Training a Neural Network in a Low-Resource Setting on Automatically Annotated Noisy Data

Michael A. Hedderich and Dietrich Klakow

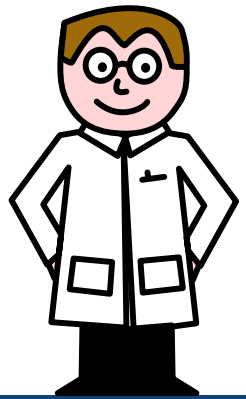
Workshop on Deep Learning Approaches for Low-Resource NLP

19.07.2018

SIC Saarland Informatics Campus

Aim

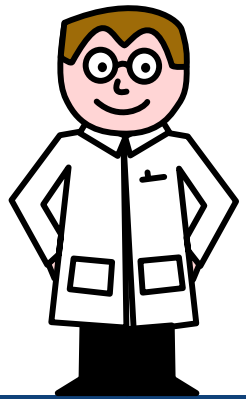
Expensive, clean labels



Aim

Expensive, clean labels

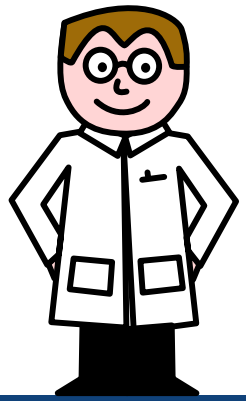
Raw, unlabeled data

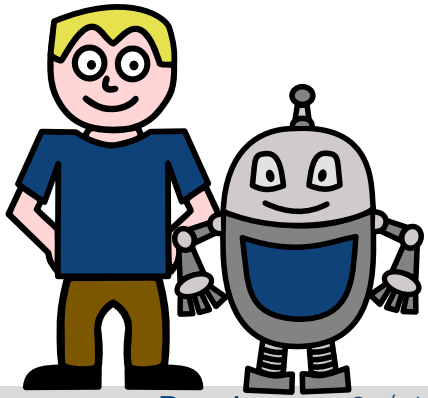


Aim

Expensive, clean labels

Raw, unlabeled data
+ cheap, noisy labels



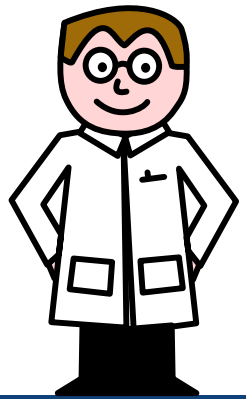


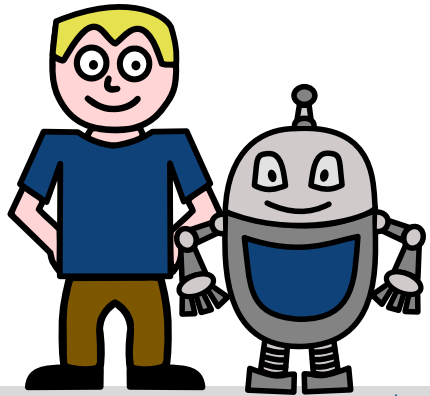
Aim

Expensive, clean labels

Raw, unlabeled data
+ cheap, noisy labels

Leveraging additional,
cheap, noisy data.



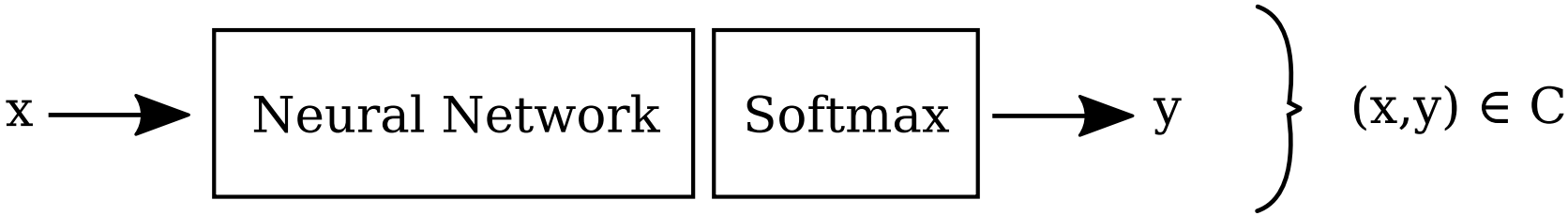


Overview

- Label Noise Model
- Training Procedure
- Automatic Annotation of Named Entities
- Evaluation & Analysis

