Understanding and Modelling Pronouns in Translation: Resources, Methods, Challenges and Insights

Christian Hardmeier

IT University of Copenhagen

2023-04-14

Anaphoric Pronouns: The Prototypical Case

The latest version released in March is equipped with ... It is sold at ...

La dernière version lancée en mars est dotée de ... Elle est vendue ...

fem.sg.

Common Assumptions

The antecedent of a referring pronoun is another noun phrase in the text.

Counterexamples:

Pleonastic pronouns:
But I think it's a tragedy when one of them doesn't see the other.

Non-nominal reference:
There's so much more information about you,
and that's an important thing [. . .]

- Evaluation of different pronoun functions
- Annotation of non-nominal coreference

ParCorFull

A *multilingual parallel* corpus with **rich annotations** of coreference.

- Predecessor: **ParCor**
  (Guillou, **Hardmeier**, Smith and Tiedemann, 2014)
  - English, German and French
  - 11 TED talks, 8 EU Bookshop docs
  - Annotations of *pronouns and their direct antecedents*
- **ParCorFull 1.0**
  (Lapshinova-Koltunski, **Hardmeier** and Krielke, 2018)
  - English and German
  - 20 TED talks, 25 news articles
  - Annotation of *nominal and non-nominal coreference*
- **ParCorFull 2.0**
  (Lapshinova-Koltunski, Ferreira, Lartaud and **Hardmeier**, 2022)
  - English, German, French and Portuguese

Coreference annotation in ParCorFull

- Anaphoric noun phrases (including split antecedents, but not singletons)
  - Nouns with modifiers, personal and demonstrative pronouns, etc. *[the new report] – [the report] – [it]*
  - Comparative reference
    same, more, less, other, . . .
  - Indefinite pronouns
    anyone, someone, . . .
  - Substitution and ellipsis
- Extratextual reference (to slides, props, etc.)
- Temporal and local adverbs
  [in the 1920s] – [then]; [in the garden] – [there]
- Event reference

Event reference

- Reference to events (expressed by verb phrases), parts of the discourse, etc.
- In the original ParCor, this was a catch-all category.

[Democracy is in trouble], no question about [that], and [it] comes in part from a deep dilemma. . .

. . . our mission is [to organize the world's information and make it universally accessible]. And people always say, is [that] really what you guys are still doing?

[And I thought, why can't we do that today]? And [that]'s how this project got going.

Translating a pronoun requires generating a matching pronoun in the target language.

Counterexample:

But the thing about tryptamines is they cannot be taken orally because they're denatured by an enzyme [. . .] in the human gut [. . .]

Par contre les tryptamines ne peuvent pas [*tryptamines cannot*] être consommées par voie orale étant dénaturé[e]s [*being denatured*] par une enzyme [. . .] dans l'intestin de l'homme [. . .]

- Recognising and categorising non-literal translation patterns

Non-Literal Translation Patterns

- Matching referring expressions across languages
  (Lapshinova-Koltunski and **Hardmeier**, DiscoMT 2017;
  Šoštarić, **Hardmeier** and Stymne, WMT 2018)
  - in manually annotated data (ParCorFull)
  - in large unannotated corpora
- Matching coreference annotations across languages
  (Lapshinova-Koltunski, Loáiciga, **Hardmeier** and Krielke, CRAC 2019)
- Methodology:
  - Automatic word alignment (GIZA++, efmaral).
  - Matching chains.
  - Finding mismatches in chains (e.g., unaligned words).

Explicitation and implicitation

- Different referring expressions because of content differences.
- One language has more information than the other.

the French banking giant Société Générale, the owner of the local Komerčni banka (Commerce Bank)

le géant français Société Générale, propriétaire de la banque tchèque Komerčni banka.

Accommodation of language differences

- Differences in grammatical systems.
  This can often be analysed as obligatory explicitation.

  EN: Those are things ∅ you have in common with your parents and with your children.

  DE: [Dinge], [die] Sie mit Ihren Eltern und Kindern gemein haben.
- Differences in linguistic preferences

  EN: A reaction to the medication the clinic gave me for my depression left me suicidal.

  DE: Die Medikamente, die sie mir in der Ambulanz gegen meine Depressionen gaben, führten bei mir zu Selbstmordgedanken.

Different interpretations of corresponding referring expressions

We can create [a global parliament of mayors]. [That]'s an idea.

[We can create a global parliament of mayors]. [That]'s an idea.

EN: Think what happens [when we collect all of that data and we can put it together in order to find patterns we wouldn't see before]$_1$. [This]$_1$, I would suggest, perhaps [it]$_1$ will take a while, but [this]$_1$ will drive . Fabulous, lots of people talk about .

DE: Was passiert, [wenn wir all diese Daten sammeln und wir sie zusammenfügen können, um Muster zu erkennen, die wir nicht vorher sehen konnten]$_1$. Vielleicht dauert [dies]$_1$ ja noch eine Weile, aber [es]$_1$ wird eine Revolution in der Medizin. Fabelhaft –– sehr viele Leute sprechen [darüber]$_1$.

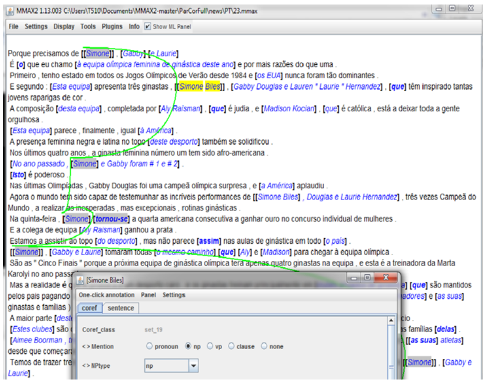
Annotation errors

- Annotation errors
- Word alignment errors
  - Statistical word alignment is a linguistically ill-defined task.
- Inconsistent interpretation of annotation guidelines across languages

Visualisation challenges

- In each language, we only see
  - one chain and
  - the properties of one markable

  at a time.
- Very easy to miss inconsistencies even in one language.
- Word alignment is not shown.



Lessons learnt

- Parallel corpus with rich coreference annotation is a very valuable resource.
- Difficult to achieve consistent annotation quality, especially over long time.
- What would we need?
  - Better corpus visualisation/navigation.
  - Word alignment with proper linguistic foundation.
  - Resources for *continuous* quality checks and annotator (re)training.

Referring pronouns agree in gender and number with their antecedent.

Counterexample:

Notional concord:
So I think Deep Mind, what's really amazing about Deep Mind is that it can actually – they're learning things in this unsupervised way.

- Studying linguistic preferences across languages and genres

Understanding Translation

How do the production and interpretation of referring expressions vary across languages?

Human production study:
Referring back to organisational named entities

Last week, Intel announced the shutdown of the factory.
In the press release, ____________

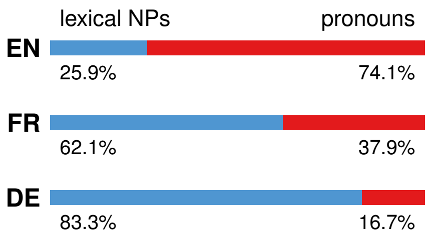
FC Barcelona won the World Cup three times since 2009.
Next year, ____________

AC/DC achieved international success in 1976.
In the next forty years, ____________

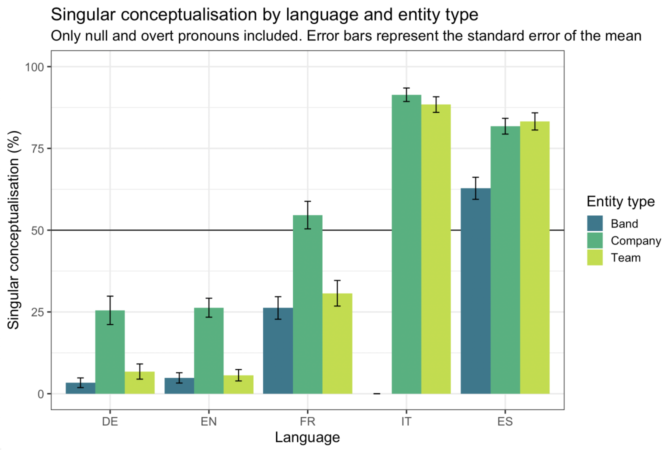
Ongoing work with Luca Bevacqua, Sharid Loáiciga and Hannah Rohde

Named Entity Reference: Results

FC Barcelona won the World Cup three times since 2009.
Next year, FC Barcelona/the club/it. . .



