Can vectors read minds better than experts? Comparing data augmentation strategies for the automated scoring of children's mindreading ability

Venelin Kovatchev[1], Phillip Smith[2], Mark Lee[2], and Rory Devine[1]
[1] School of Psychology, University of Birmingham
[2] School of Computer Science, University of Birmingham
52 Prichatts Road, Edgbaston, B15 2SB, United Kingdom
{v.o.kovatchev , P.Smith.7 , m.g.lee , R.T.Devine} @bham.ac.uk

Abstract

In this paper we implement and compare 7 different data augmentation strategies for the task of automatic scoring of children's ability to understand others' thoughts, feelings, and desires (or "mindreading").

We recruit in-domain experts to re-annotate augmented samples and determine to what extent each strategy preserves the original rating. We also carry out multiple experiments to measure how much each augmentation strategy improves the performance of automatic scoring systems. To determine the capabilities of automatic systems to generalize to unseen data, we create UK-MIND-20 - a new corpus of children's performance on tests of mindreading, consisting of 10,320 question-answer pairs.

We obtain a new state-of-the-art performance on the MIND-CA corpus, improving macro-F1-score by 6 points. Results indicate that both the number of training examples and the quality of the augmentation strategies affect the performance of the systems. The task-specific augmentations generally outperform task-agnostic augmentations. Automatic augmentations based on vectors (GloVe, FastText) perform the worst.

We find that systems trained on MIND-CA generalize well to UK-MIND-20. We demonstrate that data augmentation strategies also improve the performance on unseen data.

1 Introduction

Many state-of-the-art NLP models are limited by the availability of high quality human-annotated training data. The process of gathering and annotating additional data is often expensive and time consuming. It is especially difficult to gather data for tasks within psychology and psycholinguistics, as test administration typically requires highly trained in-domain experts, controlled environments, and large numbers of human participants.

Data augmentation is a popular technique for artificially enlarging datasets. Typically, data augmentation uses one or more predefined strategies to modify existing gold-standard examples while retaining the original label. The objectives of data augmentation are: 1) to increase the size of the dataset; 2) to introduce more variety; 3) to reduce overfitting; and 4) to improve generalizability.

Data augmentation has been used successfully in computer vision (Shorten and Khoshgoftaar, 2019) and has recently become more popular in the field of NLP (Wei and Zou, 2019; Min et al., 2020; Dai and Adel, 2020; Marivate and Sefara, 2020).

We use data augmentation to improve the performance of systems for automatic scoring of children's performance on tests of mindreading (i.e., the ability to reason about others' thoughts, feelings and desires) (Hughes and Devine, 2015). Automatic scoring of mindreading was recently introduced by Kovatchev et al. (2020). Their corpus, MIND-CA contains hand-scored data from more than 1000 children aged 7 to 14. Collecting data on children's mindreading performance is complicated, time-consuming, and expensive. It requires in-person testing sessions led by trained researchers and children's open-ended responses must be rated by trained annotators. Data augmentation could be very beneficial to improve the performance and consistency of the automated scoring systems.

In this paper we aim to measure, in a systematic way, the quality and efficiency of the different augmentation strategies. We evaluate and compare the different strategies intrinsically and extrinsically. For the intrinsic evaluation, we recruit in-domain experts to re-annotate augmented examples and determine the extent to which each strategy preserves the original label. For the extrinsic evaluation, we measure the quantitative improvement (macro-F1, F1-per-Question, Standard Deviation) of automatic systems on the MIND-CA corpus. Furthermore,

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 1196–1206
August 1–6, 2021. ©2021 Association for Computational Linguistics

we create a new corpus, UK-MIND-20, containing 10,320 question-answer pairs in English and we use it to evaluate the performance of automated systems on unseen data.

We find that the intrinsic "quality" of the augmentation strategies varies significantly, according to human raters. However, the extrinsic evaluation demonstrates that all strategies improve the performance of the automated systems. We systematically measure the importance of three factors in data augmentation: corpus size, sampling strategy, and augmentation strategy. We find corpus size to be the most important factor. However, the choice of sampling and augmentation strategies also significantly affects the performance of the automated systems. We report a correlation between the "quality" of the augmentation and the performance. With the best configuration we obtain a new state-of-the-art on MIND-CA, improving Macro-F1 score by 6 points and F1-per-Question by 10.3 points.

We demonstrate that the automated scoring systems can generalize well between MIND-CA and UK-MIND-20. These findings indicate that the methodology for administering and scoring mindreading is consistent and the automatic solutions can be adopted in practice.

