CALIBRATION FOR NATURAL LANGUAGE PROCESSING SYSTEMS

An Honor Thesis

Presented by

KHANH X. NGUYEN

Approved by:

Brendan O'Connor, School of Computer Science

Erik Learned-Miller, School of Computer Science

ABSTRACT

Title: **Calibration for natural language processing systems**

Author: **Khanh X. Nguyen**

Thesis/Project Type: **Project**

Approved By: **Brendan O'Connor, School of Computer Science**

Approved By: **Erik Learned-Miller, School of Computer Science**

Statistical natural language processing (NLP) models assign a posterior distribution to the set of possible outcomes and make their predictions based on those scores. Current performance metrics for NLP systems only take into account the model final decisions, which not only depends on the quality of the systems model but also the inference scheme, rather than explicitly reflect the quality of the prediction posterior distribution. In this thesis, I propose a metric for directly examining the quality of the posterior distribution, the calibration test. First of all, I will present the theoretical foundation for the concept of calibration. After that, I apply the test to numerous families of NLP models and show that the calibration is complementary to the traditional metrics in the sense that it provides a more comprehensive insights into the performances of NLP systems.

CHAPTER 1

INTRODUCTION

Researchers have long discovered that understanding natural language is much more difficult than simply putting the meanings of individual words together. Word meanings are affected by their surrounding contexts. The ambiguity of natural language motivated the use of the probabilistic model for tackling NLP tasks. A typical probabilistic model maps a linguistic structure to a posterior distribution over the space of all possible labels for the structure. The probability of an outcome can be interpreted as the degree of belief that the model holds for that possibility after learning the set of training instances. The model then make its decision based on an decision-making scheme.

Many NLP systems use MAP, i.e. choosing the most likely outcome, as their inference scheme. In this setting, the posterior distribution output by the model is neglected. Performance metrics such as accuracy score or F1-score measure how to the final decisions of the model align with the true labels. In complex systems that are structured as a cascade of multiple models, however, MAP inference seems to be inappropriate since the downstream models suffer from errors accumulated from upstream models. BLAH BLAH demonstrated that using the K-top prediction list yield better performance for BLAH BLAH. In a more advanced approach, Finkel et al. used MCMC to produce an approximate representation of the posterior distribution of the upstream models and passed it as input for the downstream models. Although the metrics of the final decisions of the pipeline reflects the quality of the entire system, there is a need for metrics to assess the posterior distribution of the intermediate models in order to obtain more insights into how they affect the overall performance.

Outside of the scope of NLP pipeline, knowledge about the posterior distribution of a model is also useful. Communicating the uncertainty of a model with its users is essential for calculating risk. This issue has been well-recognized in weather forecasting. Consider the task of predicting whether it will rain during a particular day or not. Suppose a weather forecasting model predicts that it will rain today with a belief score of 0.6. The delivered prediction from the model has the form There is X% chance that it will rain today. Using MAP inference with a threshold of 0.5 and reporting that It will rain today is a crude round-off and thus is more likely to lead to incorrect subsequent decisions in other tasks that rely on this piece of information.

Following BLAH BLAH, I propose a calibration-refinement framework for assessing the quality of posterior distributions of NLP models. Refinement measure is often encountered in the form of log loss or mean squared error. On the other, calibration measure receives less attention although it is complementary to refinement measure. The focus of this thesis is to develop a general procedure for measure calibration that is independent of the choice of model. The procedure takes a posterior distribution and the true data labels as input. The output can either be a visualizable calibration plot or a single calibration score depending on the need of the user.

In order to develop such procedure, first of all, a review the theoretical foundation of calibration is present. Several desirable characteristics of a well-calibrated model are shown. Next, a procedure for conducting calibration test on a model is described in details. Finally, the procedure will be applied to several families of NLP models. I show that BLAH BLAH BLAH.

CHAPTER 2

RELATED WORK

The problem of miscalibration in prediction models that employ single-best inference scheme was addressed by Draper (1995). A Bayesian approach was proposed as an alternative. Finkel et. al (2009) apply this idea to tackle the problem of cascade NLP models. In this approach, the prediction posterior distribution of one task, which is approximated by a sample of the distribution, was passed as the input for other tasks. As suggested by the author, it is more general but easier to implement than approaches using K-best list (Sutton and McCallum (2005), Wellner et al. (2004), Huang and Chiang (2005), Toutanova et al. (2005)).

