Finding a Tradeoff between Accuracy and Rater's Workload in Grading Clustered Short Answers

Andrea Horbach,[1] Alexis Palmer[1] and Magdalena Wolska[2]

[1]Saarland University, Saarbrücken, Germany
[2] LEAD Graduate School, Eberhard Karls Universität, Tübingen, Germany
(andrea|apalmer)@coli.uni-saarland.de, magdalena.wolska@uni-tuebingen.de

Abstract

In this paper we investigate the potential of answer clustering for semi-automatic scoring of short answer questions for German as a foreign language. We use surface features like word and character n-grams to cluster answers to listening comprehension exercises per question and simulate having human graders only label one answer per cluster and then propagating this label to all other members of the cluster. We investigate various ways to select this single item to be labeled and find that choosing the item closest to the centroid of a cluster leads to improved (simulated) grading accuracy over random item selection. Averaged over all questions, we can reduce a teacher's workload to labeling only 40% of all different answers for a question, while still maintaining a grading accuracy of more than 85%.

Keywords: Computer-Assisted Correction, Short Answer Scoring, Clustering

1. Introduction

This paper investigates the potential of answer clustering and label propagation as an approach to semi-automatic scoring in the domain of foreign language learning. In particular, we explore the tradeoff between grading accuracy and reduction of teacher workload, asking: can we achieve a significant reduction in the number of items a teacher needs to grade while maintaining an acceptable scoring accuracy?

Within foreign language learning, we focus on scoring of answers to **short answer questions** (usually 1 or 2 sentences) to listening comprehension exercises for learners of German as a foreign language; short answer questions are a common strategy for assessing how well a learner has understood a text. There is a growing body of work on automatic scoring of short answers like e.g. the systems by Leacock and Chodorow (2003), Pulman and Sukkarieh (2005), Mohler et al. (2011) and Meurers et al. (2011). See also Ziai et al. (2012) for an overview and our previous work (Horbach et al., 2013)). In the current paper we move away from supervised scoring models and consider a real-life scenario for manually grading short answers.

Our approach is based on the assumption that highly similar student answers are likely to receive the same grade from a teacher and thus can be grouped and graded as a single unit. To do this, we use clustering techniques based on surface features (similar to e.g. topic clustering as described in (Steinbach et al., 2000)). We simulate a grading scenario in which answers are clustered automatically, teachers label only one item per cluster, and that label is then propagated to the other items in the same cluster.

A similar approach, developed in parallel by (Basu et al., 2013), targets the grading task for short answer questions by forming clusters and evaluating the number of human actions needed to correct a set of answers. In this context, actions consist of labeling complete (sub-)clusters of answers.

We evaluate our approach directly on data from *placement tests* for learners of German as a foreign language at Saarland University. For this study, we work under the assumption that it is a tolerable outcome to have a small number of incorrectly-graded answers. This particularly fits the placement testing scenario, where the aim is to determine a base language level for students, and where the listening comprehension component is only one out of several parts of the examination. The scores from all components (which include item types that are more easily automatically graded, such as multiple choice or fill-in-the-blank exercises) are combined to come to a final placement for the student, who receives only an aggregated score and no indication of performance on individual items. It must be noted that this is as yet an untested assumption, and for high-stakes testing our tolerable-amount-of-error assumption would not hold.

Under this assumption, though, results are quite promising. Our system can achieve scoring accuracy of 85% or above (depending on precise system settings) when teachers label only 40% of learner answers.

Figure 1 illustrates the extent to which student answers can vary. These are a selection of actual answers given for two of the 21 questions in the test. In addition to variability with respect to both content and form, we see variety in spelling and grammatical errors (e.g. *er/he* vs. *sie/she*). Spelling and grammatical errors do not influence the grade given by the corrector.

QUESTION #1:
Als was arbeitet Julian?
What is Julian's occupation?

CORRECT ANSWER:
Er arbeitet als Mechaniker.
He is working as a mechanic.

LEARNER ANSWERS:
[1 point / correct]
– sie arbeitet als mechanicker
She is working as a mechanik
– julian ist mechaniker
julian is a mechanic
– mechanika
mechanike
[0.5 points / partially correct]
– mekanical
mechanical
[0 points / incorrect]
– julian arbeitet in bortschaft
julian is working for embrasy
– als sekretarin
as a secretary

QUESTION #2:
Was ist für Nitsa am wichtigsten?
What is most important to Nitsa?

