations.

which one we choose as the ground truth for later experiments. This result is not surprising as both measures quantify text similarity, albeit in a different way.

Answer correctness measures Cohen's Kappa and Pearson behave almost identically, and their correlations across item sets are 0.96 or higher. The Kappa Learning measure also behaves similarly and has high correlations with both measures dropping below 0.75 only for one item set. When compared to explanation-based measures, all three measures achieve the same result. In most cases, it is not important which of the three we choose, and the amount of available data is a much more important factor (more details in Section 3.2). This result is in contrast to previous research [7], which argued that the Kappa Learning measure brings important improvement.

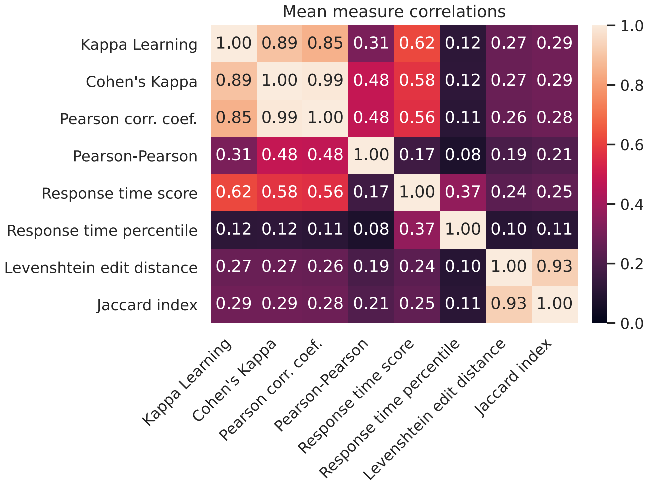

Figure 2: Heatmap of correlation among measures averaged across 68 item sets.

The second step similarity Pearson-Pearson has mostly the same or worse correlation with explanation-based measures compared to the previous three measures. It is related to Pearson and Cohen's Kappa, with correlation ranging from 0.3 to 0.8 for most item sets. The correlation with explanation-based measures is weaker compared to other measures using correctness. Thus for the used dataset, the second step does not seem useful. This observation is in contrast to previous research in another context [11].

The measures with response time do not provide any tangible benefits. When compared to explanation-based measures, they achieve either similar correlations in case of Response time score or very poor and mostly zero correlation in case of Response time percentile. A combination of answer correctness and response time in Response time score results in the best correlation for some item sets, but it is not significantly different on average. These results suggest that answer correctness might be a better indication of item similarity for our dataset.

3.2 Size of Data

Item similarity measures based on student performance are based on statistics of student performance data. All statistics need at least some amount of data to become stable and to start approximating the true statistical feature of the underlying data generating process. The question is then, how much data, i.e., answers per item, is required to obtain a good stable approximation?

In Figure 3, we have visualized the stability of performance-based measures in terms of correlation with the explanation-based measure. To simulate different numbers of answers, we have started with knowledge components with a sufficient amount of data and randomly subsampled each item's answers. We report correlation with an explanation-based measure; we report only the Jaccard index as it is highly cor-

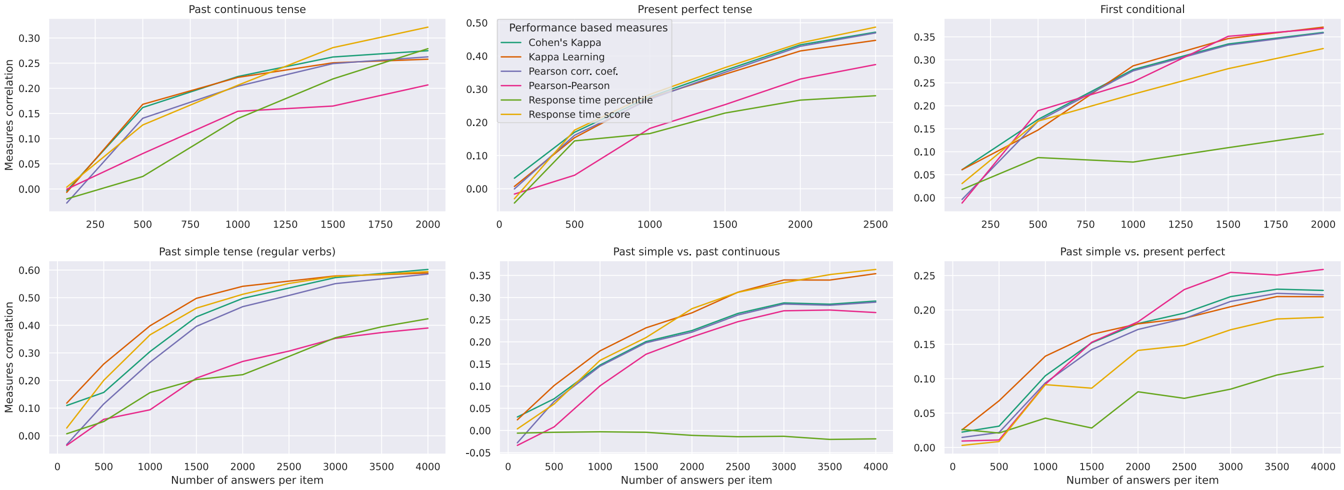

Figure 3: Correlation between performance-based measures and Jaccard index with an increasing number of answers per item across multiple knowledge components. Note that y-axis ranges differ between plots.

related with Levenshtein edit distance and has higher mean correlations with performance-based measures.

