Detecting annotation noise in automatically labelled data

Ines Rehbein & Josef Ruppenhofer

Leibniz ScienceCampus

ACL 2017

Motivation

- Many projects in the DH rely on automatically annotated data
- Quality of automatic annotations not always good enough

What we need:

- A cheap and efficient way to find errors in *automatically* labeled data

Related work

- Many studies on finding errors in manually annotated data

  (Eskin 2000; van Halteren 2000; Kveton and Oliva 2002; Dickinson and Meurers 2003; Boyd et al. 2008; Loftsson 2009; Ambati et al. 2011; Dickinson 2015; Snow et al. 2008; Bian et al. 2009; Hovy et al. 2013; . . .)

Related work

- Many studies on finding errors in manually annotated data

  (Eskin 2000; van Halteren 2000; Kveton and Oliva 2002; Dickinson and Meurers 2003; Boyd et al. 2008; Loftsson 2009; Ambati et al. 2011; Dickinson 2015; Snow et al. 2008; Bian et al. 2009; Hovy et al. 2013; . . .)

- Few studies on finding errors in automatically annotated data

  (Rocio et al. 2007; Loftsson 2009; Rehbein 2014)

Errors in automatic annotations are systematic and consistent

Related work

- Many studies on finding errors in manually annotated data

  (Eskin 2000; van Halteren 2000; Kveton and Oliva 2002; Dickinson and Meurers 2003; Boyd et al. 2008; Loftsson 2009; Ambati et al. 2011; Dickinson 2015; Snow et al. 2008; Bian et al. 2009; Hovy et al. 2013; . . .)

- Few studies on finding errors in automatically annotated data

  (Rocio et al. 2007; Loftsson 2009; Rehbein 2014)

- Our work builds on

  Hovy, Berg-Kirkpatrick, Vaswani and Hovy (2013):
  Learning Whom to Trust with MACE

MACE: Multi-Annotator Competence Estimation

Hovy et al. 2013

$word_j$	A_1	A_2	...	A_m
They	PRP	PRP	...	PRP
eat	VBP	VG	...	VBP
lots	NNS	RB	...	NN
of	IN	IN	...	IN
meat	NN	NNS	...	NN
	...	...	...	...

MACE: Multi-Annotator Competence Estimation

Hovy et al. 2013

$word_j$	A_1	A_2	...	A_m
They	PRP	PRP	...	PRP
eat	VBP	VG	...	VBP
lots	NNS	RB	...	NN
of	IN	IN	...	IN
meat	NN	NNS	...	NN
	...	...	...	...

```
1: procedure GenerateAnnot(A)
2:     for i = 1 ... I instances do
3:         T_i ~ Uniform
4:         for j = 1 ... J annotators do
5:             S_ij ~ Bernoulli(1 − θ_j)
6:             if S_ij = 0 then
7:                 A_ij = T_i
8:             else
9:                 A_ij ~ Multinomial(ξ_j)
10:            end if
11:        end for
12:    end for
13: end procedure
14: procedure UpdateParam(P(A; θ, ξ))
15:    return posterior entropies E
16: end procedure
```

MACE: Multi-Annotator Competence Estimation

Hovy et al. 2013

$word_j$	A_1	A_2	...	A_m
They	PRP	PRP	...	PRP
eat	VBP	VG	...	VBP
lots	NNS	RB	...	NN
of	IN	IN	...	IN
meat	NN	NNS	...	NN
	...	...	...	...

```
1:  procedure GenerateAnnot(A)
2:      for i = 1 ... I instances do
3:          T_i ~ Uniform
4:          for j = 1 ... J annotators do
5:              S_ij ~ Bernoulli(1 − θ_j)
6:              if S_ij = 0 then
7:                  A_ij = T_i
8:              else
9:                  A_ij ~ Multinomial(ξ_j)
10:             end if
11:         end for
12:     end for
13: end procedure
14: procedure UpdateParam(P(A; θ, ξ))
15:     return posterior entropies E
16: end procedure
```

MACE: Multi-Annotator Competence Estimation

Hovy et al. 2013

$word_j$	A_1	A_2	...	A_m
They	PRP	PRP	...	PRP
eat	VBP	VG	...	VBP
lots	NNS	RB	...	NN
of	IN	IN	...	IN
meat	NN	NNS	...	NN
	...	...	...	...

Parameters:
- θ trustworthyness of Annotator j
- ξ behaviour of j if spamming

```
1:  procedure GenerateAnnot(A)
2:      for i = 1 ... I instances do
3:          T_i ~ Uniform
4:          for j = 1 ... J annotators do
5:              S_ij ~ Bernoulli(1 − θ_j)
6:              if S_ij = 0 then
7:                  A_ij = T_i
8:              else
9:                  A_ij ~ Multinomial(ξ_j)
10:             end if
11:         end for
12:     end for
13: end procedure
14: procedure UpdateParam(P(A; θ, ξ))
15:     return posterior entropies E
16: end procedure
```

MACE: Multi-Annotator Competence Estimation

Hovy et al. 2013

$word_j$	A_1	A_2	...	A_m
They	PRP	PRP	...	PRP
eat	VBP	VG	...	VBP
lots	NNS	RB	...	NN
of	IN	IN	...	IN
meat	NN	NNS	...	NN
	...	...	...	...