Singular vs. plural conceptualisation



Two Studies

Study 1: Constructed stimuli

- Prompt sentences constructed by the authors.
- Four types of named entities: Companies, publishers, sport teams and music bands.

Study 2: Corpus stimuli

- Prompt sentences were extracted from OntoNotes and simplified.
- Continuations were constructed to increase chances of eliciting a reference to the named entity.
- Generally longer and more complex than the constructed stimuli.
- Unrelated filler items also based on corpus data.

Generating prompts from corpora

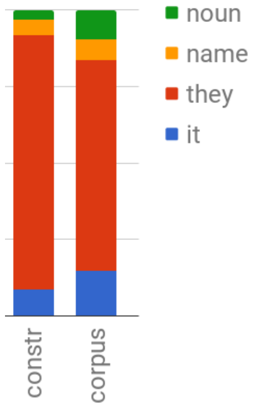
Original:
In the final trading, the House was insistent on setting aside $500 million to carry out base closings ordered to begin in fiscal 1990.

Prompt:
The House showed insistence on setting aside $500 million to carry out base closings ordered to begin in fiscal 1990.

In an amended piece of legislation, ___________

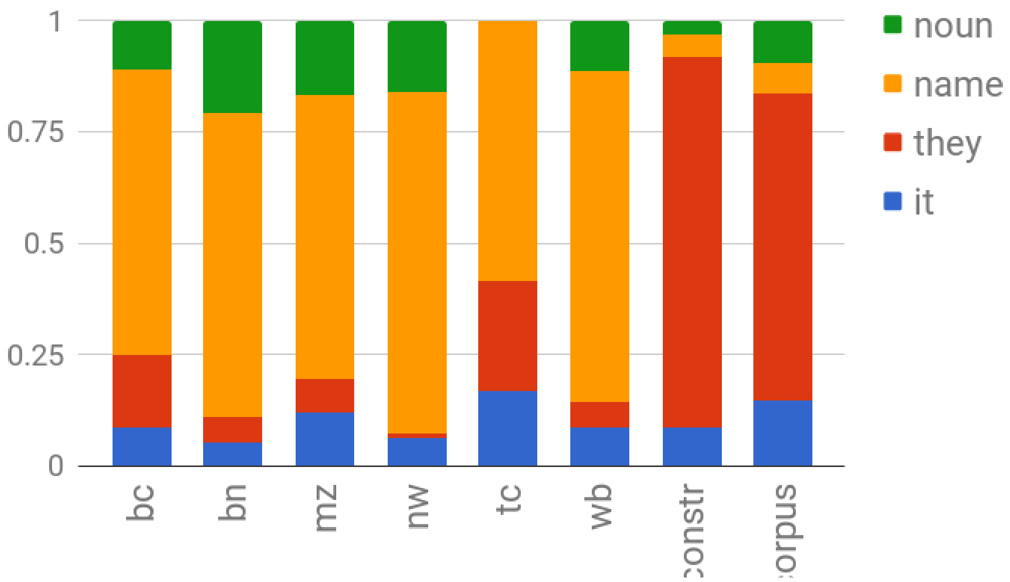
Hardmeier, Bevacqua, Loáiciga and Rohde, NEWS 2018

Continuation Studies: Results



	constructed	corpus
it	32	24
they	307	113
name	19	11
noun	12	16
total	370	164

Results including Corpus Study on OntoNotes



Conclusions

- Good pronoun translation is far more complex than enforcing gender agreement.
- Referring expression use differs significantly across languages.
  Good translation should respect target language conventions.
- Genre, register and modality also have strong effects.
- Annotation and exploration is made difficult by the lack of tools.
- ParCorFull 2.0 covers 4 European languages and can be used to study these phenomena or construct test suites.

Modelling Cross-Lingual Coreference

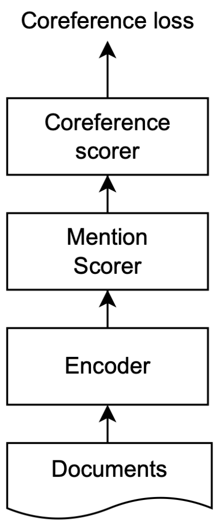
Work done with Gongbo Tang
(now at Beijing Language and Culture University)

Latent Anaphora Resolution for Cross-Lingual Pronoun Prediction

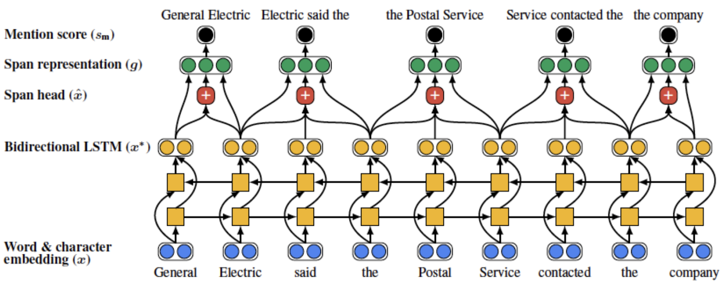
Christian Hardmeier **Jörg Tiedemann** **Joakim Nivre**
Uppsala University
Department of Linguistics and Philology
Box 635, 751 26 Uppsala, Sweden
`firstname.lastname@lingfil.uu.se`

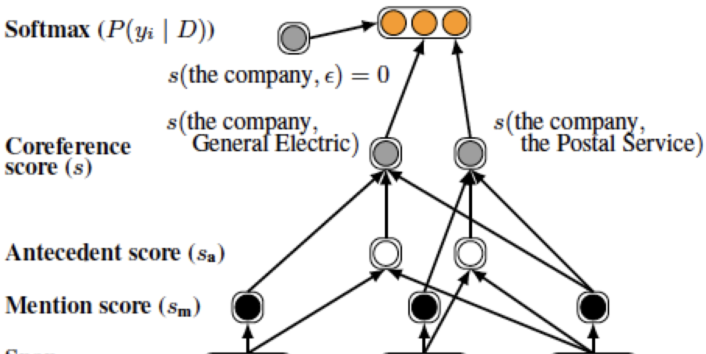
EMNLP 2013

Neural coreference models (Lee et al., 2017)



(a) Coreference Model





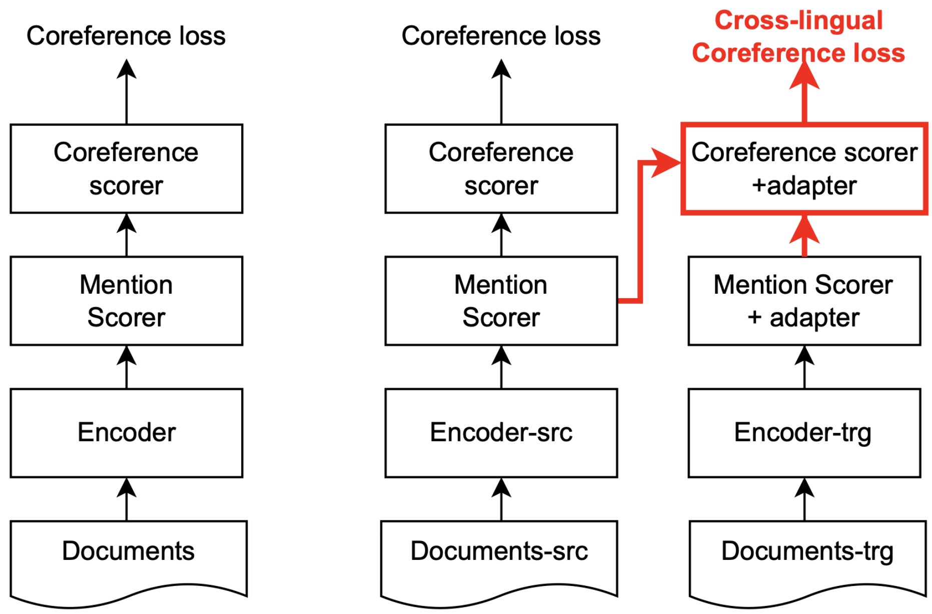
Cross-lingual coreference model

Motivation: Exploit signal from multilingual text
for better coreference resolution.

