Deep learning for sentence clustering in essay grading support

Li-Hsin Chang, Iiro Rastas, Sampo Pyysalo, and Filip Ginter
TurkuNLP Group
Department of Computing
University of Turku
{lhchan, iitara, sampyy, figint}@utu.fi

ABSTRACT

Essays test student knowledge on a deeper level than short-answer and multiple-choice questions but are more laborious to evaluate. Automatic clustering of essays, or their fragments, prior to evaluation may reduce the manual effort required. Such clustering presents numerous challenges due to the variability and ambiguity of natural language. In this paper, we introduce two datasets of undergraduate student essays in Finnish, manually annotated for salient arguments on the sentence level. Using these datasets, we evaluate several deep-learning embedding methods for their suitability to sentence clustering in support of essay grading. We find the suitable method choice to depend on the nature of the exam question and the answers, with deep-learning methods being capable of, but not guaranteeing better performance over simpler methods based on lexical overlap.

Keywords

deep learning, essay clustering, text similarity, paraphrase, grading support

1. INTRODUCTION

Essay-type questions have been shown to help with the retention of learned material [10] but are time- and labour-consuming to evaluate. Computational methods can be used to grade essays, or to assist in their evaluation. Examples of the latter approach include pre-processing to show statistics of student answers such as average answer length and keywords [11], comparing student answers to a given text [11], generating word clouds of student answers [6], and grouping student answers into clusters of similar answers [2]. Most of these systems target the pre-processing and analysis of short answers, and less effort has been dedicated to computer-aided assessment of longer essays. One approach to reducing human effort in fact-based student essay assessment computationally would be to identify similar arguments in student essays. This approach draws inspiration from qualitative research methods where interviews are first transcribed verbatim, and categories are then formed and themes are created [5]. By identifying recurring arguments across a cohort of essays, it is expected that human grading effort could be reduced, much like the analysis of interviews is made simpler after forming categories.

In this paper, we evaluate the applicability of several representative deep learning methods to the task of identifying distinctly-phrased, but semantically near-equivalent segments of student essays[1]. We approach the task from two angles. As an *information retrieval* (IR) problem, whereby given a query text, e.g. a reference answer or an essay, the task is to retrieve the matching essays from the cohort, and establish their mutual correspondence down to sentence level. The other approach is that of *clustering*, where the objective is to discover groups of sentence-long segments with same meaning in the essay cohort. We test several algorithms, including TF-IDF [7], LASER [1], BERT [4], and Sentence-BERT [13]. To evaluate these algorithms, we gather and annotate two sets of factual essays written in exams by Finnish university students.

2. DATASETS

We collected Finnish essays written by bachelor's level students as answers to exam questions. Two sets of essays replying to questions from two courses were selected for manual annotation. The annotator was a PhD student from a different discipline than the domain of the essays. The goal of the annotation was to identify similar arguments in separate essays. The data were annotated by cross-referencing the arguments found in every essay, and assigning textual labels to recurring arguments or concepts on a sentence level. Specifically, all essays were first segmented into sentences, and each sentence was then assigned zero or more textual labels representing its content. If an argument appears more than once, it is given a distinct label which is assigned to all sentences containing that argument. For an argument to be considered recurring, the two sentences are required to clearly aim to communicate the same information about a common subject matter. An example of two sentences that are considered to have the same argument (on the pros and cons of group interviews in research): "It is not the quieter and more timid individuals that come out, but the loudest ones come to the fore." and "In a group interview, there is a danger that some will talk too much and some will not have a turn to speak at all." Both of these sentences describe

[1]We refer readers to `https://arxiv.org/abs/2104.11556` for a more detailed version of the paper

Table 1: Dataset statistics

	Research methods	Accounting standards
No. of essays	47	10
Total no. of sentences	486	158
No. of labels	59	34
Avg. no. of labels per sentence	1.29	0.82

the imbalance of expression of opinions in group interviews. In the next example, however, the two sentences are considered to have different arguments, despite both of them being related to the role of trust in interviews. "In interviews, a trusting relationship must be established between the interviewee and the interviewer, which can be challenging." and "If the interviewee remains anonymous, one can also openly discuss more sensitive topics, especially when one is alone with the interviewer." This is because the two sentences make opposing arguments: the former takes a positive perspective towards the role of trust in interviews, while the latter views it as a challenge. Clearly, these communicate different information. For each dataset, the number of labels thus depends on the number of recurring arguments in the essays, and the annotation scheme differs. We estimate that the development of the annotation scheme and the annotation effort required about two person-weeks in total. We note that we do not expect to annotated all sets of essays that are to be evaluated. Instead, these two sets of annotations serve as benchmarks for testing ideas on automatically assisting essay evaluation. The two resulting datasets are introduced below. The key statistics of the two datasets are summarized in Table 1 and the distribution of the labels in the two datasets is illustrated in Figure 1 in the Appendix.

