ACL 2019: Evaluation

Deep Dominance - How to Properly Compare Deep Neural Models

Rotem Dror **Segev Shlomov** **Roi Reichart**
Faculty of Industrial Engineering and Management, Technion, IIT
rtmdrr|segevs|roiri@technion.ac.il

- Being non-convex, DNN’s convergence point depends on the random values chosen at initialization and during training

- With so many degrees of freedom governed by random and arbitrary values, when comparing two DNNs it is not possible to consider a single test-set evaluation score for each model

- We might compare just the best models that someone happened to train rather than the methods themselves

three criteria that a DNN comparison tool should meet

(a) Since we observe only a sample from the population score distribution of each model, the decision should be significant under well justified statistical assumptions.

(b) The decision mechanism should be powerful, being able to make decisions in most possible decision tasks;

(c) Since both models depend on random decisions, it is likely that none of them is promised to be superior over the other in all cases (e.g. with all possible random seeds). A powerful comparison tool should hence augment its decision with a confidence score, reflecting the probability that the superior model will indeed produce a better output

Reimers and Gurevych (2017, 2018)

1. Collection of statistics (mean, median, stdev, max, min)

“-”: statistical significance; power is limited

This approach works in 49.01% of the cases

2. Significance testing for *Stochastic Order*

“-”: it does not provide information beyond its decision if one of the distributions is stochastically dominant over the other or not; power is limited

This approach works in 0.98% of the cases

Stochastic dominance

The simplest case of stochastic dominance is **statewise dominance** (also known as **state-by-state dominance**), defined as follows:

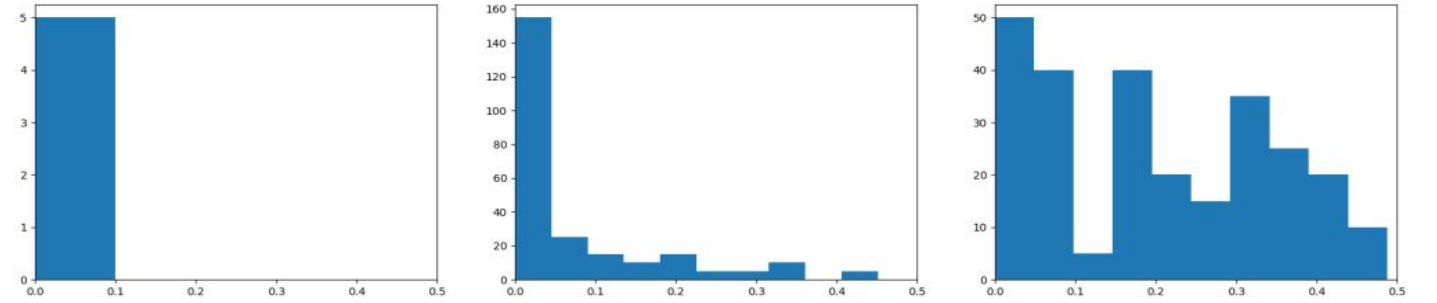
Random variable A is statewise dominant over random variable B if A gives at least as good a result in every state (every possible set of outcomes), and a strictly better result in at least one state.

(Wikipedia)

A new comparison tool

Almost Stochastic Order between two distributions (Alvarez-Esteban et al., 2017; del Barrio et al. (2018))

the test returns a variable *eps ϵ [0; 1]*, that quantifies the degree to which one algorithm is stochastically larger than the other, with *eps = 0* reflecting stochastic order



(a) Case A (b) Case B (c) Case C

Figure 1: An histogram of ϵ values of the ASO method for cases A, B and C.

We need to talk about standard splits

Kyle Gorman
City University of New York
kgorman@gc.cuny.edu

Steven Bedrick
Oregon Health & Science University
bedricks@ohsu.edu

- replication and reproduction experiments with nine part-of-speech taggers published between 2000 and 2018, each of which reports state-of-the-art performance on a widely-used “standard split”

- failure to reliably reproduce some rankings using randomly generated splits
- randomly generated splits should be used in system comparison

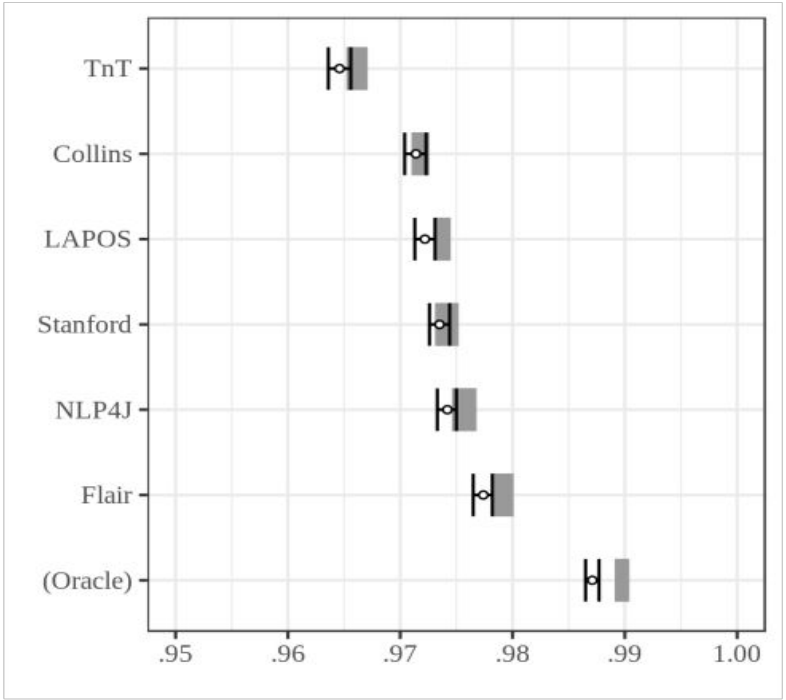


Figure 1: A visualization of Penn Treebank token accuracies in the two experiments. The whiskers shows accuracy and 95% confidence intervals in experiment 1, and shaded region represents the range of accuracies in experiment 2.

