# Towards Automatic Short Answer Assessment for Finnish as a Paraphrase Retrieval Task

**Li-Hsin Chang, Jenna Kanerva, and Filip Ginter**
TurkuNLP Group
Department of Computing
Faculty of Technology
University of Turku, Finland
{lhchan, jmnybl, figint}@utu.fi

## Abstract

Automatic grouping of textual answers has the potential of allowing batch grading, but is challenging because the answers, especially longer essays, have many claims. To explore the feasibility of grouping together answers based on their semantic meaning, this paper investigates the grouping of short textual answers, proxies of single claims. This is approached as a paraphrase identification task, where neural and non-neural sentence embeddings and a paraphrase identification model are tested. These methods are evaluated on a dataset consisting of over 4000 short textual answers from various disciplines. The results map out the suitable question types for the paraphrase identification model and those for the neural and non-neural methods.

## 1 Introduction

Computer-assisted assessment brings about the promise of alleviating the workload of teachers, allowing them to concentrate manual efforts towards more creative pedagogical tasks. Not all assessment types, however, have widely adopted fully automated or computer-assisted grading methods. Essays, for example, are a common way to evaluate student knowledge, but are resource-demanding to grade. An angle to automatic essay evaluation is to group together similar essays for batch grading, but this is complicated by the complex structure of essays. Short answers, on the other hand, often consist of only one or a few claims, and thus represent a desirable starting point for textual answer clustering. In addition to being a simplified target for studying textual answer clustering, short answers are also a common form of assessment; Very short answer questions have been shown to have desirable traits of reliable assessments, such as the scores showing a fair and balanced distribution (Puthiaparampil and Rahman, 2020).

Automated short answer assessment is used in this paper as an umbrella term to refer to computationally assisting the evaluation of short textual answers, while there is no unified definition for short textual answers (Roy et al., 2015). Whereas some impose only length restrictions on the textual answers (e.g. one phrase to one paragraph), others have additional criteria such as the answer being a natural language response, or the focus of the assessment being knowledge content instead of grammar (Burrows et al., 2015; Roy et al., 2015). In practice, the definition for short textual answers depends on the actual application, and the answers vary in terms of textual length, topic, assessment criteria, educational level of students, etc. These variations have fueled the long ongoing research on automated assessment of short textual answers. Roy et al. (2015) survey computer-assisted assessment techniques developed in the years 2000–2015 targeting short answers ranging from a sentence long to a maximum of 100 words. They suggest a matchmaking framework to guide the choice of appropriate techniques for practitioners and call for computer-assisted assessment methods that do not rely on model answers, as automated short answer grading (ASAG) systems usually do. One such alternative method is to group together semantically similar short textual answers for batch grading. This is a less explored research area but has been shown to effectively reduce the number of manual actions required for grading (Basu et al., 2013).

The essence of both ASAG and short answer grouping is how the texts are represented, and thus their research methods are influenced by the advances in semantic textual similarity (STS) and paraphrase research. Here, a typical ASAG system would measure the similarity between teacher-supplied model answer(s) and student answers, whereas short answer grouping measures and groups student answers among themselves. Apart from traditional string-based and corpus statistics-based methods, dense vector representation meth-

*Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 262 - 271
July 15, 2022 ©2022 Association for Computational Linguistics

ods based on deep learning are naturally highly applicable to the task. A typical example of such methods are Sentence-Transformers (Reimers and Gurevych, 2019) that adapt the BERT model to sentence representation by explicitly optimizing the similarity of dense-vector representation for pairs of sentences known to carry the same meaning. Such models can be applied to answer grouping in a straightforward manner by comparing the dense representations of sentences across different answers. In a different line of work, Kanerva et al. (2021b) approach paraphrase detection as a form of semantic search by training a question-answering type of a model to detect a paraphrase of a query from a given context document. This methodology can be seen as highly relevant to examining answer grouping: given an answer, or a part of an answer constituting a single claim, such model can then identify answers containing the same claim or its paraphrase among other students' answers. Such an approach would, in theory, then allow the grading teacher to retrieve all such answers and carry out a common grading action. While not eliminating manual grading work, this approach could potentially significantly reduce the need, if paired with an appropriate interface and workflow.