Parameters:
- θ trustworthyness of Annotator j
- ξ behaviour of j if spamming

```
1:  procedure GenerateAnnot(A)
2:      for i = 1 ... I instances do
3:          T_i ~ Uniform
4:          for j = 1 ... J annotators do
5:              S_ij ~ Bernoulli(1 − θ_j)
6:              if S_ij = 0 then
7:                  A_ij = T_i
8:              else
9:                  A_ij ~ Multinomial(ξ_j)
10:             end if
11:         end for
12:     end for
13: end procedure
14: procedure UpdateParam(P(A; θ, ξ))
15:     return posterior entropies E
16: end procedure
```

$$P(A;\theta,\xi) = \sum_{T,S} \Big[\prod_{i=1}^{N} P(T_i) \cdot \prod_{j=1}^{M} P(S_{ij};\theta_j) \cdot P(A_{ij}|S_{ij},T_i;\xi_j) \Big]$$

MACE: Multi-Annotator Competence Estimation

Hovy et al. 2013

$word_j$	A_1	A_2	...	A_m
They	PRP	PRP	...	PRP
eat	VBP	VG	...	VBP
lots	NNS	RB	...	NN
of	IN	IN	...	IN
meat	NN	NNS	...	NN
	...	...	...	...

Parameters:
- θ trustworthyness of Annotator j
- ξ behaviour of j if spamming

Output:
- E confidence in model predictions

```
1:  procedure GenerateAnnot(A)
2:      for i = 1 ... I instances do
3:          T_i ~ Uniform
4:          for j = 1 ... J annotators do
5:              S_ij ~ Bernoulli(1 − θ_j)
6:              if S_ij = 0 then
7:                  A_ij = T_i
8:              else
9:                  A_ij ~ Multinomial(ξ_j)
10:             end if
11:         end for
12:     end for
13: end procedure
14: procedure UpdateParam(P(A; θ, ξ))
15:     return posterior entropies E
16: end procedure
```

$$P(A;\theta,\xi) = \sum_{T,S} \Big[\prod_{i=1}^{N} P(T_i) \cdot \prod_{j=1}^{M} P(S_{ij};\theta_j) \cdot P(A_{ij}|S_{ij},T_i;\xi_j) \Big]$$

MACE: Multi-Annotator Competence Estimation

Hovy et al. 2013

$word_j$	A_1	A_2	...	A_m
They	PRP	PRP	...	PRP
eat	VBP	VG	...	VBP
lots	NNS	RB	...	NN
of	IN	IN	...	IN
meat	NN	NNS	...	NN
	...	...	...	...

Parameters:
θ trustworthyness of Annotator j
ξ behaviour of j if spamming

Output:
E confidence in model predictions

Models:
EM, Bayesian Variational Inference

```
1:  procedure GenerateAnnot(A)
2:      for i = 1 ... I instances do
3:          T_i ~ Uniform
4:          for j = 1 ... J annotators do
5:              S_ij ~ Bernoulli(1 − θ_j)
6:              if S_ij = 0 then
7:                  A_ij = T_i
8:              else
9:                  A_ij ~ Multinomial(ξ_j)
10:             end if
11:         end for
12:     end for
13: end procedure
14: procedure UpdateParam(P(A; θ, ξ))
15:     return posterior entropies E
16: end procedure
```

Estimating the reliability of *automatic* annotations

- Task: POS tagging (7 POS taggers as "annotators")
- Data: English Penn Treebank *(in-domain)*

Tagger	**Acc.**
bilstm	97.00
hunpos	96.18
stanford	96.93
svmtool	95.86
treetagger	94.35
tweb	95.99
wapiti	94.52
majority vote	**97.28**

Estimating the reliability of *automatic* annotations

- Task: POS tagging (7 POS taggers as "annotators")
- Data: English Penn Treebank *(in-domain)*

Tagger	Acc.
bilstm	97.00
hunpos	96.18
stanford	96.93
svmtool	95.86
treetagger	94.35
tweb	95.99
wapiti	94.52
majority vote	**97.28**
MACE	**97.27**

$\Rightarrow$ MACE doesn't beat the majority vote baseline

Estimating the reliability of *automatic* annotations

- Task: POS tagging (7 POS taggers as "annotators")
- Data: English Penn Treebank *(in-domain)*

Tagger	Acc.
bilstm	97.00
hunpos	96.18
stanford	96.93
svmtool	95.86
treetagger	94.35
tweb	95.99
wapiti	94.52
majority vote	**97.28**
MACE	**97.27**