The rest of this article is organized as follows. Section 2 discusses the related work. Section 3 presents the methodologies for data augmentation. Section 4 compares the quality of the augmentation strategies. Section 5 describes the machine learning experimental setup and evaluation criteria. Section 6 analyzes the effect of data augmentation on automated systems. Section 7 presents some follow-up experiments and discusses the implications of the findings. Section 8 concludes the article and proposes directions for future work.

2 Related Work

Mindreading (also known as "theory of mind") is the ability to understand others' thoughts, feelings, and desires (Hughes and Devine, 2015). For example, in the final scene of Romeo and Juliet, Romeo holds a mistaken belief that Juliet is dead. Being able to understand the state of the world ("Juliet is alive") and the mistaken belief ("Juliet is dead") is important to understand the situation and the motivation of the characters.

Individual differences in children's mindreading are linked with both social and academic outcomes and children's wellbeing (Banerjee et al., 2011; Fink et al., 2015; Devine et al., 2016). Furthermore, difficulties with mindreading are linked with a range of mental health problems and neurodevelopmental conditions (Cotter et al., 2018).

The task of automatic scoring of mindreading was first proposed by Kovatchev et al. (2020). They gathered the responses of 1066 children aged 7-14 on two standardized tests of mindreading: the Strange Story Task (Happé, 1994) and the Silent Film Task (Devine and Hughes, 2013). After digitalizing and manually scoring the responses, they created MIND-CA, a corpus of 11,311 question-answer pairs. They trained and evaluated several automated systems (i.e., SVM, BILSTM, Transformer) and obtained promising initial results.

Data augmentation is a technique for artificially increasing the size of the dataset. It can also be seen as a type of regularization at the level of the data. Data augmentation can be used to increase the number of instances of specific answer types. It can also introduce more variety, and can reduce the imbalance between classes. Data augmentation is used to improve the performance of automated systems, to reduce the risk of overfitting, and to enhance the ability of automated systems to generalize to unseen data. It is widely used in computer vision (Shorten and Khoshgoftaar, 2019).

The specifics of natural languages make it more difficult to incorporate data augmentation in NLP. A subtle change to the text can often lead to a substantial difference in meaning and a change of the label. The last two years have seen an increase in the popularity of data augmentation in NLP. Wei and Zou (2019) present a Python library that uses simple augmentation methods for improving text classification. Marivate and Sefara (2020) compare different strategies for augmentation in the context of short-text classification. Dai and Adel (2020) compare different data augmentation strategies for the task of Named Entity Recognition.

Several researchers propose more complex augmentation strategies for NLP. Hou et al. (2018) propose a sequence-to-sequence model for data augmentation. Kobayashi (2018) and Gao et al. (2019) use language models in what they call "contextual augmentation". Min et al. (2020) use syntactic augmentation to improve the performance and generalizability on Natural Language Inference.

In this paper, we take a different approach towards data augmentation. We implement and compare seven different augmentation strategies. Two

of the strategies were designed specifically for the task of automatic scoring of mindreading, while the remaining five are task agnostic. We put the emphasis on a systematic evaluation of the augmentation strategies and some key parameters of the augmentation process. We recruit and train in-domain experts to provide intrinsic human evaluation of the data augmentation. We also annotate a new corpus that can measure the performance and improvement on unseen data.

3 Data Augmentation

We used 7 different strategies for automatic data augmentation. "Dictionary" and "phrase" strategies make use of task-specific resources, created by in-domain experts. The other 5 strategies ("order", "wordnet", "ppdb", "glove", "fasttext") make use of publicly available task-agnostic resources.

For a source of the augmentation, we used the MIND-CA corpus (Kovatchev et al., 2020). It contains 11,311 question-answer pairs. There are 11 different questions, and an average of 1,028 responses per question. There are three possible labels reflecting the degree to which the response shows context-appropriate mindreading: 0 (fail), 1 (partial score), and 2 (pass). The label distribution for the full corpus is balanced, however the label distribution for the individual questions vary [1].

We sought to use data augmentation to create a well-balanced dataset in terms of questions and labels. To achieve this, we created a **policy for sampling** examples that we used in the augmentation. We split the MIND-CA corpus per question and per label, resulting in 33 question-label sub corpora. The average size of each sub-corpora is 343, and the smallest number of instances in a sub corpora is 160. We sampled 125 examples from each question-label sub-corpus, 375 from each question, for a total 4,125 examples.

Our sampling strategy ensures that each question-label combination is well represented in the augmentation process. In the original MIND-CA corpus, nine question-label pairs had less than 125 instances. As a preliminary step in the data augmentation process, our in-domain experts rewrote existing responses to improve the balance of the corpus. We used strategy similar to the one used in Hossain et al. (2020). We ran statistical and machine learning experiments to ensure that the additional examples do not introduce biases.