- Use second “copy” of coreference system in target language.
- Initialised from pretrained system, with adapter layers.
- Model scores coreference between target-language anaphors and source-language antecedents.
- Cross-lingual coreference loss:
  - Let $S = \{s_1, \ldots, s_m\}$ be the source mentions and $T = \{t_1, \ldots, t_n\}$ be the target mentions.
  - The network predicts a score s_{ij} for pairs (s_i, t_j).

$$\hat{j} = \arg\max_j s_{ij} \text{ for given } i; \qquad L = \sum_{i=1}^{m} e^{-s_{i\hat{j}}}$$

Cross-lingual coreference model



Experimental results

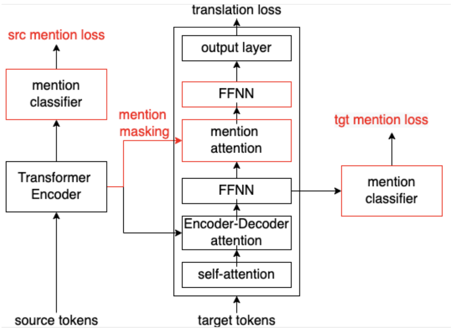
OntoNotes; TL data synthetically translated with MT systems from Facebook and Helsinki

	Mention detection F	Mention detection Δ	Coreference F	Coreference Δ
English	85.42	–	73.42	–
English–Arabic	86.13	0.71	74.58	1.16
English–Catalan	86.17	0.75	74.81	1.39
English–Chinese	86.02	0.60	74.53	1.11
English–Dutch	86.29	**0.87**	75.16	**1.74**
English–French	85.93	**0.51**	74.37	0.95
English–German	86.02	0.60	74.20	**0.78**
English–Italian	86.13	0.71	74.65	1.23
English–Russian	86.17	0.75	74.50	1.08
English–Spanish	86.21	0.79	74.50	1.08

Models with mention attention

New features:

- Mention attention module
- Mention classifiers:
  Is this part of a mention?
- Mention loss
- Mention masking:
  Only pass mention info
  to attention module



NMT modes with mention attention

Loss ratio: MT : M-src : M-tgt = 10 : 1 : 1

Experimental results

- WMT English to German (newstest2017)
- Evaluation: BLEU; APT for *it, they*

Model	BLEU	Pronouns	Ambig. pronouns
Baseline	28.01	60.1	50.4
Ours	28.23	61.2	52.2

Conclusions

- Cross-lingual data carries information relevant for coreference resolution.
- Effects on MT/coref performance are very consistent, but rather small.
- Significant cross-lingual variance in coreference structures for complex and non-obvious reasons.
- Annotating coreference involves potentially subjective *interpretation* – cross-lingual study exposes this.

Uncertainty Estimation

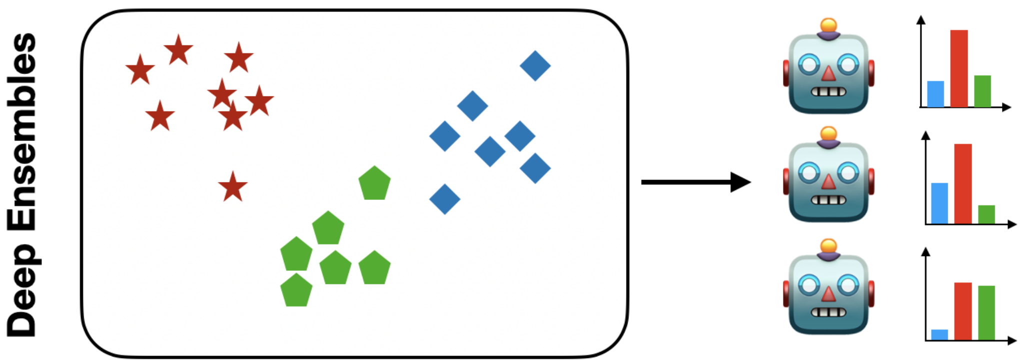
Work done with Dennis Ulmer and Jes Frellsen

Slide credit: Most of the following slides
were made by Dennis Ulmer.

Why model uncertainty?

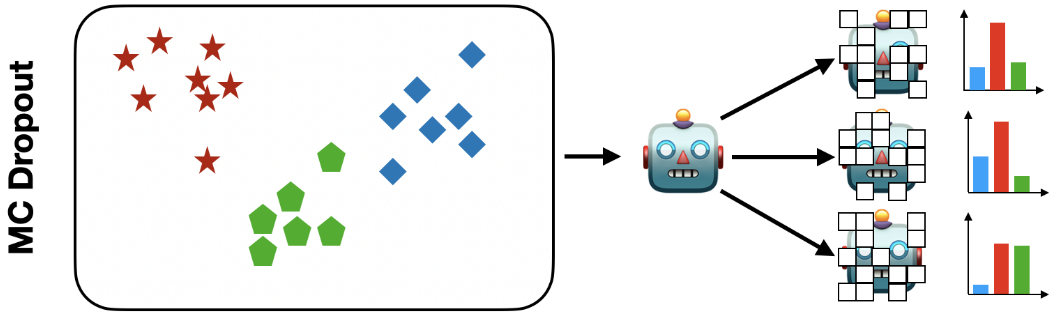
- Trustworthy AI: Systems should be open about what they know and what they don't.
- Responsible AI: Don't make decisions on an insufficient basis (stereotypes).
- Uncertainty is particularly important when an *uncertain prediction suggests a different course of action that any confident outcome*.
  - Escalate to some costlier process.
  - Refer decision to human.
  - Request further information.
  - ...

Deep Ensembles



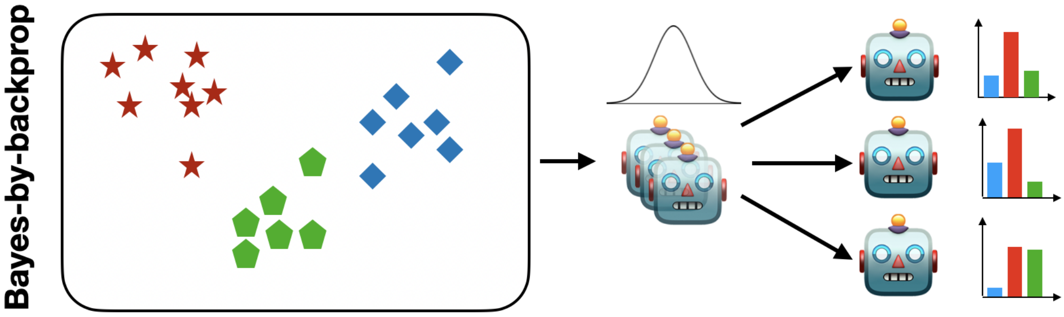
- Training multiple models allows estimating variance of predictions.
- Expensive to train.

Monte Carlo Dropout



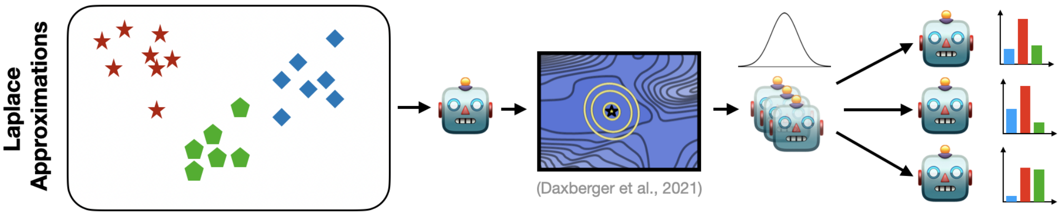
- "Ensembling" via different dropout masks.
- Easy to train, but often not a good approximation.

Bayes by Backprop



- Learn a Gaussian per parameter.
- Slower training/sampling, difficult convergence.

Laplace Approximation



(Daxberger et al., 2021)

- Gaussian approximation around MAP estimate.
- Hessian gives information about curvature.
- Difficult to compute.

Exploring Predictive Uncertainty and Calibration in NLP:
A Study on the Impact of Method & Data Scarcity

Dennis Ulmer[a] **Jes Frellsen**[b] **Christian Hardmeier**[a]
[a]Department of Computer Science, IT University of Copenhagen
[b]Department of Applied Mathematics & Computer Science, Technical University of Denmark
dennis.ulmer@mailbox.org

Findings of EMNLP, 2022

8 models, 3 languages/tasks

Models:

- LSTM and LSTM ensemble
- ST-τ LSTM: Model transitions in finite state automaton
- Variational LSTM and BERT: MC dropout
- Bayesian LSTM: Bayes-by-backprop
- SNGP BERT: Gaussian Process output layer
- DDU BERT: Fit Gaussian Mixture Model on hidden activations

Languages and Tasks:

- Danish: Named entity recognition
- Finnish: Part-of-speech tagging
- English: Intent classification

Calibration

	Model
Danish	LSTM
	Variational LSTM
	ST-τ LSTM
	Bayesian LSTM
	LSTM Ensemble
	SNGP BERT
	Variational BERT
	DDU BERT

Calibration

	Model	Task (ID/OOD) Acc.↑	Task (ID/OOD) F_1 ↑
Danish	LSTM	.93 ±.00 / .92 ±.00	.26 ±.01 / .19 ±.01
	Variational LSTM	.90 ±.02 / .90 ±.02	.08 ±.02 / .09 ±.02
	ST-τ LSTM	.92 ±.00 / .92 ±.00	.12 ±.00 / .09 ±.00
	Bayesian LSTM	.93 ±.00 / .93 ±.00	.07 ±.00 / .07 ±.00
	LSTM Ensemble	**.95** ±.00 / **.94** ±.00	**.33** ±.01 / **.25** ±.01
	SNGP BERT	.22 ±.35 / .19 ±.34	.03 ±.03 / .02 ±.02
	Variational BERT	.94 ±.00 / .89 ±.00	.29 ±.01 / .17 ±.00
	DDU BERT	.92 ±.00 / .89 ±.00	.25 ±.00 / .17 ±.00

Calibration

LSTM Ensemble beats all BERTs!