The concept of calibration was developed in the field of meteorology (Miller (1962), Murphy (1973)), referred to as validity or reliability. In Rubin (1982), it was argued that the applied statistician should Bayesian in principle and calibrated to the real world in practice. Murphy and Winkler (1982) proposed a general framework for forecast verification based on the joint distribution of forecasts and observations. They showed that the joint distribution of predictions and observations contains all of the information needed for assessing the forecast quality. They investigated it through two its Bayesian factorizations: the calibration-refinement factorization and the likelihood-base rate factorization. Their study is general in the sense that it can applied to any other type of prediction that produces a joint distribution between prediction labels and true labels. Rubin (2006) presented a method for validating software for Bayesian models using posterior quantiles. This idea will be applied in

my thesis to formulate the notion of calibration for prediction problems where the predicted variable is continuous.

TODO: work on metrics for machine learning.

CHAPTER 3

METHOD

Consider a general probabilistic prediction problem, where for each instance i there is a true value $y_i \in \Omega$, where Ω is the label space. A stastical machine learning model learns from the training set D_{train} and computes for each instance i of the testing set D_{test} a posterior distribution $q_i(k) \in P$, where $q_i(k)$ represents the degree of belief (or confidence score) on the event "$y_i = k$" and P is the probability function space. Alternatively, we can think of the posterior distribution q_i as a probabilistic label that the model assigns to instance i. All instances that are assigned the same posterior distribution belongs to the same class. In this perspective, a probabilistic problem is no difference than a classification problem.

3.1 Calibration-sharpness framework

Let y and q be particular values that y_i and $q(y_i|\alpha)$ can take. The joint distribution $P(y, q)$ offers all the information needed for analysing the quality of the predictions.

The calibration-sharpness framework for assessing probabilistic model is derived from the following factorization of the joint distributrion $P(y, q)$:

$$P(y, q) = P(y \mid q)P(q)$$

Given a fixed q, the conditional distribution $P(y \mid q)$ indicates how often different values of y have occurred among all instances that the model classifies as having the posterior distribution q. In other words, calibration verifies if the belief held by the

model matches with the realistic frequency. For example, we expect that among all the days that a weather forecaster predicts a 30% chance of rain, 30% will have rain.

The marginal distribution $P(q)$ indicates how often different values of posterior predictions are used. Going back to the weather forecast example, a forecaster who always predicts the climatological, i.e. the known long-term percentage of rain, will be perfectly calibrated but would be useless in discriminating between days with rain and days without rain. In this case, $P(q)$ is non-zero at only one point and we say that the weather forecasts are not *refined* (or sharp). Ideally, we would want a forecast that implies no uncertainty, either "100% chance of rain" or "0% chance of rain".

Note that calibration and sharpness are orthogonal and complementary to each other. Perfect prediction implies perfect calibration and perfect sharpness. In the opposite direction, maintaing calibration allows posterior predictions to be more realistic whereas having sharpness in predictions reduces uncertainty in the decision-making process.

3.2 Metrics for calibration and sharpness

For brevity, denote $p_q(k) = P(y = k \mid q)$, $\forall k \in \Omega$.

Calibration and sharpness can be quantified by two metrics as follows:

$$CalibrationScore = \frac{1}{|\Omega|} \sum_{k \in \Omega} E_q[p_q(k)(g_1(k) - g_1[p_q(k)]) + (1 - p_q(k))(g_2(k) - g_2[p_q(k)])]$$

$$SharpnessScore = \frac{1}{|\Omega|} \sum_{k \in \Omega} E_q[p_q(k)g_1(k) + (1 - p_q(k))g_2(k)]$$

The functions g_1 and g_2 are defined as follows:

$$g_1(x) = \int_0^x \alpha(t)dt$$

and

$$g_2(x) = \int_x^1 \frac{1}{1-t}\alpha(t)dt$$

where $\alpha(t)$ is a positive contiuous function on $[0, 1]$.