[1]For more details, please refer to Kovatchev et al. (2020)

For our experiments we initially chose a conservative number of examples (each augmentation increases the original corpus size by 36 %), to avoid overfitting on the underrepresented question-label pairs. We used a different random state for each augmentation strategy and we ensured that each sample is representative in terms of demographic distribution (age and gender of the participants).

In a complementary set of experiments, we applied data augmentation directly without the custom sampling strategy. We also experimented with generating larger number of augmented examples (up to 140% of the original corpus size) via oversampling (see Section 7).

In the following subsections, we discuss in more details the different augmentation strategies.

3.1 Dictionary Augmentation

The "dictionary" augmentation strategy is a task-specific synonym substitution. We automatically extract the 20 most frequent words for each of the 11 questions, a total of 220 words. We then ask trained corpus annotators to propose a list of synonyms for each word. The synonyms have the same meaning in the context of the particular question. The meaning of the contextual synonyms may not be the same outside of the context. For example, in Silent Film Question #1, "men" can be replaced with "burglars". We instruct the experts to create as many synonyms as possible for each word. Some words do not have appropriate contextual synonyms. The final synonym dictionary contains 626 synonyms for 148 words [2].

The dictionary augmentation algorithm replaces up to two words in each response with their contextual synonyms. The words and their synonyms are selected at random from the available options.

3.2 Introductory Phrase Augmentation

The task-specific "phrase" augmentation strategy adds a short phrase at the beginning of the response. The appended phrases should not modify the meaning (or score) of the response. An example for such phrase is "I think (that)". Our experts create phrases that contain mental state words, such as "think", "know", and "believe", as this category of words is important when scoring children's mindreading ability. Our corpus annotators proposed a

[2]The implementation of all augmentation strategies and all resources used (lists of synonyms and introductory phrases) can be found online at https://github.com/venelink/augment-acl21/

list of 15 such phrases. We further modify the 15 phrases with 3 optional conjunctions, resulting in 60 different combinations.The "phrase" augmentation appends a random phrase at the beginning of each response, if the response does not already begin with such a phrase.

3.3 Word Replacement Augmentation

Word replacement augmentation is a strategy that automatically replaces up to two randomly selected words with semantically similar words or phrases. The "wordnet" and "ppdb" augmentations replace the selected words with a synonym from WordNet (Fellbaum, 1998) or PPDB (Pavlick et al., 2015) respectively. The "glove" and "fasttext" augmentations replace the selected words with the most similar words (or phrases) using pre-trained GloVe (Pennington et al., 2014) or FastText (Joulin et al., 2016) word embeddings.

We implement the four "word replacement" augmentations using the NLP Augmentation python library (Ma, 2019). For this set of experiments we decided not to use BERT-based contextual word embeddings for augmentation, since we are using a DistilBERT classifier.

3.4 Change of Order Augmentation

The "order" augmentation strategy changes the position of two words in the sentence. Previous work on data augmentation for NLP (Wei and Zou, 2019; Ma, 2019) implement the "order" augmentation by changing the position of the two randomly selected words. We enforce a more stringent rule for our algorithm. Specifically, we select one word at random and change its position with one of its neighbouring words. This change is more conservative than picking two words at random. It also reflects the naturally occurring responses from 7- to 14-year-old children in the database. The reorder process is repeated up to two times.

3.5 Combining Multiple Augmentations

We also experimented with applying multiple augmentation strategies together. For example the "dictionary + phrase" augmentation first replaces up to two words with contextual synonyms and then adds a phrase at the beginning of the response. The data obtained by "combination" augmentations was included in the the "all-lq" and "all-hq" corpora.

4 Measuring Augmentation Quality

The quality of data augmentation models in NLP research is typically evaluated extrinsically, by measuring the performance of automated systems trained on augmented data. Wei and Zou (2019) propose an intrinsic evaluation inspired by the data augmentation research in computer vision. They compare the latent space representations of the original and the augmented sentences and assume that the proximity in latent space indicates that the original labels are conserved.

We argue that a direct comparison of the representation of the texts is not sufficient to determine the quality of the augmentation and the extent to which each strategy preserves the original labels. In natural language, unlike in computer vision, a minor difference in the text and the corresponding representation can cause a significant difference in the meaning of the complex expression and ultimately the label or score assigned to that answer.

We propose a manual evaluation of the different strategies. For each augmentation strategy, we selected 5 random examples from each question-label sub-corpus, adding up to 165 examples per strategy (4% of the full sample). Two trained annotators independently rate the augmented pairs for the 7 different augmentation strategies (a total of 1,155 question-answer pairs). To ensure a fair evaluation, the annotators receive a single file with the examples for all augmented strategies shuffled at random. The inter-annotator agreement was 87% with a Cohen's Kappa of .83.

