The Alan Turing Institute

WARWICK
THE UNIVERSITY OF WARWICK

Utrecht University

Aiming beyond the Obvious: Identifying Non-Obvious Cases in Semantic Similarity Datasets

Nicole Peinelt

ACL, Italy, 30 July 2019

Authors

Nicole Peinelt

Dr. Maria Liakata

The Alan Turing Institute, UK

University of Warwick, UK

Dr. Dong Nguyen

The Alan Turing Institute, UK

Utrecht University, The Neverlands

Semantic Similarity Detection for English

– Given a sentence pair, predict binary relatedness

Dataset	Task
MSRP	Paraphrase detection
Quora	Question paraphrase detection
STS	Similarity scoring
SemEval 2017 Task C	(A) CQA answer ranking (B) CQA paraphrase ranking (C) CQA answer ranking

Examples from Quora

Positive label (paraphrase)	Negative label (not a paraphrase)
Which is the best way to learn coding? --- How do you learn to program?	What is meant by ‘*e*’ in mathematics? --- What is meant by mathematics?

Examples from Quora

Positive label (paraphrase)	Negative label (not a paraphrase)
Which is the best way to learn coding? --- How do you learn to program?	What is meant by '*e*' in mathematics? --- What is meant by mathematics?
What's the origin of the word o'clock? --- What is the origin of the word o'clock?	What are the range of careers in Biotechnology in Indonesia? --- How do you tenderize beef stew meat?

Examples from Quora

Hard

Easy

Positive label (paraphrase)	Negative label (not a paraphrase)
Which is the best way to learn coding? --- How do you learn to program?	What is meant by ‘*e*’ in mathematics? --- What is meant by mathematics?
What’s the origin of the word o’clock? --- What is the origin of the word o’clock?	What are the range of careers in Biotechnology in Indonesia? --- How do you tenderize beef stew meat?

But equal weight for evaluation

Background

- Scrutiny of model and dataset biases in NLU
  (Wadhwa et al. 2018; Rajpurkar et al., 2018)
- High model performance even without crucial task information
  (Kaushik and Lipton 2018)
- Model performance inflated by annotation artefacts
  (Gururangan et al. 2018)