CORRECT ANSWER:
Für Nitsa ist es am wichtigsten, sich mit ihren Freunden zu treffen.
It is most important to her to meet her friends.

LEARNER ANSWERS:
[1 point / correct]
– freunde treffen
to meet friends
– sie mag ihre freunde treffen
she likes meeting her friends
– sie akzeptiert dass fernseher ein wichtiges medium aber sie bevorzugt am abends mit ihren freunden zu treffen
she accepts that TV is an important medium but prefers to meet her friends in the evening
[0.5 points / partially correct]
– mit freundin
with girlfriend
[0 points / incorrect]
– sehr wichtich
very importat
– mit freunden ins kino gehen
goiing with friends to the cinema

Figure 1: Sample of answers given to two individual questions. Note variety in form, even amount correct answers. English translations (including misspellings) added by the authors.

2. Data and Features

In this section we describe in more detail the particular listening comprehension task, the data, and the features extracted from that data.

2.1. Task and Data

The listening comprehension task is one component of a placement test for German as a foreign language courses at Saarland University (the other two parts being a grammar test and a c-test, both in the form of gap-filling exercises). Students listen to a prerecorded audio segment and answer questions about the text they have heard. Data collection is done online via a web-based language learning platform, Laempel.[1] The questions asked are of various types, ranging from those looking for a single word answer (e.g. *Where is she from?*) to questions asking for longer answers, such as questions requesting explanations (e.g. *Why does she have to leave?*). Students are not required to answer questions with complete sentences.

In listening comprehension – as well as for other types of short answer comprehension exercises – the most important factor for identifying correct answers is semantic content. The aim is to measure whether the student has understood the meaning of the text. This is different from grammar tests, in which small spelling differences are often crucial to distinguish, for example, correct verb forms incorrect forms. In listening/reading comprehension, only the semantics of the student answer are important, and spelling or grammar errors are ignored. Answers with high lexical overlap are likely to contain the same or similar semantic content, and thus to receive the same grade from a teacher. This motivates our approach of clustering and scoring answers on the basis of surface features.

We use 1668 individual answers to 21 different questions about 3 individual audio texts, collected from 98 students. These are data from a placement test administered in August 2013. Together with the textual content produced by the learners, the data consists of

[1] http://laempel.coli.uni-saarland.de

target answers (sample solutions) provided by teachers for each question and teacher-assigned grades for each learner answer. Grades are given as point values (usually: 0.0 for incorrect answers, 0.5 for partially correct answers, and 1.0 for correct answers), and we treat each grade as an individual, discrete label.

One way data from listening comprehension tasks differs from reading comprehension data is the higher frequency of spelling errors. When answering *reading* comprehension questions, it is well known that language learners often directly copy relevant material from the text into their answers. This strategy is known as *lifting* in the second language acquisition community. Lifting leads to a high overlap of both lexical material and orthography between the text and the learner answer. Learners answering listening comprehension questions make less frequent use of material from the audio input; this leads to a much higher degree of orthographic variability in learner answers. To determine the extent of the variability, we run the data through a German spelling correction system,[2] with the result that 18.6% of words are unknown. Our clustering approach needs to account for this variability.

Another difference between the listening comprehension data we use and typical short answer settings is length. Where most work on short answer scoring reports answers that are 1 – 3 sentences in length, the average length of learner answers in the data we use is 4.8 tokens, not counting punctuation.

2.2. Features and Modeling

As punctuation or capitalization errors are irrelevant in the scoring of short answer questions, we remove sentence punctuation and lowercase all learner answers. After these preprocessing steps, we merge string-identical answers. This step already reduces the number of different items to be graded by 25%. In this way we build a set of answer types based on string identity (e.g. *in Berlin* is a different answer type from *in the north of Berlin*, but the same as *in berlin*). All further clustering is done on answer types, and all reported statistics are also over these types.

We extract features based on word n-grams, character n-grams and keywords. For word feature extraction we first lemmatize all words using Treetagger (Schmid, 1994). We then extract word uni-, bi- and trigrams as well as skip-bi- and trigrams (i.e. pairs and triples of words with an arbitary number of other words in between).

To handle spelling errors, we use character bi- to fourgrams extracted from the unlemmatized text. For example, the two answers in (1) below are clearly conveying the same material, but if we only consider word overlap in clustering, the only shared lemma would be the pronoun *she*, which is unlikely to result in the two answers sharing a cluster at any meaningful level of clustering. Character n-grams, on the other hand, are able to capture such similarities.