In this paper, we pursue this direction, approaching answer grouping from an information retrieval (IR) perspective, i.e. given an answer, or a claim from one answer, the task is to identify other answers containing the same claim or its paraphrase, not relying on the availability of model answers. The objective here is to retrieve similar answers for a given query to support e.g. batch grading. While we do not want to limit our methods to short answer assessment only, full long essays are likely too long as retrieval candidates. Rather than retrieving on essay level, the natural unit for the retrieval would be to do it on the claim level, looking for similar claims inside the essays. However, for the time being we lack any manual annotation for individual claims posed in the essays, making the evaluation of such claim-level retrieval methods difficult. Therefore, we approach the problem by using short answers only, where the answer typically includes only one or a few claims. The overall score assigned for the answer can then be used as a proxy of claim similarity, as all answers with high scores can be assumed to contain similar claims, even if using different wordings. We therefore formulate the overall task setup as such: Given one claim as a query (in the form of a short answer), how well the experimented models are able to retrieve a similar claim among all candidates answering the same prompt (here "prompt" refers to the question posed by the teacher to which the students are answering) when judging the similarity based on the scores assigned to the answers. We use a dataset of over 4,000 teacher-graded short answers from actual university examinations of 24 distinct courses. We test non-neural and neural sentence embedding methods as well as the above-mentioned question answering -based paraphrase retrieval model, and map which types of questions are suitable for what types of answer grouping methods.

## 2 Related work

The most researched direction for automated evaluation of short textual answers is automatic short answer grading (ASAG). This research field has seen the application of rule-based, machine learning, and deep learning methods (Burrows et al., 2015; Roy et al., 2015; Bonthu et al., 2021). ASAG is typically modelled as a supervised learning task and seen as either a classification or a regression task, where a student answer is compared to a model answer, and the output label or score is based on their similarity. Consequently, model answers are usually required for these systems. Camus and Filighera (2020) test the performance of various Transformer-based (Vaswani et al., 2017) language models on the SemEval-2013 dataset (Dzikovska et al., 2013), one of the most common ASAG dataset. They find that a Robustly Optimized BERT Pretraining Approach (RoBERTa)-large model (Liu et al., 2019) fine-tuned on the Multi-Genre Natural Language Inference (MNLI) dataset (Williams et al., 2018) performs best.

Short answer grouping is a less explored research direction, where short textual answers are grouped together based on their similarity. Basu et al. (2013) use a feature-based similarity metric to group short textual answers into hierarchical clusters. Their features include i.a. difference in length, fraction of words with matching base forms, and cosine-similarity of TFIDF vectors. They show that such clustering can effectively reduce the number of actions required for grading. Hämäläinen et al. (2018) use the Hyperlink-Induced Topic Search (HITS) algorithm (Kleinberg, 1999) to cluster open-ended questionnaire answers from students. Applying this method to both English and Finnish datasets, they obtain satisfactory results on the English dataset but less ideal results on the Finnish dataset, potentially due to the Finnish answers being longer in length. Both the study of Basu et al. (2013) and Hämäläinen et al. (2018) predate the era of deep neural network-based methods of meaning representation.

## 3 Data

Our experiments are based on a large scale dataset of over 261K anonymized textual answers from different university-level examinations. However, for the purpose of this study, the dataset is heavily filtered in order to obtain a subset including only examples considered as short answers suitable for the study. We aim to find prompts looking for concise fact-based descriptions, which are likely to contain only a single claim and therefore have an increased likelihood that two answers with a high score are likely to be paraphrases of each other (although that naturally cannot be guaranteed without manual annotation). Such suitable prompts ask for example term definitions, listings of the components of certain concepts, explanation of the workings of a process or device, explanations why e.g. a German noun is of a certain gender, or basically anything targeting to a short semantic-focused answer. In addition to the proper answer content, we also need the prompts to fit to our retrieval task setting, meaning that for each unique prompt, we need to have several student answers as retrieval candidates. One such example prompt together with few graded student answers for it is given in Table 1.

The original dataset is a collection[1] of 261K student answers gathered across various disciplines in the University of Turku, Finland. Together with the textual answers, the data include the course identifier, question prompt, assigned score, and possible score range for each answer. The textual answers are written by mainly undergraduate students, and the most common languages are Finnish and English. Figure 1 illustrates the data filtering process. The filtering criteria for identifying a suitable short answer subset for this study are as follows: the prompt length must be under 10 tokens and the answer length under 30 tokens as determined based on the FinBERT model tokenizer[2], and the language of the answer must be Finnish. All answers with 0 as the highest possible score are excluded, as these are often dummy prompts related to course feedback, assignment submission, or attendance rather than being actual exam questions. Additionally, due to the retrieval task setup used, each prompt included in the subset must have at least 10 answers passing the above-mentioned filtering in order to have enough retrieval candidates in the experiments.

After the automatic filtering, some amount of manual cleaning is also used to remove answers and prompts unsuitable for the experiments. These mostly include prompts from language courses targeting to grammatical correctness rather than semantics (therefore including very little variation), prompts asking the students to name parts of a figure, or occasional dummy prompts that passed the zero score filter.

