Region-dependent temperature scaling for certainty calibration and application to class-imbalanced token classification

Hillary Dawkins
University of Guelph, Canada
Vector Institute, Toronto, Canada
hdawkins@uoguelph.ca

Isar Nejadgholi
National Research Council Canada
Ottawa, Canada
isar.nejadgholi@nrc-cnrc.gc.ca

Abstract

Certainty calibration is an important goal on the path to interpretability and trustworthy AI. Particularly in the context of human-in-the-loop systems, high-quality low to mid-range certainty estimates are essential. In the presence of a dominant high-certainty class, for instance the non-entity class in NER problems, existing calibration error measures are completely insensitive to potentially large errors in this certainty region of interest. We introduce a region-balanced calibration error metric that weights all certainty regions equally. When low and mid certainty estimates are taken into account, calibration error is typically larger than previously reported. We introduce a simple extension of temperature scaling, requiring no additional computation, that can reduce both traditional and region-balanced notions of calibration error over existing baselines.

1 Introduction

Calibrating the certainty estimates of neural networks is of the utmost importance for interpretability of results and building trust in AI systems. Ideally, if a model outputs some prediction with an associated probability, we would like to interpret that quantity as the probability of a correct prediction (i.e. as a meaningful certainty estimate) (Zadrozny and Elkan, 2001; Niculescu-Mizil and Caruana, 2005). However, contemporary models are consistently over-confident in their output probabilities (Guo et al., 2017).

Guo et al. (2017) demonstrates that over-confident models can arise by overfitting to the Negative Log-Likelihood (NLL) loss, without overfitting to the classification accuracy. Many calibration methods involve modulating the output logits somehow, according to a prescribed functional form. The parameters of the modulation function are learned on the associated *validation* set by minimizing the NLL loss (thereby correcting the overfit). Guo et al. (2017), as well as many subsequent studies (e.g. Müller et al., 2019; Gupta et al., 2021), showcase the surprising effectiveness of temperature scaling, a single-parameter modulation function.

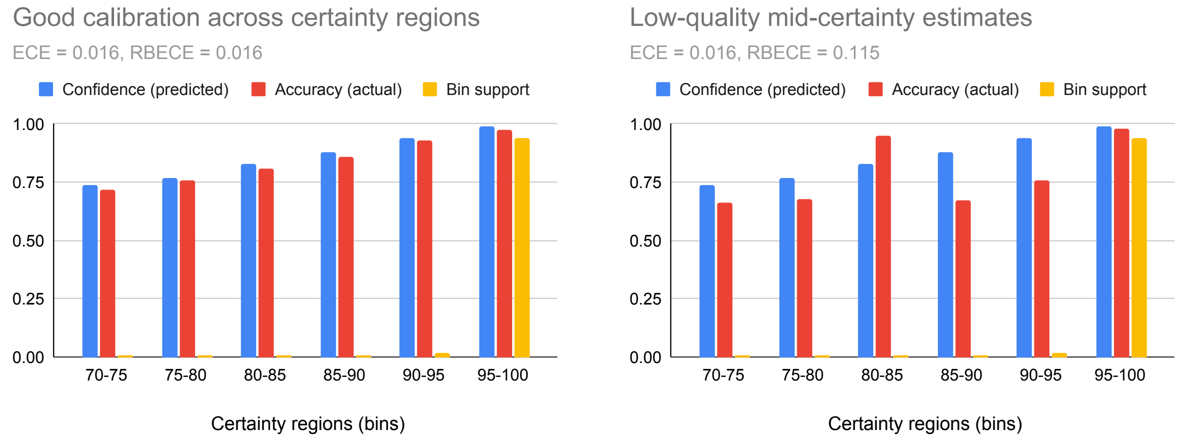
The calibration error is reported as a single quantity computed on the associated test set. Typically, the error is composed of a sum of observed errors across the certainty landscape, visualized using a reliability diagram (DeGroot and Fienberg, 1983; Niculescu-Mizil and Caruana, 2005). However, not all regions contribute equally, especially in the case of class-imbalanced datasets. Consider an output with a predicted certainty of 99.9% vs. an expected actual certainty of 99.8%. In terms of human interpretability and intervention, this difference is negligible. Now consider 79% predicted certainty vs. 71% expected certainty. Clearly the second case is one we should care more about correcting. However, as we will discuss in the following section, the presence of a dominant high-certainty class can cause the first discrepancy to contribute more to the reported calibration error than the second. High quality mid-certainty estimates are most impactful for human-in-the-loop applications, yet current error measures are not sensitive to this region.