Augmentation	Quality	Invalid
Phrase	96	1
Order	94.5	3.5
Dictionary	94	2
WordNet	83	10
FastText	77	10
PPDB	73	12
GloVe	68	17

Table 1: Expert comparison of augmentation strategies. Quality - % of pairs where the label does not change. Invalid - % of pairs where the augmented instance is semantically incoherent and cannot be scored.

Table 1 shows the results of the re-annotation for each augmentation strategy. We define "quality" as the % of examples where the re-annotated label was the same as the original label. We also measure the % of "invalid" examples, where both annotators

agreed not to assign a label due to a semantically incoherent response. An example for an incoherent response can be seen in (1).

(1) *Why did the men hide ?*
so telling does'nt get told his .

Based on the analysis, we distinguish between "high quality" augmentation strategies ("phrase", "order", and "dictionary") and "low quality" augmentations ("wordnet", "fasttext", "ppdb", and "glove"). The "high quality" augmentations preserve the label in over 94% of the instances and contain less than 4% invalid responses. The "low quality" augmentations preserve the label in less than 83% of the instances and contain more than 10% invalid responses. According to our raters, GloVe is the worst of all augmentation strategies with 68% quality and 17% invalid.

The expert analysis indicates that, at least in our data, there is a substantial difference in the quality of the different augmentation strategies. The task-specific strategies perform much better than the task-agnostic ones, with the exception of "change of order" augmentation. In the following sections, we perform a number of machine learning experiments to determine if the quality of the data affects the performance of the automated systems.

Corpus	Size	Corpus Contents
orig	11,311	The MIND-CA corpus
uk-20	10,320	The UK-MIND-20 corpus
phrase	15,436	MIND-CA + "phrase"
dict	15,436	MIND-CA + "dictionary"
order	15,436	MIND-CA + "order"
wordnet	15,436	MIND-CA + "wordnet"
fasttext	15,436	MIND-CA + "fasttext"
ppdb	15,436	MIND-CA + "ppdb"
glove	15,436	MIND-CA + "glove"
ab-lq	27,811	MIND-CA + "wordnet", "fasttext", "ppdb", and "glove"
all-lq	44,311	MIND-CA + "wordnet", "fasttext", "ppdb", and "glove" + all 4 synonym substitutions combined with reorder
ab-hq	23,686	MIND-CA + "phrase", "dictionary" and "order"
all-hq	40,186	MIND-CA + "phrase", "dictionary", and "order" + all four possible combinations of the three strategies

Table 2: All Augmented Training Sets

5 Evaluating Automated Systems

In our experiments, we used the two best systems reported by Kovatchev et al. (2020) - a BiLSTM neural network and a DistilBERT transformer. These systems obtained good results on the original MIND-CA corpus and at the same time were lightweight enough to be implemented in a practical end-to-end application for automatic scoring. We used the same configuration and hyperparameters as reported by Kovatchev et al. (2020). We modified the existing classes to incorporate and keep track of data augmentation and to implement additional evaluation on UK-MIND-20. All of our code and data are available online [3].

5.1 Automated Systems. Training setup.

We trained each of the automated systems on 13 different training sets, shown in Table 2. Each set includes the original corpus (MIND-CA) and a number of augmented samples. For example, the **phrase** dataset contained the 11,311 examples from MIND-CA + 4,125 from the "phrase" augmentation, for a total of 15,436 examples.

In addition to the 7 "basic" augmented training sets (one for each augmentation strategy), we also created 4 larger training sets, containing augmented samples from multiple different strategies. The "All Bassic HQ" (**ab-hq**) dataset contains the 11,311 examples from MIND-CA + 4,125 from "phrase" + 4,125 from "dictionary" + 4,125 from "order" for a total of 23,686 examples. Similarly, the "All Basic LQ" (**ab-lq**) dataset contains 27,811 examples from MIND-CA + "wordnet", "fasttext", "ppdb", and "glove".

The two largest datasets, the **all-lq** and the **all-hq** datasets contain the corresponding "all basic" datasets and additional examples obtained by consecutively applying more than one augmentation strategy to the same original data (the "combined" augmentations described in Section 3.5). We kept the "low quality" and the "high quality" data separated, so we can measure the correlation between the "quality" and the performance of the automated systems.

[3] https://github.com/venelink/augment-acl21/

5.2 The UK-MIND-20 Corpus

One of the objectives behind data augmentation is to improve the capabilities of automated systems to generalize to unseen data. However, finding unseen data for the same task is often non-trivial, so researchers typically use train-test split or 10-fold cross validation to evaluate the models.