Statistics of the final filtered subset are summarized in Table 2, the final dataset including prompts from 24 different courses and 12 different disciplines. In total, there are 4,082 student answers. The disciplines of the courses are otherwise evenly distributed, except for life sciences, which has 9 courses with 93 prompts and 2523 answers, accounting for more than half of the obtained short answers. On average, each prompt has about 24 different answers. The maximum number of answers a prompt has is 75, while 22 prompts pass the filter with the minimum of 10 answers. Since the highest possible score varies across courses and prompts, the assigned scores of each answer are normalized to a range of 0–1 with respect to the highest possible score. For pass-fail questions, scores of passed answers are converted to 1 and the failed ones 0. The normalized score distribution of the short answers is shown in Table 3.

## 4 Experiments

The grouping of semantically similar answers is approached from an IR point of view. For each answer, the answer itself is considered the query and all the other answers to the same prompt are considered the documents. This is repeated for every answer of a prompt. Three methods are tested for retrieval: TFIDF, Sentence-Transformers, and the paraphrase span detection model (Kanerva et al., 2021a). The grade is used as a proxy allowing for method comparison: intuitively, a correct retrieval i.e. an answer which paraphrases the answer used

[1] The nature of student examination answer data unfortunately precludes its free distribution.

[2] https://huggingface.co/TurkuNLP/bert-base-finnish-cased-v1

**Prompt: digital legacy**

| Score | Answer |
|---|---|
| 1.0 | A digital legacy is all the files and data about a person that remain on the internet or the digital world after the death of the person. |
| 1.0 | The trace that we leave behind digitally when we die (e.g. files, digital photos and usernames). |
| 1.0 | All the digital material that remains of a person after death. Digital legacies include for examples passwords, usernames and photos of the deceased person. |
| 0.5 | Traces left by the user of a computer or other technological device. What websites they have visited and what software they have on their device. |
| 0.5 | Any data a person leaves behind on the Internet or other computer systems. |
| 0.0 | All the things that were born in the digital form. |
| 0.0 | Digital legacy means electronic waste, often exported to the third world. |
| 0.0 | The evolutionary trajectory of digital devices. |

Table 1: An illustrative example of one example prompt together with few student answers for it translated into English.

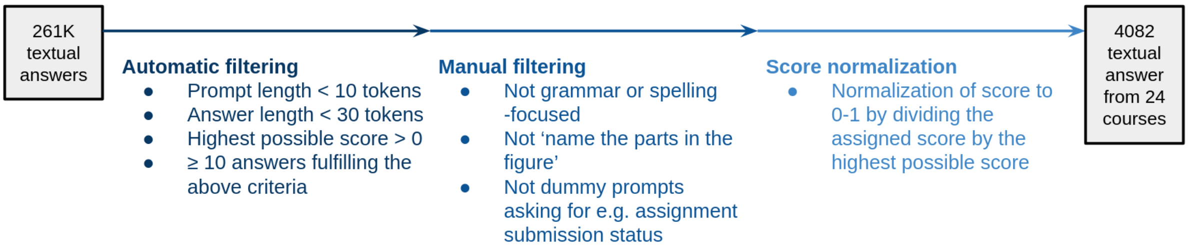
261K textual answers
Automatic filtering
Prompt length < 10 tokens
Answer length < 30 tokens
Highest possible score > 0
≥ 10 answers fulfilling the above criteria
Manual filtering
Not grammar or spelling -focused
Not 'name the parts in the figure'
Not dummy prompts asking for e.g. assignment submission status
Score normalization
Normalization of score to 0-1 by dividing the assigned score by the highest possible score
4082 textual answer from 24 courses

Figure 1: Illustrative diagram of the data filtering process.

| | Courses | Prompts | Answers |
|---|---|---|---|
| **Full dataset** | | | |
| Total | 1,745 | — | 261K |
| **Filtered subset** | | | |
| Communication | 1 | 1 | 10 |
| Computer sciences | 1 | 14 | 393 |
| Economics | 1 | 3 | 37 |
| Educational sciences | 2 | 6 | 62 |
| German | 1 | 14 | 172 |
| Information systems science | 1 | 11 | 437 |
| Life sciences | 9 | 93 | 2523 |
| Media research | 1 | 1 | 10 |
| Medicine | 2 | 3 | 33 |
| Philology | 1 | 6 | 86 |
| Philosophy | 1 | 5 | 65 |
| Psychology | 3 | 14 | 254 |
| Total | 24 | 171 | 4,082 |

Table 2: Statistics of the filtered short answers dataset used in this study.