Guide Variational Inference model with human feedback from active learning

Combining Bayesian Inference with Active Learning

- **Selection strategy 1 (Baseline)**: *Query-by-Committee* (QBC)

  Use disagreements in the predictions to identify errors:

  1. compute entropy over predicted labels M:

$$H = -\sum_{m=1}^{M} P(y_i = m) \log P(y_i = m)$$

  2. select N instances with highest entropy $\Rightarrow$ *potential errors*
  3. replace predicted label with true label

- Evaluate accuracy for QBC after updating N instances ranked highest for entropy

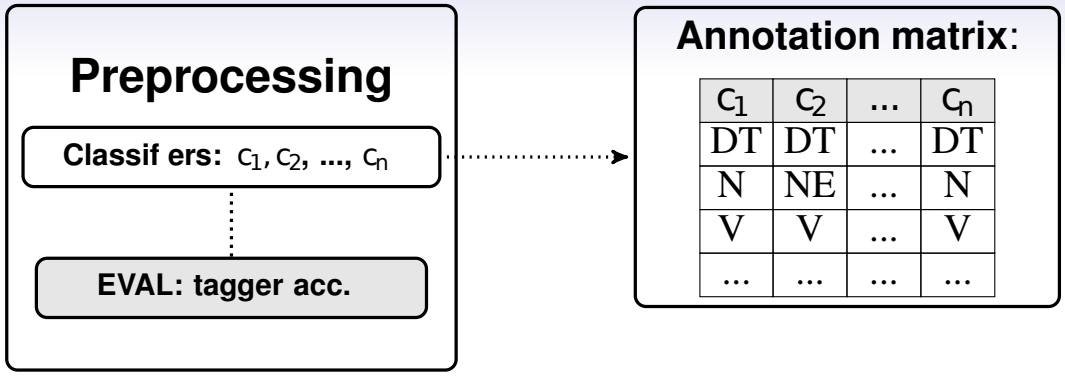
Combining Bayesian Inference with Active Learning

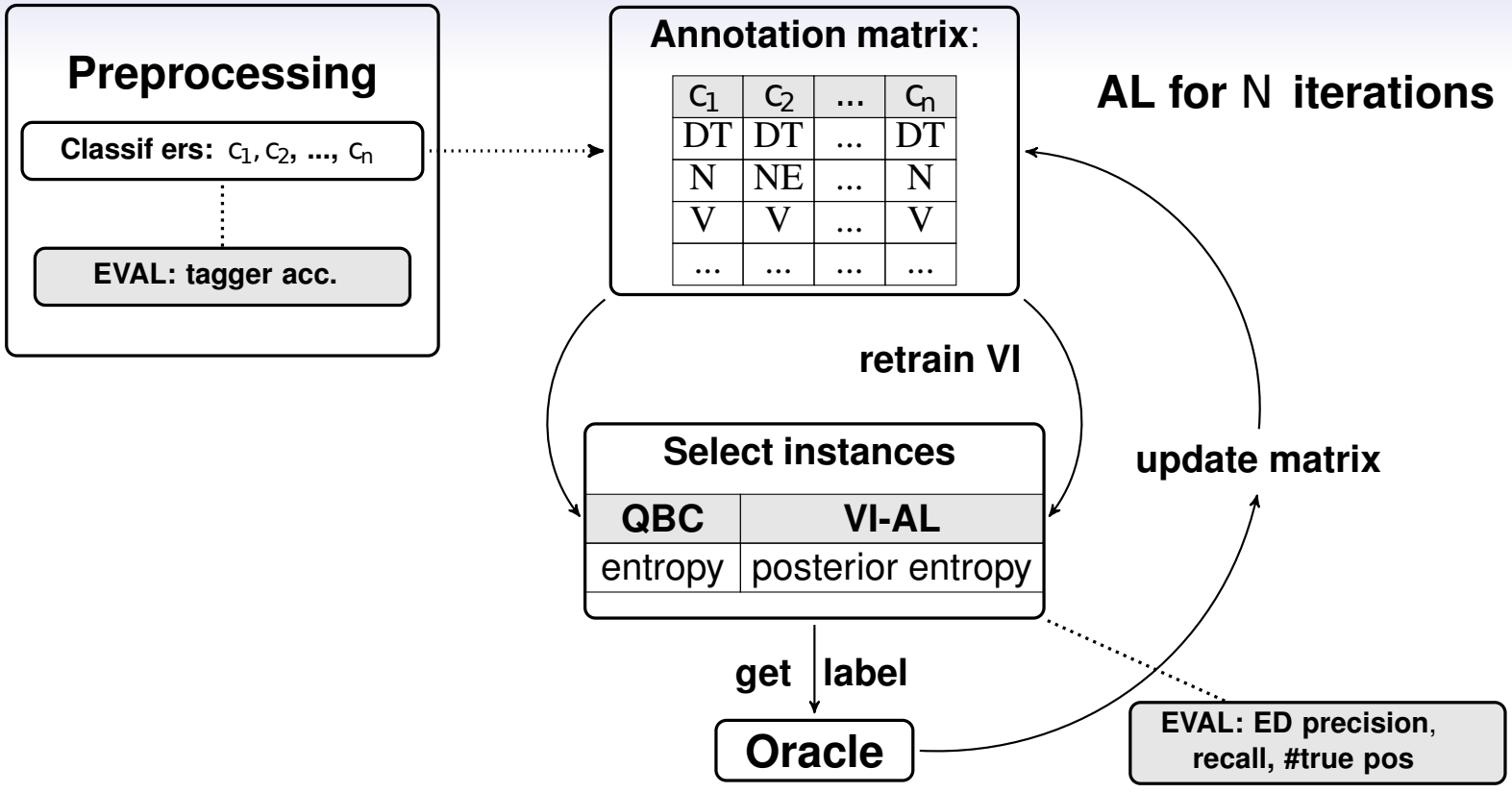
- **Selection strategy 2**: *Variational Inference & AL* (VI-AL)

