Estimating effect size across datasets

Anders Søgaard
Center for Language Technology
University of Copenhagen
soegaard@hum.ku.dk

Abstract

Most NLP tools are applied to text that is different from the kind of text they were evaluated on. Common evaluation practice prescribes significance testing across data points in available test data, but typically we only have a single test sample. This short paper argues that in order to assess the robustness of NLP tools we need to evaluate them on diverse samples, and we consider the problem of finding the most appropriate way to estimate the true effect size across datasets of our systems over their baselines. We apply *meta-analysis* and show experimentally – by comparing estimated error reduction over observed error reduction on held-out datasets – that this method is significantly more predictive of success than the usual practice of using macro- or micro-averages. Finally, we present a new parametric meta-analysis based on non-standard assumptions that seems superior to standard parametric meta-analysis.

1 Introduction

NLP tools and online services such as the Stanford Parser or Google Translate are used for a wide variety of purposes and therefore also on very different kinds of data. Some use the Stanford Parser to parse literature (van Cranenburgh, 2012), while others use it for processing social media content (Brown, 2011). The parser, however, was not necessarily evaluated on literature or social media content during development. Still, users typically expect reasonable performance on any natural language input. This paper asks what we as developers can do to estimate the effect of a change to our system – not on the labeled test data that happens to be available to us, but on future, still unseen datasets provided by our end users.

The usual practice in NLP is to evaluate a system on a small sample of held-out labeled data. The observed effect size on this sample is then validated by significance testing across data points, testing whether the observed difference in performance means is likely to be due to mere chance. The preferred significance test is probably the non-parametric paired bootstrap (Efron and Tibshirani, 1993; Berg-Kirkpatrick et al., 2012), but many researchers also resort to Student's t-test for dependent means relying on the assumption that their metric scores are normally distributed.

Such significance tests tell us nothing about how likely our change to our system is to lead to improvements on new datasets. The significance tests all rely on the assumption that our datapoints are sampled i.i.d. at random. The significance tests only tell us how likely it is that the observed difference in performance means would change if we sampled a bigger test sample the same way we sampled the one we have available to us right now.

In standard machine learning papers a similar situation arises. If we are developing a new perceptron learning algorithm, for example, we are interested in how likely the new learning algorithm is to perform better than other perceptron learning algorithms *across* datasets, and we may for that reason evaluate it on a large set of repository datasets.

Demsar (2006) presents motivation for using non-parametric methods such as the Wilcoxon signed

Proceedings of NAACL-HLT 2013, pages 607–611,
Atlanta, Georgia, 9–14 June 2013. ©2013 Association for Computational Linguistics

rank test to estimate significance across datasets. The t-test is based on means, and typically results across datasets are not commensurable. The t-test is also extremely sentitive to outliers. Notice also that typically we do not have enough datasets to do paired bootstrapping (van den Noortgate and Onghena, 2005).

In this paper we will assume that the Wilcoxon signed rank test provides a reasonable estimate of the significance of an observed difference in performance means across datasets, or of the significance of observed error reductions, but note that this still depends on the assumption that datasets are sampled i.i.d. at random. More importantly, a non-parametric test across data sets does not provide an actual estimate of the effect size. Estimating effect size is important, e.g. when there is a trade-off between performance gains and computational efficiency.

In evaluations across datasets in NLP we typically use the macro-average as an estimate of effect size, but in other fields such as psychology or medicine it is more common to use a weighted mean obtained using what is known as the **fixed effects model** or the **random effects model** for meta-analysis.

The experiments reported on in this paper focus on estimating error reduction and show that meta-analysis is generally superior to macro- and micro-average in terms of predicting future error reductions. Parametric meta-analysis, however, overparameterizes the distribution of error reductions, leading to some instability. While meta-analysis is generally superior to macro-average, it is sometimes off by a large margin. We therefore introduce a new parametric meta-analysis that seems better suited to predicting error reductions. In our experiments test set sizes are balanced, so micro-averages will be near-identical to macro-averages.

2 Meta-analysis

Meta-analysis is the statistical analysis of the effect sizes of several studies and is very popular in fields such as psychology or medicine. Meta-analysis has not been applied very often to NLP. In NLP most people work on applying *new* methods to *old* datasets, and meta-analysis is designed to analyze series of studies applying *old* methods to *new* datasets, e.g. running the same experiments on new subjects. However, meta-analysis *is* applicable to experiments with multiple datasets.

In psychology or medicine you often see studies running similar experiments on different samples with very different results. Meta-analysis stems from the observation that if we want to estimate an effect from a large set of studies, the average effect across all the studies will put too much weight on results obtained on small datasets in which you typically see more variance. The most popular approaches to meta-analysis are the fixed effects and the random effects model. The fixed effects model is applicable when you assume a true effect size (estimated by the individual studies). If you cannot make that assumption because the studies may differ in various aspects, leading the within-study estimates to be estimates of slightly different effect sizes, you need to use the random effects model. Both approaches to meta-analysis are parametric and rely on the effect sizes to be normally distributed.