| Normalized score | Occurrence |
|---|---|
| 0.0 | 754 |
| 0.25 | 53 |
| 0.5 | 298 |
| 0.75 | 137 |
| 1.0 | 2792 |

Table 3: Occurrences of the normalized score of 4082 short answers. 15 values between the range of 0-1 are omitted in the table due to low (<10) occurrences.

as the query, should have the same grade. Consequently, a method which is better at the retrieval task should, on average, be more likely to retrieve answers with the same score as the query than a method which is worse at the retrieval task. As we are mostly interested in relative method performance, we measure and report the success of the retrieval by top-1 accuracy and R-precision. The relevance of the retrieval is binary, meaning that retrieval with matching grade to the query is counted as "correct", and otherwise "incorrect".

### 4.1 TFIDF

The term frequency–inverse document frequency (TFIDF) represents a commonly used family of IR metrics based on lexical overlap. TFIDF estimates the importance of a word in a document by the number of times it appears in the document, and the inverse of the number of documents the word appears in a document collection. It generates sparse high-dimensional vectors without inherent similarity between words.

For our experiments, TFIDF representation is generated for every short answer. The TFIDF representation of an answer is calculated from the entire collection of over 201K Finnish textual answers. The features used are the ngrams (n=2–5) of character within word boundaries. The short answers are used as-is, without stop word removal or lemmatization because these processing did not improve the results in our preliminary experiments.

### 4.2 Sentence-Transformers

Sentence-Transformers are trained from language models such as BERT or XLM-R (Conneau et al., 2020) using Siamese or triplet networks to induce sentence encoders whose representation can be compared using cosine similarity (Reimers and Gurevych, 2019). The resulting representations are dense, low-dimensional, and context-sensitive. For our experiments, two Sentence-Transformer models available on HuggingFace (Wolf et al., 2019) are tested: `sbert-cased-finnish-paraphrase` and `paraphrase-xlm-r-multilingual-v1` (thereafter SBERT-Finn and XLM-R←SBERT-para). The SBERT-Finn model is based on the FinBERT-base-cased model (Virtanen et al., 2019), fine-tuned for an epoch on the Finnish Paraphrase Corpus (Kanerva et al., 2021a), as well as 500K of positive and 5M of negative automatically collected paraphrase pair candidates[3], with mean pooling and a classification objective. The XLM-R←SBERT-para is fine-tuned from the XLM-RoBERTa-base model (Conneau et al., 2020) to mimic the embeddings of the English Sentence-BERT (Reimers and Gurevych, 2020). The fine-tuning uses a teacher–student framework and parallel data of over 50 languages. The resulting model was reported to outperform multiple competitive baselines on the multilingual semantic textual similarity 2017 dataset (Cer et al., 2017).

### 4.3 Span detection model

Treating paraphrase recognition as a span detection task, Kanerva et al. (2021b) train FinBERT models to paraphrase recognition taking inspiration from the question answering task, where given a query, a question answering model retrieves a span out of a given document as the answer to the query. Instead of retrieving answers, the paraphrase span detection model takes in a query and identifies a span from the given document that paraphrases the query. The models are trained on the Finnish Paraphrase Corpus, which includes not only the paraphrase pairs but also their context documents where each paraphrase statement originally occurred. They train two flavors of models, one with only positive examples always selecting a span from the given document, and the other being able to produce a null span, indicating that no paraphrase of the query can be detected from the given document.

For our experiments, an answer of a prompt is used as the query and all other answers from the same prompt are concatenated as the context document, as shown in Figure 2. The model produces candidate spans that it detects as paraphrases of the query, and the most likely prediction is selected as the final retrieval. The full model that also predicts null spans is used as there may not always be other answers that are semantically similar to an answer. The model produces several (start-of-span, end-of-span) candidates sorted based on an assigned probability score for each. The model is modified so that the probability is always calculated for a whole answer, instead of arbitrary spans. The retrieved spans can be considered as all the predictions ranked before the null span.

### 4.4 Evaluation metrics

Top 1 accuracy measures if the first retrieved document (an answer to the same prompt as a query) is correct, i.e. if it has the same grade/score as the query. Top 1 accuracy allows for quick understanding of how well the method roughly works, though it does not take into account the expected value of a random retrieval (e.g. if all the answers to the prompt score the same, the accuracy is high no matter what the model retrieves), nor how close numerically the score of the retrieval is to that of the query, if they are not equal. The course-wise top 1 accuracy is reported, which is calculated as the arithmetic average of the prompt-wise top 1 accuracy. The prompt-wise top 1 accuracy is in turn calculated from the arithmetic average of the top 1 accuracy of all the queries answering the same prompt. For the span detection model, a null prediction is ignored for the calculation of top 1 accuracy. That is, the first non-null prediction is taken if the first prediction is null.