Here we take NER (Grishman and Sundheim, 1996; Yadav and Bethard, 2018; Li et al., 2020) as a case study for class-imbalanced token classification. Naturally, the "outside" or non-entity class dominates the dataset. In the following section, we introduce a region-balanced calibration error. We then introduce region-dependent temperature scaling, a calibration method that further reduces error over traditional temperature scaling, across various NER scenarios, without additional computation.

2 Region-balanced expected calibration error

The most popular calibration error metric is the expected calibration error (ECE) (Naeini et al., 2015). A test set is partitioned into certainty bins, each

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 538 - 544
May 22-27, 2022 ©2022 Association for Computational Linguistics



(a) Sample reliability diagram for the case of consistently good certainty estimates across all regions.

(b) Sample reliability diagram for the case of low-quality certainty estimates in the mid-certainty region.

Figure 1: Reliability diagrams contrasting two cases with equal ECE values. Both cases have the same support distribution (yellow), where 90% of all samples have an estimated certainty above 0.95. In each bin, the confidence (blue) is defined as the mean certainty of samples in the bin (i.e. the predicted certainty). The accuracy (red) is the proportion of samples with a correct prediction (i.e. the actual certainty). The calibration error per bin is the difference in predicted and actual certainty. In case (a), calibration error is consistently low across all certainty regions. In case (b), calibration error is high across the mid-certainty regions. However, because of the dominant support in the highest certainty bin, this error is undetected by the ECE measure.

containing samples with a certainty score h within the bin boundaries. The uncalibrated certainty h for a given sample is simply the output probability associated with the predicted class for that sample. Within each bin, we compare the actual and predicted certainty:

$$ECE = \sum_i \frac{n_i}{N} |\mathrm{acc}(B_i) - \mathrm{conf}(B_i)| \qquad (1)$$

where $\mathrm{conf}(B_i)$ is the predicted confidence score (the mean h of samples in bin B_i), and $\mathrm{acc}(B_i)$ is the actual accuracy (proportion of correct predictions in bin B_i). Each bin error is weighted by the bin support, where n_i is the number of samples in B_i. If a very high proportion of all samples have a high certainty estimate, only the final bin error has a non-negligible contribution to the overall ECE. Refer to Figure 1 for an illustrated example.

One extension of ECE is to find bin partitions adaptively (Nixon et al., 2020), such that each bin contains an equal number of samples, and each bin contributes equally to the overall error. The result is that many more bins exist in the high certainty region, each of which are narrower in width. Essentially, adaptive-ECE reports the exact same error quantity as ECE in theory, but estimates the quantity using a finer-toothed comb. Neither metric is informative on lower or mid-certainty regions if support is dominated by a high-certainty class.

Maximum expected calibration error (MECE) (Naeini et al., 2015) partially tells the story of low-certainty regions by reporting the maximum bin error. However, MECE is overly sensitive to outlier bins. For example, if a single sample happens to fall in the 0-5% certainty bin, and it has the correct predicted class, we have MECE $> .95$, which is clearly an unusable characterization of the calibration error as a whole.

Here we consider Region-balanced ECE (RBECE) as a way to characterize calibration error weighted evenly across certainty regions. Simply,

$$RBECE = \frac{1}{|\Theta|} \sum_{B_i \in \Theta} |\mathrm{acc}(B_i) - \mathrm{conf}(B_i)|. \qquad (2)$$

The error in each bin B_i contributes to the error equally, subject to some threshold support requirement $n_i > \theta$ (to ensure $\mathrm{acc}(B_i)$ is well-defined). The set of bins that meet this requirement is denoted by Θ.

