Posterior calibration and exploratory analysis for natural language processing models

Khanh Nguyen (Univ. of Maryland, College Park)
Brendan O'Connor (Univ. of Massachusetts, Amherst)
EMNLP 2015

What is calibration? When a model knows it's wrong.

Everyone knows it's impossible for NLP systems to resolve all ambiguity. That's why we have probabilistic models.
Ambiguities should be passed down the pipeline. Why do we only evaluate the most-probable structure?
Models output probabilities, and good probabilities ought to match frequencies on test data.
We propose to evaluate **calibration**, as an alternative to single-structure accuracy.

Definitions

q: a posterior probability (prediction)
$q_i \equiv P(y_i = 1 \mid x_i, \theta)$

p: an empirical frequency of the label
$p_q \equiv \text{Frac. of } y_i = 1 \text{ among all } i \text{ where } q_i = q$

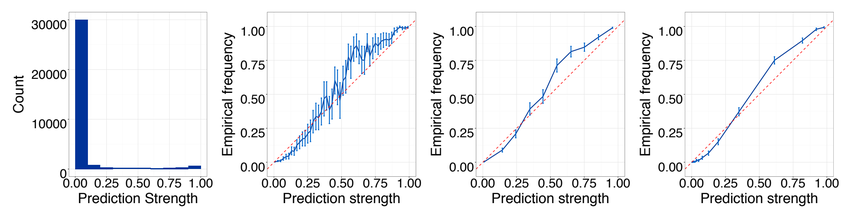


Figure 1: (a) A skewed distribution of predictions on whether a word has the NN tag (§4.2.2). Calibration curves produced by equally-spaced binning with bin width equal to 0.02 (b) and 0.1 (c) can have wide confidence intervals. Adaptive binning (with 1000 points in each bin) (d) gives small confidence intervals and also captures the prediction distribution. The confidence intervals are estimated as described in §3.1.

We contribute a new(?) adaptive binning method, since *q* distributions are very skewed in NLP

Calibration error

Calibration-refinement (bias/variance) breakdown of squared error

$$\underbrace{\frac{1}{N}\sum_i (y_i - q_i)^2}_{\text{Brier score (L2)}} = \underbrace{\mathbb{E}_q[q - p_q]^2}_{\text{Calibration MSE}} + \underbrace{\mathbb{E}_q[p_q(1-p_q)]}_{\text{Refinement}}$$

$$CalibErr = \sqrt{\mathbb{E}_q[q - \mathbb{P}(y = 1 \mid q)]^2}$$

Calibration does not imply perfect accuracy
Calibration could exist at any level of accuracy
Perfect accuracy implies calibration

Sentiment classification

Naive Bayes, Logistic Regression

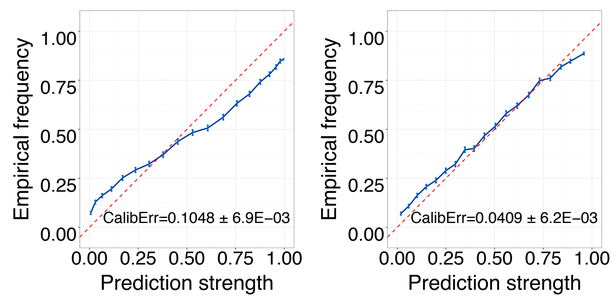


Figure 2: Calibration curve of (a) Naive Bayes and (b) logistic regression on predicting whether a tweet is a "happy" tweet.

Tagging HMM, CRF

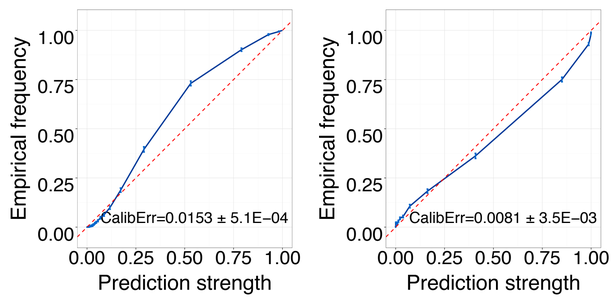


Figure 3: Calibration curves of (a) HMM, and (b) CRF, on predictions over all POS tags.

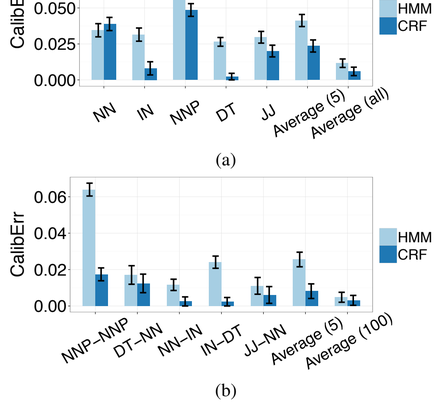


Figure 4: Calibration errors of HMM and CRF on predicting (a) single-word tags and (b) two-consecutive-word tags.