2.1 Fixed effects model

In the fixed effects model we weight the effect sizes $T_1, \ldots, T_M$ – or error reductions, in our case – by the inverse of the variance v_i in the study, i.e. $w_i = \frac{1}{v_i}$. The combined effect size T is then:

$$\hat{T} = \frac{\sum_{i\geq 1}^{M} w_i T_i}{\sum_{i\geq 1}^{M} w_i}$$

The variance of the combined effect is now:

$$v = \frac{1}{\sum_{i\geq 1}^{M} w_i}$$

and the 95% confidence interval is then $\hat{T} \pm 1.96\sqrt{v}$.

2.2 Random effects model

In the random effects model we replace the variance v_i with the variance plus between-studies variance τ^2:

$$\tau^2 = \frac{\sum_{i\geq 1}^{k} w_i T_i^2 - \frac{(\sum_{i\geq 1}^{k} w_i T_i)^2}{\sum_{i\geq 1}^{k} w_i} - df}{\sum_{i\geq 1}^{k} w_i - \frac{\sum_{i\geq 1}^{k} w_i^2}{\sum_{i\geq 1}^{k} w_i}} \qquad (1)$$

with $df = N - 1$, except all negative values are replaced by 0.

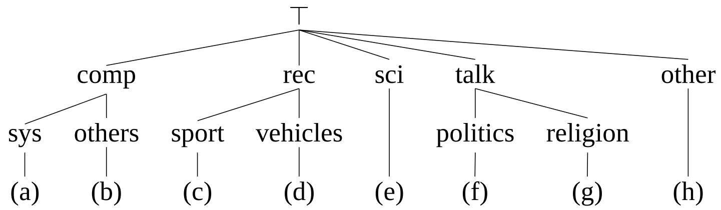

Figure 1: Hierarchical structure of 20 Newsgroups. (a) IBM, MAC, (b) GRAPHICS, MS-WINDOWS, X-WINDOWS, (c) BASEBALL, HOCKEY, (d) AUTOS, MOTORCYCLES, (e) CRYPTOGRAPHY, ELECTRONICS, MEDICINE, SPACE, (f) GUNS, MIDEAST, MISCELLANEOUS, (g) ATHEISM, CHRISTIANITY, MISCELLANEOUS, (h) FORSALE

	macro-av	fixed	random	gumbel
$k = 5$				
err.	-0.1656	**-0.0350**	-0.0428	-0.0400
p-value	-	< 0.001	< 0.001	< 0.001
$k = 10$				
err.	-0.1402	**-0.0329**	-0.0413	-0.0359
p-value	-	< 0.001	< 0.001	< 0.001
$k = 15$				
err.	-0.0809	-0.0799	-0.0804	**-0.0704**
p-value	-	< 0.001	< 0.001	< 0.001

Figure 2: Using macro-average and meta-analysis to predict error reductions on document classification datasets based on k observations. The scores are averages across 20 experiments. The p-values were computed using Wilcoxon signed rank tests.

The random effects model is obviously more conservative in its confidence intervals, and often we will not be able to obtain significance across datasets using a random effects model. If, for example, we apply a fixed effects model to test whether Bernoulli naive Bayes (NB) fairs better than a perceptron (P) model on 25 randomly extracted cross-domain document classification problem instances from the 20 Newsgroups dataset (see Sect. 4), the 95% confidence interval is $[3.9\%, 5.2\%]$. The weighted mean is 4.6% (macro-average 3.9%). Using a random effects model on the same 25 datasets, the 95% confidence interval becomes $[-6.5\%, 6.6\%]$. The weighted mean estimate is also slighly different from that of a fixed effects model. The first question we ask is which of these models provides the best estimate of effect size as observed on future datasets?

2.3 The error reductions distribution

Both the fixed effects and the random effects model assume that effect sizes are normally distributed. We can apply Darling-Anderson tests to test whether error reductions in 20 Newsgroups are in fact normally distributed. Even a small sample of ten 20 Newsgroups datasets provides enough evidence to reject the hypothesis that error reductions (of NB over P) are normally distributed. The Darling-Anderson tests consistently tell us that the chance that our sample distribtutions of error reductions are normally distributed is below 1%. The over-paramaterization means that the estimates we get are unstable. While both models are superior to macro-average estimates, they may provide ’far-off’ estimates.

Using Darling-Anderson tests we could also reject the hypothesis that error reductions were logistically distributed, but we did not find evidence for rejecting the hypothesis that error reductions are Gumbel-distributed.[1] Gumbel distributions are used to model error distributions in the latent variable formulation of multinomial logit regression. A parametric meta-analysis model based on the assumption that error reductions are Gumbel distributed is an interesting alternative to non-parametric meta-analysis (Hedges and Olkin, 1984; van den Noortgate and Onghena, 2005), since there seems to be little consensus in the literature about the best way to approach non-parametric meta-analysis.