Alternative threshold requirements such as variance in $\mathrm{conf}(B_i)$ vs. bin size could be explored in the future. Another possible extension is custom bin-weighting according to a certainty region of interest for your application (e.g. for human-in-the-loop systems with an intervention criterion).

3 Region-dependent temperature scaling

The idea underlying all calibration methods is generally to modulate overconfident predictions. In traditional temperature scaling (TS), a higher temperature means stronger modulation. Temperature is taken to be a constant, meaning all samples are treated with the same modulation strength.

The idea underlying region-dependent temperature scaling (RD-TS) is simply that the most confident predictions likely need greater modulation than less confident predictions, and therefore temperature should depend on the uncalibrated certainty. If we consider the hypothetical limit of a 0% confidence score, it is intuitive that this does not need any modulation. To investigate this idea empirically, we apply TS to subsets of the OntoNotes dataset, partitioned according to uncalibrated confidence scores. For each confidence region, the ideal temperature is shown in Figure 2. As expected, temperature increases as a function of confidence. A linear fit sufficiently describes the dependence. Within uncertainty, the intercept is equal to the expected value of 1 ($T(h = 0) = 1$, corresponding to no modulation).

To apply RD-TS, uncalibrated logits $\vec{a}$ are scaled as $\vec{q} = \vec{a}/T(h)$ to obtain calibrated logits $\vec{q}$. Temperature is now a function of confidence $T(h) = mh + 1$, where $h = \max(\text{softmax}(\vec{a}))$ is the probability estimate for the predicted class on each sample. The slope m is the single parameter controlling modulation strength.

To estimate m, one could repeat temperature scaling on multiple data subsets, collect data points, and fit the slope as in Figure 2. However, this method increases computational overhead. Instead, let us estimate m from the original TS constant T_0 and some knowledge of the validation dataset which was used to compute T_0. Each sample in the validation set has an ideal temperature, here taken to be in the form $T_i = mh_i + 1$. Assuming each sample contributed to the found T_0 equally, $T_0 = \frac{1}{N}\sum_i^N(mh_i + 1)$. Given access to the validation set, this sum can be computed exactly to find m. However, we can further approximate the sum by loosely assuming that the data has a high proportion of samples (say $\approx 90\%$) with very high certainty estimates (say $\approx .99$ on average). Then the sum is dominated by the first leading term, $T_0 \approx .9(.99m + 1)$. This quick sketch is sufficient to achieve good error reduction over the baseline TS method. The numerical exactness is not too important, but rather the general signature of a high proportion of high-certainty samples is sufficient. We take this further approximation to gain the advantage that nothing specifically needs to be known about the calibration dataset. I.e. If a large pre-trained model has been calibrated on a large or private dataset, and the corresponding temperature T_0 is known, RD-TS can be applied to your model outputs without access to the calibration data or further computation.

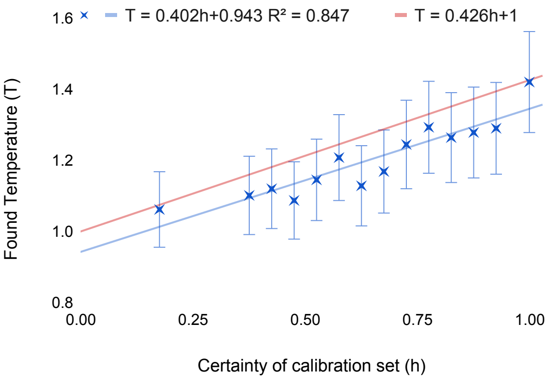


Figure 2: The OntoNotes 5.0 validation set is split into 14 bins according to uncalibrated confidence scores h. For each subset, regular temperature scaling is applied to find the ideal T_0 as a function of average confidence. Blue: Linear regression fit of empirical data ($m = .402 \pm .108$, $b = .943 \pm .073$ with a 95% confidence interval). Red: Region-dependent temperature scaling parameter $T(h)$ as determined by our protocol (see points 1-3). Both methods produce equivalent results within the uncertainty.

