Paper Sharing

Zhu Liu

2022.12.02

Outline

Nibbling at the Hard Core of Word Sense Disambiguation

Marco Maru[1], **Simone Conia**[1], **Michele Bevilacqua**[1], and **Roberto Navigli**[2]

Sapienza NLP Group
[1]Department of Computer Science
[2]Department of Computer, Control and Management Engineering
Sapienza University of Rome
firstname.lastname@uniroma1.it

Calibration of Pre-trained Transformers

Shrey Desai and **Greg Durrett**
Department of Computer Science
The University of Texas at Austin
shreydesai@utexas.edu gdurrett@cs.utexas.edu

Nibbling at the Hard Core of Word Sense Disambiguation

Marco Maru[1], **Simone Conia**[1], **Michele Bevilacqua**[1], and **Roberto Navigli**[2]

Sapienza NLP Group
[1]Department of Computer Science
[2]Department of Computer, Control and Management Engineering
Sapienza University of Rome
firstname.lastname@uniroma1.it

ACL 2022; Cited by 1

https://aclanthology.org/2022.acl-long.324

Introduction

- SOTA models have achieved the F1 score of 80%. (estimated upper bound)
- These models benefit from: 1) transformer architecture; 2) transfer learning

Model	M_F1 for ALL
ESC (NAACL'21)	80.7
MLWSD (EACL'21)	80.2
EWISER (ACL'20)	80.1

Introduction

- SOTA models have achieved the F1 score of 80%. (estimated upper bound)
- These models benefit from: 1) transformer architecture; 2) transfer learning

context: The banks battling against a strong wind in the USA several years later. Investors and regulators (. . .)

gold: A tendency or force that influences events.

ESCHER: Air moving (. . .) from an area of high pressure to an area of low pressure.

context: I was just sitting down to meet with some new therapy clients, a couple, and the building started shaking (. . .)

gold: A pair of people who live together.

Conia and Navigli (2021): A small indefinite number.

Some trivial errors by SOTA models

Introduction

- SOTA models have achieved the F1 score of 80%. (estimated upper bound)
- These models benefit from: 1) transformer architecture; 2) transfer learning
- Is a high F1 score enough?
- It is still necessary to extract erroneous cases, analyze the reasons and provide new test beds.

Contributions

- A detailed quantitative and qualitative analysis of common misidentified instances
- To propose two new test sets:

1) an amended version of the English all-words WSD evaluation benchmarks.
2) A distribution-shifted (both domain and sense) dataset: 42D (pron. [for·ti·tude]).

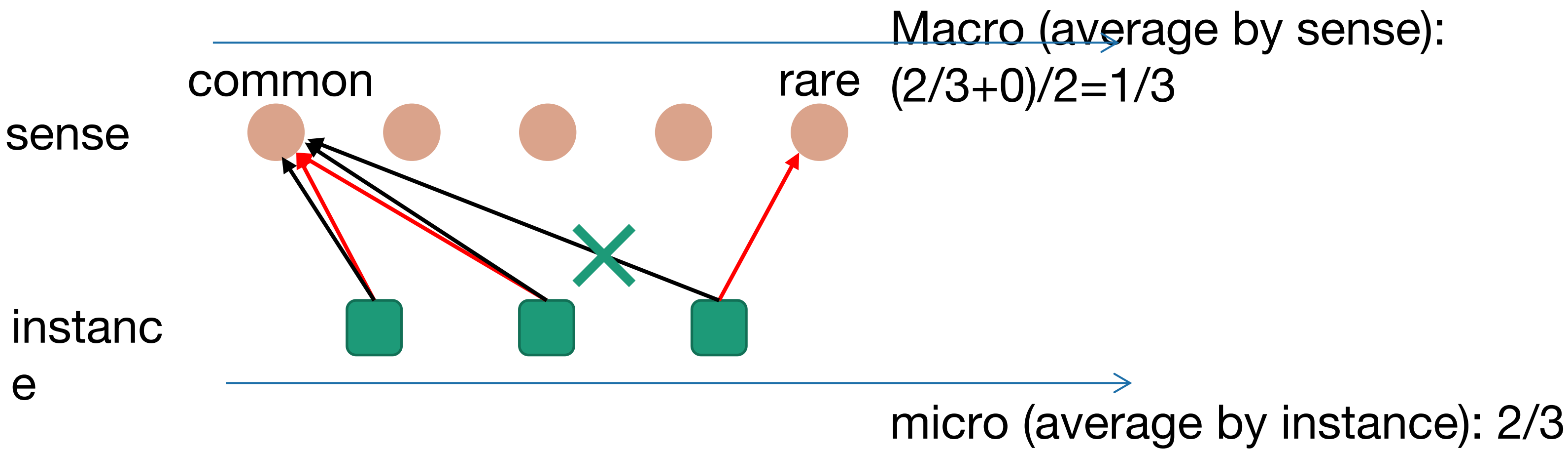
- To use Macro-averaged F1 score instead of traditional micro-averaged F1 score to measure the accuracy