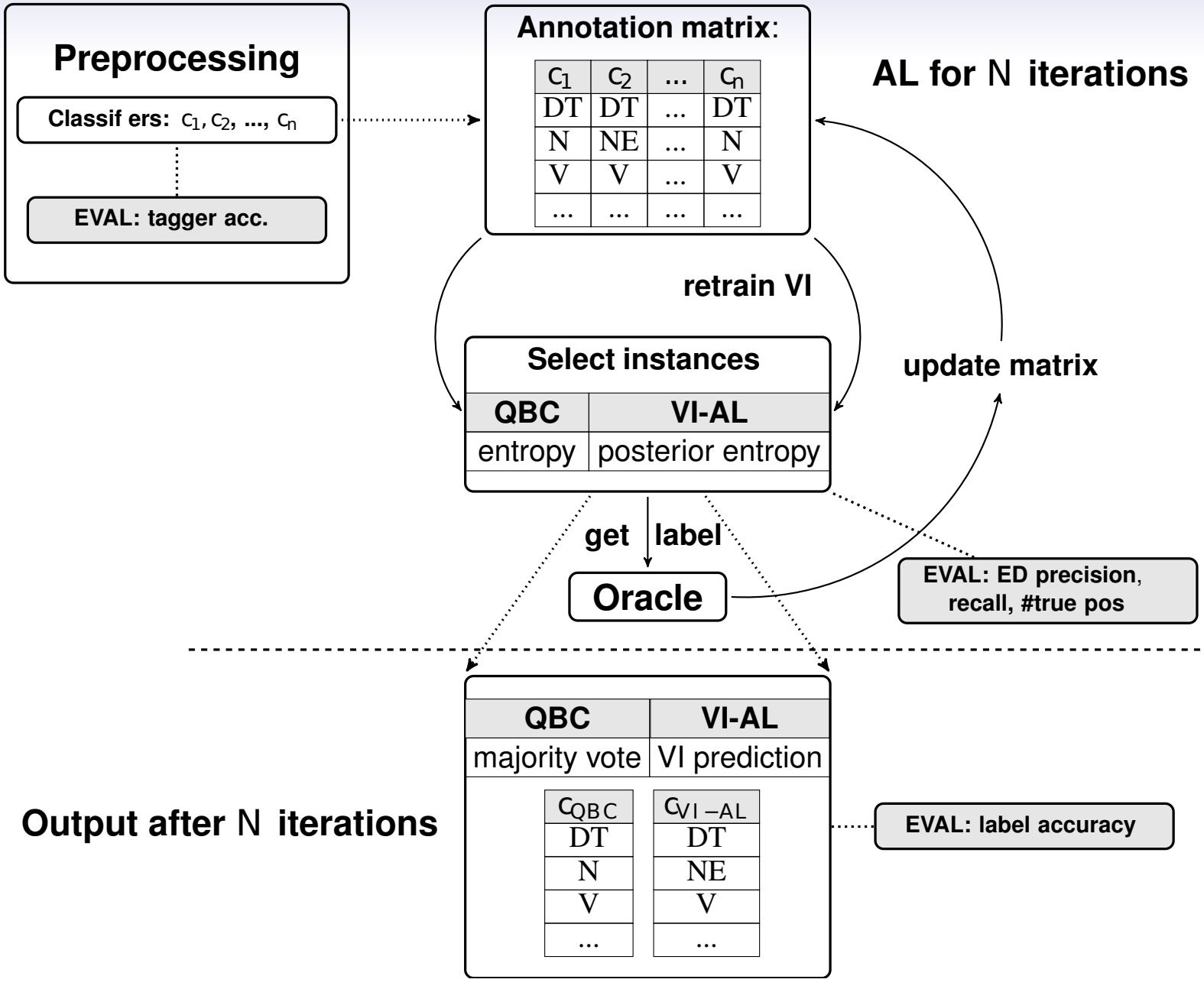
  Maximize the probability of the observed data, using the variational model:

  1. compute *posterior entropy* over predicted labels M
  2. select N instances with highest entropy $\Rightarrow$ *potential errors*
  3. replace randomly selected predicted label with true label
  4. compute new probabilities, based on the updated labels

- Evaluate accuracy of VI-AL after updating N instances ranked highest for entropy


Preprocessing
Classif ers: $c_1, c_2, ..., c_n$
EVAL: tagger acc.
Annotation matrix:
c_1 c_2 ... c_n
DT DT ... DT
N NE ... N
V V ... V
... ...