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

Nicole Peinelt[1,2] and **Maria Liakata**[1,2] and **Dong Nguyen**[1,3]
[1]The Alan Turing Institute, London, UK
[2]University of Warwick, Coventry, UK
[3]Utrecht University, Utrecht, The Netherlands

id	case	documents
160174	P_o	what‘s the origin of the word o‘clock? what is the origin of the word o‘clock?
115695	P_n	which is the best way to learn coding? how do you learn to program?
193190	N_o	what are the range of careers in biotechnology in indonesia? how do you tenderize beef stew meat?
268368	N_n	what is meant by ‘e‘ in mathematics? what is meant by mathematics?

Table 1: Examples for difficulty cases from the development set of the Quora dataset. o=obvious, n=non-obvious, N=negative label, P=positive label

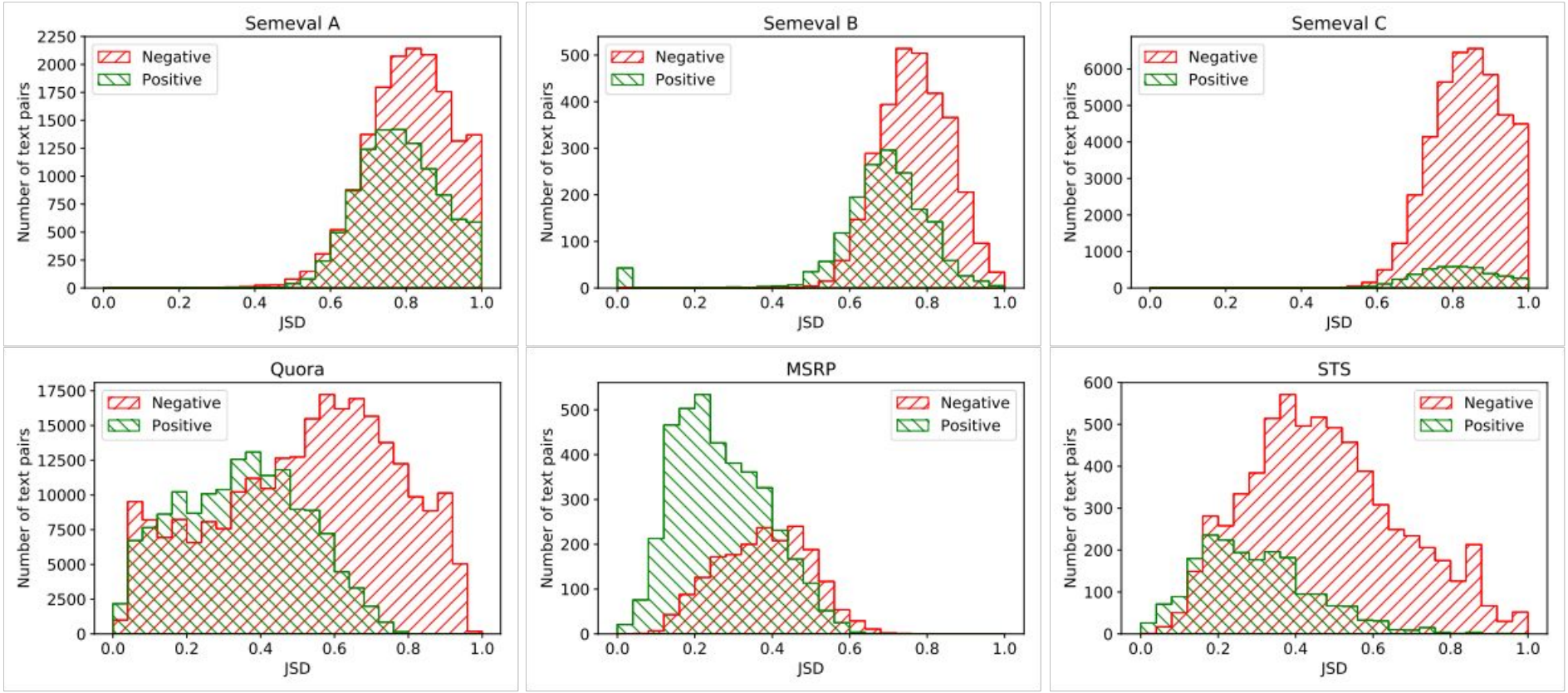


Figure 1: Lexical divergence distribution by labels across datasets. JSD=Jensen-Shannon divergence.

	positive label	negative label
low div	obvious pos (P_o)	non-obvious neg (N_n)
high div	non-obvious pos (P_n)	obvious neg (N_o)

id	case	documents
160174	P_o	what‘s the origin of the word o‘clock? what is the origin of the word o‘clock?
115695	P_n	which is the best way to learn coding? how do you learn to program?
193190	N_o	what are the range of careers in biotechnology in indonesia? how do you tenderize beef stew meat?
268368	N_n	what is meant by ‘e‘ in mathematics? what is meant by mathematics?

Table 1: Examples for difficulty cases from the development set of the Quora dataset. o=obvious, n=non-obvious, N=negative label, P=positive label

there are more obvious positives (P_o) than non-obvious positives (P_n) and more obvious negatives (N_o) than non-obvious negatives (N_n).

All obvious cases combined (P_o+N_o) make up more than 50% of pairs across all 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
o	56	63	52	66	65	64
m	0.80	0.79	0.82	0.53	0.52	0.52

Table 5: Difficulty case splits across datasets (train, dev and test combined). o=obvious, m=median JSD.