To provide a fair evaluation benchmark for generalizability, we created a new corpus of children's mindreading ability, the UK-MIND-20 corpus. The data for the corpus is part of our own research on children's mindreading in large-scale study involving 1020 8- to 13-year-old children (556 girls, 453 boys, 11 not disclosed) from the United Kingdom. Children completed three mindreading tasks during whole-class testing sessions led by trained research assistants: Strange Stories task (Happé, 1994), Silent Film task (Devine and Hughes, 2013), and Triangles Task (Castelli et al., 2000).

Each child answered 14 questions: five from the Strange Story Task, six from the Silent Film Task, and three from the Triangles Task. We do not use the responses for the Triangles task for the evaluation of data augmentation, since that task is not part of the MIND-CA corpus. We obtained a total of 10,320 question-answer pairs for the Strange Stories and the Silent Film portion of the corpus. Similar to MIND-CA, UK-MIND-20 also includes the age and gender of the participants and responses to a standardized verbal ability test (Raven, 2008).

The children's responses were scored by two trained research assistants, the same assistants that measured the augmentation quality in Section 4. Each response was scored by one annotator. The inter-annotator agreement was measured on a held-out set of questions. We report an inter-annotator agreement of 94% and a Fleiss Kappa score of .91.

When creating UK-MIND-20, we used the same procedures for administering, scoring, and digitalizing the children responses as the ones used by Kovatchev et al. (2020). The data for the UK-MIND-20 corpus is gathered in a different time-frame (Oct 2019 – Feb 2020) and from different locations than MIND-CA (2014 – 2019).

5.3 Evaluation Criteria

The task defined by Kovatchev et al. (2020) consists of scoring the children's mindreading abilities based on the open-text responses to 11 different questions from the Strange Stories Task and the Silent Film Task using three categories (i.e., fail, partial, pass). A single automated system has to score all 11 questions. In this paper we evaluate the system performance in three ways:

Overall F1: The macro-F1 on the full test set, containing all 11 questions, shuffled at random.

F1-per-Q: We split the test set on 11 parts, one for each question. We obtain the macro-F1 score on each question and calculate the average.

STD-per-Q: Similar to F1-per-Q, we obtain the macro-F1 for each question and then calculate the standard deviation of the performance per question.

The **Overall F1** measures the performance of the system on the full task. **F1-per-Q** and **STD-per-Q** measure the consistency of the system across the different questions. A practical end-to-end system needs to obtain good results in both. The additional data facilitates the statistical analysis of the system performance. This evaluation methodology was proposed by Kovatchev et al. (2019).

For each system we performed a 10-fold cross validation using each corpus from Table 2. For each fold, we evaluated on both the corresponding test set and on the full UK-MIND-20 corpus. Our code dynamically removes from the current training set any augmented examples that are based on the current test set to ensure a fair evaluation. All test sets contain only gold-standard human-labeled examples and do not include any augmented data.

6 Results

Table 3 presents the results of the 13 different training configurations with the DistilBERT transformer, using both question and answer as input[4]. The numbers are the average across 10-fold cross validation. For reference, we also include the results obtained by training the system on UK-MIND-20 and testing on MIND-CA.

The DistilBERT architecture is the best performing system from Kovatchev et al. (2020). The baseline system, trained on the original data already obtained very good results: .925 F1 and .877 F1-per-Q on the MIND-CA corpus and .889 F1 and .839 F1-per-Q on the UK-MIND-20 corpus. We demonstrate that systems trained on either of the two datasets can generalize well on the other one

[4]We carried out 4 different sets of experiments: two classifiers (BILSTM and DistilBERT) and two different input setups (i.e., only the answer or both question and answer). Due to space restrictions, we report only the results for the best system, DistilBERT (question + answer). The findings apply to all sets of experiments. The code and results for all experiments are available online at https://github.com/venelink/augment-acl21/

Training Set	Test-F1	Test-F1-Q	Test-STD	UK20-F1	UK20-F1-Q	UK20-STD
Orig (baseline)	.925	.877	.059	.889	.839	.063
UK-MIND-20	.893	.844	.058	.890	.839	.063
Phrase	.946	.930	.031	**.893**	**.854**	.024
Dictionary	**.947**	**.936**	.028	.892	.853	.024
Order	.947	.933	**.025**	.891	.852	**.022**
FastText	.942	.924	.030	.890	.851	.023
GloVe	.942	.925	**.028**	.891	.849	**.021**
PPDB	.946	.929	.030	.893	.851	.022
WordNet	**.947**	**.932**	.033	**.894**	**.853**	.023
AB-LQ	.967	.957	.021	.895	.855	.021
AB-HQ	.972	.963	.022	.897	.858	**.020**
All-LQ	.978	.973	.015	.895	.957	.021
All-HQ	**.985**	**.980**	**.011**	**.898**	**.858**	.023

Table 3: Performance of a DistilBERT classifier using different augmented sets for training. We report F1, F1-per-Question and standard deviation (per question) on two corpora: Test (MIND-CA), and UK20 (UK-MIND-20).