Preprocessing
Classif ers: $c_1, c_2, ..., c_n$
EVAL: tagger acc.
Annotation matrix:
c_1 c_2 ... c_n
DT DT ... DT
N NE ... N
V V ... V
... ...
AL for N iterations
retrain VI
Select instances
QBC
entropy
VI-AL
posterior entropy
update matrix
get label
Oracle
EVAL: ED precision, recall, #true pos


Preprocessing
Classif ers: $c_1, c_2, ..., c_n$
EVAL: tagger acc.
Annotation matrix:
c_1 c_2 ... c_n
DT DT ... DT
N NE ... N
V V ... V
... ...
AL for N iterations
retrain VI
update matrix
Select instances
QBC
VI-AL
entropy
posterior entropy
get label
Oracle
EVAL: ED precision, recall, #true pos
QBC
VI-AL
majority vote
VI prediction
c_{QBC}
c_{VI-AL}
DT DT
N NE
V V
... ...
EVAL: label accuracy
Output after N iterations

Experiments

- We test our approach
  - on 2 different tasks → *POS tagging, NER*
  - on 2 different languages → *English, German*
  - on in-domain data → *Penn Treebank*
  - on out-of-domain data → *web data, EuroParl*
  - in AL simulations → *Experiments 1-3*
  - and in a real-world setting → *Experiment 4*

Experiment 1 – In-domain POS tagging

- Large training set (WSJ), in-domain

Experiment 1 – In-domain POS tagging

- Large training set (WSJ), in-domain

MACE

10% of data

	QBC		VI-AL	
N	**label acc**	**ED prec**	**label acc**	**ED prec**
0	97.58	-	97.56	-
100	97.84	13.0	98.42	**41.0**
200	97.86	7.0	98.90	**33.0**
300	97.90	5.3	99.16	**26.3**
400	97.82	3.0	99.26	**21.0**
500	97.92	3.4	99.34	**17.6**

Table : **Label accuracies** on 5,000 tokens of WSJ text after *N* iterations, and **precision for error detection** (ED prec).

Experiment 1 – In-domain POS tagging

- Large training set (WSJ), in-domain

		QBC		VI-AL	
	N	**label acc**	**ED prec**	**label acc**	**ED prec**
MACE	0	97.58	-	97.56	-
	100	97.84	13.0	98.42	**41.0**
	200	97.86	7.0	98.90	**33.0**
	300	97.90	5.3	99.16	**26.3**
	400	97.82	3.0	99.26	**21.0**
10% of data	500	97.92	3.4	99.34	**17.6**

Table : **Label accuracies** on 5,000 tokens of WSJ text after N iterations, and **precision for error detection** (ED prec).

Experiment 1 – In-domain POS tagging

- Large training set (WSJ), in-domain

		QBC		VI-AL	
	N	**label acc**	**ED prec**	**label acc**	**ED prec**
MACE	0	97.58	-	97.56	-
	100	97.84	13.0	98.42	**41.0**
	200	97.86	7.0	98.90	**33.0**
	300	97.90	5.3	99.16	**26.3**
	400	97.82	3.0	99.26	**21.0**
10% of data	500	97.92	3.4	99.34	**17.6**

Table : **Label accuracies** on 5,000 tokens of WSJ text after N iterations, and **precision for error detection** (ED prec).

Experiment 1 – Errors we were not able to detect

freq.	gold	predicted	freq.	gold	predicted
18	**JJ**	**VBN**	1	NNP	JJ
2	IN	CC	1	NNP	NN
2	NN	NNP	1	PRP	PRP$
2	RBR	JJR	1	RP	IN
1	CD	DT	1	VBD	VBN
1	JJR	JJ	1	VBN	VBD
1	NN	JJ			

(1) companies were **closed**$_{JJ/VBN}$ yesterday
adjective or past participle?

Experiment 1 – Errors we were not able to detect

freq.	gold	predicted	freq.	gold	predicted
18	**JJ**	**VBN**	1	NNP	JJ
2	IN	CC	1	NNP	NN
2	NN	NNP	1	PRP	PRP$
2	RBR	JJR	1	RP	IN
1	CD	DT	1	VBD	VBN
1	JJR	JJ	1	VBN	VBD
1	NN	JJ			

(1) companies were **closed**$_{JJ/VBN}$ yesterday

adjective or past participle?