3.3 Calibration for discrete probabilistic task

In this setting, Ω is a discrete space. Calibration is defined as follows:

Definition 3.1. A predictive model is said to be *perfectly calibrated* if and only if:

$$P(y = k \mid q) = q(k) \qquad \forall q \in P,\ k \in \Omega.$$

The left hand side of the equation is the realistic frequency of an event and the right hand side is the posterior predtions for that event. In practice, $P(y = k \mid q)$ is unknown so we have to estimate it using an emperical ratio $\hat{P}(y = k \mid q)$ computed from a sample of data.

Moreover, if the task is to predict a binary variable then we can visualize the degree of calibration of a model from a *reliabilitycurve* or *empiricalcalibrationcurve* as shown in Figure BLAH. The procedure of constructing such curve will be discussed in BLAH.

Unfortunately, for the multilcass case, we would a curve for each type of label. Presenting calibration in such a way is not helpful for drawing conclusions. Alternatively, we can summarize miscalibration in a single scalar, the expected square error between the empirical frequency and the posterior predictions:

$$TheorCalibMSE = \frac{1}{|\Omega|}\sum_{k\in\Omega} E_q[q(k) - P(y = k \mid q)]^2$$

CHAPTER 4

A INTRODUCTION TO SHEEP

Is there life afters:m:w sheep? [1] Yes, I say there is.

4.1 Pulling the wool over your eyes

Sheep are fabulou creatures. The noises they make are truly stupendous [2]. We also want to refer to figure 4.1 here. Here' some verbatim text to screw us up:

```
xxx := y;
xy := x;
```

4.1.1 All about sheep noises

Lots of text here just to fill up some space so we can be sure that we really are double-spacing and doing all the other things that might be necessary in formatting a dissertation to U.Mass. guidelines. We're also going to have another figure here, figure 4.2, just for fun, and to make sure that the list of figures is formatted correctly. Now it's time for table 4.1. We really are going to need a third figure, figure 4.3, two more tables, table 4.2 and table 4.3 and a fourth figure, figure 4.4, just to really make sure.

Table 4.1. Some numbers.

Type of Animal	Minimum Observed	Average Observed	Maximum Observed
Cats	12	20	24
Dogs	20	20	20

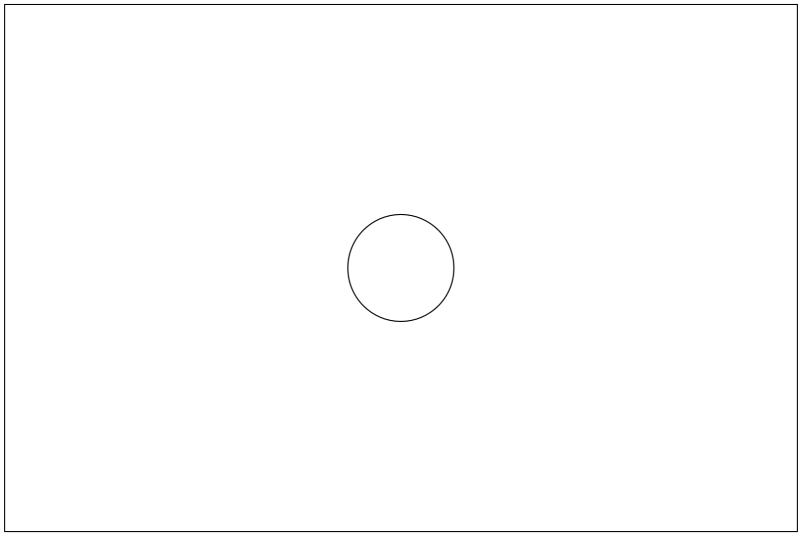

Figure 4.1. A circle in a square.