(1) *She lives in Berlin*
She livs Berlim

We decide against correcting learner answers with a spellchecker in order to allow for the case that a misspelled form is a misunderstanding rather than an orthographic mistake.

QM condition. As an optional preprocessing step we exclude from the answer strings lexical material contained in the question. In this condition, we want to treat as equivalent answers which reiterate the theme of the question and those which simply state the rheme. For example, for the question *Where does she live?* the two answers seen in (2) would be treated as belonging to the same answer type.

(2) *She lives in Berlin*
Berlin

We refer to this as the exclude-question-material option (*QM+*).

KEY condition. Finally, we define keywords for each question based on the target answer given by the teachers. These specify the minimal requirement of lexemes that should be present in a correct answer and consist mainly of the nouns in the target answers. Consider the following (translated) question and target answer pair:

(3) Q: *Why does she have to leave?*
A: *Her German language course starts soon.*

In this case, there is one relevant keyword: *Deutschkurs*; for English data we would include the two phrases *German course* and *German language course*. Some orthographic variation is allowed for detecting keywords in learner answers. The *KEY* feature is implemented by determining, for each keyword, whether or not it is present within an answer type. To give this feature greater weight, it is repeated 100 times in the feature vectors.

[2] German version of aspell: `http://aspell.net/`.