2.1 Research methods dataset

The first dataset is created from student essays from the course "Research process and qualitative research methods" (henceforth *Research methods*). The essays answer the question, "Consider the positive and negative aspects of interviews". Several main points are frequently mentioned by students: for example, almost all students discussed how time consuming interviews can be (label `time_consuming`). 93% of the dataset sentences have at least one label, indicating that the great majority of sentences involve at least one argument repeated in other essays.

2.2 Accounting standards dataset

The second dataset consists of student essays from the course titled "IAS/IFRS accounting standards" (henceforth *Accounting standards*). The essay prompt is "What are the components of IFRS financial statements? Consider the significance of the various components in the light of the qualitative criteria for the financial statement information". The label distribution of this dataset is more even, and almost a third of the sentences do not have a label. This may be due to the fact that there are fewer essays in this dataset. This implies that given one main argument, it is less likely that the argument is also mentioned by somebody else.

3. SENTENCE REPRESENTATIONS

To identify sentences with similar arguments, we consider a set of methods for representing each sentence with a vector, which allows efficient computation of sentence similarity via the similarity of their vectors. As baselines, TF-IDF vectors and average of word embeddings are used for sentence representation. For deep learning methods, the encoders LASER, BERT, and Sentence-BERT are tested. The distance measure used is the cosine similarity between two sentence vectors, a standard metric applied also in previous studies.

3.1 TF-IDF

Term frequency–inverse document frequency (TF-IDF) is a family of popular IR metrics that estimate the importance of a given word in a document from a document collection based on the number of times the word appears in the document (term frequency) and the inverse of the number of documents the word appears in (document frequency) [7]. TF-IDF can be applied with words or character sequences. For this baseline, all the tokens in a sentence are first lemmatized using the Universal Lemmatizer [8]. Character ngrams, specifically bigrams, trigrams, 4-grams and 5-grams, are created out of text inside word boundaries. We note that the TF-IDF encoding generates sparse high-dimensional vectors where there is no inherent similarity between words.

3.2 Average of word embeddings

This baseline represents each sentence using the average of the vector representations of the words in the sentence. We use the Finnish word embeddings created by Kanerva et al. [9] and refer readers to this paper for further details of the embeddings. These embedding were induced using the implementation of the skip-gram algorithm [12] in the `word2vec` software package on Finnish Common Crawl data. The average of word embeddings produces dense, comparatively low-dimensional representations that can capture the similarity between words, but the representation of words is independent of the context they appear in.

3.3 LASER

The Language-Agnostic SEntence Representations (LASER) released by Facebook is a sentence embedding method that aims to achieve universality with respect to language and NLP task. The encoder can encode 93 languages, all of which share a byte-pair encoding [14] vocabulary. The encoder consists of a BiLSTM with max-pooling operation, coupled with an LSTM layer during training on parallel corpora [1]. LASER produces dense, low-dimensional representations that can capture the contextual meaning of words.

3.4 BERT

Bidirectional Encoder Representations from Transformers (BERT) introduced by Google is a deep contextual language representation model [4]. The training objectives of BERT make them cross-encoders, i.e. the model takes in a pair of sentences at a time. However, we encode one sentence at a time and use the mean-pooling of the resulting outputs as the sentence representation. We use the uncased variant of FinBERT, a monolingual Finnish BERT Base model that has been demonstrated to provide better performance in Finnish text processing tasks than multilingual BERT [16]. Like LASER, BERT produces dense, low-dimensional representations that account for context.