(F1 of .89 and F1-per-Q of .84). This indicates that the two corpora are compatible and that automated systems can generalize to unseen data.

It is evident in the table that all of the augmentation strategies successfully improved the performance of the automated systems across all evaluation criteria. For the MIND-CA corpus: F1 improved between 1.7 points (FastText) and 6 points (All-HQ); F1-per-Qiestion improved between 4.7 points (FastText) and 10.3 points (All-HQ); STD-per-Question was reduced by between 1.6 points (WordNet) and 4.8 points (All-HQ). For the UK-MIND-20 corpus: F1 improved between 0.1 point (FastText) and 0.9 point (All-HQ); F1-per-Question improved between 1 point (GloVe) and 1.9 points (All-HQ); STD-per-Question was reduced between 3.9 points (dictionary) and 4.2 points (AB-HQ).

Based on these results, we can draw two conclusions. **First**, data augmentation can successfully be used to improve the performance of the systems on the MIND-CA corpus. **Second**, data augmentation also improves the performance of the automated systems on the unseen examples from UK-MIND-20. While the improvement is not as substantial as seen on MIND-CA, the improvement on all three criteria on UK-MIND-20 indicates that the systems are not just overfitting to MIND-CA.

We use the Autorank Python library (Herbold, 2020) to carry out a statistical analysis on the results and compare the performance gain from each of the augmentation strategies. We use the data from both algorithms and input formats, a total of 480 machine learning models, 40 for each dataset. Based on the provided data, Autorank determines that the most appropriate statistical test is the Friedman-Nemeyni test (Demšar, 2006). The Friedman test reports that there is a statistically significant difference between the median values of the populations. That means that some training sets are consistently performing better (or worse) than others. The post-hoc Nemenyi test can be used to determine and visualise which training sets are better and which are worse.

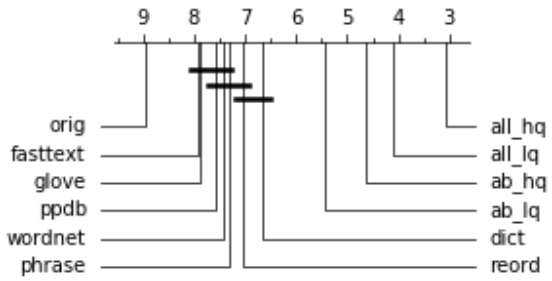

Figure 1: Critical Difference Diagram (all).
Average ranking of training sets (lower is better).
Connected with a line =>not statistically significant.

Figure 1 shows the Critical Difference diagram of the post-hoc Nemenyi test for all training sets. Each set is plotted with its average ranking across all systems. The difference between systems connected with a line is not statistically significant. The original corpus is the worst performing of all datasets with an average rank of 9. The 7 "basic" training sets are grouped in the middle (rank 6.5 to 8). That is, they are all better than the original corpus, but worse than the combined training sets.

There is a significant difference between “All-HQ”, “All-LQ”, “AB-HQ”, and “AB-LQ”. Collectively they are also better than the original training set and the “basic” training sets.

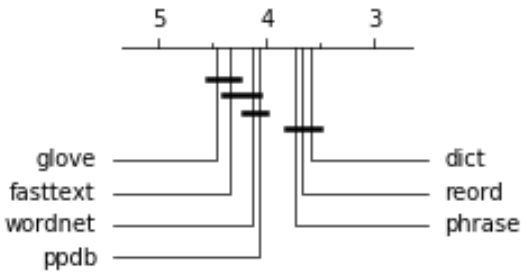

Figure 2: Critical Difference Diagram (basic). Average ranking of training sets (lower is better). Connected with a line =>not statistically significant.

Figure 2 shows the Critical Difference diagram of the post-hoc Nemenyi test applied only to the 7 “basic” augmentations. After removing the outliers (the original corpus and the collections of multiple augmentation), we can observe a clear, statistically significant distinction between “high quality” augmentations (“dictionary”, “phrase”, and “order”) and “low quality” augmentations (“glove”, “fasttext”, “wordnet”, and “ppdb”).