Manning (2011): Error categorisation

$\Rightarrow$ *underspecified/unclear*

Experiment 2 – Ouf-of-domain POS tagging

- No in-domain training data, taggers trained on WSJ
- Target domain: English Web Treebank (Bies et al., 2012)
- New tags in the target domain

Experiment 2 – Ouf-of-domain POS tagging

- No in-domain training data, taggers trained on WSJ
- Target domain: English Web Treebank (Bies et al., 2012)
- New tags in the target domain

Tagger accuracies for different web genres

	answer	**email**	**newsg.**	**review**	**weblog**
bilstm	85.5	84.2	86.5	86.9	89.6
hun	88.5	87.4	89.2	89.7	92.2
stan	**89.0**	88.1	**89.9**	90.7	93.0
svm	87.4	86.1	88.2	88.8	91.3
tree	86.8	85.6	87.1	88.7	87.4
tweb	88.2	87.1	88.5	89.3	92.0
wapiti	85.2	82.4	84.6	86.5	87.3
major.					
MACE					

Experiment 2 – Ouf-of-domain POS tagging

- No in-domain training data, taggers trained on WSJ
- Target domain: English Web Treebank (Bies et al., 2012)
- New tags in the target domain

Tagger accuracies for different web genres

	answer	**email**	**newsg.**	**review**	**weblog**
bilstm	85.5	84.2	86.5	86.9	89.6
hun	88.5	87.4	89.2	89.7	92.2
stan	**89.0**	88.1	**89.9**	90.7	93.0
svm	87.4	86.1	88.2	88.8	91.3
tree	86.8	85.6	87.1	88.7	87.4
tweb	88.2	87.1	88.5	89.3	92.0
wapiti	85.2	82.4	84.6	86.5	87.3
major.	87.4	**88.8**	89.1	90.9	93.8
MACE					

Experiment 2 – Ouf-of-domain POS tagging

- No in-domain training data, taggers trained on WSJ
- Target domain: English Web Treebank (Bies et al., 2012)
- New tags in the target domain

Tagger accuracies for different web genres

	answer	**email**	**newsg.**	**review**	**weblog**
bilstm	85.5	84.2	86.5	86.9	89.6
hun	88.5	87.4	89.2	89.7	92.2
stan	**89.0**	88.1	**89.9**	90.7	93.0
svm	87.4	86.1	88.2	88.8	91.3
tree	86.8	85.6	87.1	88.7	87.4
tweb	88.2	87.1	88.5	89.3	92.0
wapiti	85.2	82.4	84.6	86.5	87.3
major.	87.4	**88.8**	89.1	90.9	93.8
MACE	87.4	88.6	89.1	**91.0**	**93.9**

Experiment 2 – Ouf-of-domain POS tagging

	N	**answer**	**email**	**newsg**	**review**	**weblog**
MACE	0	87.4	88.6	89.1	91.0	93.9
	100	88.9	90.0	90.4	92.2	95.2
	200	90.3	91.1	91.3	93.4	96.2
	300	91.6	92.2	92.0	94.4	97.2
	400	92.9	93.3	92.8	95.4	97.5
10% of data	500	93.9	94.0	93.5	96.0	97.8
	600	94.8	94.9	93.9	96.5	97.9
	700	95.6	95.6	94.1	96.9	98.0
	800	96.2	95.9	94.7	97.3	98.4
	900	96.7	96.2	94.9	97.7	98.6
20% of data	1000	97.0	96.8	95.1	97.9	98.6

Table : Increase in POS **label accuracy** on the web genres (5,000 tokens) after N iterations of error correction with VI-AL.

Experiment 2 – Ouf-of-domain POS tagging

	N	**answer**	**email**	**newsg**	**review**	**weblog**
MACE	0	87.4	88.6	89.1	91.0	93.9
	100	88.9	90.0	90.4	92.2	95.2
	200	90.3	91.1	91.3	93.4	96.2
	300	91.6	92.2	92.0	94.4	97.2
	400	92.9	93.3	92.8	95.4	97.5
10% of data	500	93.9	94.0	93.5	96.0	97.8
	600	94.8	94.9	93.9	96.5	97.9
	700	95.6	95.6	94.1	96.9	98.0
	800	96.2	95.9	94.7	97.3	98.4
	900	96.7	96.2	94.9	97.7	98.6
20% of data	1000	97.0	96.8	95.1	97.9	98.6

Table : Increase in POS **label accuracy** on the web genres (5,000 tokens) after N iterations of error correction with VI-AL.

Experiment 2 – Ouf-of-domain POS tagging

	N	answer	email	newsg	review	weblog
MACE	0	87.4	88.6	89.1	91.0	93.9
	100	88.9	90.0	90.4	92.2	95.2
	200	90.3	91.1	91.3	93.4	96.2
	300	91.6	92.2	92.0	94.4	97.2
	400	92.9	93.3	92.8	95.4	97.5
10% of data	500	93.9	94.0	93.5	96.0	97.8
	600	94.8	94.9	93.9	96.5	97.9
	700	95.6	95.6	94.1	96.9	98.0
	800	96.2	95.9	94.7	97.3	98.4
	900	96.7	96.2	94.9	97.7	98.6
20% of data	1000	97.0	96.8	95.1	97.9	98.6

Table : Increase in POS **label accuracy** on the web genres (5,000 tokens) after N iterations of error correction with VI-AL.