Systems at Issue

Model	Author	Note
ARES (EMNLP'20)	Bianca Scarlini, Tommaso Pasini and Roberto Navigli	Semi-supervised approach of contextualized sense embedding; 1nn
BEM (ACL'20)	Terra Blevins and Luke Zettlemoyer	Bi-encoder for context and gloss
ESCHER (ESR, NAACL'21)	Edoardo Barba, Tommaso Pasini, Roberto Navigli	A span extraction task
EWISER (EWR, ACL'20)	Michele Bevilacqua and Roberto Navigli	Exploiting relational information in Wordnet
Generationary (GLN, EMNLP'20)	Bevilacqua Michele; Maru Marco; Navigli Roberto	Generative; seq2seq
GlossBERT (GLB, EMNLP'19)	Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang	Gloss knowledge
SyntagRank (SYN	Federico Scozzafava, Marco Maru, Fabrizio Brignone, Giovanni Torrisi, and Roberto Navigli	Knowledge-based

“Hard Core”

- Test set (ALL): a unified evaluation benchmark [Raganato et al. (2017a)]
- HardCore set (ALL_HC): instances that are wrongly disambiguated by all the SOTA models (thus 0.0% F1 score)
- Macro-averaged F1 v. micro-averaged F1 (traditional)



Accuracy & MFS bias

dataset	#inst (#mono)	ARES		BEM		ESR		EWR		GEN		GLB		SYN	
		M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
ALL	7,253 (1,301)	72.9	77.9	73.9	79.0	**76.4**	80.7	73.3	78.3	70.7	76.3	71.3	76.9	64.1	71.7
ALL_{HC}	541 (0)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

accurac
y

dataset	#inst	#mono	ARES	BEM	ESR	EWR	GEN	GLB	SYN	gold
ALL	7,253	1,301	71.3%	72.6%	71.2%	72.7%	**69.0%**	74.8%	81.1%	65.2%
ALL_{HC}	541	0	64.7%	71.0%	68.6%	67.8%	**62.7%**	70.6%	80.2%	2.0%

Table 1: Times (%) systems predict the MFS in WordNet, i.e., WN1st (top), or a sense occurring at least once in SemCor (bottom). Left to right: dataset, number of instances (#inst), number of monosemous instances (#mono), system percentages (ARES, BEM, ESR, EWR, GEN, GLB, SYN), gold standard percentages (gold). **Bold** is closer to gold.

MFS
bias

Training dataset bias

dataset	#inst (#mono)	ARES		BEM		ESR		EWR		GEN		GLB		SYN	
		M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
ALL	7,253 (1,301)	72.9	77.9	73.9	79.0	**76.4**	80.7	73.3	78.3	70.7	76.3	71.3	76.9	64.1	71.7
ALL_{no1st}	2,525 (0)	45.7	50.1	47.8	50.5	**54.2**	55.2	46.8	49.0	45.3	48.4	42.4	45.0	26.9	29.5
ALL_{noSC}	1,138 (448)	60.3	65.3	63.7	67.1	**71.0**	75.0	58.6	64.0	65.5	68.6	57.4	62.2	55.1	61.0
ALL_{HC}	541 (0)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

accurac
y

dataset	#inst	#mono	ARES	BEM	ESR	EWR	GEN	GLB	SYN	gold
ALL	7,253	1,301	88.2%	87.4%	86.3%	88.8%	**85.9%**	88.6%	88.8%	84.3%
ALL_{HC}	541	0	96.9%	96.7%	96.5%	98.0%	**95.0%**	97.2%	98.3%	67.1%

Table 1: Times (%) systems predict the MFS in WordNet, i.e., WN1st (top), or a sense occurring at least once in SemCor (bottom). Left to right: dataset, number of instances (#inst), number of monosemous instances (#mono), system percentages (ARES, BEM, ESR, EWR, GEN, GLB, SYN), gold standard percentages (gold). **Bold** is closer to gold.

Training dataset bias

Qualitative Analysis

- Given the sizable wrong instances and possible biases, what may cause the misclassification?
- Aleatoric (from data itself): Is there something wrong in the annotations?
- Epistemic (OOD testing): Is it because the model fails to see enough data in various domains?
- Several new benchmarks

Dataset Amendment

- A human linguistic to tag each instance in the original test set, for
- *unchanged*
- fine-grained
- error: token-lemma
- error: pos
- <u>error: sense</u>
- <u>error: inventory</u>