	Model	Task (ID/OOD) Acc.↑	Task (ID/OOD) F_1 ↑
Danish	LSTM	.93 ±.00 / .92 ±.00	.26 ±.01 / .19 ±.01
	Variational LSTM	.90 ±.02 / .90 ±.02	.08 ±.02 / .09 ±.02
	ST-τ LSTM	.92 ±.00 / .92 ±.00	.12 ±.00 / .09 ±.00
	Bayesian LSTM	.93 ±.00 / .93 ±.00	.07 ±.00 / .07 ±.00
	LSTM Ensemble	**.95** ±.00 / **.94** ±.00	**.33** ±.01 / **.25** ±.01
	SNGP BERT	.22 ±.35 / .19 ±.34	.03 ±.03 / .02 ±.02
	Variational BERT	.94 ±.00 / .89 ±.00	.29 ±.01 / .17 ±.00
	DDU BERT	.92 ±.00 / .89 ±.00	.25 ±.00 / .17 ±.00

Calibration

LSTM Ensemble beats all BERTs!

	Model	Task (ID/OOD) Acc.↑	Task (ID/OOD) F_1 ↑	Calibration (ID/OOD) ECE↓	Calibration (ID/OOD) ACE↓	Calibration (ID/OOD) %Cov.↑	Calibration (ID/OOD) ⌀Width↓
Danish	LSTM	.93 ±.00 / .92 ±.00	.26 ±.01 / .19 ±.01				
	Variational LSTM	.90 ±.02 / .90 ±.02	.08 ±.02 / .09 ±.02				
	ST-τ LSTM	.92 ±.00 / .92 ±.00	.12 ±.00 / .09 ±.00				
	Bayesian LSTM	.93 ±.00 / .93 ±.00	.07 ±.00 / .07 ±.00				
	LSTM Ensemble	**.95** ±.00 / **.94** ±.00	**.33** ±.01 / **.25** ±.01				
	SNGP BERT	.22 ±.35 / .19 ±.34	.03 ±.03 / .02 ±.02				
	Variational BERT	.94 ±.00 / .89 ±.00	.29 ±.01 / .17 ±.00				
	DDU BERT	.92 ±.00 / .89 ±.00	.25 ±.00 / .17 ±.00				

Calibration

	Model	Task (ID/OOD) Acc.↑	Task (ID/OOD) F_1 ↑	Calibration (ID/OOD) ECE↓	Calibration (ID/OOD) ACE↓	Calibration (ID/OOD) %Cov.↑	Calibration (ID/OOD) ⌀Width↓
Danish	LSTM	.93 ±.00 / .92 ±.00	.26 ±.01 / .19 ±.01	17.18 ±.00 / 17.17 ±.00	.16 ±.01 / .10 ±.01	**1.00** ±.00 / **1.00** ±.00	19.00 ±.00 / 19.00 ±.00
	Variational LSTM	.90 ±.02 / .90 ±.02	.08 ±.02 / .09 ±.02	16.74 ±.03 / 16.72 ±.03	.26 ±.02 / .17 ±.01	.99 ±.01 / .98 ±.01	6.62 ±.37 / 6.68 ±.33
	ST-τ LSTM	.92 ±.00 / .92 ±.00	.12 ±.00 / .09 ±.00	16.67 ±.00 / 16.63 ±.01	.24 ±.01 / .15 ±.01	1.00 ±.00 / .99 ±.00	7.10 ±.07 / 7.03 ±.08
	Bayesian LSTM	.93 ±.00 / .93 ±.00	.07 ±.00 / .07 ±.00	16.81 ±.00 / 16.79 ±.00	.25 ±.01 / .18 ±.01	1.00 ±.00 / 1.00 ±.00	1.68 ±.04 / 1.70 ±.05
	LSTM Ensemble	**.95** ±.00 / **.94** ±.00	**.33** ±.01 / **.25** ±.01	16.37 ±.00 / **16.35** ±.00	.18 ±.01 / .13 ±.01	.98 ±.00 / .97 ±.00	**1.62** ±.00 / **1.58** ±.01
	SNGP BERT	.22 ±.35 / .19 ±.34	.03 ±.03 / .02 ±.02	17.19 ±.01 / 17.18 ±.01	**.08** ±.01 / **.06** ±.01	1.00 ±.00 / 1.00 ±.00	18.84 ±.32 / 18.83 ±.34
	Variational BERT	.94 ±.00 / .89 ±.00	.29 ±.01 / .17 ±.00	**16.36** ±.00 / 16.43 ±.00	.20 ±.00 / .22 ±.00	.99 ±.00 / .98 ±.00	2.25 ±.01 / 3.86 ±.08
	DDU BERT	.92 ±.00 / .89 ±.00	.25 ±.00 / .17 ±.00	16.41 ±.00 / 16.44 ±.00	.19 ±.01 / .21 ±.01	.99 ±.00 / .99 ±.00	3.48 ±.01 / 4.04 ±.03

LSTM Ensemble beats
all BERTs!

Calibration

	Model	Task (ID/OOD) Acc.↑	Task (ID/OOD) F_1 ↑	Calibration (ID/OOD) ECE↓	Calibration (ID/OOD) ACE↓	Calibration (ID/OOD) %Cov.↑	Calibration (ID/OOD) ⌀Width↓
Danish	LSTM	.93 ±.00 / .92 ±.00	.26 ±.01 / .19 ±.01	17.18 ±.00 / 17.17 ±.00	.16 ±.01 / .10 ±.01	**1.00** ±.00 / **1.00** ±.00	19.00 ±.00 / 19.00 ±.00
	Variational LSTM	.90 ±.02 / .90 ±.02	.08 ±.02 / .09 ±.02	16.74 ±.03 / 16.72 ±.03	.26 ±.02 / .17 ±.01	.99 ±.01 / .98 ±.01	6.62 ±.37 / 6.68 ±.33
	ST-τ LSTM	.92 ±.00 / .92 ±.00	.12 ±.00 / .09 ±.00	16.67 ±.00 / 16.63 ±.01	.24 ±.01 / .15 ±.01	1.00 ±.00 / .99 ±.00	7.10 ±.07 / 7.03 ±.08
	Bayesian LSTM	.93 ±.00 / .93 ±.00	.07 ±.00 / .07 ±.00	16.81 ±.00 / 16.79 ±.00	.25 ±.01 / .18 ±.01	1.00 ±.00 / 1.00 ±.00	1.68 ±.04 / 1.70 ±.05
	LSTM Ensemble	**.95** ±.00 / **.94** ±.00	**.33** ±.01 / **.25** ±.01	16.37 ±.00 / **16.35** ±.00	.18 ±.01 / .13 ±.01	.98 ±.00 / .97 ±.00	**1.62** ±.00 / **1.58** ±.01
	SNGP BERT	.22 ±.35 / .19 ±.34	.03 ±.03 / .02 ±.02	17.19 ±.01 / 17.18 ±.01	**.08** ±.01 / **.06** ±.01	1.00 ±.00 / 1.00 ±.00	18.84 ±.32 / 18.83 ±.34
	Variational BERT	.94 ±.00 / .89 ±.00	.29 ±.01 / .17 ±.00	**16.36** ±.00 / 16.43 ±.00	.20 ±.00 / .22 ±.00	.99 ±.00 / .98 ±.00	2.25 ±.01 / 3.86 ±.08
	DDU BERT	.92 ±.00 / .89 ±.00	.25 ±.00 / .17 ±.00	16.41 ±.00 / 16.44 ±.00	.19 ±.01 / .21 ±.01	.99 ±.00 / .99 ±.00	3.48 ±.01 / 4.04 ±.03

LSTM Ensemble beats
all BERTs!

