# Semi-Supervised Learning and Domain Adaptation in Natural Language Processing

6 Evaluating under Bias

6.1 What is Language?

6.2 Significance Across Corpora

小野寺喜行

# What is Language?

In NLP we often talk about having "a good parser of German”
or whether it is harder to parse Chinese than to parse English.

↓ however

We never seem to answer those questions.

Example)

In order to test whether a parser A is a better parser of German than a parser B,
we would have to evaluate A and B on a representative sample of German.

↓ but

Such samples probably do not exist.

# What is Language?

If we can establish significant correlations
between data characteristics and performance

We may be able to say something.

About whether A will perform better than B
on new datasets with specific characteristics

Even in the absence of a representative sample of database

# Significance Across Corpora

Weaknesses that running a paired t-test
over the results of two classifiers on m datasets

(a) scores need to be commensurable for averaging
to make sense.

(b) unless we have about thirty or more datasets,
the performances across datasets need
to be normally distributed, which they probably are not.

(c) The t-test is very sensitive to outliers.

Instead

Recommended doing a Wilcoxon signed rank test
over multiple datasets.

(By Demsar)