Dataset Amendment - Example

tag (id)	**fine-grained** (semeval2010.d003.s043.t001)
ctx_tgt	See Map 1 for the boundaries of the realms
old	boundary%1:15:00:: the line or plane indicating the limit or extent of something
new	+ boundary%1:25:00:: a line determining the limits of an area
tag (id)	**error:pos** (senseval3.d001.s022.t007)
ctx_tgt	[...] have become virtually immune to defeat.
old	defeat%2:33:00:: win a victory over (VERB)
new	defeat%1:11:00:: an unsuccessful ending to a struggle or contest (NOUN)
tag (id)	**error:sense** (semeval2013.d003.s013.t002)
ctx_tgt	[...] which have cultivated close ties with the Iraqi Oil Ministry [...]
old	tie%1:11:00:: the finish of a contest in which the score is tied and the winner is undecided
new	tie%1:26:01:: a social or business relationship
tag (id)	**error:inventory** (semeval2010.d003.s059.t001)
ctx_tgt	Mangroves provide nurseries for 85 per cent of commercial fish species [...]
old	nursery%1:06:00:: a building with glass walls and roof; for the cultivation and exhibition of plants [...]
new	(*no suitable word sense featured in WordNet for "nursery"*)
tag (id)	**error:token-lemma** (semeval2015.d002.s021.t005)
ctx_tgt	[...] Italy, the Netherlands and the *United Kingdom*.
old_1	kingdom%1:14:01:: a monarchy with a king or queen as head of state
old_2	kingdom%1:15:01:: a country with a king as head of state
new	united_kingdom%1:15:00:: a monarchy in northwestern Europe occupying most of the British isles [...]

Dataset Amendment - Statistics

dataset	#inst	unch.	fine	token	pos	sense	inv.
ALL-	5,523	72.6	9.4	2.9	**0.3**	8.0	6.8
ALL_{NS}-	5,023	**75.4**	8.3	2.9	0.0	7.0	6.1
ALL_{HC}-	500	44.6	**20.4**	3.0	0.0	**17.8**	14.2
S10-	1,251	62.4	7.6	**4.7**	0.0	8.2	**17.1**

Table 4: Times (%) a label type is assigned to test set instances during the qualitative evaluation. **Bold** is highest.

ALL-: All except monosemous words and SemEval-2007 instances
ALL (HC-): hard core subset
ALL (NS-): ALL- minus HC-
S10-: SemEval2010 Task 17 data except monosemous words

New benchmark 1 – All_NEW and S10_NEW

- The same linguistic re-labeling sense correct the instance with the tag: fine-grained and errors.
- ALL_NEW: 4917 polysemous words (4917/5523?)
- S10_NEW : 955 polysemous words. (955/1251?)

- hardEN: all misclassified instances by SOTAs; 476
- softEN: ALL after removing hardEN; 5766

New benchmark 2 – 42D

- built from scratch by manually annotating paragraphs taken from the British National Corpus
- 42 domains defined in BabelNet 4.0.
- OOD bias-aware.

  For each of the instances, the GT:

1) does not occur in SemCor
2) is not the first sense in Wordnet.

- 370 words

New benchmarks – Evaluation

dataset	#inst	ARES		BEM		ESR		EWR		GEN		GLB		SYN	
		M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
ALL$*$	4,917	69.3	75.5	69.9	76.2	**73.1**	78.3	70.0	76.0	66.1	73.1	67.7	74.4	57.9	66.9
ALL$_{NEW}$	4,917	75.2	79.0	75.6	79.5	**78.7**	81.6	75.6	79.2	72.2	76.7	73.2	77.4	61.4	68.5
S10$_{NEW}$	955	77.9	81.4	77.1	82.2	**78.0**	82.1	76.1	81.1	72.3	77.0	75.8	80.4	64.0	66.7
42D	370	41.8	37.8	53.2	47.8	**58.9**	54.1	43.9	40.8	50.2	48.9	45.7	41.9	32.8	28.1
softEN	5,766	78.7	83.3	80.3	84.5	**83.7**	86.8	79.2	85.0	76.4	82.3	77.1	82.0	63.4	71.3
hardEN	476	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 5: F1 scores for the reported systems on the datasets described in Section 5. Left to right: dataset/subdataset (dataset), number of instances (#inst), system performances (ARES, BEM, ESR, EWR, GEN, GLB, SYN) measured using both macro (M-F1) and micro F1 (m-F1). **Bold** is M-F1 best. $*$ indicates the subset of ALL (Raganato et al., 2017b) that includes only those instances that are also featured in ALL$_{NEW}$.

Where to go?

- Joint forces (model ensembling)

-> Ensembling strategies:

1) Uniform E.: Majority Voting;
2) Ranked E.: with weights ranked according to its performance rank on ALL_NEW

dataset	ESCHER		Uniform E.		Ranked E.	
	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
ALL_{NEW}	78.7	81.6	77.8	81.6	**78.8**	82.3
$S10_{NEW}$	78.0	82.1	79.5	83.7	**80.7**	84.9
42D	**58.9**	54.1	50.9	46.8	53.2	48.9
softEN	**83.7**	86.8	82.7	87.6	83.4	88.3
hardEN	0.0	0.0	0.0	0.0	0.0	0.0

Table 6: Macro- (M-F1) and micro-averaged F1 (m-F1) scores of our Uniform and Ranked ensemble strategies compared against the best performing systems, ESCHER. Best macro-averaged F1 scores are in **bold**.

