Work Smart –
Reducing Human Effort in Short-Answer Grading

Margot Mieskes, Hochschule Darmstadt
Ulrike Padó, Hochschule für Technik Stuttgart

Introduction

- Testing is an integral part of (language) teaching
- Specifically in focus: Tests with Short-Answer Questions (SAQs) for language or content assessment

Short-Answer Questions

Example from CREE corpus, Meurers et al. (2011b): Reading Comprehension

- Read text "Television and Children"
- Question: How is violence portrayed in cartoons according to the article?
- Student Answer: *There are underlying themes of justice and punishment, that is, the "bad guys" do not usually win.*
- Grader 1: correct, grader 2: correct

Our Goal

- Grading SAQs takes time, especially for large cohorts or frequently repeated tests

Our Goal

- Grading SAQs takes time, especially for large cohorts or frequently repeated tests
- Reduce human grading effort!

Our Goal

- Grading SAQs takes time, especially for large cohorts or frequently repeated tests
- Reduce human grading effort!
- Specifically: Machines do some of the work and humans step in where machines fail

Our Goal

- Grading SAQs takes time, especially for large cohorts or frequently repeated tests
- Reduce human grading effort!
- Specifically: Machines do some of the work and humans step in where machines fail
- Determine likely machine failure through measures like Accuracy and Fleiss' κ

Our Goal

- Grading SAQs takes time, especially for large cohorts or frequently repeated tests
- Reduce human grading effort!
- Specifically: Machines do some of the work and humans step in where machines fail
- Determine likely machine failure through measures like Accuracy and Fleiss' κ
- This means **not every student answer will be human-graded**
- Appropriate for placement testing etc., where the overall grade is reported

Outline of Talk

- Machine Graders: Data, features and evaluation

Outline of Talk

- Machine Graders: Data, features and evaluation
- Study 1: Human performance

Outline of Talk

- Machine Graders: Data, features and evaluation
- Study 1: Human performance
- Study 2: Combining machine graders

Data Sets

Corpus	**#Questions/ #Answers**	**Language**
ASAP (www.kaggle.com/c/asap-sas)	5/8182	EN
SEB (Dzikovska et al., 2013)	135/4969	
Beetle (Dzikovska et al., 2013)	47/3941	
Mohler (Mohler et al., 2011)	81/2273	
CREE (Meurers et al., 2011a)	61/566	
CREG (Meurers et al., 2011b)	85/543	GER
CSSAG (Padó and Kiefer, 2015)	31/1926	

Models and Features

- Three learning algorithms: Random Forest, Support Vector Machine, Decision Tree
- Features designed to cover the feature types used in the literature: N-Grams, text similarity measures, dependency parses and deep semantic representations, textual entailment (Padó, 2016)

Evaluation Measures

- Comparing predictions and gold annotation
  - Accuracy: Which percentage of the answers has been labelled correctly?

Evaluation Measures

- Comparing predictions and gold annotation
  - Accuracy: Which percentage of the answers has been labelled correctly?
- Comparing parallel annotations: Fleiss' κ
  - Do the annotators agree more than they would by chance?

Evaluation Measures

- Comparing predictions and gold annotation
  - Accuracy: Which percentage of the answers has been labelled correctly?
- Comparing parallel annotations: Fleiss' κ
  - Do the annotators agree more than they would by chance?
  - Compare multiple annotators, for multiple target grades, down to the individual answer

Outline of Talk

- Machine Graders: Data, features and evaluation
- Study 1: Human performance
- Study 2: Combining machine graders

Human Performance

Measure		ASAP	CREG	CSSAG	Mohler
Human	Acc	93.7	85.8	89.9	83.5
Human	κ	0.82	0.64	0.54	0.41

- Easiest case: Correct-incorrect decision

Human Performance

Measure		ASAP	CREG	CSSAG	Mohler
Human	Acc	93.7	85.8	89.9	83.5
Human	κ	0.82	0.64	0.54	0.41

- Easiest case: Correct-incorrect decision
- Doubly-annotated corpora show large variation between high-volume and ad-hoc testing

Human Performance

Measure		ASAP	CREG	CSSAG	Mohler
Human	Acc	93.7	85.8	89.9	83.5
Human	κ	0.82	0.64	0.54	0.41

- Easiest case: Correct-incorrect decision
- Doubly-annotated corpora show large variation between high-volume and ad-hoc testing
- Accuracies around 85% have been accepted: $\sim$ 15% error

Our Goal

- Reduce human grading effort!
- Ideally, at the same error levels as before
- Idea: Machines do some of the work and humans step in where machines fail

Strategy

- Train several classifiers and collect their predictions: Ensemble learning
- As long as each ensemble learner is better than chance, the ensemble is guaranteed to improve over the individual learners

Strategy

- Train several classifiers and collect their predictions: Ensemble learning
- As long as each ensemble learner is better than chance, the ensemble is guaranteed to improve over the individual learners
- Also, now there are multiple annotations! Use κ to determine ensemble agreement