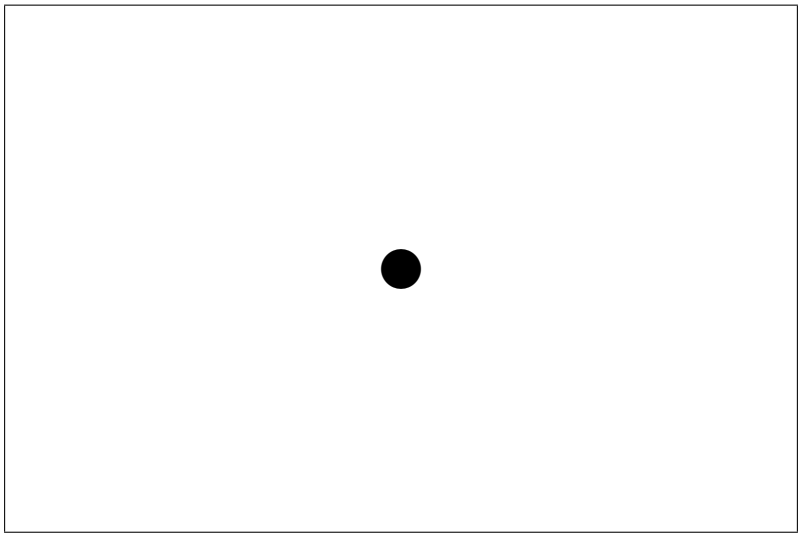

Figure 4.2. A disc in a square.

Table 4.2. More numbers.

Type of Animal	Arms	Legs	Ears
Person	2	2	2
Dog	0	4	2

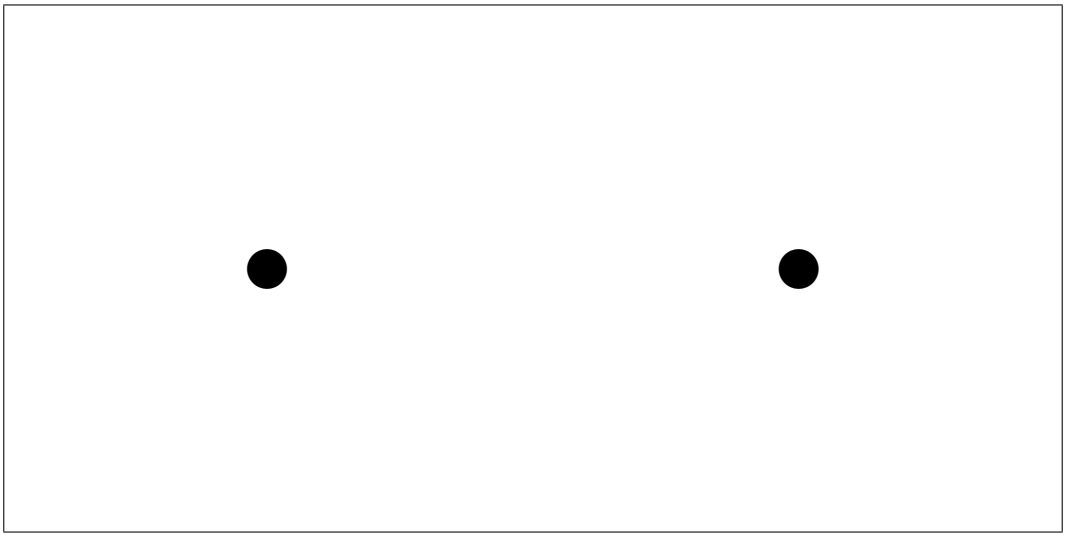

Figure 4.3. Two discs in a rectangle.

Table 4.3. Even more numbers; together with a caption long enough to ensure that multi-line caption formatting works correctly. If you want a shorter caption to appear in the Table of Figures you're going to have to put the shorter caption in the `[]` as shown in this example.