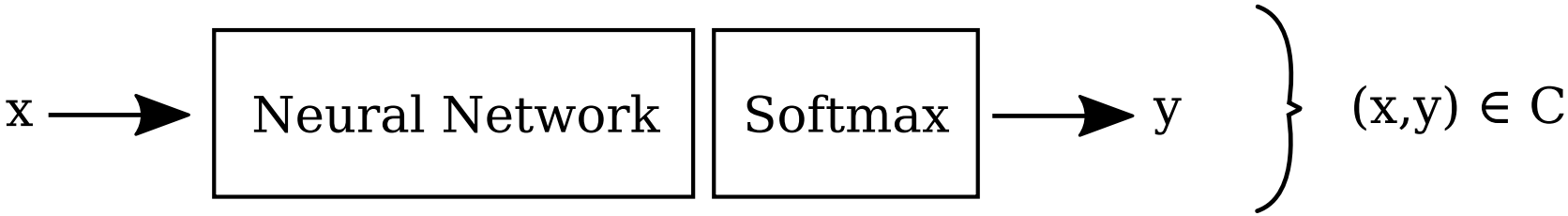
Setting

- Small, clean dataset $(x, y) \in C$
- Multi-class classification: $p(y = i|x; w)$



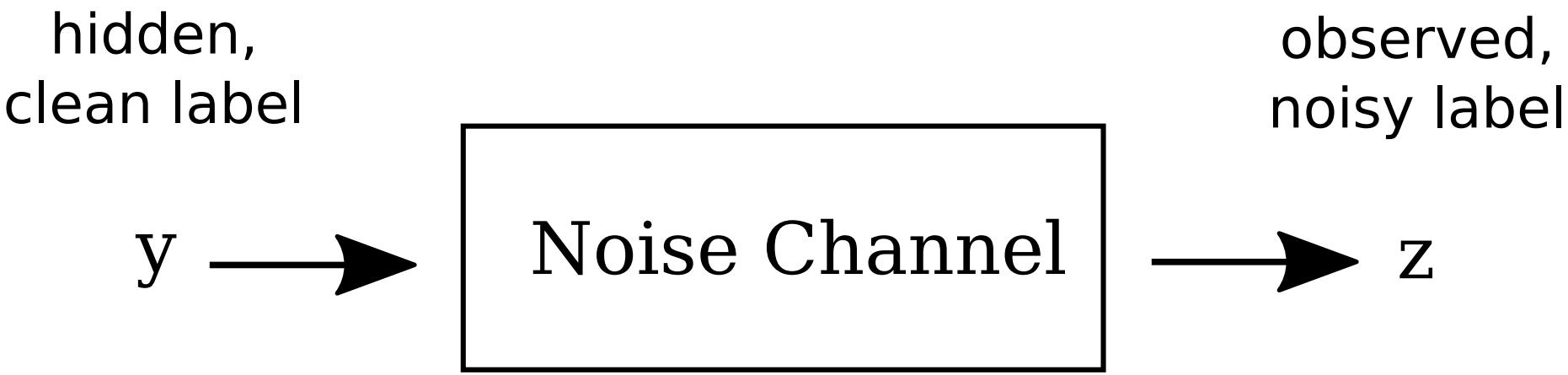
Setting

- Small, clean dataset $(x, y) \in C$
- Multi-class classification: $p(y = i|x; w)$