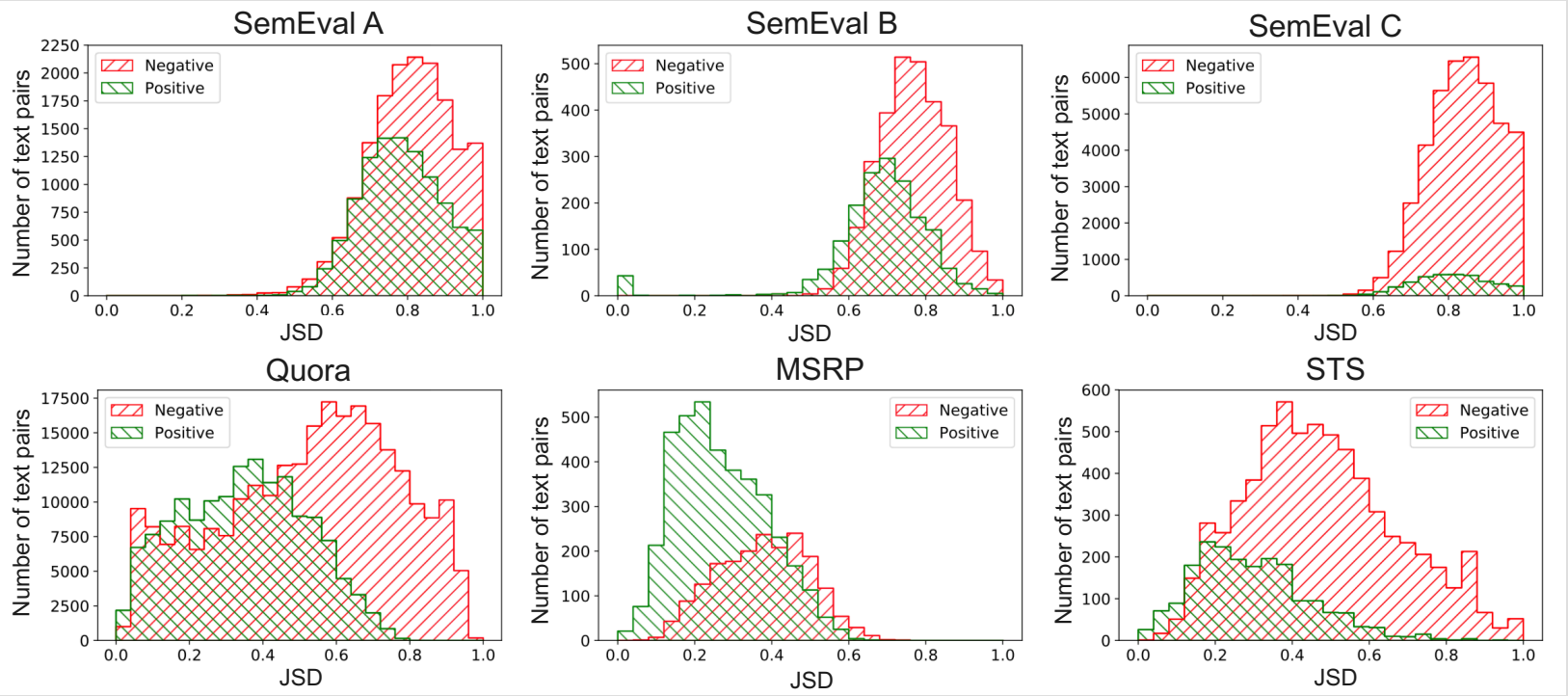
Questions

1. How to automatically distinguish obvious vs. non-obvious examples?

2. How common are obvious examples?

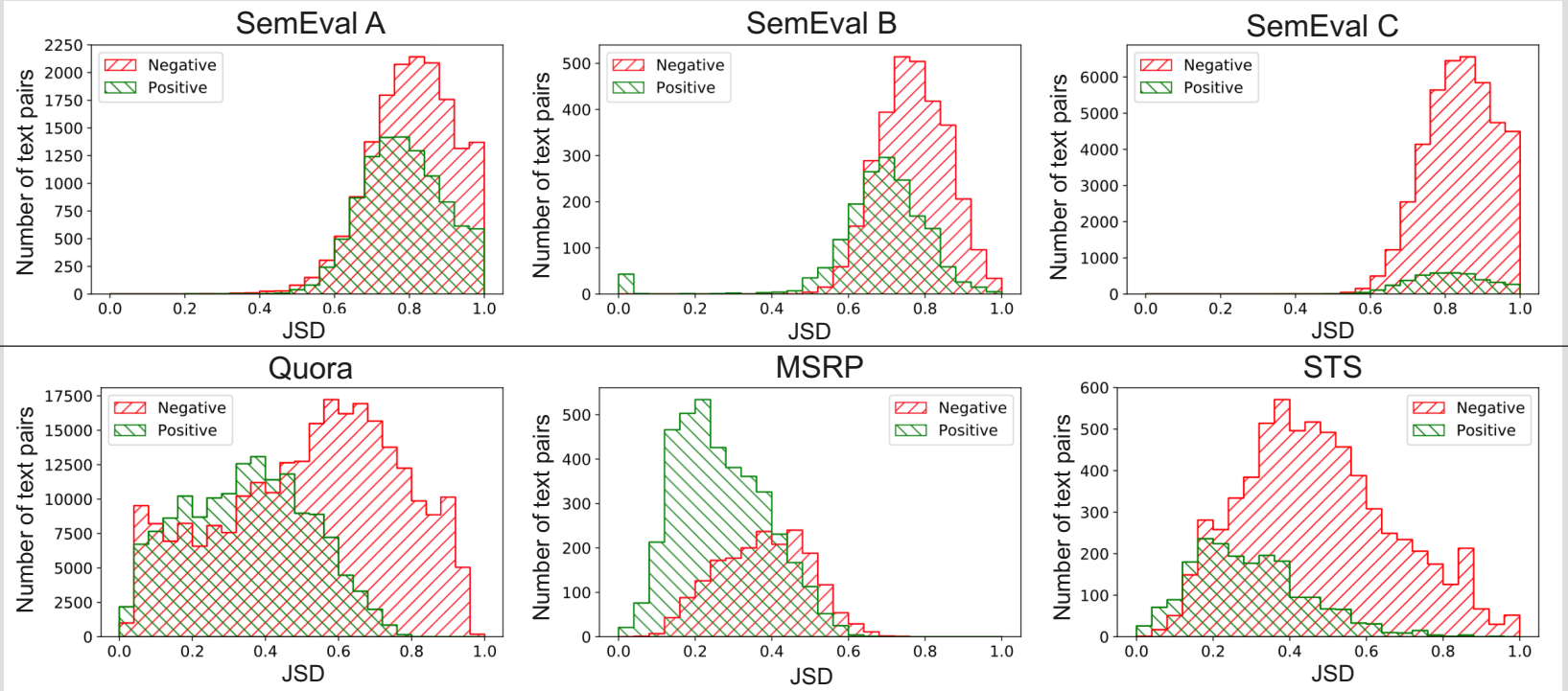
3. How to take this into account for evaluation?