Based on the statistical analysis, we can draw two additional conclusions. **Third**, we found that the most important factor affecting the system performance is the number of training examples. We obtain the best results by combining the examples from various different augmentation strategies. **Fourth**, we demonstrated that when the training size is comparable, the high quality augmentations improve the performance more than the low quality ones. The difference is significant and is consistent both in “basic” datasets and in “combined” datasets. Vector based augmentations (GloVe and FastText) are performing worse than augmentations based on task-specific or task-agnostic knowledge bases.

7 Discussion and Further Experiments

The intrinsic and extrinsic evaluation presented in Section 4 and Section 6 answered the main research questions posed in this paper. We demonstrated that data augmentation can improve the performance of automated systems including on novel, unseen data. We found that the data augmentation strategies vary in preserving the original label and in how much they improve the machine learning systems trained on them. We also showed that automated scoring systems can generalize well from MIND-CA corpus to UK-MIND-20 and the other way around. All these findings are important for further research on mindreading. At the same time, our data augmentation strategies and evaluation methodology can also be extended to other tasks and domains, contributing to the research of Data Augmentation in NLP in general.

In this section we present additional experiments and an analysis of the impact of several different factors in the process of data augmentation [5].

Corpus Size Our experiments indicated that the most important factor for improving the system performance is the corpus size. In Table 3 the systems that perform best are trained on the largest possible amount of data (all-lq/all-hq). To further explore the impact of corpus size, we ran an additional set of experiments. We sampled 500 examples for each question-label subcorpora instead of the original 125, increasing the corpus size four times. For each augmentation strategy this resulted in a corpus approximately the same size as ab-lq.

As expected, the performance of each system increased with corpus size. The ranking of the individual systems remained similar to the one reported with 125 base examples. “High quality” augmentations still performed better than “low quality” ones. The F1, F1-per-Q, and STD-per-Q for the “basic low quality” strategies was approximately the same as the performance for ab-lq. The F1, F1-per-Q, and STD-per-Q for the “basic high quality” strategies was approximately the same as the performance for ab-hq.

This new set of experiments confirmed the importance of corpus size. Even strategies that human experts perceive as “low quality” are improving the performance of the automated systems. And while the ranking consistently favors the “high quality” augmentations, the absolute difference is relatively small. This is in line with the findings on noisy learning which show that machine learning models can be very noise-tolerant (Natarajan et al., 2013). We performed one final experiment by combining the all-lq and all-hq data together, but found no increase or decrease of performance compared with using only the all-hq data.

Sampling Strategy In our experiments, we designed a sampling strategy to ensure that each

[5] Due to space restrictions, we only discuss the overall tendencies. The actual results are available online.

question-response combination appears in the training data with sufficient frequency. In a complementary set of experiments, we evaluated the importance of the sampling. For each augmentation strategy, we created an augmented dataset with 1500 examples for each question, using a standard sampling that keeps the original ratio of the responses. The size of the dataset is the same as sampling 500 examples for each of the 3 labels. We found that for all strategies, the sampling improves Test-F1-Q between .6 and 1 point and reduces STD-per-Q by 1 point. This finding validates our choice of sampling strategy.

Augmentation Strategy In Section 6 we demonstrated that when all parameters (sampling, corpus size) are equal the "high-quality" strategies rank higher than the "low-quality" ones. While the absolute difference in F1 and STD is relatively small on our datasets, the consistency of the performance of the "high-quality" strategies has to be taken into consideration. Furthermore, the quantitative performance is only one factor that has to be considered when choosing a strategy for data augmentation. Reducing the noise in the training data can be a desirable characteristic when interpreting the performance of the neural network models, or when working with sensitive data, such as (e.g.) in the health domain. The task-specific augmentations that we proposed and used may require in-domain experts, however the design is rather simple and the process is not time or labour intensive. After the task-specific resource (dictionary, list of phrases) is created, it can be reused for multiple examples and scales very well with corpus size.

8 Conclusions and Future Work

We presented a systematic comparison of multiple data augmentation strategies for the task of automatic scoring of children's mindreading ability. We argued that the nature of natural language requires a more in-depth analysis of the quality and performance of the different data augmentation strategies. We recruited in-domain experts and incorporated them in the process of evaluation.

We demonstrated that, for some of the augmentation strategies ("glove", "fasttext", "ppdb") there is a substantial portion of the examples (over 20%) where the rating changes or cannot be assigned due to semantically incoherent text. These differences in the datasets cannot be captured trivially via the visualisation techniques that are typically used for intrinsic evaluation. We also found that the difference in augmentation quality corresponds to a difference in the performance of automated systems trained on the data. To the best of our knowledge, this is the first evaluation of data augmentation in NLP that involves both expert evaluation and automatic metrics and the first study that demonstrates the connection between the two.