LSTM Ensemble on par with
pre-trained models

Calibration

LSTM spreads prob. many
classes

	Model	Task (ID/OOD) Acc.↑	Task (ID/OOD) F_1 ↑	Calibration (ID/OOD) ECE↓	Calibration (ID/OOD) ACE↓	Calibration (ID/OOD) %Cov.↑	Calibration (ID/OOD) ⌀Width↓
Danish	LSTM	.93 ±.00 / .92 ±.00	.26 ±.01 / .19 ±.01	17.18 ±.00 / 17.17 ±.00	.16 ±.01 / .10 ±.01	**1.00** ±.00 / **1.00** ±.00	19.00 ±.00 / 19.00 ±.00
	Variational LSTM	.90 ±.02 / .90 ±.02	.08 ±.02 / .09 ±.02	16.74 ±.03 / 16.72 ±.03	.26 ±.02 / .17 ±.01	.99 ±.01 / .98 ±.01	6.62 ±.37 / 6.68 ±.33
	ST-τ LSTM	.92 ±.00 / .92 ±.00	.12 ±.00 / .09 ±.00	16.67 ±.00 / 16.63 ±.01	.24 ±.01 / .15 ±.01	1.00 ±.00 / .99 ±.00	7.10 ±.07 / 7.03 ±.08
	Bayesian LSTM	.93 ±.00 / .93 ±.00	.07 ±.00 / .07 ±.00	16.81 ±.00 / 16.79 ±.00	.25 ±.01 / .18 ±.01	1.00 ±.00 / 1.00 ±.00	1.68 ±.04 / 1.70 ±.05
	LSTM Ensemble	**.95** ±.00 / **.94** ±.00	**.33** ±.01 / **.25** ±.01	16.37 ±.00 / **16.35** ±.00	.18 ±.01 / .13 ±.01	.98 ±.00 / .97 ±.00	**1.62** ±.00 / **1.58** ±.01
	SNGP BERT	.22 ±.35 / .19 ±.34	.03 ±.03 / .02 ±.02	17.19 ±.01 / 17.18 ±.01	**.08** ±.01 / **.06** ±.01	1.00 ±.00 / 1.00 ±.00	18.84 ±.32 / 18.83 ±.34
	Variational BERT	.94 ±.00 / .89 ±.00	.29 ±.01 / .17 ±.00	**16.36** ±.00 / 16.43 ±.00	.20 ±.00 / .22 ±.00	.99 ±.00 / .98 ±.00	2.25 ±.01 / 3.86 ±.08
	DDU BERT	.92 ±.00 / .89 ±.00	.25 ±.00 / .17 ±.00	16.41 ±.00 / 16.44 ±.00	.19 ±.01 / .21 ±.01	.99 ±.00 / .99 ±.00	3.48 ±.01 / 4.04 ±.03

LSTM Ensemble beats
all BERTs!

LSTM Ensemble on par with
pre-trained models

Calibration

LSTM spreads prob. many classes

	Model	Task (ID/OOD) Acc.↑	Task (ID/OOD) F_1 ↑	Calibration (ID/OOD) ECE↓	Calibration (ID/OOD) ACE↓	Calibration (ID/OOD) %Cov.↑	Calibration (ID/OOD) ⌀Width↓
Danish	LSTM	.93 ±.00 / .92 ±.00	.26 ±.01 / .19 ±.01	17.18 ±.00 / 17.17 ±.00	.16 ±.01 / .10 ±.01	**1.00** ±.00 / **1.00** ±.00	19.00 ±.00 / 19.00 ±.00
	Variational LSTM	.90 ±.02 / .90 ±.02	.08 ±.02 / .09 ±.02	16.74 ±.03 / 16.72 ±.03	.26 ±.02 / .17 ±.01	.99 ±.01 / .98 ±.01	6.62 ±.37 / 6.68 ±.33
	ST-τ LSTM	.92 ±.00 / .92 ±.00	.12 ±.00 / .09 ±.00	16.67 ±.00 / 16.63 ±.01	.24 ±.01 / .15 ±.01	1.00 ±.00 / .99 ±.00	7.10 ±.07 / 7.03 ±.08
	Bayesian LSTM	.93 ±.00 / .93 ±.00	.07 ±.00 / .07 ±.00	16.81 ±.00 / 16.79 ±.00	.25 ±.01 / .18 ±.01	1.00 ±.00 / 1.00 ±.00	1.68 ±.04 / 1.70 ±.05
	LSTM Ensemble	**.95** ±.00 / **.94** ±.00	**.33** ±.01 / **.25** ±.01	16.37 ±.00 / **16.35** ±.00	.18 ±.01 / .13 ±.01	.98 ±.00 / .97 ±.00	**1.62** ±.00 / **1.58** ±.01
	SNGP BERT	.22 ±.35 / .19 ±.34	.03 ±.03 / .02 ±.02	17.19 ±.01 / 17.18 ±.01	**.08** ±.01 / **.06** ±.01	1.00 ±.00 / 1.00 ±.00	18.84 ±.32 / 18.83 ±.34
	Variational BERT	.94 ±.00 / .89 ±.00	.29 ±.01 / .17 ±.00	**16.36** ±.00 / 16.43 ±.00	.20 ±.00 / .22 ±.00	.99 ±.00 / .98 ±.00	2.25 ±.01 / 3.86 ±.08
	DDU BERT	.92 ±.00 / .89 ±.00	.25 ±.00 / .17 ±.00	16.41 ±.00 / 16.44 ±.00	.19 ±.01 / .21 ±.01	.99 ±.00 / .99 ±.00	3.48 ±.01 / 4.04 ±.03

LSTM Ensemble beats all BERTs!

Ensemble and BERTs are confidently correct

LSTM Ensemble on par with pre-trained models

Uncertainty Quality

Uncertainty Quality

	Model	Task (ID/OOD) Acc.↑	Task (ID/OOD) F_1 ↑
Danish	LSTM	.93 ±.00 / .92 ±.00	.26 ±.01 / .19 ±.01
	Variational LSTM	.90 ±.02 / .90 ±.02	.08 ±.02 / .09 ±.02
	ST-τ LSTM	.92 ±.00 / .92 ±.00	.12 ±.00 / .09 ±.00
	Bayesian LSTM	.93 ±.00 / .93 ±.00	.07 ±.00 / .07 ±.00
	LSTM Ensemble	**.95** ±.00 / **.94** ±.00	**.33** ±.01 / **.25** ±.01
	SNGP BERT	.22 ±.35 / .19 ±.34	.03 ±.03 / .02 ±.02
	Variational BERT	.94 ±.00 / .89 ±.00	.29 ±.01 / .17 ±.00
	DDU BERT	.92 ±.00 / .89 ±.00	.25 ±.00 / .17 ±.00

Uncertainty Quality

🐘🏠: How to evaluate uncertainty quality w/o gold labels?

	Model	Task (ID/OOD)	
		Acc.↑	F_1 ↑
Danish	LSTM	.93 ±.00 / .92 ±.00	.26 ±.01 / .19 ±.01
	Variational LSTM	.90 ±.02 / .90 ±.02	.08 ±.02 / .09 ±.02
	ST-τ LSTM	.92 ±.00 / .92 ±.00	.12 ±.00 / .09 ±.00
	Bayesian LSTM	.93 ±.00 / .93 ±.00	.07 ±.00 / .07 ±.00
	LSTM Ensemble	**.95** ±.00 / **.94** ±.00	**.33** ±.01 / **.25** ±.01
	SNGP BERT	.22 ±.35 / .19 ±.34	.03 ±.03 / .02 ±.02
	Variational BERT	.94 ±.00 / .89 ±.00	.29 ±.01 / .17 ±.00
	DDU BERT	.92 ±.00 / .89 ±.00	.25 ±.00 / .17 ±.00

Uncertainty Quality

🐘🏠: How to evaluate uncertainty quality w/o gold labels?