Since the grades of all the answers are available,

[3] https://turkunlp.org/paraphrase.html

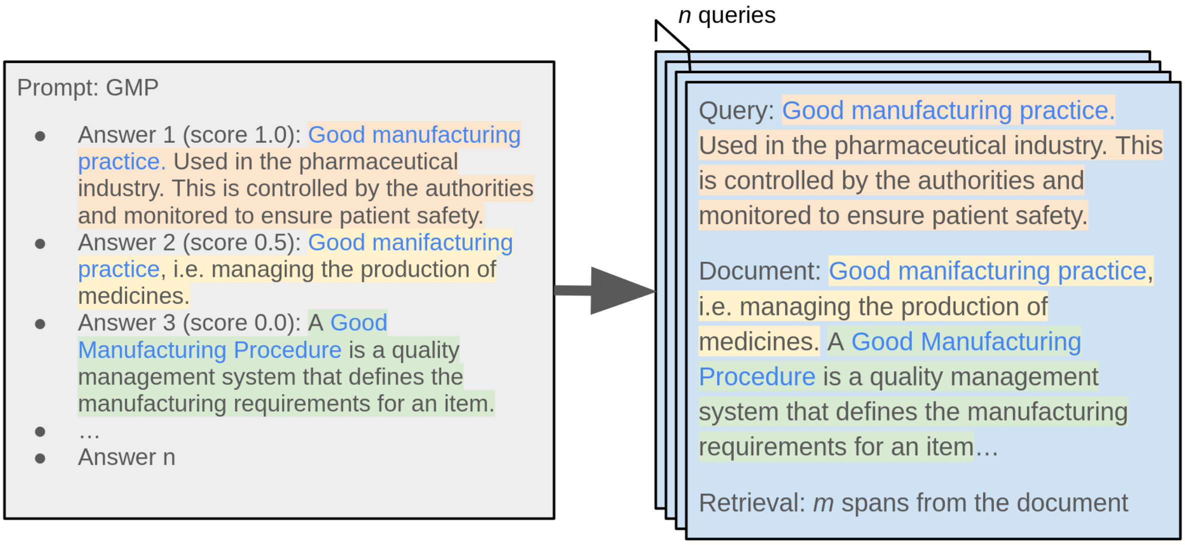
n queries
Prompt: GMP
Answer 1 (score 1.0): Good manufacturing practice. Used in the pharmaceutical industry. This is controlled by the authorities and monitored to ensure patient safety.
Answer 2 (score 0.5): Good manifacturing practice, i.e. managing the production of medicines.
Answer 3 (score 0.0): A Good Manufacturing Procedure is a quality management system that defines the manufacturing requirements for an item.
…
Answer n
Query: Good manufacturing practice. Used in the pharmaceutical industry. This is controlled by the authorities and monitored to ensure patient safety.
Document: Good manifacturing practice, i.e. managing the production of medicines. A Good Manufacturing Procedure is a quality management system that defines the manufacturing requirements for an item…
Retrieval: m spans from the document

Figure 2: Illustration of the span detection setup. The blue text is in its original language while the black text has been translated from Finnish.

the total number of relevant documents is known. This allows for the calculation of R-precision, the number of relevant documents in the first $R$ retrievals, where $R$ is the total number of relevant documents for a query. R-precision is also equal to recall with $R$ as cutoff. As with the top 1 accuracy, null spans are ignored for the calculation of R-precision in the case of the span detection model.

## 5 Results

The top 1 accuracy and R-precision of the methods across the 24 courses are shown in Tables 4 and 5 respectively. Courses numbers 8 and 17 have values of 1.0 on both metrics for all methods because both of them have 10 answers to a prompt where all students answer correctly. Excluding these two courses, the span detection model scores the best or equally the best with another method on 11 and 12 courses respectively on top 1 accuracy and R-precision, outperforming the other methods. SBERT-Finn performs well in terms of R-precision on the life sciences discipline, performing the best on 8 out of 9 courses. The numerical differences of the accuracy values among these four methods are oftentimes minimal, and we investigate the ones with bigger differences to establish whether certain kinds of prompts are particularly suitable for a given method. We observe that the neural representation is advantageous when the prompts are challenging, which leads to the students inventing plausible answers using the keywords. An example of a query from a prompt where the TFIDF method underperforms the neural method by a large margin (0.4 vs. 0.7) is shown in Table 6. This prompt is challenging not only because it requires the recollection of certain principles, but also that there are multiple key points the students have to make to obtain a full score.