Strategy

- Train several classifiers and collect their predictions: Ensemble learning
- As long as each ensemble learner is better than chance, the ensemble is guaranteed to improve over the individual learners
- Also, now there are multiple annotations! Use κ to determine ensemble agreement
- Assumption: The better ensemble agreement is on a prediction, the more reliable it is

Strategy

- Train several classifiers and collect their predictions: Ensemble learning
- As long as each ensemble learner is better than chance, the ensemble is guaranteed to improve over the individual learners
- Also, now there are multiple annotations! Use κ to determine ensemble agreement
- Assumption: The better ensemble agreement is on a prediction, the more reliable it is
- **Human checks of the machine labels are only needed for unreliable predictions**

Verifying the Assumption

Classes	ASAP	CREE	CREG	CSSAG	Mohler	Beetle	SEB
Binary	**10%**	**15%**	**12%**	24%	**9%**	**17%**	25%

- Percentage of incorrect predictions made in full agreement

Verifying the Assumption

Classes	ASAP	CREE	CREG	CSSAG	Mohler	Beetle	SEB
Binary	**10%**	**15%**	**12%**	24%	**9%**	**17%**	25%
Multi	18%	–	–	38%	30%	–	–

- Percentage of incorrect predictions made in full agreement
- For most corpora, decisions made in full agreement are as reliable as human annotators
- The task is noticeably harder for more than two grade levels

Verifying the Assumption

Classes	ASAP	CREE	CREG	CSSAG	Mohler	Beetle	SEB
Binary	**10%**	**15%**	**12%**	24%	**9%**	**17%**	25%
Multi	18%	–	–	38%	30%	–	–

- Percentage of incorrect predictions made in full agreement
- For most corpora, decisions made in full agreement are as reliable as human annotators
- The task is noticeably harder for more than two grade levels

Identifying Unreliable Predictions

- Clearly: Any answers the ensemble couldn't label (no agreement; multiclass case only)

Identifying Unreliable Predictions

- Clearly: Any answers the ensemble couldn't label (no agreement; multiclass case only)
- Next: Any answers the ensemble didn't label in full agreement

Effort and Remaining Error: Binary Case

		ASAP	CREE	CREG	CSSAG	Mohler	Beetle	SEB
NA	Effort	0	0	0	0	0	0	0
only	Error	16%	**15%**	16%	29%	**11%**	23%	30%

- Binary case: Pass-fail decision
- First: Any answers the ensemble couldn't label (none here!)

Effort and Remaining Error: Binary Case

		ASAP	CREE	CREG	CSSAG	Mohler	Beetle	SEB
NA	Effort	0	0	0	0	0	0	0
only	Error	16%	**15%**	16%	29%	**11%**	23%	30%
all	Effort	20%	19%	12%	27%	7%	24%	28%
PartA	Error	**8%**	**9%**	**9%**	17%	**8%**	**13%**	18%

- Binary case: Pass-fail decision
- First: Any answers the ensemble couldn't label (none here!)
- Next: Any answers the ensemble didn't label in full agreement: Remaining error below human levels at 20-30% of answers graded

Effort and Remaining Error: Multiclass Case

		ASAP	CSSAG	Mohler
NA	Effort	7%	4%	9%
only	Error	28%	44%	41%

- Multiclass case: 5 to 10-way decision
- First: Revise answers the ensemble couldn't label – more work clearly needed

Effort and Remaining Error: Multiclass Case

		ASAP	CSSAG	Mohler
NA	Effort	7%	4%	9%
only	Error	28%	44%	41%
all	Effort	39%	50%	59%
PartA	Error	**11%**	19%	**15%**

- Multiclass case: 5 to 10-way decision
- First: Revise answers the ensemble couldn't label – more work clearly needed
- Second: Revise cases of partial agreement

Effort and Remaining Error: Multiclass Case

		ASAP	CSSAG	Mohler
NA	Effort	7%	4%	9%
only	Error	28%	44%	41%
all	Effort	39%	50%	59%
PartA	Error	**11%**	19%	**15%**

- Multiclass case: 5 to 10-way decision
- First: Revise answers the ensemble couldn't label – more work clearly needed
- Second: Revise cases of partial agreement
- Acceptable error levels, but more manual work than in the binary case

Did we reach our goal?

- Human effort can be reduced while error levels remain stable

Did we reach our goal?

- Human effort can be reduced while error levels remain stable
- Our approach works better for the learner corpora: Binary decisions, reliable machine learners

Did we reach our goal?

- Human effort can be reduced while error levels remain stable
- Our approach works better for the learner corpora: Binary decisions, reliable machine learners
- For the multiclass case, similar efficiency as reported in Horbach et al. (2014) (60% effort saved, 15% remaining error); much better for pass-fail

What else to consider?

- When planning ensemble-supported grading:
  - Know your requirements

What else to consider?

- When planning ensemble-supported grading:
  - Know your requirements
- When creating corpora:
  - Few classes
  - Well-trained annotators
  - Size matters (somewhat)

What else to consider?

- When planning ensemble-supported grading:
  - Know your requirements
- When creating corpora:
  - Few classes
  - Well-trained annotators
  - Size matters (somewhat)
- Future work
  - Run a user study: Get feedback on usefulness and usability