	Model	Task (ID/OOD)		Uncertainty (ID/OOD)			
		Acc.↑	F_1 ↑	AUROC↑	AUPR↑	Token τ ↑	Seq. τ ↑
Danish	LSTM	.93 ±.00 / .92 ±.00	.26 ±.01 / .19 ±.01				
	Variational LSTM	.90 ±.02 / .90 ±.02	.08 ±.02 / .09 ±.02				
	ST-τ LSTM	.92 ±.00 / .92 ±.00	.12 ±.00 / .09 ±.00				
	Bayesian LSTM	.93 ±.00 / .93 ±.00	.07 ±.00 / .07 ±.00				
	LSTM Ensemble	**.95** ±.00 / **.94** ±.00	**.33** ±.01 / **.25** ±.01				
	SNGP BERT	.22 ±.35 / .19 ±.34	.03 ±.03 / .02 ±.02				
	Variational BERT	.94 ±.00 / .89 ±.00	.29 ±.01 / .17 ±.00				
	DDU BERT	.92 ±.00 / .89 ±.00	.25 ±.00 / .17 ±.00				

Uncertainty Quality

How well can uncertainty distinguish ID / OOD?

🐘🏠: How to evaluate uncertainty quality w/o gold labels?

	Model	Task (ID/OOD)		Uncertainty (ID/OOD)			
		Acc.↑	F_1 ↑	AUROC↑	AUPR↑	Token τ ↑	Seq. τ ↑
Danish	LSTM	.93 ±.00 / .92 ±.00	.26 ±.01 / .19 ±.01				
	Variational LSTM	.90 ±.02 / .90 ±.02	.08 ±.02 / .09 ±.02				
	ST-τ LSTM	.92 ±.00 / .92 ±.00	.12 ±.00 / .09 ±.00				
	Bayesian LSTM	.93 ±.00 / .93 ±.00	.07 ±.00 / .07 ±.00				
	LSTM Ensemble	**.95** ±.00 / **.94** ±.00	**.33** ±.01 / **.25** ±.01				
	SNGP BERT	.22 ±.35 / .19 ±.34	.03 ±.03 / .02 ±.02				
	Variational BERT	.94 ±.00 / .89 ±.00	.29 ±.01 / .17 ±.00				
	DDU BERT	.92 ±.00 / .89 ±.00	.25 ±.00 / .17 ±.00				

Uncertainty Quality

🐘🏠: How to evaluate uncertainty quality w/o gold labels?

How well can uncertainty distinguish ID / OOD?

How indicative is uncertainty of model loss?

	Model	Task (ID/OOD) Acc.↑	Task (ID/OOD) F_1 ↑
Danish	LSTM	.93 ±.00 / .92 ±.00	.26 ±.01 / .19 ±.01
	Variational LSTM	.90 ±.02 / .90 ±.02	.08 ±.02 / .09 ±.02
	ST-τ LSTM	.92 ±.00 / .92 ±.00	.12 ±.00 / .09 ±.00
	Bayesian LSTM	.93 ±.00 / .93 ±.00	.07 ±.00 / .07 ±.00
	LSTM Ensemble	**.95** ±.00 / **.94** ±.00	**.33** ±.01 / **.25** ±.01
	SNGP BERT	.22 ±.35 / .19 ±.34	.03 ±.03 / .02 ±.02
	Variational BERT	.94 ±.00 / .89 ±.00	.29 ±.01 / .17 ±.00
	DDU BERT	.92 ±.00 / .89 ±.00	.25 ±.00 / .17 ±.00

Uncertainty Quality

🐘🏠: How to evaluate uncertainty quality w/o gold labels?

How well can uncertainty distinguish ID / OOD?

How indicative is uncertainty of model loss?

	Model	Task (ID/OOD) Acc.↑	Task (ID/OOD) F_1 ↑	Uncertainty (ID/OOD) AUROC↑	Uncertainty (ID/OOD) AUPR↑	Uncertainty (ID/OOD) Token τ ↑	Uncertainty (ID/OOD) Seq. τ ↑
Danish	LSTM	.93 ±.00 / .92 ±.00	.26 ±.01 / .19 ±.01	.50○ ±.02	.14○ ±.01	.50○ ±.01 / .47○ ±.00	−.26✚ ±.02 / −.28○ ±.05
	Variational LSTM	.90 ±.02 / .90 ±.02	.08 ±.02 / .09 ±.02	.60✚ ±.04	.21✚ ±.02	.23○ ±.06 / .23○ ±.05	−.04✖ ±.02 / −.02□ ±.05
	ST-τ LSTM	.92 ±.00 / .92 ±.00	.12 ±.00 / .09 ±.00	.54✚ ±.01	.15✚ ±.01	.50○ ±.00 / .48○ ±.00	−.05□ ±.03 / −.01□ ±.05
	Bayesian LSTM	.93 ±.00 / .93 ±.00	.07 ±.00 / .07 ±.00	.65⬠ ±.17	.31⬠ ±.30	.53○ ±.01 / **.55**○ ±.01	−.01□ ±.07 / −.02✚ ±.04
	LSTM Ensemble	**.95** ±.00 / **.94** ±.00	**.33** ±.01 / **.25** ±.01	.60□ ±.02	.18□ ±.01	.44□ ±.00 / .45□ ±.00	−.19✚ ±.01 / −.28□ ±.01
	SNGP BERT	.22 ±.35 / .19 ±.34	.03 ±.03 / .02 ±.02	.86△ ±.06	.49△ ±.12	.17□ ±.09 / .26□ ±.14	**.29**✖ ±.03 / **.44**□ ±.11
	Variational BERT	.94 ±.00 / .89 ±.00	.29 ±.01 / .17 ±.00	.86✚ ±.01	.46✚ ±.02	.42○ ±.00 / .17⬠ ±.00	−.35□ ±.01 / −.41□ ±.01
	DDU BERT	.92 ±.00 / .89 ±.00	.25 ±.00 / .17 ±.00	.86○ ±.01	.39○ ±.02	**.56**○ ±.00 / .25○ ±.01	−.24○ ±.01 / −.38○ ±.03

Uncertainty Quality

🐘🏠: How to evaluate uncertainty quality w/o gold labels?

How well can uncertainty distinguish ID / OOD?

How indicative is uncertainty of model loss?

	Model	Task (ID/OOD) Acc.↑	Task (ID/OOD) F_1 ↑	Uncertainty (ID/OOD) AUROC↑	Uncertainty (ID/OOD) AUPR↑	Uncertainty (ID/OOD) Token τ ↑	Uncertainty (ID/OOD) Seq. τ ↑
Danish	LSTM	.93 ±.00 / .92 ±.00	.26 ±.01 / .19 ±.01	.50○ ±.02	.14○ ±.01	.50○ ±.01 / .47○ ±.00	−.26✚ ±.02 / −.28○ ±.05
	Variational LSTM	.90 ±.02 / .90 ±.02	.08 ±.02 / .09 ±.02	.60✚ ±.04	.21✚ ±.02	.23○ ±.06 / .23○ ±.05	−.04✖ ±.02 / −.02□ ±.05
	ST-τ LSTM	.92 ±.00 / .92 ±.00	.12 ±.00 / .09 ±.00	.54✚ ±.01	.15✚ ±.01	.50○ ±.00 / .48○ ±.00	−.05□ ±.03 / −.01□ ±.05
	Bayesian LSTM	.93 ±.00 / .93 ±.00	.07 ±.00 / .07 ±.00	.65⬠ ±.17	.31⬠ ±.30	.53○ ±.01 / **.55**○ ±.01	−.01□ ±.07 / −.02✚ ±.04
	LSTM Ensemble	**.95** ±.00 / **.94** ±.00	**.33** ±.01 / **.25** ±.01	.60□ ±.02	.18□ ±.01	.44□ ±.00 / .45□ ±.00	−.19✚ ±.01 / −.28□ ±.01
	SNGP BERT	.22 ±.35 / .19 ±.34	.03 ±.03 / .02 ±.02	.86△ ±.06	.49△ ±.12	.17□ ±.09 / .26□ ±.14	**.29**✖ ±.03 / **.44**□ ±.11
	Variational BERT	.94 ±.00 / .89 ±.00	.29 ±.01 / .17 ±.00	.86✚ ±.01	.46✚ ±.02	.42○ ±.00 / .17⬠ ±.00	−.35□ ±.01 / −.41□ ±.01
	DDU BERT	.92 ±.00 / .89 ±.00	.25 ±.00 / .17 ±.00	.86○ ±.01	.39○ ±.02	**.56**○ ±.00 / .25○ ±.01	−.24○ ±.01 / −.38○ ±.03

Pre-trained models most sensitive to OOD

Uncertainty Quality

🐘🏠: How to evaluate uncertainty quality w/o gold labels?

How well can uncertainty distinguish ID / OOD?

How indicative is uncertainty of model loss?