Figure 3 shows that performance-based measures are data-hungry. There are nontrivial differences in correlations until 2000 answers per item, and some improvement can be observed even for more data. The general shape of the curves is mostly similar across multiple knowledge components and final achieved correlations. There are a few changes in the relative ordering of measure, but these could be partly attributed to random noise for low data quantities. Different answer correctness measures have similar correlations regardless of data available. Response time score measures utilize more information from the data, and thus we expected them to converge faster. This, however, does not happen.

3.3 Differences among Knowledge Components

There are significant differences in the best achieved correlations among knowledge components. The best correlation achieved between any performance-based measure and explanation-based measure for a given knowledge component ranges from 0.06 to 0.67. Even if we filter out item sets with fewer than 2000 answers per item, the best correlation achieved are still between 0.25 and 0.67. Moreover, the ordering of performance-based measures in terms of achieved correlation with explanation measures differs between knowledge components. For example, Response time score with Levenshtein edit distance has the best correlation 0.61 for *Present simple tense* but the same pair has the worst correlation 0.06 for *Passive voice*. Therefore, the choice of knowledge component is more significant than the choice of similarity measures.

There is a multitude of factors causing these differences. We have identified some of these factors and give examples of their effect on correlations. The identified factors are features of the knowledge component, differences in student populations, and biases in data caused by the addition of content to the system.

Features of knowledge components describe how students use the knowledge component to answer an item. One such feature is how much the component is rule-based. There are more factual components, e.g., *Past simple tense of irregular verbs*, and more rule-based components, e.g., *Past simple tense of regular verbs*. In our data, more rule-based components achieve higher correlations on average. For example, *Past simple tense of regular verbs* achieved a correlation of 0.63 while *Past simple tense of irregular verbs* achieved only a correlation of 0.32.

The difference in student populations is especially important in systems that target a wider audience. The audience of item sets in our dataset range from grades 4 to 10, and thus the student population solving each item set differ. Simpler item sets for grades 4 to 7 achieve a better correlation of performance and explanation-based measures, while more advanced item sets for grades 8 to 10 achieve lower correlations.

Our dataset comes from a system that continuously evolves and has its content modified. These modifications also include the addition of new items among existing items. This poses a challenge for measuring similarity from performance data. Groups of items with varying amounts of collected data can make recently added items artificially different from the rest. For example, item set *Past tense: questions and negative* has 63 items with around 1700 answers per item and 20 newly added items with only around 800 answers per item. The best correlation between performance- and explanation-based measures rises from 0.3 to 0.36 when we filter out newly added items.

4. DISCUSSION

In this work, we have evaluated previously proposed measures for quantifying educational items' similarity based on students' performance. We have used a large dataset from a widely used learning system. The results provide important warnings for both practitioners and researchers.

Many educational data mining techniques require a large size of data for good performance. However, research papers often do not provide any indication of what size of data is good enough. Our results show that performance-based measures are data-hungry and may require upwards of 2000 answers per item before converging. Results reported on smaller datasets thus may be misleading in some aspects. Note that even a large university class would mean only around 200 answers per item which is still an order of magnitude smaller than the required 2000.

Another understudied issue is the generalizability of results across knowledge components. Our dataset is in many aspects very homogeneous: we consider only alternate-choice questions for English grammar. Nevertheless, there are nontrivial differences between the knowledge components (rule-based vs. fact-based, simple vs. advanced), and we have observed significant differences in results depending on the choice of a knowledge component. This observation raises a question of the generalizability of results reported on just a few knowledge components.

5. REFERENCES

[1] V. Aleven, E. A. McLaughlin, R. A. Glenn, and K. R. Koedinger. Instruction based on adaptive learning technologies. *Handbook of research on learning and instruction*, pages 522–560, 2016.

[2] M. Banko and E. Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pages 26–33, 2001.

[3] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[4] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

[5] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

[6] Q. Liu, Z. Huang, Y. Yin, E. Chen, H. Xiong, Y. Su, and G. Hu. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 33(1):100–115, 2019.

[7] T. Nazaretsky, S. Hershkovitz, and G. Alexandron. Kappa learning: A new item-similarity method for clustering educational items from response data. *International Educational Data Mining Society*, 2019.

[8] R. Pelánek. The details matter: methodological nuances in the evaluation of student models. *User Modeling and User-Adapted Interaction*, 28(3):207–235, 2018.

[9] R. Pelánek. Measuring similarity of educational items: An overview. *IEEE Transactions on Learning Technologies*, 13:354–366, 2020.

[10] R. Pelánek, T. Effenberger, M. Vaněk, V. Sassmann, and D. Gmiterko. Measuring item similarity in introductory programming. In *Proc. of Learning at Scale*. ACM, 2018.

[11] J. Řihák and R. Pelánek. Measuring similarity of educational items using data on learners' performance. In *Educational Data Mining*, pages 16–23, 2017.

[12] S. Zhao, C. Wang, and S. Sahebi. Modeling knowledge acquisition from multiple learning resource types. *arXiv preprint arXiv:2006.13390*, 2020.