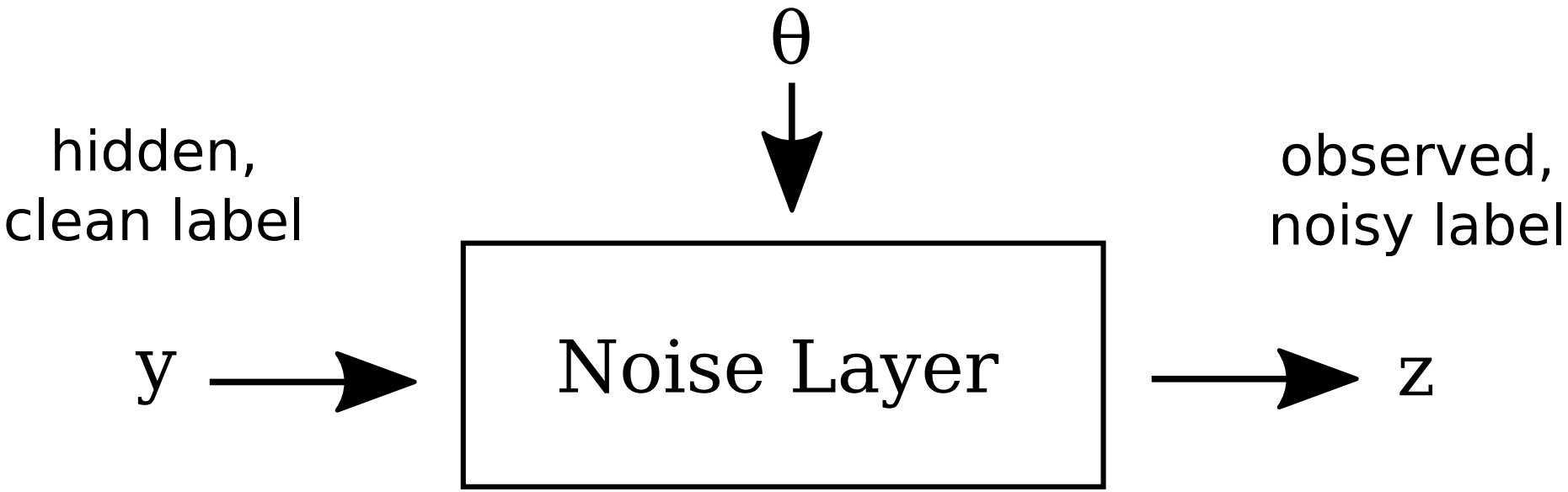
- Large, noisy dataset $(x, z) \in N$

Noise Channel



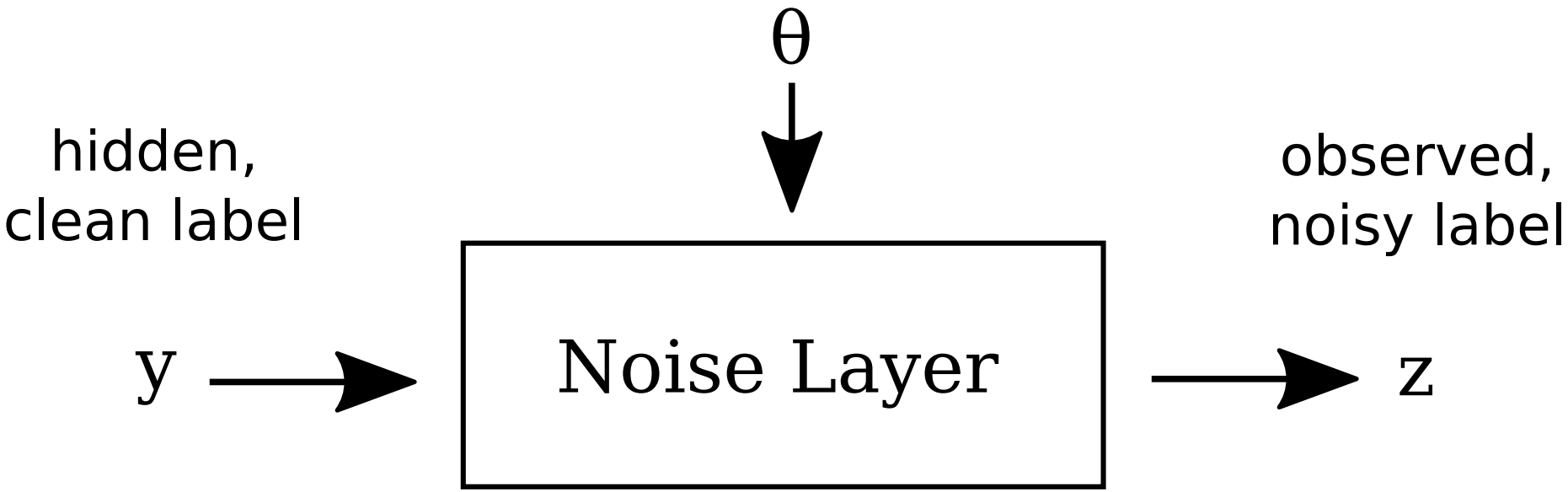
Goldberger and Ben-Reuven: *Training Deep Neural-Networks Using a Noise Adaptation Layer*. Int. Conference on Learning Representations (2017).

Noise Channel



Goldberger and Ben-Reuven: *Training Deep Neural-Networks Using a Noise Adaptation Layer*. Int. Conference on Learning Representations (2017).