x	1	1	1
y	2	2	2
z	3	3	3

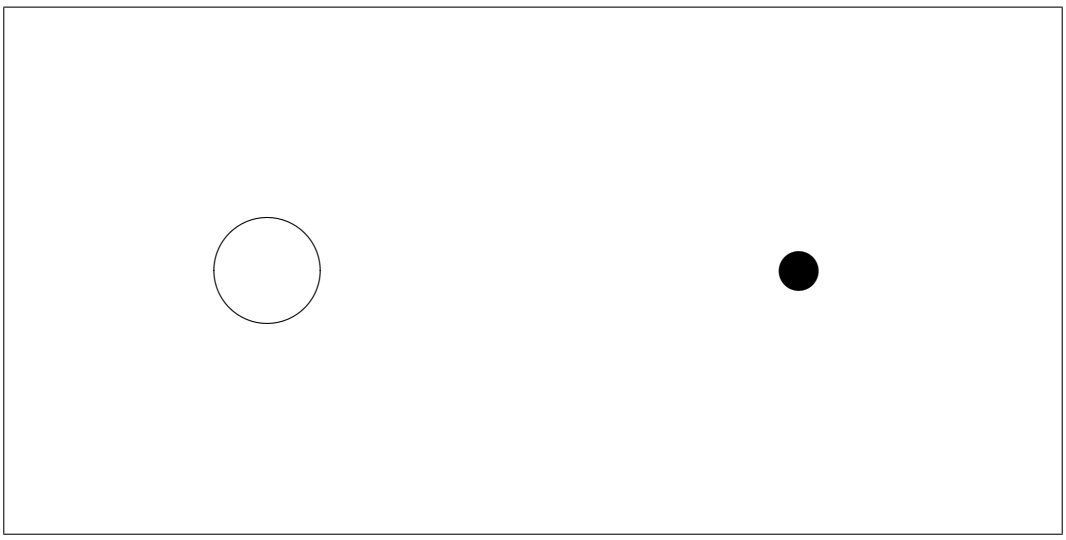

Figure 4.4. A circle and a disc in a square. We want this caption to be very long to ensure that the formatting of very long captions is handled correctly. The case of short captions has already been dealt with.

4.1.1.1 Baahs

4.1.2 Even more about sheep noises

4.1.3 And yet more about sheep noises

4.2 What about wolves?

What about wolves?[1]

4.3 What about shepherds?

What about shepherds? I don't really know, but I want some text here to fill things in so that I can verify that everything is OK.[2]

---

[1] To be fair, some wolves are probably nice...

[2] Some shepherds are good, some are bad. The reader is referred to Mary and The Boy Who Cried Wolf for further insight into this much-debated issue. (This needs to be a very long footnote so we can test the spacing between lines on a footnote.)

4.3.1 A subsection

This is a subsection of the subsection about shepherds.

4.3.2 Another subsection

This is another subsection of that section.

4.3.2.1 A subsubsection

This is a subsubsection of that subsection that will in turn havae a paragraph with a pair of subparagraphs. I am aware that I shouldn't have only one subsubsection in the subsection...

4.3.2.1.1 A Paragraph This is the text associated with this paragraph. I really want enough text to make it look like a paragraph. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah.

4.3.2.1.1.1 A Subparagraph This is the text associated with this subparagraph. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah.

4.3.2.1.1.2 Another Subparagraph Better not have subparagraphs without text in them. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah.

4.3.2.1.2 Another Paragraph Baah, baah, baah.

Baah, baah, baah. Baah, baah, baah.

4.3.2.2 Another Subsubsection

With some text. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah. Baah, baah, baah.

CHAPTER 5

SHEEP AND GRASS

5.1 Introduction

Grass is a wonderful food... Baah, baah, baah.

CHAPTER 6

A WONDERFULLY LONG CHAPTER TITLE THAT IS THIS LONG IN ORDER TO TEST THE CHAPTER HEADING STUFF

Note that we shouldn't really have a chapter heading with no body, so here is a body for this chapter. Baah, baah, baah.

6.1 The antidisestablishmentarainism supercalifragilisticexpialidocious longlonglonglonglongword

A `quotation`:

> Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut nibh orci, molestie non vehicula ac, ultricies quis purus. Nunc euismod metus vel nulla sodales quis tempus nisi varius. Sed ornare pulvinar bibendum. Ut egestas mollis nisi vel cursus.

. . . and a `quote`:

> Ut dolor libero, blandit tristique accumsan non, viverra a magna. Sed pretium sollicitudin neque, sit amet ornare lorem convallis ac. Fusce mollis gravida aliquam. Nullam vulputate turpis vitae orci porttitor auctor. Donec in auctor erat.

APPENDIX A

THE FIRST APPENDIX TITLE

...

APPENDIX B

THE SECOND APPENDIX TITLE

...

BIBLIOGRAPHY

[1] Barrett, Daniel J., Ridgway, John V. E., and Wileden, Jack C. Why there are no sheep in our work. In *Proceedings of the Third Sheep Conference* (Edinburgh, Scotland, Jan. 1997), Ian McPherson Sheepish, Ed., American Sheepherders Society, Sheepdip and Associates, pp. 39–45.

[2] Scrooge, Ebenezer, and Shepherd, Alan. On the growth of green in space. *Journal of Astrophysical Economics 3*, 4 (August 1992), 47–89.