Compared to the other methods, the span detection model performs well on retrieving relevant answers, but it also assigns relatively high probabilities to null spans. When using the position of the null span as cutoff instead of the number of relevant documents, we observe that the span detection model scores the best or equally the best on only 6 out of 24 courses, whereas TFIDF, SBERT-Finn and XLM-R←SBERT-para achieve 10, 14, and 7 respectively[4].

## 6 Discussion

In this paper, the span detection model is forced to only predict the probabilities of whole documents being paraphrases of the query. If this restriction is removed, the span detection model is capable of predicting arbitrary spans as the paraphrases of the query. This becomes relevant when obtaining the full score requires mentioning of multiple claims. For example, if a prompt asks students to explain abbreviations, a full scoring answer requires the student to provide the full form of an abbreviation and explain what it means. In our initial experi-

[4] This result is not shown, since the cutoff is only meaningful for the span detection model, and its application to the other methods is merely for comparison purposes

| Course ID | Discipline | TFIDF | SBERT-Finn | XLM-R←SBERT-para | Span detection | No. prompts | No. queries |
|---|---|---|---|---|---|---|---|
| 1 | communication | **0.6** | **0.6** | 0.3 | **0.6** | 10 | 1 |
| 2 | computer sciences | 0.619 | 0.622 | **0.654** | 0.649 | 393 | 3 |
| 3 | economics | 0.30 | **0.39** | **0.39** | 0.31 | 37 | 2 |
| 4 | educational sciences | 0.3 | 0.3 | 0.1 | **0.4** | 21 | 1 |
| 5 | educational sciences | 0.49 | **0.56** | 0.44 | **0.56** | 41 | 2 |
| 6 | German | **0.84** | 0.81 | 0.76 | 0.82 | 172 | 2 |
| 7 | information systems science | 0.471 | 0.474 | 0.458 | **0.489** | 437 | 3 |
| 8 | life sciences | **1.0** | **1.0** | **1.0** | **1.0** | 10 | 1 |
| 9 | life sciences | **0.97** | **0.97** | 0.91 | 0.91 | 32 | 2 |
| 10 | life sciences | 0.49 | 0.46 | 0.36 | **0.54** | 33 | 2 |
| 11 | life sciences | 0.855 | **0.867** | 0.859 | 0.864 | 748 | 3 |
| 12 | life sciences | 0.852 | 0.841 | 0.853 | **0.864** | 365 | 3 |
| 13 | life sciences | 0.89 | 0.88 | 0.88 | **0.90** | 198 | 2 |
| 14 | life sciences | **0.83** | 0.76 | 0.75 | 0.75 | 114 | 2 |
| 15 | life sciences | 0.788 | 0.794 | 0.779 | **0.800** | 990 | 3 |
| 16 | life sciences | **0.74** | 0.70 | 0.68 | 0.65 | 33 | 2 |
| 17 | media research | **1.0** | **1.0** | **1.0** | **1.0** | 10 | 1 |
| 18 | medicine | 0.6 | **1.0** | 0.8 | **1.0** | 12 | 1 |
| 19 | medicine | 0.5 | **0.7** | 0.6 | 0.5 | 21 | 1 |
| 20 | philology | 0.73 | 0.65 | 0.68 | **0.76** | 86 | 2 |
| 21 | philosophy | 0.44 | **0.47** | 0.45 | 0.45 | 65 | 2 |
| 22 | psychology | **0.58** | 0.44 | 0.52 | 0.50 | 94 | 2 |
| 23 | psychology | 0.66 | 0.68 | **0.75** | **0.75** | 68 | 2 |
| 24 | psychology | 0.54 | **0.67** | 0.55 | 0.66 | 92 | 2 |
|  | Number of best or equal best | 8 | 11 | 5 | **13** | — | — |

Table 4: Top 1 accuracy by course. *No. prompts* refers to the number of prompts, or exam questions, in a course. *No. queries* refers to the total number of short answers in a course.

ments, we observe that the span detection model can retrieve a span out of the full answer which is semantically equivalent to a partial answer. The evaluation of such retrievals, however, is not possible given our current data without manual annotations because a full scoring answer has a different score than a partial answer, nor is there a way to attribute which sub-spans of the full answer contribute how much to the final score. The exploration of how the span detection model can be applied to answers consisting of multiple claims may pave the way to eventually automatically evaluating essays. A potential way is to combine the answers of related prompts as queries and documents. We leave this to future work.