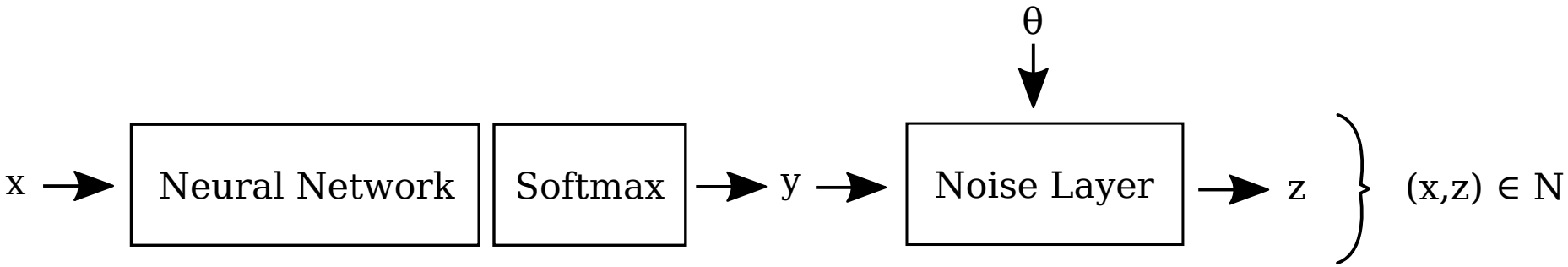
Noise Channel



$$\theta(i,j) = p(z = j|y = i) = \frac{\exp(b_{ij})}{\sum_{l=1}^{k} \exp(b_{il})}$$

Goldberger and Ben-Reuven: *Training Deep Neural-Networks Using a Noise Adaptation Layer*. Int. Conference on Learning Representations (2017).