In summary, the RD-TS method is performed as follows:

1. Perform regular temperature scaling to obtain T_0, or obtain a previously published T_0 for your model.

2. Find the linear dependence parameter $m = (T_0 - .9)/.89$.

3. Apply calibration to logits $\vec{a}$ as $\vec{q} = \vec{a}/T(h)$, $T = mh + 1$.

RD-TS is a simple extension of temperature scaling which requires no additional training. Like temperature scaling, RD-TS cannot change the predicted class or model accuracy (unlike some other generalizations, vector and matrix scaling).

Scenario	**Uncal.**	**TS**	**VS**	**MS**	**WTS**	**RD-TS**
Classic	.09328	.02543 ($T_0 = 1.28$)	.07040	.06940	.05236	**.02151** ($m = .426$)
Rare & emerging	.09878	.05777 ($T_0 = 1.39$)	.07490	.04932	.11559	**.03549** ($m = .550$)
Fine-grained	.05333	.02179 ($T_0 = 1.12$)	.03440	.04628	.03278	**.01263** ($m = .243$)
Specialized	.07088	.04147 ($T_0 = 1.29$)	.03844	.03590	.03820	**.02781** ($m = .439$)
Sparse training	.09683	.07820 ($T_0 = 1.10$)	.11653	.09528	.06279	**.04110** ($m = .229$)
Differing sources	.05730	.05960 ($T_0 = 1.09$)	.10824	.08470	.05551	**.04019** ($m = .214$)

Table 1: Region-balanced expected calibration error (RBECE); refer to eq. 2.

Scenario	**Uncal.**	**TS**	**VS**	**MS**	**WTS**	**RD-TS**
Classic	.02001	.00862 ($T_0 = 1.28$)	.01359	.01083	.00962	**.00155** ($m = .426$)
Rare & emerging	.04278	.02323 ($T_0 = 1.39$)	.02585	.01580	.04712	**.00949** ($m = .550$)
Fine-grained	.02287	**.00783** ($T_0 = 1.12$)	.01587	.01786	.01462	.00839 ($m = .243$)
Specialized	.01555	.00617 ($T_0 = 1.29$)	.00608	**.00573**	.00631	.00651 ($m = .439$)
Sparse training	.03267	.02190 ($T_0 = 1.10$)	.03113	.02599	**.01645**	.01798 ($m = .229$)
Differing sources	.00950	.00723 ($T_0 = 1.09$)	.01211	.01344	.01020	**.00383** ($m = .214$)

Table 2: Expected calibration error (ECE); refer to eq. 1.

Dataset	$h \vert (P = .9)$	$P \vert (h = .99)$
OntoNotes	.998	.964
W-NUT 17	.997	.953
Few-nerd	.972	.801
BC2GM	.997	.968
OntoNotes (tc)	.999	.978

Table 3: The mean certainty h of the top .9 most certain samples, $h|(P = .9)$, and the proportion of samples we need to take such that the mean certainty is .99, $P|(h = .99)$. All datasets refer to the corresponding validation set, which is used for calibration. As shown, all datasets have the general signature of a high proportion of high-certainty samples, yet the exact numerical values can deviate from our sketch.

4 Experimental results

4.1 Baseline methods

As RD-TS is a simple extension of regular temperature scaling, we focus comparison on similar post-training parametric calibration methods:

Temperature scaling (TS): Uncalibrated logits $\vec{a}$ are scaled by a single constant T_0 (as $\vec{q} = \vec{a}/T_0$) before softmax is applied to obtain calibrated probability estimates over all classes (Guo et al., 2017).

Vector (generalized Platt) scaling (VS): A generalization of TS such that logits are scaled by $2k$ learned parameters, $\vec{q} = \vec{v} \circ \vec{a} + \vec{b}$, where k is the number of classes (Platt, 1999; Niculescu-Mizil and Caruana, 2005; Guo et al., 2017).