A challenge for experimental design is the selection of suitable metrics. Top 1 accuracy has the advantage of being easily understandable and interpretable, but its calculation ignores the expected value of random retrievals. R-precision mitigates the randomness to some extend, since it takes into account the top $R$ retrievals where $R$ is the number of relevant documents. When all the documents are relevant, R-precision is always 1 and it is not immediately obvious if the model performs meaningful prediction, though this can arguable be regarded as unsuitable data for retrieval, or, from a practical point of view, the retrievals will always be relevant. The design of R-precision is not completely compatible with the nature of the span detection model, as the model predicts null, which has to be taken into account if it ranks among the top $R$. The null prediction can either be regarded as an irrelevant prediction, or ignored altogether as we have done so in this paper. The use of binary relevance means a retrieved document is either relevant if it has the same score as the query, or irrelevant if it does not. This does not take advantage of some of the scores being of higher granularity. For example, if the query scores 1 and model A retrieves a document scoring 0.7 and model B a document scoring 0.3, the retrieval of model A is likely better than that of model B. An ideal metric would thus take into account the numerical difference between the scores of the query and the retrieval, as well as the informativeness of the set of documents available for retrieval.

A class of metrics we have explored but did not eventually use is normalized discounted cumulative gain (NDCG). NDCG is a class of commonly used IR metrics, where the discounted cumulative gain, which sums the relevance of the query and retrieval (which can be graded instead of binary) discounted by the ranked position, is normalized by the ideal

| **Course ID** | **Discipline** | **TFIDF** | **SBERT-Finn** | **XLM-R←SBERT-para** | **Span detection** | **No. prompts** | **No. queries** |
|---|---|---|---|---|---|---|---|
| 1 | communication | **0.5** | 0.4 | **0.5** | **0.5** | 10 | 1 |
| 2 | computer sciences | 0.589 | 0.608 | 0.600 | **0.615** | 393 | 3 |
| 3 | economics | 0.29 | 0.38 | 0.35 | **0.44** | 37 | 2 |
| 4 | educational sciences | 0.4 | 0.4 | 0.3 | **0.5** | 21 | 1 |
| 5 | educational sciences | 0.42 | 0.48 | 0.42 | **0.54** | 41 | 2 |
| 6 | German | **0.86** | 0.82 | 0.78 | 0.80 | 172 | 2 |
| 7 | information systems science | 0.387 | **0.403** | 0.393 | 0.400 | 437 | 3 |
| 8 | life sciences | **1.0** | **1.0** | **1.0** | **1.0** | 10 | 1 |
| 9 | life sciences | 0.93 | **0.96** | 0.91 | 0.84 | 32 | 2 |
| 10 | life sciences | 0.59 | **0.63** | 0.59 | 0.62 | 33 | 2 |
| 11 | life sciences | 0.790 | **0.802** | 0.792 | 0.799 | 748 | 3 |
| 12 | life sciences | 0.809 | **0.829** | 0.817 | 0.822 | 365 | 3 |
| 13 | life sciences | **0.89** | **0.89** | 0.88 | 0.88 | 198 | 2 |
| 14 | life sciences | 0.76 | 0.74 | 0.76 | **0.77** | 114 | 2 |
| 15 | life sciences | 0.737 | **0.741** | 0.738 | 0.734 | 990 | 3 |
| 16 | life sciences | 0.60 | **0.64** | 0.59 | 0.54 | 33 | 2 |
| 17 | media research | **1.0** | **1.0** | **1.0** | **1.0** | 10 | 1 |
| 18 | medicine | 0.6 | 0.6 | 0.5 | **0.7** | 12 | 1 |
| 19 | medicine | **0.6** | **0.6** | **0.6** | **0.6** | 21 | 1 |
| 20 | philology | 0.58 | 0.58 | 0.56 | **0.60** | 86 | 2 |
| 21 | philosophy | 0.43 | 0.43 | 0.42 | **0.44** | 65 | 2 |
| 22 | psychology | 0.42 | **0.45** | 0.40 | 0.42 | 94 | 2 |
| 23 | psychology | 0.65 | 0.68 | 0.70 | **0.71** | 68 | 2 |
| 24 | psychology | 0.54 | 0.56 | 0.54 | **0.59** | 92 | 2 |
|  | Number of best or equal best | 6 | 12 | 4 | **14** | — | — |

Table 5: R-precision by course. *No. prompts* refers to the number of prompts, or exam questions, in a course. *No. queries* refers to the total number of short answers in a course.

| **Query** | **0.5** | **The central principle of processing level theories is that the quality of information is thought to be more important than its duration.** |
|---|---|---|
| **Model** | **Score** | **Top 1 retrieval** |
| TFIDF | 0.0 | In processing level theory, stimuli are processed in parts, at different levels. |
| SBERT-Finn | 0.5 | The theory is that the more information you process, the better it is remembered. The quality of processing is more important than the duration. |
| XLM-R←SBERT-para | 0.5 | The most important thing in information processing is quality, not duration. |
| Span detection | 0.5 | The most important thing in information processing is quality, not duration. |