We carried out further experiments measuring the importance of factors such as corpus size and sampling strategy. Our findings on the quality and efficiency of data augmentation strategies and on the use of task-specific resources are relevant for researchers in the area of data augmentation, specifically in domains where the quality of the training gold examples is important or where the amount of data is very limited.

For the purpose of evaluation, we also created a new corpus: UK-MIND-20. It is the second corpus for automatic scoring of mind reading in children. We demonstrated that systems trained on MIND-CA generalize well on UK-MIND-20. We also showed that data augmentation improves the performance on unseen data. These findings are promising both for the task of scoring children's mindreading and for the use of data augmentation in NLP. To the best of our knowledge, this is the first work where augmentation is evaluated on novel, unseen data for the same task.

This work opens several directions of future work. As a direct continuation of this research, we will incorporate the best performing automated systems and data augmentation techniques in the work of developmental psychologists. This will facilitate a large-scale studies on mindreading in children and adolescents. We are also exploring the possibility of using NLP to address other time and labour intensive problems within psychology. Open-ended short text responses are widely-used within psychological research and the good results obtained in this paper can be replicated in other similar tasks.

Acknowledgements

We would like to thank Imogen Grumley Traynor and Irene Luque Aguilera for the annotation and the creation of the lists of synonyms and phrases. We also want to thank the anonymous reviewers for their feedback and suggestions. This project was funded by a grant from Wellcome to R. T. Devine.

Ethical Statement

The study was approved by the University of Birmingham STEM Research Ethics Committee and complies with the British Psychological Society Code of Human Research Ethics (2014). Parents and caregivers were provided with detailed information about the study at least one week in advance of data collection and given the opportunity to opt out of the study. Children were also permitted to opt out of the study on the day of data collection without consequence. Data were anonymous at source as children did not provide names or contact information to the research team.

References

Robin Banerjee, Dawn Watling, and Marcella Caputi. 2011. Peer relations and the understanding of faux pas: Longitudinal evidence for bidirectional associations. *Child Development*, 82(6):1887–1905.

F Castelli, F Happe, U Frith, and C Frith. 2000. Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage*, 12(3):314 – 325.

Jack Cotter, Kiri Granger, Rosa Backx, Matthew Hobbs, Chung Yen Looi, and Jennifer H. Barnett. 2018. Social cognitive dysfunction as a clinical marker: A systematic review of meta-analyses across 30 clinical conditions. *Neuroscience and Biobehavioral Reviews*, 84:92 – 99.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.

Rory T Devine, Naomi White, Rosie Ensor, and Claire Hughes. 2016. Theory of mind in middle childhood: Longitudinal associations with executive function and social competence. *Developmental psychology*, 52(5):758—771.

R.T. Devine and C. Hughes. 2013. Silent films and strange stories: theory of mind, gender, and social experiences in middle childhood. *Child Development*, 82:989–1003.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Elian Fink, Sander Begeer, Candida C. Peterson, Virginia Slaughter, and Marc de Rosnay. 2015. Friendlessness and theory of mind: A prospective longitudinal study. *British Journal of Developmental Psychology*, 33(1):1–17.

Fei Gao, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu. 2019. Soft contextual data augmentation for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5539–5544, Florence, Italy. Association for Computational Linguistics.

FG Happé. 1994. An advanced test of theory of mind: understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and developmental disorders*, 24(2):129—154.

Steffen Herbold. 2020. Autorank: A python package for automated ranking of classifiers. *Journal of Open Source Software*, 5(48):2173.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1234–1245, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Claire Hughes and Rory Devine. 2015. Individual differences in theory of mind from preschool to adolescence: Achievements and directions. *Child Development Perspectives*, 9.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Venelin Kovatchev, M. Antonia Marti, Maria Salamo, and Javier Beltran. 2019. A qualitative evaluation framework for paraphrase identification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 568–577, Varna, Bulgaria. INCOMA Ltd.

Venelin Kovatchev, Phillip Smith, Mark Lee, Imogen Grumley Traynor, Irene Luque Aguilera, and Rory Devine. 2020. "what is on your mind?" automated scoring of mindreading in childhood and early adolescence. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6217–6228, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Vukosi Marivate and Tshephisho Sefara. 2020. Improving short text classification through global augmentation methods. In *Machine Learning and Knowledge Extraction*, pages 385–399, Cham. Springer International Publishing.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online. Association for Computational Linguistics.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, page 1196–1204, Red Hook, NY, USA. Curran Associates Inc.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevich, Benjamin Van Durme, and Chris Callison-Burch. 2015. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

John C. Raven. 2008. Raven's standard progressive matrices and mill hill vocabulary scale.

Connor Shorten and T. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:1–48.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.