Matrix scaling (MS): A further generalized linear transformation such that logits are scaled by $k^2 + k$ learned parameters, $\vec{q} = M\vec{a} + \vec{b}$ (Guo et al., 2017).

Weighted temperature scaling (WTS): TS using a class-weighted NLL loss during convergence (Obadinma et al., 2021).

4.2 Datasets

We take the NER task as a case study. Datasets represent several important scenarios in token classification settings more broadly:

Classic: The OntoNotes 5.0 NER dataset (Weischedel et al., 2013) represents a baseline “classic” scenario involving plentiful training and calibration data from robust sources.

Rare and emerging named entities: The W-NUT NER dataset[1] (Derczynski et al., 2017) is gathered from noisy social media data which contains difficult entities (e.g. “kktny") due to informal and evolving language.

Fine-grained and few-shot: Few-nerd[2] (Ding et al., 2021) is a challenging few-shot NER dataset with 66 fine-grained entity types (e.g. “art-film").

Specialized language: The BioCreative II Gene Mention Recognition (BC2GM) dataset[3] (Smith et al., 2008) is composed of scientific text where named entities are gene mentions.

[1] huggingface.co/datasets/wnut_17

[2] huggingface.co/datasets/dfki-nlp/few-nerd

[3] huggingface.co/datasets/bc2gm_corpus

Sparse training data: OntoNotes telephone call data is used for training while the full OntoNotes dataset is used for calibration and evaluation. The telephone call data subset is a sparse representation since it is very heavily skewed to the non-entity outside class, and entity mentions are concentrated on “person" and “location", compared to the full OntoNotes dataset (generally containing much richer entity mentions from news sources).

Differing language sources: OntoNotes broadcast news data is used for training, and telephone call data is used for calibration and evaluation. Broadcast news language is professional and grammatically correct. Telephone call language is casual, fragmented and incoherent at times.

4.3 Implementation notes

All NER models use DistilBERT[4] (Sanh et al., 2019) as the base pre-trained model, fine-tuned for NER using the train dataset for each scenario as described above. Further details and performance on the NER task are provided in Appendix A.

Calibration is performed using the uncalibrated logits of the associated validation set as model inputs. Calibration parameters are learned by minimizing the NLL (or weighted NLL) loss for 50 epochs (using SGD with 0.01 learning rate, and 0.9 momentum). Calibration error is computed on the associated test set. To compute both ECE (eq. 1) and RBECE (eq. 2), the number of bins is set to 20. To compute RBECE, the threshold for support per bin is set to $\theta = 40$. The code needed to reproduce these results is made publicly available[5]. All datasets are publicly available with preset train/validation/test data splits.

4.4 Results

Experimental results are summarized in Tables 1 and 2. When low and mid-certainty regions are taken into account by the RBECE, calibration error is larger than previously thought (as reported by ECE). In all scenarios, RD-TS produces the smallest RBECE (in many cases quite substantially). Additionally, RD-TS improves the traditional ECE in the majority of scenarios. The results show that RD-TS is an effective extension of TS across a range of temperature (T_0) values.

Recall in Section 3, we sketch a way to estimate the modulation parameter m, and this approximation follows from assuming that a high proportion of all samples in the calibration set (say $\approx .9$) have a high certainty estimate (say $\approx .99$ on average). We claim that the numerical exactness of these values is not too important (and therefore RD-TS outperforms TS across a range of datasets). This claim is supported empirically (Table 3).

5 Discussion and Conclusion

Good quality mid-range certainty estimates are essential for productive human-model interactions. Despite this, existing calibration error measures can be insensitive to all but the highest certainty regions. We propose a region-balanced error metric to probe this unreported information. When low and mid-certainty regions are taken into account, greater calibration errors are revealed.