Gumbel distributions take the following form:

$$\frac{1}{\beta}e^{z-e^{-z}}$$

where $z = \frac{x-\alpha}{\beta}$ with α the location, and β the scale. We fit a Gumbel distribution to our weighted error reductions (w_iT_i) and compute the combined

[1] Abidin et al. (2012) has shown that Darling-Anderson is superior to other goodness-of-fit tests for Gumbel distributions.

	macro-av	fixed	random	gumbel
$k = 5$				
err.	0.0531	0.0525	0.0526	**0.0489**
p-value	-	~ 0.98	~ 0.98	~ 0.79
$k = 7$				
err.	0.0928	**0.0852**	**0.0852**	0.0858
p-value	-	< 0.001	< 0.001	< 0.001
$k = 9$				
err.	0.0587	0.05743	0.05743	**0.0532**
p-value	-	~ 0.68	~ 0.68	~ 0.13

Figure 3: Using macro-average and meta-analysis to predict error reductions in cross-lingual dependency parsing. See text for details.

effect

$$\hat{T} = \frac{\alpha + 0.57721\beta}{\frac{1}{M}\sum_{i\geq 1}^{M} w_i}$$

where 0.57721 is the Euler-Mascheroni constant, and the variance of the combined effect $v = \frac{\pi^2}{6}\beta^2$.

3 Experiments in document classification and dependency parsing

Our first experiment makes use of the 20 Newsgroups document classification dataset.[2] The topics in 20 Newsgroups are hierarchically structured, which enables us to extract a large set of binary classification problems with considerable bias between source and target data (Chen et al., 2009; Sun et al., 2011). See the hierarchy in Figure 1. We extract 20 high-level binary classification problems by considering all pairs of top-level categories, e.g. COMPUTERS-RECREATIVE (comp-rec). For each of these 20 problems, we have different possible datasets, e.g. IBM-BASEBALL, MAC-MOTORCYCLES, etc. A *problem instance* takes training and test data from two *different* datasets belong to the same high-level problem. For example a problem instance could be learning to distinguish articles about Macintosh and motorcycles MAC-MOTORCYCLES (evaluated on the 20 Newsgroups test section) using labeled data from IBM-BASEBALL (the training section). In total we have 288 available problem instances in the 20 Newsgroups dataset.

In our first experiment we are interested in predicting the error reductions of a naive Bayes learner over a perceptron model. We use publicly available implementations with default parameters.[3] In each experiment we randomly select k datasets and estimate the true effect size using macro-average, a fixed effects model, a random effects model, and a corrected random effects model. In order to estimate the within-study variance we take 50 paired bootstrap samples of the system outputs. We evaluate our estimates against the observed average effect across 5 new randomly extracted datasets. For each k we repeat the experiment 20 times and report average error. We vary k to see how many observations are needed for our estimates to be reliable.

The results are presented in Figure 2. We note that meta-analysis provides much better estimates than macro-averages across the board. Our parametric meta-analysis based on the assumption that error reductions are Gumbel distributed performs best with more observations.

Our second experiment repeats the same procedure using available data from cross-lingual dependency parsing. We use the submitted results by participants in the CoNLL-X shared task (Buchholz and Marsi, 2006) and try to predict the error reduction of one system over another given k many observations. Given that we only have 12 submissions per system we use $k \in \{5, 7, 9\}$ randomly extracted datasets for observations and test on another five randomly extracted datasets. While results (Figure 3) are only statistically significant with $k = 7$, we see that meta-analysis estimates effect size across data sets better than macro-average in all cases.

4 Conclusions

We have argued that evaluation across datasets is important for developing robust NLP tools, and that meta-analysis can provide better estimates of effect size across datasets than macro-average. We also noted that parametric meta-analysis over-parameterizes error reduction distributions and suggested a new parametric method for estimating effect size across datasets.

Acknowledgements

Anders Søgaard is funded by the ERC Starting Grant LOWLANDS No. 313695.

[2] http://people.csail.mit.edu/jrennie/20Newsgroups/

[3] http://scikit-learn.org/stable/

References

Nahdiya Abidin, Mohd Adam, and Habshah Midi. 2012. The goodness-of-fit test for Gumbel distribution: a comparative study. *MATEMATIKA*, 28(1):35–48.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *EMNLP*.

Gregory Brown. 2011. An error analysis of relation extraction in social media documents. In *ACL*.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *CoNLL*.

Bo Chen, Wai Lam, Ivor Tsang, and Tak-Lam Wong. 2009. Extracting discriminative concepts for domain adaptation in text mining. In *KDD*.

Janez Demsar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Bradley Efron and Robert Tibshirani. 1993. *An introduction to the bootstrap*. Chapman & Hall, Boca Raton, FL.

Larry Hedges and Ingram Olkin. 1984. Nonparametric estimators of effect size in meta-analysis. *Psychological Bulletin*, 96:573–580.

Qian Sun, Rita Chattopadhyay, Sethuraman Panchanathan, and Jieping Ye. 2011. Two-stage weighting framework for multi-source domain adaptation. In *NIPS*.

Andreas van Cranenburgh. 2012. Literary authorship attribution with phrase-structure fragments. In *Workshop on Computational Linguistics for Literature, NAACL*.

Wim van den Noortgate and Patrick Onghena. 2005. Parametric and nonparametric bootstrap methods for meta-analysis. *Behavior Research Methods*, 37:11–22.