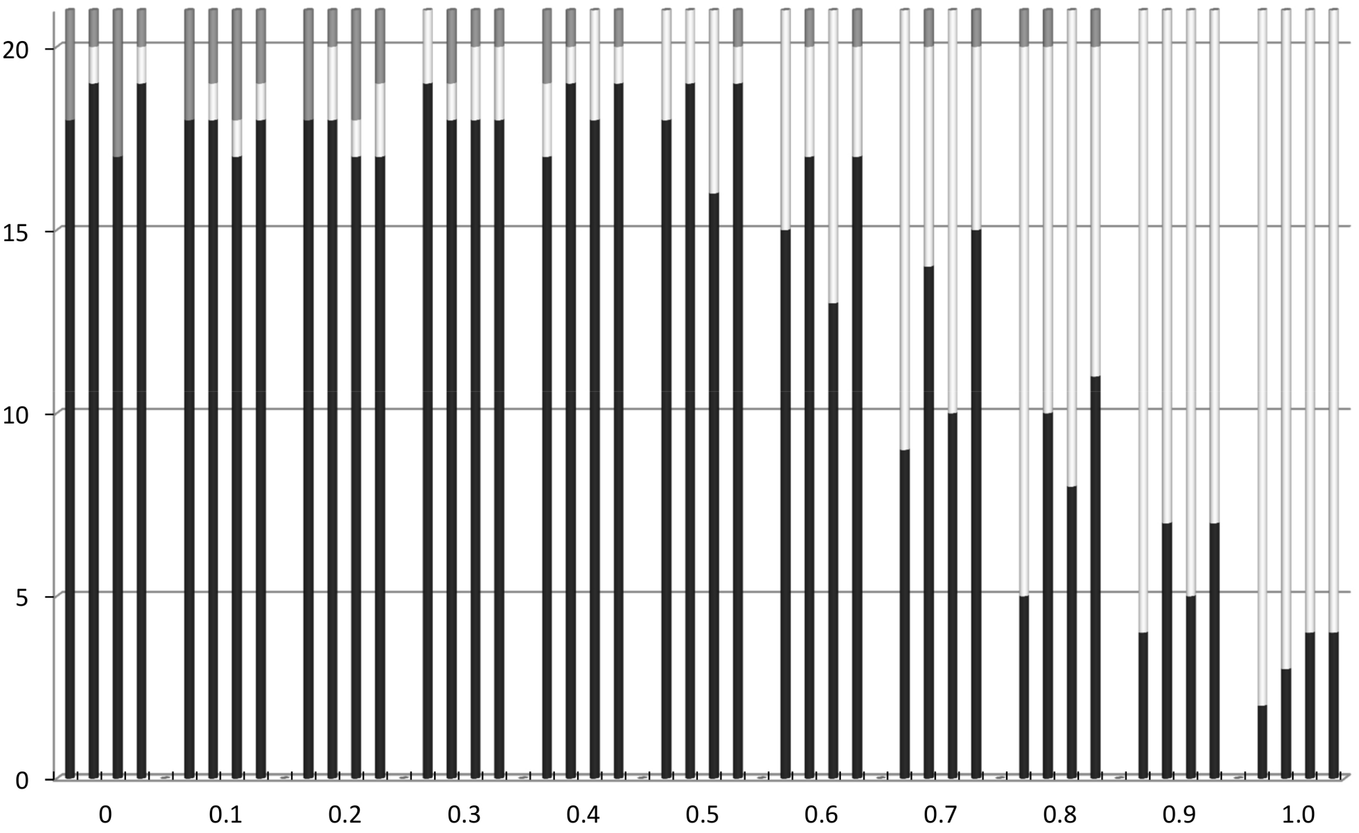

Figure 2: This graph shows, for each threshold (x-axis), the relative performances of our two item selection methods (centroid-based and random). For the 21 questions in our data set, we indicated on the y-axis how often (in # of questions) centroid-based selection is better than random (bottom/black), how often they are equally good (middle/white), and how often random selection does better (top/grey). The four columns for each threshold are (left to right): KEY- QM-, KEY+ QM-, KEY- QM+, KEY+ QM+

3. Clustering and Experiments

In this section we discuss parameters used for clustering and metrics for cluster evaluation, followed by presentation of experiments and experimental results.

3.1. Clustering and Evaluation Metrics

We use single-pass clustering. For each item, we calculate cosine similarity between its feature vector and the centroid of each existing cluster. If the highest similarity value is above the specified threshold, we add the item to the cluster whose centroid it is most similar to. Otherwise we establish a new cluster based on the item in question. Clustering is done individually per question.

We consider 4 conditions resulting from 2 binary options for feature extraction: whether or not to use the keyword feature (KEY+ and KEY-) and the exclude-question-material option (QM+ and QM-) We also vary the similarity threshold from 0.0 to 1.0 (in steps of 0.1). At one extreme, using a similarity threshold of 0.0 results in all items being placed in a single cluster. At the other extreme, enforcing a threshold of 1.0 means that each items ends up in its own single-item cluster.

In addition to standard cluster evaluation metrics as described in (Amigó et al., 2009), we want to assess the usefulness of clustering to a teacher in a simulated grading scenario, where a teacher grades just one item per cluster. Therefore, we compute the accuracy achieved if the label assigned by a teacher to this one exemplary item is propagated to all other items in the cluster. The number of clusters is then a suitable approximation of teacher workload.

Here we consider three different conditions with respect to selecting the item to be graded. As a baseline, we randomly select one answer type from each cluster. In a second, more informed, method, we select the single answer type that is closest to the centroid of the cluster. This method aims to choose for labeling the item that is most representative of its cluster. Finally, to simulate the best possible grading accuracy given a particular clustering, we explore a third option, an oracle condition in which we assume that we are always able to select an item from the majority class of a cluster. This measure represents best reachable grading accuracy for a given clustering.

We compute accuracies for these three conditions for different thresholds, with those different thresholds leading to different numbers of clusters.

In order to avoid effects from applying the single-pass clustering on items in a particular order, we run the clustering always 20 times on different random orderings of our items. Unless indicated otherwise, we report all results as the average over these 20 runs.

Note that clusters are re-computed for each threshold and are thus not hierarchically organized. That means that while a higher threshold leads to a higher number of clusters and thus more items being labeled by a teacher, it does not necessarily always lead to a higher grading accuracy. For a few "noisy" items, it can happen that having more clusters actually leads to a decrease in accuracy.

3.2. Experiment 1: Comparison of Item Selection Methods

The first experiment compares methods for selecting the one-item-per-cluster to be labeled. We compare *random* item selection to *centroid-based* item selection.

We see (Figure 2, previous page) that in most conditions, scoring based on centroid-based item selection leads to a higher accuracy than random selection. For higher clustering thresholds, the random and centroid-based selection are often equally good, because there are more clusters with fewer items. This has the effect that both random and informed selection more frequently result in the majority class label of the oracle condition (trivially so for single-item clusters).