Where to go?

- Data augmentation

-> to generate training examples automatically (Exemplification modeling) by [Barba et al., 2021b]

- K1: trained only with one automatically generated example per sense (K1)

dataset	SemCor		K1		SemCor+K1	
	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
ALL_{NEW}	**78.7**	81.6	61.0	60.8	75.9	80.0
$S10_{NEW}$	**78.0**	82.1	68.5	67.4	76.2	80.1
42D	58.9	54.1	63.0	60.3	**65.2**	60.5
softEN	**83.7**	86.8	65.1	64.3	80.4	84.6
hardEN	0.0	0.0	**35.3**	33.6	16.8	14.5

Table 7: Macro- (M-F1) and micro-averaged F1 (m-F1) scores of ESCHER: trained only on SemCor, only on K1 (automatically-generated dataset containing one example per sense), and on SemCor + K1. Improving on hardEN decreases scores on softEN. Best macro-averaged F1 scores are in **bold**.

Conclusion

- What: errors made by a heterogeneous set of seven SOTA systems
- Why: (1) distribution shift: two biases (MFS, Training)

(2) “measurement error”/ non-system-dependent issues: label noise

- How: several new benchmarks (42D, hardEN, ALL_new…)
- What interests me:

How to use these benchmarks to estimate/measure uncertainty (aleatoric uncertainty v. epistemic uncertainty)

Calibration of Pre-trained Transformers

Shrey Desai and **Greg Durrett**
Department of Computer Science
The University of Texas at Austin
shreydesai@utexas.edu gdurrett@cs.utexas.edu

EMNLP 2020; Cited by 96

https://aclanthology.org/2020.emnlp-main.21.pdf

Introduction

- Neural networks have seen wide adoption but are frequently criticized for being black boxes: why and what's wrong?
- One step towards interpretation: to what extent is the model can be trusted? (whether they are calibrated)
- Specifically, do these models' posterior probabilities provide an accurate empirical measure of how likely the model is to be correct on a given example?

Calibration

Calibration: a frequentist perspective

- Suppose our confidence is the predicted probability of correctness.
- We say our model is **calibrated** if:

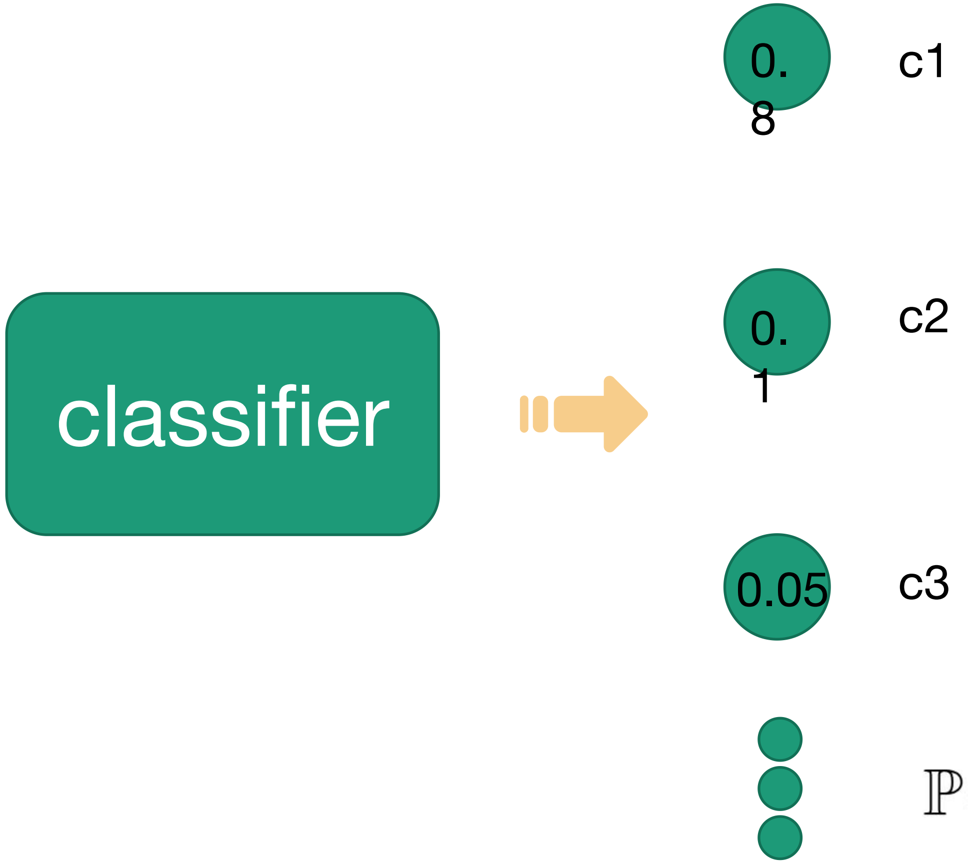
$$\mathbb{P}(\text{model is correct} \mid \text{confidence is } \alpha) = \alpha$$

- In other words, ***α*-fraction** of all predictions with confidence α should be **correct**.