Experiment 2 – Ouf-of-domain POS tagging

	N	answer	email	newsg	review	weblog
MACE	0	87.4	88.6	89.1	91.0	93.9
	100	88.9	90.0	90.4	92.2	95.2
	200	90.3	91.1	91.3	93.4	96.2
	300	91.6	92.2	92.0	94.4	97.2
	400	92.9	93.3	92.8	95.4	97.5
10% of data	500	93.9	94.0	93.5	96.0	97.8
	600	94.8	94.9	93.9	96.5	97.9
	700	95.6	95.6	94.1	96.9	98.0
	800	96.2	95.9	94.7	97.3	98.4
	900	96.7	96.2	94.9	97.7	98.6
20% of data	1000	97.0	96.8	95.1	97.9	98.6

Table : Increase in POS **label accuracy** on the web genres (5,000 tokens) after N iterations of error correction with VI-AL.

Experiment 3 – Out-of-domain NER on German

- New language (German)
- Out-of-domain test data (EuroParl)
- Small label set, skewed distribution

Experiment 3 – Out-of-domain NER on German

- New language (German)
- Out-of-domain test data (EuroParl)
- Small label set, skewed distribution

- We were able to identify >35% of all errors by querying less than 1% of the data

Experiment 4 – AL error detection in a realistic scenario

- Out-of-domain POS tagging with a real human annotator

		VI-AL *with human annotator*					
		answers			*weblog*		
	N	**# tp**	**ED prec**	**rec**	**# tp**	**ED prec**	**rec**
	100						
	200						
	300						
	400						
	500						
Simulation	500	282	56.4	48.8	196	39.2	64.5

Experiment 4 – AL error detection in a realistic scenario

- Out-of-domain POS tagging with a real human annotator

		VI-AL *with human annotator*					
		answers			*weblog*		
	N	**# tp**	**ED prec**	**rec**	**# tp**	**ED prec**	**rec**
	100	71	68.0	10.8	62	62.0	20.3
	200						
	300						
	400						
	500						
Simulation	500	282	56.4	48.8	196	39.2	64.5

Experiment 4 – AL error detection in a realistic scenario

- Out-of-domain POS tagging with a real human annotator

		VI-AL *with human annotator*					
		answers			*weblog*		
	N	**# tp**	**ED prec**	**rec**	**# tp**	**ED prec**	**rec**
	100	71	68.0	10.8	62	62.0	20.3
	200	103	63.5	20.2	112	56.0	36.7
	300						
	400						
	500						
Simulation	500	282	56.4	48.8	196	39.2	64.5

Experiment 4 – AL error detection in a realistic scenario

- Out-of-domain POS tagging with a real human annotator

		VI-AL *with human annotator*					
		answers			*weblog*		
	N	**# tp**	**ED prec**	**rec**	**# tp**	**ED prec**	**rec**
	100	71	68.0	10.8	62	62.0	20.3
	200	103	63.5	20.2	112	56.0	36.7
	300	177	58.0	27.6	156	52.0	51.1
	400						
	500						
Simulation	500	282	56.4	48.8	196	39.2	64.5

Experiment 4 – AL error detection in a realistic scenario

- Out-of-domain POS tagging with a real human annotator

		VI-AL *with human annotator*					
		answers			*weblog*		
	N	**# tp**	**ED prec**	**rec**	**# tp**	**ED prec**	**rec**
	100	71	68.0	10.8	62	62.0	20.3
	200	103	63.5	20.2	112	56.0	36.7
	300	177	58.0	27.6	156	52.0	51.1
	400	224	55.3	35.1	170	42.5	55.7
	500						
Simulation	500	282	56.4	48.8	196	39.2	64.5

Experiment 4 – AL error detection in a realistic scenario

- Out-of-domain POS tagging with a real human annotator

	VI-AL *with human annotator*					
	answers			*weblog*		
N	**# tp**	**ED prec**	**rec**	**# tp**	**ED prec**	**rec**
100	71	68.0	10.8	62	62.0	20.3
200	103	63.5	20.2	112	56.0	36.7
300	177	58.0	27.6	156	52.0	51.1
400	224	55.3	35.1	170	42.5	55.7
500	259	51.2	40.6	180	36.0	59.0
Simulation 500	282	56.4	48.8	196	39.2	64.5