Figure 3 (next page) shows the magnitude of this performance difference, averaging over questions. Different styles of boxes are used for the four different conditions. We see that with lower similarity thresholds, there is a greater increase in performance from using informed item selection. Figure 4 compares centroid-based scoring to the oracle condition in which we score each cluster according to the majority class label. For each threshold, we plot the number of questions (out of 21) for which centroid-based scoring reaches the same accuracy we achieve if we score according to the majority class. The general trend across thresholds and clustering conditions is that for more than 15 of the 21 questions, centroid-based item selection does as well as the oracle condition.

percentage of answer types needed for 90% accuracy	#questions within that range
0 -19.99	5
20-39.99	7
40-59.99	3
60-79.99	1
80-100	5

Table 1: Amount of data (in percent of answer types per question) needed to reach 90% accuracy

	40% or more	40% or less
KEY-, QM-	0.907 (SD=0.0061)	0.847 (SD=0.0108)
KEY+, QM-	0.899 (SD=0.0053)	0.859 (SD=0.0110)
KEY-, QM+	0.917 (SD=0.0069)	0.859 (SD=0.0065)
KEY+, QM+	0.904 (SD=0.0067)	0.870 (SD=0.0043)

Table 2: Mean Accuracies and Standard Deviations obtained when choosing a threshold so that either at least or at most 40 % of the data is labeled

3.3. Experiment 2: Reducing Workload while Keeping an Acceptable Accuracy

This second experiment is motivated by the assumption that scoring with some degree of error (90% accuracy) would still be useful to a teacher in a real-life scenario, provided it comes with a significant reduction in workload. We evaluated exemplarily on one run of our clustering experiments, how many answer types need to be labeled in order to obtain this accuracy. The main result for our scenario is that, averaging over all questions, accuracy near 90% can be achieved by labeling on average only 40% of all answer types per question, using centroid-based item selection.

However, the scoring accuracy achieved with this amount of labeled data varies considerably from question to question. Table 1 shows the range of variation for the same exemplary experimental setting (centroid-based selection, similarity threshold of 0.4, KEY+, QM+). We see, for example, that for 5 questions, labeling maximally 20% of answer types is sufficient to reach 90% accuracy; for another 5 questions, upward of 80% of answer types need to be labeled. It is clear that question type plays a role in the effectiveness of such semi-automatic scoring strategies; further exploration of the influence of question type is needed if such a strategy is to be successfully implemented in a non-simulation setting.

Putting the influence of question types to the side, we ask what would actually happen if we were to apply this 40% labeling strategy to all questions in our placement test scenario; results are shown in Table 2. First we measure the overall scoring accuracy if, for each