37

https://sites.google.com/view/uncertainty-nlp

Calibration



We could say: P(c1 is the correct label) = 0.8.

But, what do we mean by saying the probability/confidence of “0.8”?

1) Stochastic models: Given the same input, 80 times out of 100 the model will choose c1 as correct.
2) <u>Deterministic models: Given the same level of confidence for multiple samples, 80 samples out of 100 is</u>

$$\mathbb{P}(\text{model is correct} \mid \text{confidence is } \alpha) = \alpha$$

Introduction

- Neural networks have seen wide adoption but are frequently criticized for being black boxes: why and what's wrong?
- One step towards interpretation: to what extent is the model can be trusted? (whether they are calibrated)
- The paper evaluates the calibration of two pre-trained models, BERT and RoBERTa on three tasks: natural language inference, paraphrase detection and commonsense reasoning.
- Corresponding techniques to improve calibration in the settings of in domain and out of domain.

Background

- These background slides come from the COLING 2022 tutorial.
- https://sites.google.com/view/uncertainty-nlp

Experiments – Tasks and Datasets

- Natural language inference: to determine whether a hypothesis is entailed, contradicted by, or neutral with respect to a premise

  Corpus: The Stanford Natural Language Inference (SNLI)

  OOD: Multi-Genre Natural Language Inference (MNLI)
- Paraphrase detection: **whether** two sentences are semantically equal?

  Corpus: Quora Question Pairs (QQP)

  OOD: TwitterPPDB (TPPDB)
- Commonsense reasoning: models must select the most plausible continuation of a sentence among **four** candidates

  Corpus: Situations With Adversarial Generations (SWAG)

  OOD: HellaSWAG (HSWAG)

Experiments – Systems

Model	Parameters	Architecture	Pre-trained
DA	382K	LSTM	✗
ESIM	4M	Bi-LSTM	✗
BERT	110M	Transformer	✓
RoBERTa	110M	Transformer	✓

Table 1: Models in this work. Decomposable Attention (DA) (Parikh et al., 2016) and Enhanced Sequential Inference Model (ESIM) (Chen et al., 2017) use LSTMs and attention on top of GloVe embeddings (Pennington et al., 2014) to model pairwise semantic similarities. In contrast, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) are large-scale, pre-trained language models with stacked, general purpose Transformer (Vaswani et al., 2017) layers.

Results – OOB Calibration

OOB: out-of-box

1. Non-pre-trained models exhibit an inverse relationship between complexity and calibration.

2. However, pre-trained models are generally more accurate and calibrated.

3. Using RoBERTa always improves in-domain calibration over BERT.

Model	Accuracy		ECE	
	ID	OD	ID	OD
Task: SNLI/MNLI				
DA	84.63	57.12	**1.02**	8.79
ESIM	88.32	60.91	1.33	12.78
BERT	90.04	73.52	2.54	7.03
RoBERTa	**91.23**	**78.79**	1.93	**3.62**
Task: QQP/TwitterPPDB				
DA	85.85	83.36	3.37	9.79
ESIM	87.75	84.00	3.65	**8.38**
BERT	90.27	**87.63**	2.71	8.51
RoBERTa	**91.11**	86.72	**2.33**	9.55
Task: SWAG/HellaSWAG				
DA	46.80	32.48	5.98	40.37
ESIM	52.09	32.08	7.01	19.57
BERT	79.40	34.48	2.49	12.62
RoBERTa	**82.45**	**41.68**	**1.76**	**11.93**

Table 2: Out-of-the-box calibration results for in-domain (SNLI, QQP, SWAG) and out-of-domain (MNLI, TwitterPPDB, HellaSWAG) datasets using the models described in Table 1. We report accuracy and expected calibration error (ECE), both averaged across 5 fine-tuning runs with random restarts.

Post-hoc Calibration

Post-hoc methods:

1. temperature scaling

- Higher T: softens probabilities
- Lower T: sharpens probabilities

$$p(y_i \mid x) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

Label smoothing

Soft label: placing a 1 – α fraction of probability mass on the gold label and α/(|Y|–1) fraction of mass on each other label, where α ∈ (0, 1) is a hyperparameter.

e.g., [1, 0, 0] -> [0.9, 0.05, 0.05], when α = 0.1

- To train the model MLE or LS using the in-domain training set
- To learn an optimal T using the in-domain development set