Table 2: Results of the IR evaluation

Accounting standards	**Avg First**	**Avg Med**	**Avg Mean**	**Avg Last**	**MRR**	**MAP**
TF-IDF	**4%**	**9%**	**11%**	**24%**	0.47	**0.48**
word2vec	6%	17%	20%	40%	0.47	0.34
LASER	**4%**	13%	15%	33%	**0.53**	0.42
BERT	5%	15%	17%	37%	**0.53**	0.41
SBERT	5%	11%	14%	31%	0.46	0.42
Research methods	**Avg First**	**Avg Med**	**Avg Mean**	**Avg Last**	**MRR**	**MAP**
TF-IDF	**1%**	18%	24%	72%	0.46	**0.28**
word2vec	2%	26%	31%	79%	0.34	0.19
LASER	2%	19%	26%	73%	0.42	0.23
BERT	**1%**	**17%**	23%	70%	**0.49**	**0.28**
SBERT	2%	**17%**	**22%**	**65%**	0.43	**0.28**

3.5 Sentence-BERT

Sentence-BERT (SBERT) trains BERT models using Siamese and/or triplet networks to induce a single-sentence encoder specialized for cosine-similarity comparison [13]. We obtain machine translated versions of the SNLI [3] and MNLI [17] corpora using the English to Finnish Opus-MT model [15]. Finnish SBERT is subsequently trained from FinBERT-base-uncased using these two natural language inference corpora. Specifically, the model is fine-tuned for an epoch with learning rate 2e-5 and batch size of 16, with mean pooling as the pooling method. The representations produced by SBERT are dense, low-dimensional, and context-sensitive, like those of LASER and BERT.

4. EVALUATION

The sentence representations are evaluated from IR and clustering perspectives. Six evaluation metrics are used for the IR approach: two well-known metrics **mean reciprocal rank (MRR)** and **mean average precision (MAP)**, and four metrics tailored to our specific task setting. **average of highest rank (Avg first)**, **average of median rank (Avg med)**, **average of mean rank (Avg mean)**, and **average of lowest rank (Avg last)** measure the rank of the highest, median, mean, and lowest rank of the relevant items respectively, as percentage of the whole (0% first rank, 100% last rank), averaged over all items. These four metrics give more insight into the distribution of the relevant retrievals by measuring where, on average, the first, median, mean, and last relevant items are ranked. Since some sentences have more than one label, sentences with at least one overlapping label are considered relevant retrievals for all metrics.

The clustering evaluation measures how well the clustering induced by the vector embeddings corresponds to the clustering induced by the sentence labels. **Cluster accuracy** is based on the most frequent label of a cluster: for each cluster, the majority label is obtained from the ground truth annotations of the sentences in the cluster. A sentence is considered to be correctly clustered if it has the majority label of its cluster as one of its labels. The number of correctly and incorrectly clustered sentences can then be interpreted as an accuracy percentage. We note that random baseline performance varies drastically between different datasets with this metric, so accuracy values are not directly comparable between datasets. **Adjusted Rand index** and **adjusted mutual information** are established clustering metrics. We use sampling to work around the multi-label nature of the annotations: For each sentence with multiple labels, one label is randomly chosen. Then the clusters are evaluated against these labels with the two metrics. This process is repeated 50 times and the values of the metrics are subsequently averaged. The resulting scores are between -1 and 1, and they are adjusted for chance, so that a random clustering has a score close to zero. We use the agglomerative clustering algorithm with ward linkage. Sentences that have no labels, i.e. containing a unique argument, are each given a unique label for the purposes of the clustering evaluation, effectively each forming one singleton cluster. The resulting true number of clusters (60 for the research methods dataset and 95 for the accounting standards dataset) is the clustering model input.

Table 3: Results of the two clustering evaluation methods. Average adjusted Rand (Avg adj. Rand), Average adjusted mutual information (Avg adj. mutual info.), Cluster accuracy (Clus. acc.), Standard deviation (Std dev).

Accounting standards	**Avg adj. Rand**	**Std dev**	**Avg adj. mutual info.**	**Std dev**	**Clus. acc.**
TF-IDF	**0.31**	0.02	**0.33**	0.02	**73%**
word2vec	0.18	0.02	0.23	0.02	69%
LASER	0.21	0.01	0.27	0.01	72%
BERT	0.21	0.01	0.27	0.02	72%
SBERT	0.28	0.01	**0.33**	0.02	**73%**
Research methods	**Avg adj. Rand**	**Std dev**	**Avg adj. mutual info.**	**Std dev**	**Clus. acc.**
TF-IDF	**0.12**	0.01	0.22	0.01	**55%**
word2vec	0.05	0.00	0.13	0.01	41%
LASER	0.08	0.00	0.17	0.01	46%
BERT	0.11	0.01	**0.23**	0.01	50%
SBERT	0.11	0.01	0.22	0.01	51%