Further, we explore the idea of a certainty-dependent temperature. While previous generalizations of TS, such as vector and matrix scaling, allow certainty dependence by increasing the number of learned parameters, these methods are generally outperformed by TS (Guo et al., 2017). Rather than allowing a complicated certainty dependence, we enforce a simple linear dependence (motivated by intuition and an empirical example) without introducing any learnable parameters. Unlike vector and matrix scaling, RD-TS cannot change the relative ranking of logits, and therefore model accuracy is retained (in single-label settings). One line of future work could be to apply RD-TS on top of weighted temperature scaling, a method known to decrease variance in calibration error among classes (Obadinma et al., 2021). Another line of work would be to investigate whether improved certainty estimates can increase model accuracy (in multi-label settings where predictions are applied by meeting a certainty threshold), especially in out-of-domain problems.

Finally, it is important to note that our discussion of a region-balanced error measure, as well as our sketch derivation of the RD-TS method, have been generally applicable to **any problem with a dominant proportion of high-certainty predictions**. This situation does arise in any token classification problem with a dominant “easy” class, as is the case in NER, however this situation can equally occur in class-balanced situations. Therefore, region-dependent temperature scaling can find utility beyond NER, token classification, or class-imbalanced situations.

[4] huggingface.co/transformers/model_doc/distilbert.html
[5] github.com/hillary-dawkins/RegDepTempScaling

Ethical Considerations

We proposed a novel method to calibrate class-imbalanced token classifiers, and demonstrated the method for NER models. This calibration method is a step toward responsible use of AI by offering a measure of reliability, but also has risks that should be considered from an ethical point of view. Calibrated scores are a measure of transparency, and users can interpret a well-calibrated model better. However, all transparency methods expose AI systems to malicious attacks by providing more information about the internal workings of the system. This risk should be taken into account in sensitive tasks, e.g. when an NER model is used to extract personally identifiable information for privacy reasons. Also, users should be warned that a low calibration error does not guarantee robustness in out-of-domain settings. Therefore, in the case of safety-critical tasks such as medical applications of NER, a low calibration error should be interpreted with caution.

Further, low calibration errors should not be used to justify inherently unethical tasks or those out of the scope of the capabilities of NLP technologies. Every task should be evaluated in terms of feasibility and ethical use regardless of reliability and transparency of trained models. It is also important to keep in mind that a well-calibrated model can become miscalibrated as the data changes, and continuous calibration is needed to deal with the ever-changing nature of language.

References

Morris H. DeGroot and Stephen E. Fienberg. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2):12–22.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-NERD: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3198–3213, Online. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org.

Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. 2021. Calibration of neural networks using splines. In *International Conference on Learning Representations*.

J. Li, A. Sun, J. Han, and C. Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge & Data Engineering*, (01):1–1.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2901–2907. AAAI Press.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 625–632, New York, NY, USA. Association for Computing Machinery.

Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. 2020. Measuring calibration in deep learning.

Stephen Obadinma, Hongyu Guo, and Xiaodan Zhu. 2021. Class-wise calibration: A case study on covid-19 hate speech. *Proceedings of the Canadian Conference on Artificial Intelligence*. Https://caiac.pubpub.org/pub/vd3v9vby.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Larry Smith, Lorraine K. Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I.-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M. Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A. Struble, Richard J. Povinelli, Andreas Vlachos, William A. Baumgartner, Lawrence Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel Maña-López, Jacinto Mata, and W. John Wilbur. 2008. Overview of BioCreative II gene mention recognition. *Genome Biol.*, 9(2):1–19.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes Release 5.0.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 609–616, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

A NER performance

NER models were obtained by fine-tuning DistilBERT, using the default configuration, for 3 epochs (with learning rate of 2e-5, and weight decay of 0.01). The performance of all NER models is provided in Table A.1 for reference.

Dataset	**P**	**R**	**F**	**A**
OntoNotes	.778	.621	.691	.976
W-NUT 17	.543	.234	.327	.938
Few-nerd	.639	.679	.659	.906
BC2GM	.802	.844	.822	.965
OntoNotes (bc)	.711	.753	.732	.973

Table A.1: For all datasets that were used to train an NER model, we report the precision (P), recall (R), F-score (F) and accuracy (A) of the model on the corresponding test set.