Results

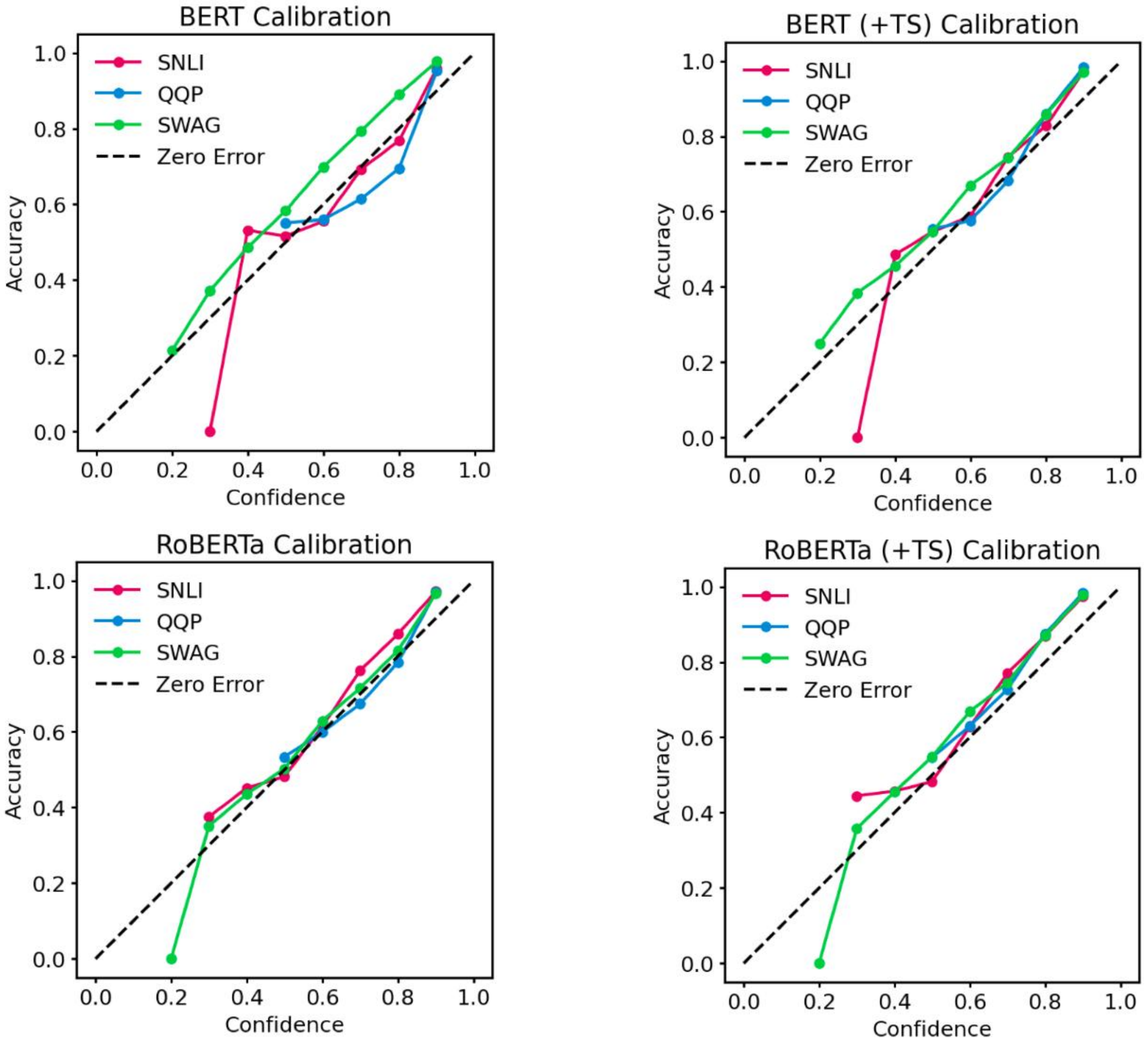
- MLE models with temperature scaling achieve low in-domain calibration error.

Method	In-Domain						Out-of-Domain					
	SNLI		QQP		SWAG		MNLI		TPPDB		HSWAG	
	MLE	LS	MLE	LS	MLE	LS	MLE	LS	MLE	LS	MLE	LS
Model: BERT												
Out-of-the-box	2.54	7.12	2.71	6.33	2.49	10.01	7.03	3.74	8.51	6.30	12.62	5.73
Temperature scaled	1.14	8.37	0.97	8.16	0.85	10.89	3.61	4.05	7.15	5.78	12.83	5.34
Model: RoBERTa												
Out-of-the-box	1.93	6.38	2.33	6.11	1.76	8.81	3.62	4.50	9.55	8.91	11.93	2.14
Temperature scaled	0.84	8.70	0.88	8.69	0.76	11.4	1.46	5.93	7.86	5.31	11.22	2.23

Table 3: Post-hoc calibration results for BERT and RoBERTa on in-domain (SNLI, QQP, SWAG) and out-of-domain (MNLI, TwitterPPDB, HellaSWAG) datasets. Models are trained with maximum likelihood estimation (MLE) or label smoothing (LS), then their logits are post-processed using temperature scaling (§4.4). We report expected calibration error (ECE) averaged across 5 runs with random restarts. Darker colors imply lower ECE.