5. RESULTS

The IR evaluation results are shown in Table 2. We find that there is no single method that systematically outperforms the others. Surprisingly, for the accounting standards dataset, the advanced methods fail to outperform the TF-IDF baseline, which achieves the highest results for all metrics except MRR. This indicates that while TF-IDF is not the most competitive in consistently ranking relevant items at the highest ranks, it is able to concentrate relevant items towards higher ranks in general. This is particularly evident for the average of the lasts metric, where TF-IDF scores 7% points higher than the second best performer, SBERT. Here the number 24% indicates that, for the accounting standards dataset, TF-IDF on average ranks all the relevant items within rank 24 out of 100. The high performance of TF-IDF on this dataset may be attributed partly to the essay prompt requiring students to list the correct keywords. The elements of the IFRS financial statements are only so many, and these items cannot be paraphrased. Methods that compare strings directly thus outperform methods that use dense vector representations that approximate their meaning.

The research methods dataset, however, does not have such a strong emphasis on exact keyword matching: there are no fixed numbers of keywords that have to be mentioned in the answers. Rather, the pros and cons of interviews as a research method are described, and thus sentences that

describe the same concept using different words are more likely to occur. On this dataset, considering the retrieval of the first relevant item, both TF-IDF and BERT perform best on the average of the firsts metric, while BERT performs best on the mean reciprocal rank. Since the average of the firsts metric is more lenient on lower rankings of first relevant items, we can infer that BERT performs more consistently on the retrieval of the first relevant item. Overall, BERT-based methods obtain better results, with SBERT in particular outperforming the other methods by 5% points on the retrieval of the last relevant items. BERT and SBERT both obtain the highest results on four out of six metrics.

The results of the clustering evaluation are summarized in Table 3. These results clearly tend towards the TF-IDF baseline, while the word2vec-based approach is the weakest, tallying with the IR evaluation. Of the neural methods, SBERT is particularly strong in the accounting standards dataset, while being in line with BERT in the research methods dataset. Of the two sentence embedding methods, SBERT outperforms LASER in all tests. We are surprised to find that the TF-IDF model seems to be better suited to the clustering objective than the neural methods, and will examine this in future work.

Overall, we find that the comparative ranking of the methods varies strikingly depending on the dataset, evaluation setting, and metric. The dataset dependence may partly be explained by the nature of the arguments made: if the argument is required to contain certain specific terms, TF-IDF can be a very strong method. On the other hand, if the argument involves more abstract concepts that can be expressed in many ways, neural methods may have an advantage over methods that are based on exact string matching. While deep neural methods have led to breakthroughs in many NLP tasks, the gain they show here over the simple TF-IDF baseline is quite small even in the cases where they outperform it. This may indicate challenges specific to the task and domain beyond those we have identified here, and calls for further research into the topic. This includes searching for more suitable encoding methods, improved evaluation methods, and also study of how data should best be annotated to develop methods serving the needs of essay graders.

6. DISCUSSION

Our annotation makes at least two assumptions that call for further investigation: the sentence is the unit of annotation, and the labels are categorical and non-overlapping. Figure 1 shows that approximately 57% and 64% of the sentences in the Accounting standards and Research methods datasets (respectively) have exactly one label. Another 33% and 7% (resp.) of sentences do not have any labels. Since labels are only assigned if a main argument appears more than once, these sentences can be seen as singleton clusters with a label that occurs exactly once. With the current annotation granularity, the annotation is best applicable to cases where each sentence conveys a single main argument. However, the annotation statistics indicate that sentence may not always be the most suitable unit of annotation. These include cases where an argument is made across several sentences, and where a sentence makes several arguments.

In addition to issues related to the sentence as a unit of annotation, there is also a degree of subjectivity to their labeling. For example, in the Research methods dataset, the two labels `workload` and `time_consuming`, which state that interviews are labor-intensive and time-consuming respectively, could arguably be merged. For such boundary decisions to be helpful for essay graders, the marking criteria play a central role and there is no universal cut-off. As an alternative to disjoint categorical labels, one could consider that the arguments (and the labels that represent them) can be organized hierarchically. For instance, in the research methods dataset, the label `interviewer_influence` represents the argument that the stance of the interviewer may affect the research results, and the label `unnatural_performance` describes the affect of the interview situation on the performance of interviewees. On a higher level, both of the labels convey the research results being negatively affected by artificial factors. For these two datasets, the boundary decisions also depend on the sample size: if there are more essays, chances are that a small number of students make the exact same argument, in which case the boundary is unambiguous, or could be seen as a subcluster of a bigger cluster. We hope to address these and related challenges in future work.