	Model	Task (ID/OOD) Acc.↑	F_1 ↑	Uncertainty (ID/OOD) AUROC↑	AUPR↑	Token τ ↑	Seq. τ ↑
	LSTM	.93 ±.00 / .92 ±.00	.26 ±.01 / .19 ±.01	.50○ ±.02	.14○ ±.01	.50○ ±.01 / .47○ ±.00	−.26✚ ±.02 / −.28○ ±.05
	Variational LSTM	.90 ±.02 / .90 ±.02	.08 ±.02 / .09 ±.02	.60✚ ±.04	.21✚ ±.02	.23○ ±.06 / .23○ ±.05	−.04✖ ±.02 / −.02□ ±.05
	ST-τ LSTM	.92 ±.00 / .92 ±.00	.12 ±.00 / .09 ±.00	.54✚ ±.01	.15✚ ±.01	.50○ ±.00 / .48○ ±.00	−.05□ ±.03 / −.01□ ±.05
Danish	Bayesian LSTM	.93 ±.00 / .93 ±.00	.07 ±.00 / .07 ±.00	.65⬠ ±.17	.31⬠ ±.30	.53○ ±.01 / **.55**○ ±.01	−.01□ ±.07 / −.02✚ ±.04
	LSTM Ensemble	**.95** ±.00 / **.94** ±.00	**.33** ±.01 / **.25** ±.01	.60□ ±.02	.18□ ±.01	.44□ ±.00 / .45□ ±.00	−.19✚ ±.01 / −.28□ ±.01
	SNGP BERT	.22 ±.35 / .19 ±.34	.03 ±.03 / .02 ±.02	.86△ ±.06	.49△ ±.12	.17□ ±.09 / .26□ ±.14	**.29**✖ ±.03 / **.44**□ ±.11
	Variational BERT	.94 ±.00 / .89 ±.00	.29 ±.01 / .17 ±.00	.86✚ ±.01	.46✚ ±.02	.42○ ±.00 / .17⬠ ±.00	−.35□ ±.01 / −.41□ ±.01
	DDU BERT	.92 ±.00 / .89 ±.00	.25 ±.00 / .17 ±.00	.86○ ±.01	.39○ ±.02	**.56**○ ±.00 / .25○ ±.01	−.24○ ±.01 / −.38○ ±.03

Pre-trained models most sensitive to OOD

Bayesian LSTM performs best on ID & OOD

Uncertainty Quality

🐘🏠: How to evaluate uncertainty quality w/o gold labels?

How well can uncertainty distinguish ID / OOD?

How indicative is uncertainty of model loss?

	Model	Task (ID/OOD) Acc.↑	F_1 ↑	Uncertainty (ID/OOD) AUROC↑	AUPR↑	Token τ ↑	Seq. τ ↑
	LSTM	.93 ±.00 / .92 ±.00	.26 ±.01 / .19 ±.01	.50○ ±.02	.14○ ±.01	.50○ ±.01 / .47○ ±.00	−.26✚ ±.02 / −.28○ ±.05
	Variational LSTM	.90 ±.02 / .90 ±.02	.08 ±.02 / .09 ±.02	.60✚ ±.04	.21✚ ±.02	.23○ ±.06 / .23○ ±.05	−.04✖ ±.02 / −.02□ ±.05
	ST-τ LSTM	.92 ±.00 / .92 ±.00	.12 ±.00 / .09 ±.00	.54✚ ±.01	.15✚ ±.01	.50○ ±.00 / .48○ ±.00	−.05□ ±.03 / −.01□ ±.05
Danish	Bayesian LSTM	.93 ±.00 / .93 ±.00	.07 ±.00 / .07 ±.00	.65⬠ ±.17	.31⬠ ±.30	.53○ ±.01 / **.55**○ ±.01	−.01□ ±.07 / −.02✚ ±.04
	LSTM Ensemble	**.95** ±.00 / **.94** ±.00	**.33** ±.01 / **.25** ±.01	.60□ ±.02	.18□ ±.01	.44□ ±.00 / .45□ ±.00	−.19✚ ±.01 / −.28□ ±.01
	SNGP BERT	.22 ±.35 / .19 ±.34	.03 ±.03 / .02 ±.02	.86△ ±.06	.49△ ±.12	.17□ ±.09 / .26□ ±.14	**.29**✖ ±.03 / **.44**□ ±.11
	Variational BERT	.94 ±.00 / .89 ±.00	.29 ±.01 / .17 ±.00	.86✚ ±.01	.46✚ ±.02	.42○ ±.00 / .17⬠ ±.00	−.35□ ±.01 / −.41□ ±.01
	DDU BERT	.92 ±.00 / .89 ±.00	.25 ±.00 / .17 ±.00	.86○ ±.01	.39○ ±.02	**.56**○ ±.00 / .25○ ±.01	−.24○ ±.01 / −.38○ ±.03

Except for SNGP, seq.-level correlations are **negative**! 🤯
(but SNGP training very brittle)

Pre-trained models most sensitive to OOD

Bayesian LSTM performs best on ID & OOD

Uncertainty Quality

🐘🏠: How to evaluate uncertainty quality w/o gold labels?

No superior uncertainty metric!

How well can uncertainty distinguish ID / OOD?

How indicative is uncertainty of model loss?

	Model	Task (ID/OOD) Acc.↑	F_1 ↑	Uncertainty (ID/OOD) AUROC↑	AUPR↑	Token τ ↑	Seq. τ ↑
	LSTM	.93 ±.00 / .92 ±.00	.26 ±.01 / .19 ±.01	.50○ ±.02	.14○ ±.01	.50○ ±.01 / .47○ ±.00	−.26✚ ±.02 / −.28○ ±.05
	Variational LSTM	.90 ±.02 / .90 ±.02	.08 ±.02 / .09 ±.02	.60✚ ±.04	.21✚ ±.02	.23○ ±.06 / .23○ ±.05	−.04✖ ±.02 / −.02□ ±.05
	ST-τ LSTM	.92 ±.00 / .92 ±.00	.12 ±.00 / .09 ±.00	.54✚ ±.01	.15✚ ±.01	.50○ ±.00 / .48○ ±.00	−.05□ ±.03 / −.01□ ±.05
Danish	Bayesian LSTM	.93 ±.00 / .93 ±.00	.07 ±.00 / .07 ±.00	.65⬠ ±.17	.31⬠ ±.30	.53○ ±.01 / **.55**○ ±.01	−.01□ ±.07 / −.02✚ ±.04
	LSTM Ensemble	**.95** ±.00 / **.94** ±.00	**.33** ±.01 / **.25** ±.01	.60□ ±.02	.18□ ±.01	.44□ ±.00 / .45□ ±.00	−.19✚ ±.01 / −.28□ ±.01
	SNGP BERT	.22 ±.35 / .19 ±.34	.03 ±.03 / .02 ±.02	.86△ ±.06	.49△ ±.12	.17□ ±.09 / .26□ ±.14	**.29**✖ ±.03 / **.44**□ ±.11
	Variational BERT	.94 ±.00 / .89 ±.00	.29 ±.01 / .17 ±.00	.86✚ ±.01	.46✚ ±.02	.42○ ±.00 / .17⬠ ±.00	−.35□ ±.01 / −.41□ ±.01
	DDU BERT	.92 ±.00 / .89 ±.00	.25 ±.00 / .17 ±.00	.86○ ±.01	.39○ ±.02	**.56**○ ±.00 / .25○ ±.01	−.24○ ±.01 / −.38○ ±.03

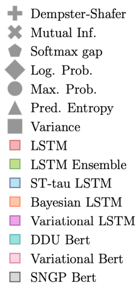
Except for SNGP, seq.-level correlations are **negative**! 🤯
(but SNGP training very brittle)