Results

- MLE models with temperature scaling achieve low in-domain calibration error.

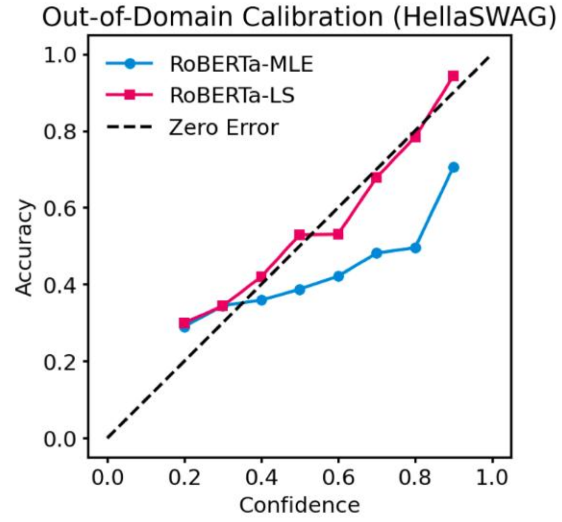


Results

- MLE models with temperature scaling achieve low in-domain calibration error.
- However, out-of-domain, label smoothing is generally more effective.

Method	In-Domain						Out-of-Domain					
	SNLI		QQP		SWAG		MNLI		TPPDB		HSWAG	
	MLE	LS	MLE	LS	MLE	LS	MLE	LS	MLE	LS	MLE	LS
Model: BERT												
Out-of-the-box	2.54	7.12	2.71	6.33	2.49	10.01	7.03	3.74	8.51	6.30	12.62	5.73
Temperature scaled	1.14	8.37	0.97	8.16	0.85	10.89	3.61	4.05	7.15	5.78	12.83	5.34
Model: RoBERTa												
Out-of-the-box	1.93	6.38	2.33	6.11	1.76	8.81	3.62	4.50	9.55	8.91	11.93	2.14
Temperature scaled	0.84	8.70	0.88	8.69	0.76	11.4	1.46	5.93	7.86	5.31	11.22	2.23

Table 3: Post-hoc calibration results for BERT and RoBERTa on in-domain (SNLI, QQP, SWAG) and out-of-domain (MNLI, TwitterPPDB, HellaSWAG) datasets. Models are trained with maximum likelihood estimation (MLE) or label smoothing (LS), then their logits are post-processed using temperature scaling (§4.4). We report expected calibration error (ECE) averaged across 5 runs with random restarts. Darker colors imply lower ECE.



Results

- Optimal temperature scaling values are bounded within a small interval.

-> It suggests the degree of distribution shift and magnitude of T may be closely re

Model	In-Domain			Out-of-Domain		
	SNLI	QQP	SWAG	MNLI	TPPDB	HSWAG
BERT	1.20	1.34	0.99	1.41	2.91	3.61
RoBERTa	1.16	1.39	1.10	1.25	2.79	2.77

Table 4: Learned temperature scaling values for BERT and RoBERTa on in-domain (SNLI, QQP, SWAG) and out-of-domain (MNLI, TwitterPPDB, HellaSWAG) datasets. Values are obtained by line search with a granularity of 0.01. Evaluations are very fast as they only require rescaling cached logits.