Table 6: Example retrievals of the four methods to a query answering the prompt "Key principles of the theory of processing levels". Example of a full-scoring answer is "The quality of a process means more than its duration. The processing of meanings improves memory retention."

discounted cumulative gain (Wang et al., 2013). It is not suitable for this task, however, as the task differs from typical IR scenarios in that we have a small number of answers where the retrieval of all relevant answers are important, whereas in e.g. web search the focus is on ranking the most relevant document as high as possible.

The multilingual sentence embedding model does not outperform the non-neural baseline. This is somewhat surprising, as some of the short answers contain code-switching, such as the examples in Figure 2. This shows that language-specific sentence embeddings and models are still more suitable for this task.

The task setup is only an approximation. The same grade does not imply the query and document being paraphrases, not for high grades nor for low grades, unless the grading criteria is semantically stringent, in the cases of e.g. translation studies. However, the hope is that the noise can be mitigated by using a large dataset and some signals can be seen as to whether the models are able to retrieve semantically documents. Our results show that they indeed can.

## 7 Conclusion

In this work, we explored several methods for grouping student answers to exam prompt. In addition to the standard setup whereby whole short answers are represented as either sparse (TFIDF) or dense (Transformer) vectors and compared to one another, we also tested a more retrieval-style approach, whereby we formed documents by concatenating a number of answers to the same prompt and testing to what extent the model is able to retrieve similar answers from such documents. This approach models the case of matching individual claims in longer answers.

Unsurprisingly, we find that the dense representations are more suitable to the task. Interestingly, we find that a span detection model trained on Finnish paraphrase data performs better than sentence-level embedding comparison methods. It might therefore be fruitful to pursue models which are not restricted to apriori given sentence boundaries, and which are capable of finding individual claims in collections of potentially longer essay-style answers.

While the study is based on real exam answers from a number of courses, the data lacks manual annotation of the semantic equivalence of answers, which is challenging to produce. Further, to be able to use the grades as a proxy to retrieval evaluation, we had to restrict ourselves to short, fact-checking questions, only using a small portion of the over 200,000 answers we have at our disposal. A natural further study would expand the use of the retrieval model to longer answers and employ teachers to evaluate the retrievals provided by the model and establish the overall benefit of such approach.

## Acknowledgements

We warmly thank Totti Tuhkanen and Kaapo Seppälä for administrative support and data collection. Computational resources were provided by *CSC — the Finnish IT Center for Science*. This research was supported by the Academy of Finland and the DigiCampus consortium. We thank the anonymous reviewers for their useful comments.

## References

Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. 2013. Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, 1:391–402.

Sridevi Bonthu, S. Rama Sree, and M. H. M. Krishna Prasad. 2021. Automated short answer grading using deep learning: A survey. In *Machine Learning and Knowledge Extraction*, pages 61–78.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25:60–117.

Leon Camus and Anna Filighera. 2020. Investigating transformers for automatic short answer grading. In *Artificial Intelligence in Education*, pages 43–48, Cham. Springer International Publishing.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.

Wilhelmiina Hämäläinen, Mike Joy, Florian Berger, and Sami Huttunen. 2018. Clustering students' open-ended questionnaire answers. *CoRR*, abs/1809.07306.

Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Hanna-Mari Kupari, Jemina Kilpeläinen, Jenna Saarni, Maija Sevón, and Otto Tarkka. 2021a. Finnish paraphrase corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*.

Jenna Kanerva, Hanna Kitti, Li-Hsin Chang, Teemu Vahtola, Mathias Creutz, and Filip Ginter. 2021b. Semantic search as extractive paraphrase span detection. arXiv:2112.04886.

Jon M Kleinberg. 1999. Hubs, authorities, and communities. *ACM computing surveys (CSUR)*, 31(4es):5–es.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692.

Thomas Puthiaparampil and Md Mizanur Rahman. 2020. Very short answer questions: a viable alternative to multiple choice questions. *BMC Medical Education*, 20.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Shourya Roy, Yadati Narahari, and Om Deshmukh. 2015. A perspective on computer assisted assessment techniques for short free-text answers. pages 96–109.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. arXiv:1912.07076.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Wei Chen, and Tie-Yan Liu. 2013. A theoretical analysis of NDCG ranking measures. In *Proceedings of the 26th annual conference on learning theory (COLT 2013)*, volume 8, page 6. Citeseer.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. arXiv:1910.03771.