Lexical Divergence by Labels



Lexical Divergence by Labels

similar distributions



distinct distributions

Obvious vs. Non-Obvious Examples

	Positive label	Negative label
Low lexical divergence (JSD≤median)	What's the origin of the word o'clock? --- What is the origin of the word o'clock?	What is meant by '*e*' in mathematics? --- What is meant by mathematics?
High lexical divergence (JSD>median)	Which is the best way to learn coding? --- How do you learn to program?	What are the range of careers in Biotechnology in Indonesia? --- How do you tenderize beef stew meat?

Obvious vs. Non-Obvious Examples

	Positive label	Negative label
Low lexical divergence (JSD≤median)	What's the origin of the word o'clock? --- What is the origin of the word o'clock? **P_o**	What is meant by '*e*' in mathematics? --- What is meant by mathematics?
High lexical divergence (JSD>median)	Which is the best way to learn coding? --- How do you learn to program?	What are the range of careers in Biotechnology in Indonesia? --- How do you tenderize beef stew meat? **N_o**

Obvious vs. Non-Obvious Examples

	Positive label	Negative label
Low lexical divergence (JSD≤median)	What's the origin of the word o'clock? **P_o** What is the origin of the word o'clock?	What is meant by '*e*' in mathematics? **N_n** What is meant by mathematics?
High lexical divergence (JSD>median)	Which is the best way to learn coding? **P_n** How do you learn to program?	What are the range of careers in Biotechnology in Indonesia? **N_o** How do you tenderize beef stew meat?

Human Annotation Statistics on Quora

	Fleiss' Kappa	Avg. time per pair	Instances
P_o	0.6429	11.58s	35
P_n	0.0878	11.68s	15
N_o	0.3886	12.50s	34
N_n	0.0892	13.83s	16
Total			100

Obvious Cases in Datasets

	SemEval			Quora	MSRP	STS
	A	B	C			
P_o	5893	1162	2492	107612	2398	1597
P_n	4428	531	1590	41691	1502	409
N_o	8842	1843	22155	160410	1398	3900
N_n	7377	1213	21253	94632	503	2719
Obvious	56%	63%	52%	66%	65%	64%
Median	0.80	0.79	0.82	0.53	0.52	0.52

Evaluation

- Report non-obvious F1
- **Performance decrease & different ranking across SemEval tasks**

	IIT UHH	Bunji	KeLP	EICA
F1	0.144 (2nd)	**0.197** (1st)	0.121 (3rd)	0.008
$F1_n$	0.047 (2nd)	0.028 (3rd)	**0.054** (1st)	0.000

Evaluation for SemEval Task C

Conclusion

1. Criterion to distinguish obvious vs. non-obvious items
2. Existing datasets contain >50% of obvious examples
3. Non-obvious F1 score to evaluate on difficult cases

Thank you

@wuningxi
n.peinelt@warwick.ac.uk