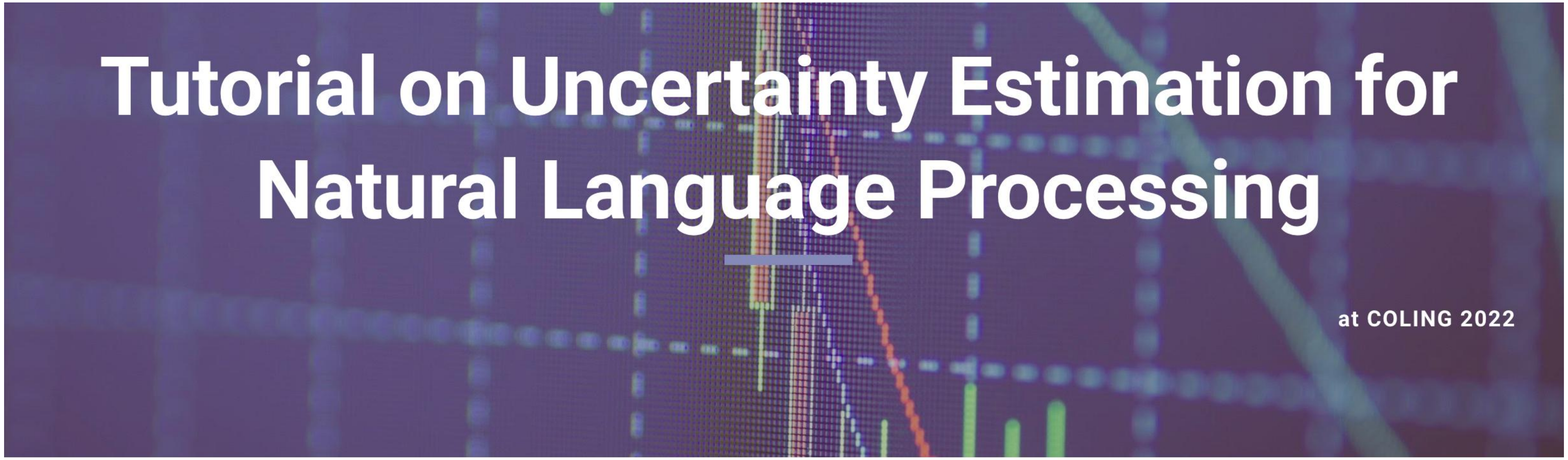
Conclusions

- The paper examines the calibration of pre-trained Transformers in both in-domain and out-of-domain settings
- Results show BERT and RoBERTa coupled with temperature scaling achieve low ECEs in-domain, and when trained with label smoothing, are also competitive out-of-domain.

Q & A

THANK YOU

Coling'22



Overview

Speakers

Outline

Slides

https://sites.google.com/view/uncertainty-nlp

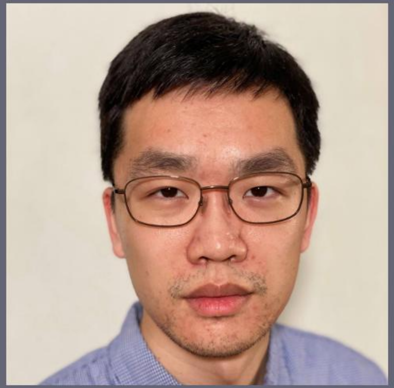
Coling'22

Accurate estimates of uncertainty are important for many difficult or sensitive prediction tasks in natural language processing (NLP). Though large-scale pre-trained models have vastly improved the accuracy of applied machine learning models throughout the field, there still are many instances in which they fail. The ability to precisely quantify uncertainty while handling the challenging scenarios that modern models can face when deployed in the real world is critical for reliable, consequential-decision making. This tutorial is intended for both academic researchers and industry practitioners alike, and provides a comprehensive introduction to uncertainty estimation for NLP problems---from fundamentals in probability calibration, Bayesian inference, and confidence set (or interval) construction, to applied topics in modern out-of-distribution detection and selective inference.

Speakers

Adam Fisch

MIT



Robin Jia

USC

Tal Schuster

Google Research

https://sites.google.com/view/uncertainty-nlp

Coling'22

Outline

(1) **Introduction**
- Understanding uncertainty.
- How do we express it? Use it?
- Examples in NLP applications.

(2) **Probability Calibration**
- A frequentist definition.
- Measuring calibration.
- Simple re-calibration methods.

(3) **Bayesian Approaches**
- Probabilistic models.
- Bayesian NNs, ensembles & dropout.
- Uses in active learning.

(4) **Conformal Prediction**
- Set-valued predictions with guarantees.
- Nonconformity scores to sets.
- Extensions and applications.

(5) **Selective Prediction & OOD Detection**
- Choosing to abstain.
- Training selection mechanisms.
- Distinguishing in-domain vs. out-domain.

(6) **Conclusion**
- Review of core concepts.
- Different views for uncertainty.
- Active areas of relevant research.

https://sites.google.com/view/uncertainty-nlp

NLI

Natural language inference

Natural language inference is the task of determining whether a “hypothesis” is true (entailment), false (contradiction), or undetermined (neutral) given a “premise”.

Example:

Premise	Label	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction	The man is sleeping.
An older and younger man smiling.	neutral	Two men are smiling and laughing at the cats playing on the floor.
A soccer game with multiple males playing.	entailment	Some men are playing a sport.

http://nlpprogress.com/english/natural_language_inference.html

PD

(P)	What is ultimate **purpose** of **life**?
(Q)	What is the **purpose** of **life** , if not money?
(P')	What is ultimate **measure** of **value**?
(Q')	What is the **measure** of **value** , if not money?
Label	*Positive*
Output	*Positive* (99.4%) → *Negative* (85.2%)
(P)	How can I get my **Gmail account** back ?
(Q)	What is the best **school management** software ?
(P')	How can I get my **credit score** back ?
(Q')	What is the best **credit score** software ?
Label	*Negative*
Output	*Negative* (100.0%) → *Positive* (68.3%)

Figure 1: Examples with labels *positive* and *negative* respectively, originally from Quora Question Pairs (QQP) (Iyer et al., 2017). "(P)" and "(Q)" are original sentences while "(P')" and "(Q')" are modified. Modified words are highlighted in bold. "Output" indicates the change of output labels by BERT (Devlin et al., 2018), where the percentage numbers are confidence scores.