Experiment 4 – AL error detection in a realistic scenario

- Out-of-domain POS tagging with a real human annotator

	VI-AL *with human annotator*					
	answers			*weblog*		
N	**# tp**	**ED prec**	**rec**	**# tp**	**ED prec**	**rec**
100	71	68.0	10.8	62	62.0	20.3
200	103	63.5	20.2	112	56.0	36.7
300	177	58.0	27.6	156	52.0	51.1
400	224	55.3	35.1	170	42.5	55.7
500	259	51.2	40.6	180	36.0	59.0
Simulation 500	282	56.4	48.8	196	39.2	64.5

- Label accuracies: answers: 87.4 → 92.5% (93.9%)
  weblog: 93.9 → 97.5% (97.8%)

Sum-up

- Method for error detection in *automatically* annotated data: Guide Variational Inference model with human feedback from active learning

$\Rightarrow$ Error detection with high precision *and* recall

Sum-up

- Method for error detection in *automatically* annotated data: Guide Variational Inference model with human feedback from active learning

$\Rightarrow$ Error detection with high precision *and* recall

- Advantages of our method
  - language-agnostic
  - no need to retrain classifiers (advantage for AL)
  - can deal with new, unknown target labels

Sum-up

- Method for error detection in *automatically* annotated data: Guide Variational Inference model with human feedback from active learning

$\Rightarrow$ Error detection with high precision *and* recall

- Advantages of our method
  - language-agnostic
  - no need to retrain classifiers (advantage for AL)
  - can deal with new, unknown target labels

Future work

- Extend model for non-sequential annotations (trees)

Thanks for listening!

Questions?

Thanks to Julius Steen for implementing the GUI

Code: https://github.com/julmaxi/MACE-AL

Experiment 2 – Ouf-of-domain POS tagging

	QBC			VI-AL		
	# tp	**ED prec**	**rec**	**# tp**	**ED prec**	**rec**
answer	282	56.4	44.8	323	**64.6**	**51.3**
email	264	**52.8**	**47.1**	261	52.2	46.6
newsg.	195	39.0	36.0	214	**42.8**	**39.6**
review	227	45.4	49.7	255	**51.0**	**55.8**
weblog	166	33.2	54.6	196	**39.2**	**64.5**

Table : No. of true positives (# tp), precision (ED prec) and recall for error detection on 5,000 tokens after 500 iterations on all web genres.

Experiment 2 – Ouf-of-domain POS tagging

	QBC			VI-AL		
N	**# tp**	**ED prec**	**rec**	**# tp**	**ED prec**	**rec**
100	85	**85.0**	**13.5**	75	75.0	11.9
200	148	**74.0**	**23.5**	146	73.0	23.2
300	198	66.0	31.4	212	**70.7**	**33.6**
400	239	59.7	37.9	278	**69.5**	**44.1**
500	282	56.4	44.8	323	**64.6**	**51.3**
600	313	52.2	49.7	374	**62.3**	**59.4**
700	331	47.3	52.5	412	**58.9**	**65.4**
800	355	44.4	56.3	441	**55.1**	**70.0**
900	365	40.6	57.9	465	**51.7**	**73.8**
1000	371	37.1	58.9	484	**48.4**	**76.8**

Table : No. of true positives (# tp), precision (ED prec) and recall for error detection on 5,000 tokens from the *answers* set after *N* iterations.

Experiment 3 – Out-of-domain NER on German

- Small label set, skewed distribution
- New language (German), out-of-domain test data

	QBC			VI-AL		
N	# tp	ED prec	rec	# tp	ED prec	rec
100	54	54.0	3.1	76	**76.0**	**4.7**
200	113	56.5	6.4	155	**77.5**	**9.6**
300	162	54.0	9.2	217	**72.3**	**13.4**
400	209	52.2	11.9	297	**74.2**	**18.2**
500	274	54.8	15.6	352	**70.4**	**22.3**
600	341	56.8	19.4	409	**68.2**	**25.5**
700	406	58.0	23.1	452	**64.6**	**27.8**
800	480	60.0	27.3	483	**60.4**	**29.8**
900	551	**61.2**	31.4	512	56.9	**31.9**
1000	617	**61.7**	35.1	585	58.5	**35.8**
1000	remaining errors:1,139			remaining errors:1,043		

Table : Error detection results on the GermEval 2014 NER testset after *N* iterations (true positives, ED precision and recall).

Experiment 4 – AL error detection in a realistic scenario

- Out-of-domain POS tagging with a real human annotator

	VI-AL *with human annotator*					
	answers			*weblog*		
N	**# tp**	**ED prec**	**rec**	**# tp**	**ED prec**	**rec**
100	71	68.0	10.8	62	62.0	20.3
200	103	63.5	20.2	112	56.0	36.7
300	177	58.0	27.6	156	52.0	51.1
400	224	55.3	35.1	170	42.5	55.7
500	259	51.2	40.6	180	36.0	59.0

- High error detection precision and recall also for real human annotator
- Label acc. answers: 92.5% (93.9%), weblog: 97.5% (97.8%)

Time requirements for correction:

- 500 instances from answers, annotator 1: 135 minutes
- 500 instances from weblog, annotator 2: 157 minutes