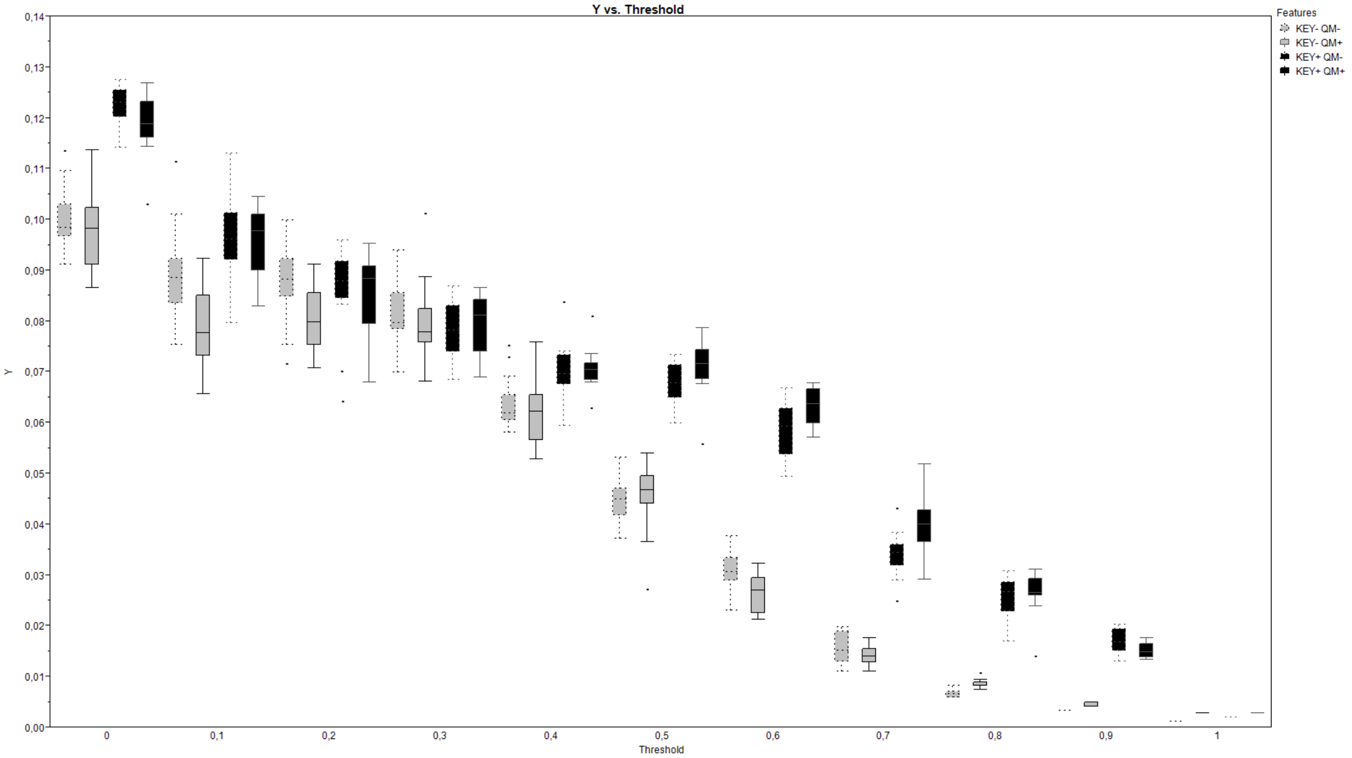

Figure 3: Accuracy gain if we use centroid-based instead of random item selection

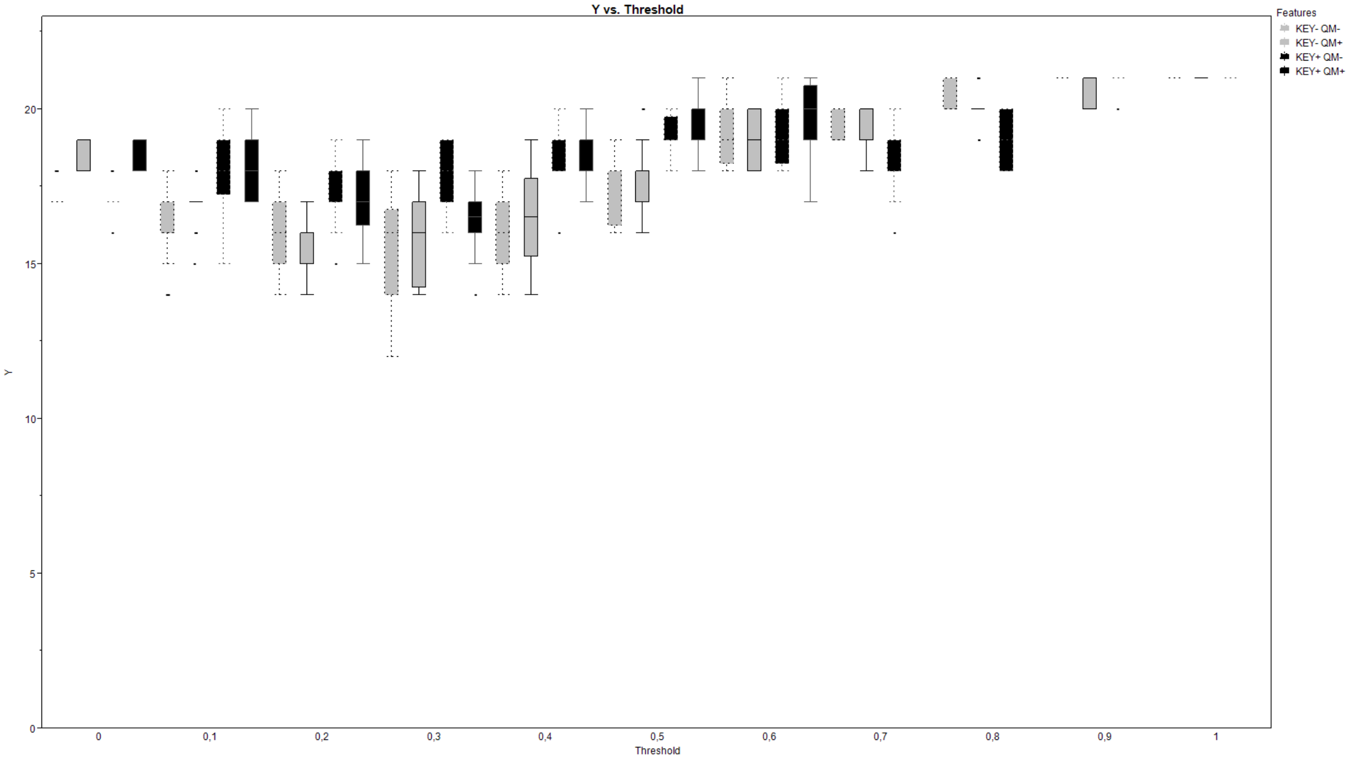

Figure 4: How often does centroid-based item selection reach the same accuracy as the oracle condition?