Pre-trained models most sensitive to OOD

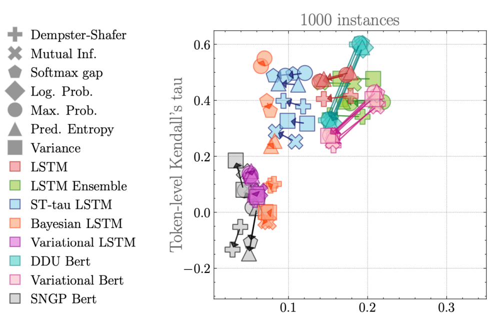
Bayesian LSTM performs best on ID & OOD

Influence of Training Set Size

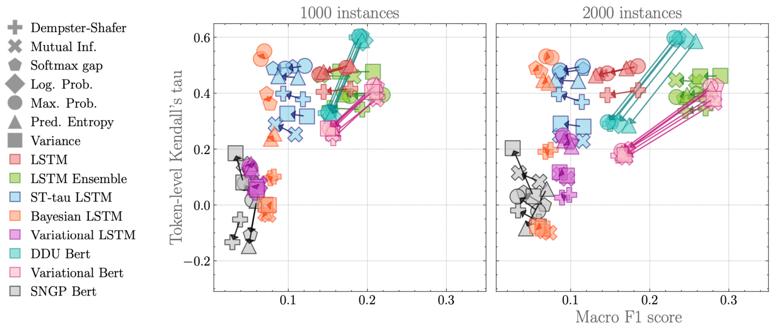
Influence of Training Set Size



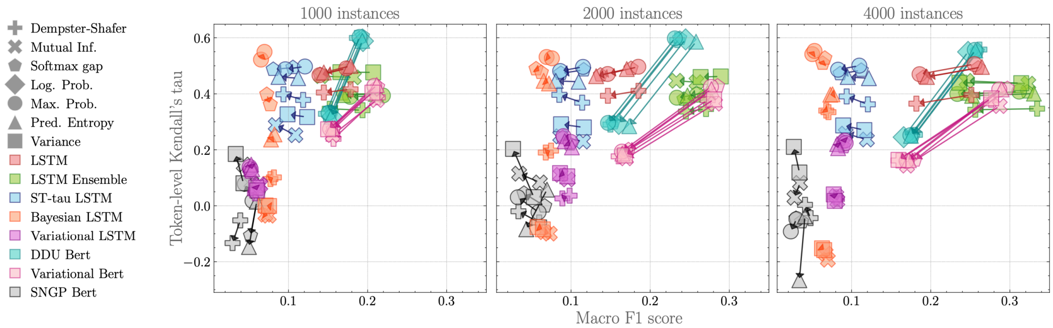
Influence of Training Set Size



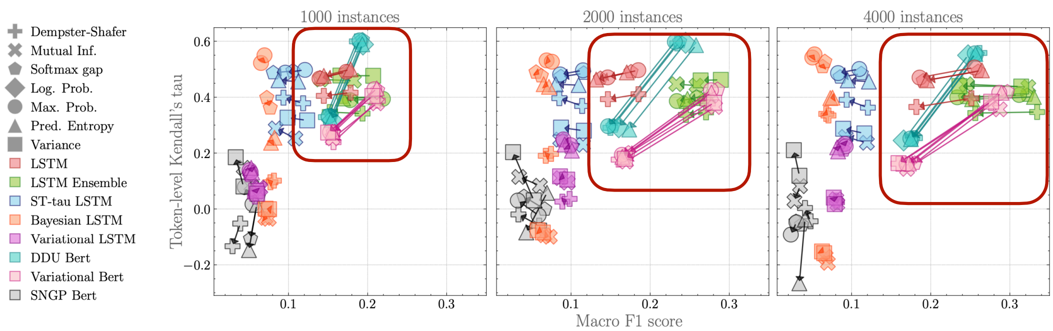
Influence of Training Set Size



Influence of Training Set Size



Influence of Training Set Size



The more training data we add, the higher the BERT gap on OOD performance!

Evidential Deep Learning

- Instead of training/approximating ensemble,
  directly parameterise a distribution over outputs.
- Get uncertainty estimates in a single pass, without MC etc.
- Model represents the accumulation of *evidence* in the training data.
- For a categorical output distribution (classification),
  the model's output is a Dirichlet distribution (conjugate prior).

Published in Transactions on Machine Learning Research (04/2023)

Prior and Posterior Networks: A Survey on Evidential Deep Learning Methods For Uncertainty Estimation

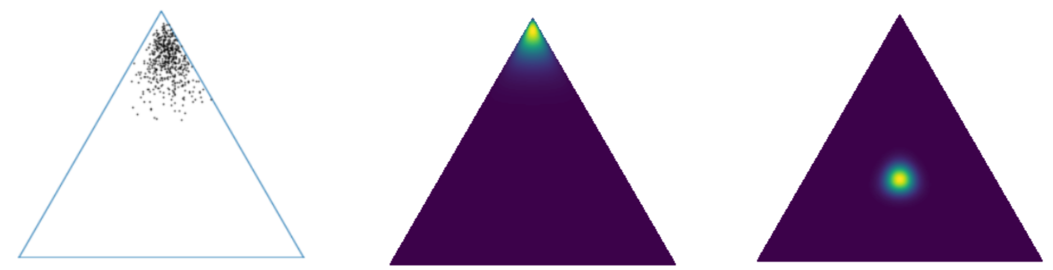
Dennis Ulmer[a, c] *dennis.ulmer@mailbox.org*
Christian Hardmeier[a, c] *chrha@itu.dk*
Jes Frellsen[b, c] *jefr@dtu.dk*

[a] *IT University of Copenhagen,* [b] *Technical University of Denmark,* [c] *Pioneer Centre for Artificial Intelligence*

Types of uncertainty

Data (aleatoric) uncertainty
Uncertainty inherent in the data (e.g., true ambiguity, annotation error, etc.).
Not reducible by acquiring more data.

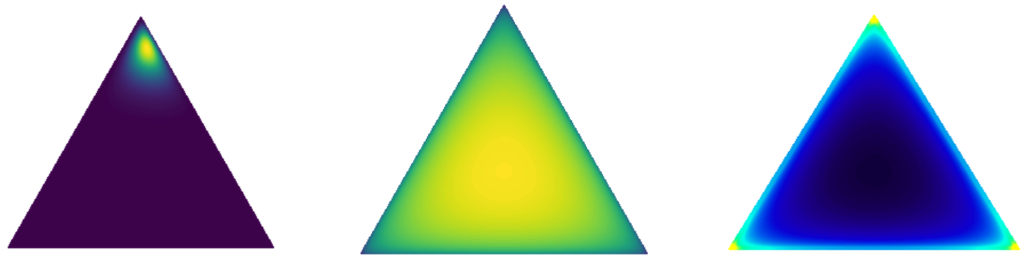
Model (epistemic) uncertainty
Uncertainty due to the model not having enough information.
Adding more data should reduce this.



(a) Categorical distributions predicted by a neural ensemble on the probability simplex.

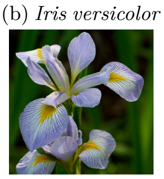
(b) Probability simplex for a confident prediction, for with the density concentrated in a single corner.

(c) Dirichlet distribution for a case of data uncertainty, with the density concentrated in the center.

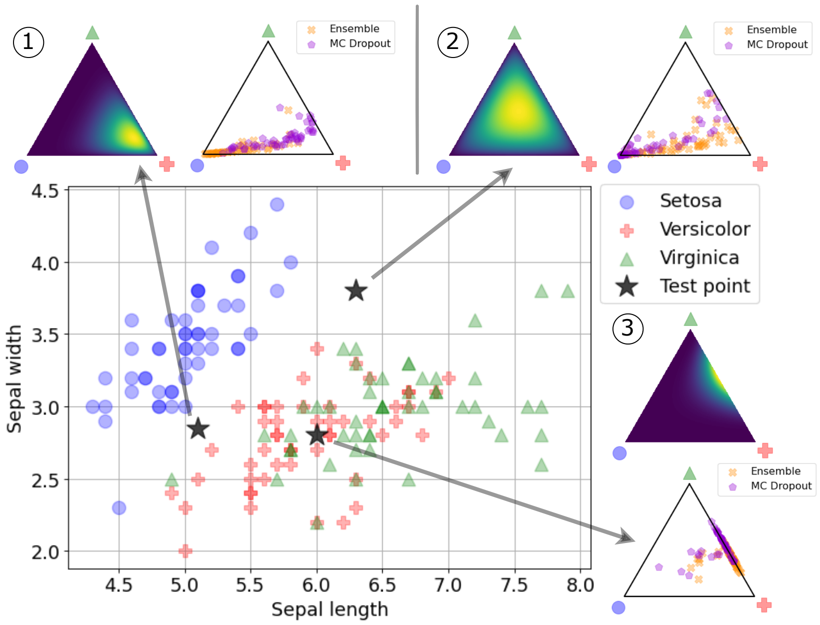


(a) *Iris setosa*

(b) *Iris versicolor*



(c) *Iris virginica*



Uncertainty estimation in NLP

- Some methods have been adapted for NLP.
  - Ensembling
  - Bayes-by-backprop
  - MC dropout
- Few NLP-specific works.
- But. . .
  - Mostly proposed for classification tasks.
  - Not applicable to really big models (GPT3 ensembles???)
  - Not tailored towards NLP-specific challenges
    (sequences, sparsity, low-resource languages)