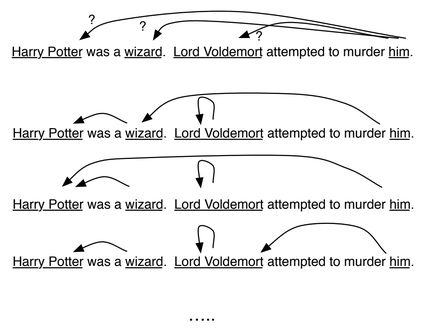
Coreference ambiguity

Slight modification of Berkeley Coreference model/system yields an exact sampling algorithm

Definition 2 (Antecedent coreference model and sampling algorithm)**.**

- For $i = 1..N$, sample $a_i \sim \frac{1}{Z_i}\exp(\mathbf{w}^\mathsf{T}\mathbf{f}(i, a_i, x))$
- Calculate the entity clusters as $\mathbf{e} := CC(\mathbf{a})$, the connected components of the antecedent graph having edges (i, a_i) for i where $a_i \neq$ NEW.

Rapidly sample from posterior over all possible clusterings



Calibration is surprisingly? good

$$P(\ell_{ij} = 1 \mid x) = \sum_{\mathbf{e}} 1\{(i,j) \in \mathbf{e}\} P(\mathbf{e} \mid x)$$

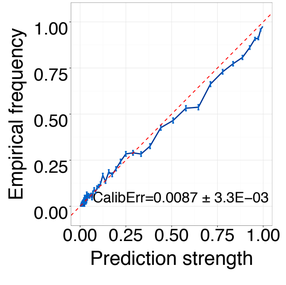


Figure 5: Coreference calibration curve for predicting whether two mentions belong to the same entity cluster.

International relations event extraction

We want ***confidence intervals for event extraction***.
Test case for coreference-dependent event extraction: international relations revents

Russian *troops were sighted … and* ***they attacked***

Entity affiliated with a country name is the agent of an "attack".
Coreference propagates dependencies between noun phrase mentions.
Re-run extractor on every coreference sample => integrates out coreference uncertainty.

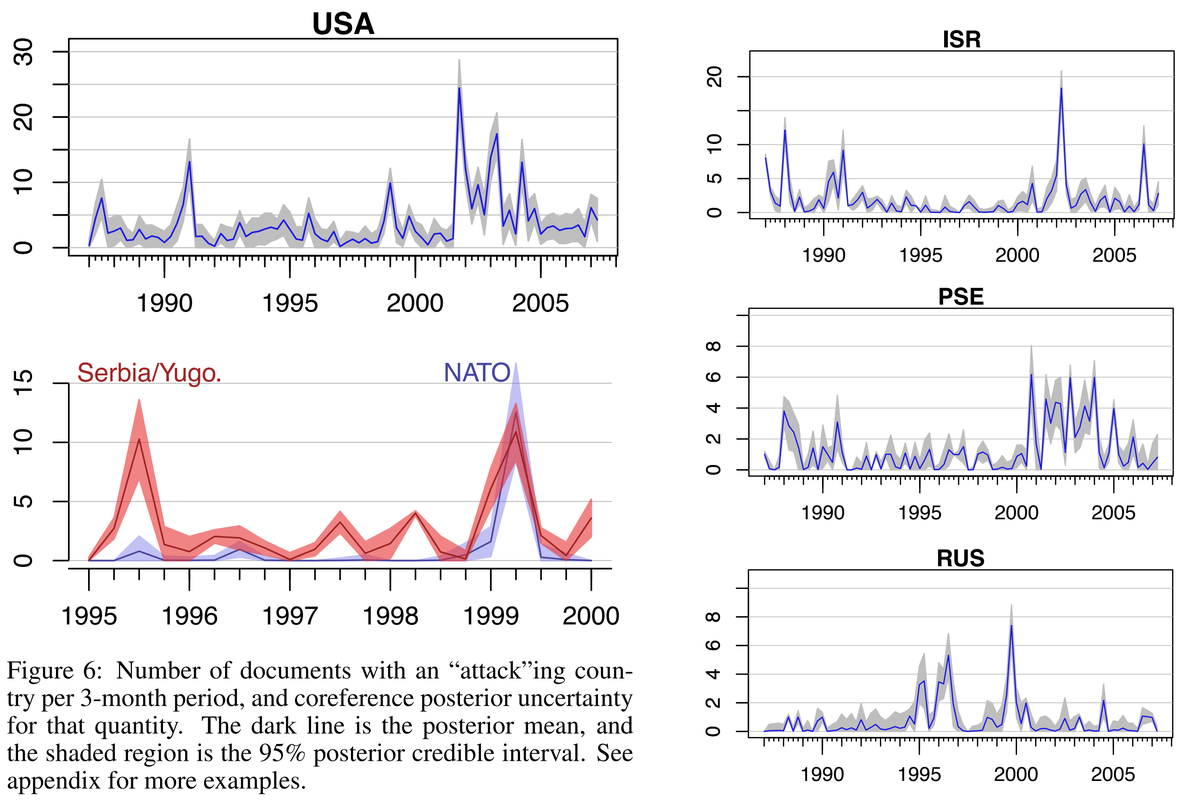


Figure 6: Number of documents with an "attack"ing country per 3-month period, and coreference posterior uncertainty for that quantity. The dark line is the posterior mean, and the shaded region is the 95% posterior credible interval. See appendix for more examples.