question, we apply our centroid-based scoring strategy to the clustering produced by the lowest threshold that results in labeling 40% or more of the answer types. Next, we choose instead the *highest* threshold that leads to 40% *or less* of the answer types being labeled. We report mean accuracies and standard deviation from averaging over 20 runs of the clustering.

We see here that, despite differences due to question type, a labeling accuracy between 85 and 90% (depending on settings for KEY and QM) can be achieved by labeling only 40% of answer types. We don't observe interesting differences between the four individual conditions.

4. Conclusions and Future Work

We have shown in this work that an answer clustering and label propagation strategy can be used to reduce a teacher's grading workload while still maintaining a grading accuracy near 90%. We have shown this using learner answers to German short answer questions, simulating a strategy of labeling just one item out of each cluster.

The aim of reducing grading workload, which could also be construed as labeling workload, raises the natural possibility of using approaches based on active learning (Settles, 2010) in the same scenario. With active learning, strategies employ various methods for selecting items for labeling that will be maximally informative to the classifier or machine learner. This is one avenue we plan to investigate in the near future.

Another important question to investigate is the so-far untested assumption that a grading accuracy of 90% is a useful and acceptable accuracy for certain grading scenarios. As a first approximation, we could say that, if errors are evenly distributed over all students and all test items, and the likelihood of misgrading any individual item is 10% for each of the roughly 20 items on the test, the probability of mis-scoring 4 or more items for any single student can be estimated at below 15%. We calculate this using a Bernoulli distribution to determine for what percentage of students the test produces at most a certain number of errors. However, it is clear that some items are harder than others, and the effectiveness of our semi-automatic scoring strategy varies considerably across questions. In future work, we will investigate how to improve the quality of clustering by considering the nature of the question. Factors that seem to contribute to differences in difficulty are the expected length of the answer, the type of information asked for (e.g. a single noun phrase, an entire sentence, or several different pieces of information), and the question type (e.g. reproduction or inference). We will explore how to make use of those properties in determining clustering settings; it may be that some items should be graded via clustering and some should rather be graded by hand.

Finally, it would be interesting to explore how to detect elements of a cluster that have the minority label of this cluster, and thus would get the wrong label even under a well-performing cluster-label-propagate strategy. This could be done, for example, by considering linguistic cues such as modality and negation, or by distinguishing between true misspellings and cases of similar words with different semantics (e.g. for a target answer of *Angola*, some spelling variants (*Angla*, *Engloa*, etc) may be acceptable, while words like *England* would not, despite containing similar character n-grams).

Acknowledgments

We thank the three anonymous reviewers for their helpful comments.

This work was partially funded by the Cluster of Excellence "Multimodal Computing and Interaction" of the German Excellence Initiative.

Magdalena Wolska's work is supported by the Institutional Strategy of the University of Tuebingen (Deutsche Forschungsgemeinschaft, ZUK 63)

5. References

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, 12(4):461–486, August.

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

Andrea Horbach, Alexis Palmer, and Manfred Pinkal. 2013. Using the text to evaluate short answers for reading comprehension exercises. In *Proceedings of *SEM 2013: Second Joint Conference on Lexical and Computational Semantics*.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405.

Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. 2011. Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 1–9, Edinburgh, Scottland, UK.

Michael Mohler, Razvan C. Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *ACL*, pages 752–762.

Stephen G. Pulman and Jana Z. Sukkarieh. 2005. Automatic short answer marking. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, EdAppsNLP 05, pages 9–16.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, United Kingdom.

Burr Settles. 2010. Active learning literature survey. *University of Wisconsin, Madison*, 52:55–66.

Michael Steinbach, George Karypis, and Vipin Kumar. 2000. A comparison of document clustering techniques. In *In KDD Workshop on Text Mining*.

Ramon Ziai, Niels Ott, and Detmar Meurers. 2012. Short answer assessment: Establishing links between research strands. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7)*, Montreal, Canada.