One focus of our ongoing work is the practical use of the clusters. An approach to capitalizing on these clusters would be to make them manually adjustable, i.e. examiners can adjust the contents of the clusters, create new clusters, and delete clusters. These clusters can then be color-coded or annotated with text, indicating whether the presence of a certain cluster is desirable in an essay. In addition, if reference answers are available, essays with more overlapping clusters with the reference answers can be automatically identified.

7. CONCLUSIONS AND FUTURE WORK

We focused on the task of computer-assisted assessment of comparatively long essays through the perspectives of IR and clustering. We have created two datasets based on two exam questions from different fields, on which we tested several deep-learning methods with respect to their ability to retrieve and cluster sentences containing the same arguments paraphrased. We found no method to be universally best; rather, the results depend on the nature of the essays under assessment. Overall, the difference between the state-of-the-art deep learning methods and the much simpler TF-IDF baseline is not numerically large, leaving clear room for further development and application of more advanced methods for embedding meaning. Developing such methods, as well as further practical testing of the approach constitute our future work.

8. ACKNOWLEDGMENTS

The research presented in this paper was partially supported by the European Language Grid project through its open call for pilot projects. The European Language Grid project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 825627 (ELG). The research was also supported by the Academy of Finland and the DigiCampus project coordinated by the EXAM consortium. Computational resources were provided by *CSC — the Finnish IT Center for Science*. We thank Kaapo Seppälä and Totti Tuhkanen for administrative support and data collection.

9. REFERENCES

[1] M. Artetxe and H. Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 09 2019.

[2] S. Basu, C. Jacobs, and L. Vanderwende. Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, 1:391–402, 2013.

[3] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[5] C. L. Erlingsson and P. Brysiewicz. A hands-on guide to doing content analysis. In *African journal of emergency medicine : Revue africaine de la medecine d'urgence*, 2017.

[6] S. Jayashankar and R. Sridaran. Superlative model using word cloud for short answers evaluation in elearning. *Education and Information Technologies*, 22:2383–2402, 2016.

[7] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972.

[8] J. Kanerva, F. Ginter, and T. Salakoski. Universal Lemmatizer: A sequence to sequence model for lemmatizing Universal Dependencies treebanks. *Natural Language Engineering*, pages 1–30, 2020.

[9] J. Kanerva, M. Luotolahti, V. Laippala, and F. Ginter. Syntactic N-gram collection from a large-scale corpus of internet Finnish. In *Proceedings of the Sixth International Conference Baltic HLT 2014*, pages 184–191. IOS Press, 2014.

[10] J. D. Karpicke and H. Roediger. The critical importance of retrieval for learning. *Science*, 319 5865:966–8, 2008.

[11] J. McDonald and A. C. M. Moskal. Quantext: Analysing student responses to short-answer questions. *Me, Us, IT*, pages 133–137, 2017.

[12] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013.

[13] N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP-IJCNLP*, pages 3982–3992, 2019.

[14] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *ACL*, 2016.

[15] J. Tiedemann and S. Thottingal. OPUS-MT — Building open translation services for the World. In *EAMT*, 2020.

[16] A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, and S. Pyysalo. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*, 2019.

[17] A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *ACL*, pages 1112–1122, 2018.

APPENDIX

A. LABEL DISTRIBUTION

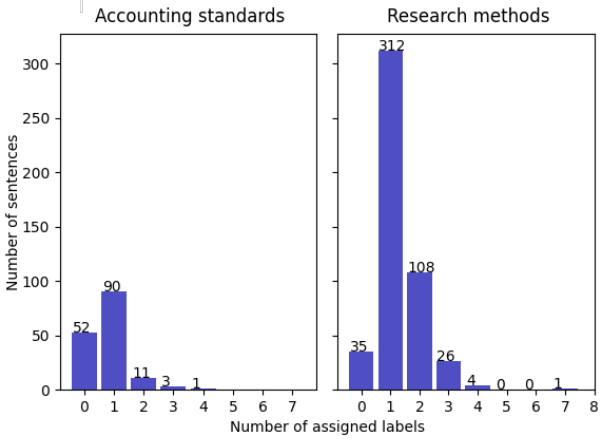

Figure 1: Number of labels per sentence