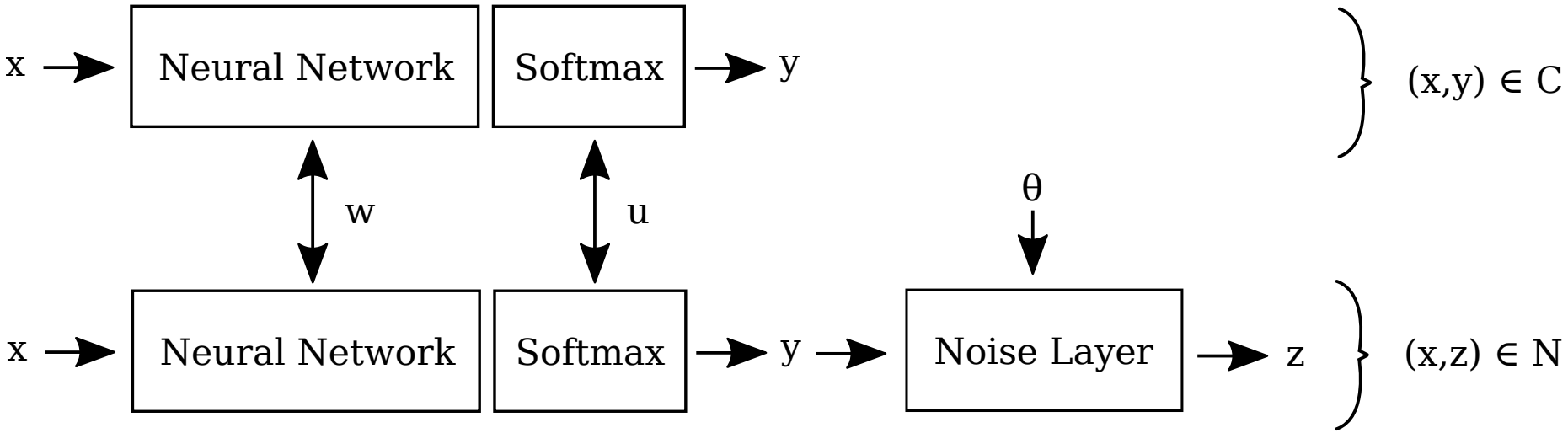
Model Structure



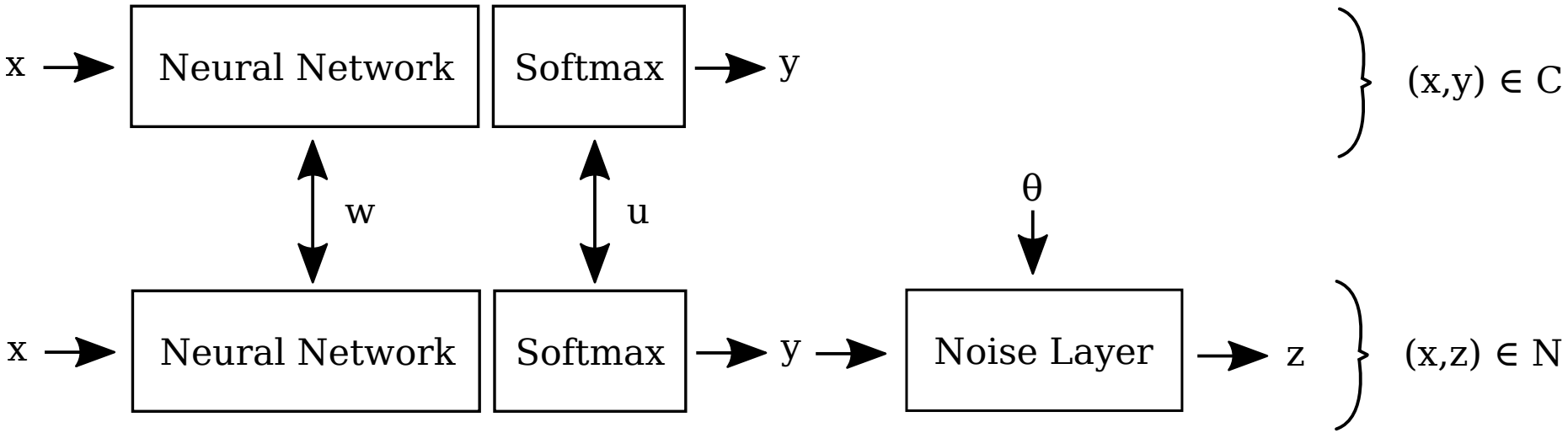
$$p(z = j|x; w; \theta) = \sum_{i=1}^{k} p(z = j|y = i; \theta) p(y = i|x; w)$$

Goldberger and Ben-Reuven: *Training Deep Neural-Networks Using a Noise Adaptation Layer*. Int. Conference on Learning Representations (2017).

Model Structure

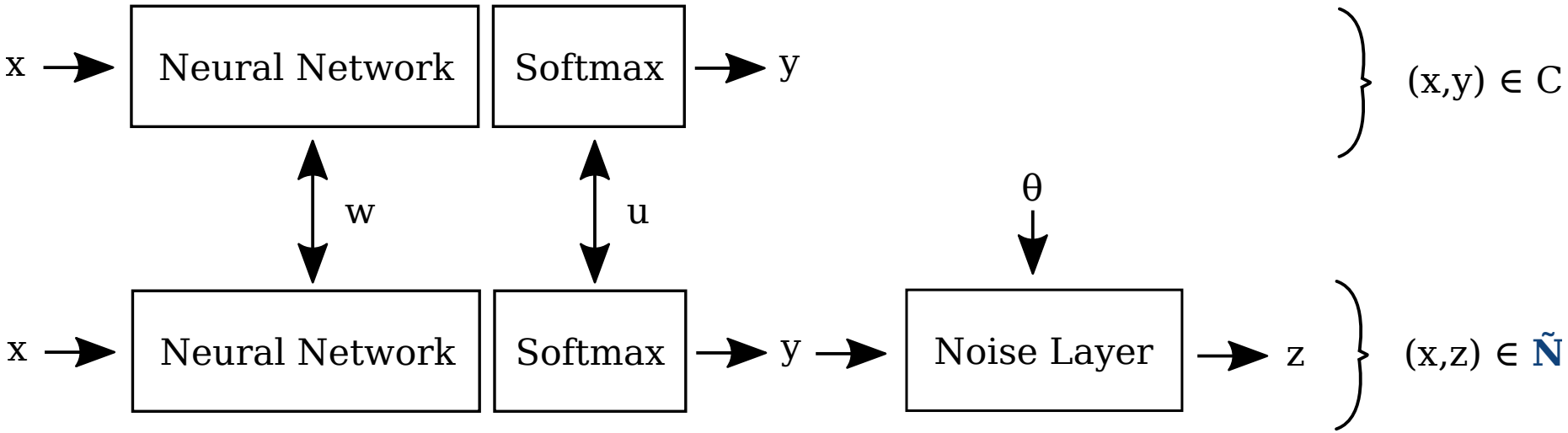


Model Structure



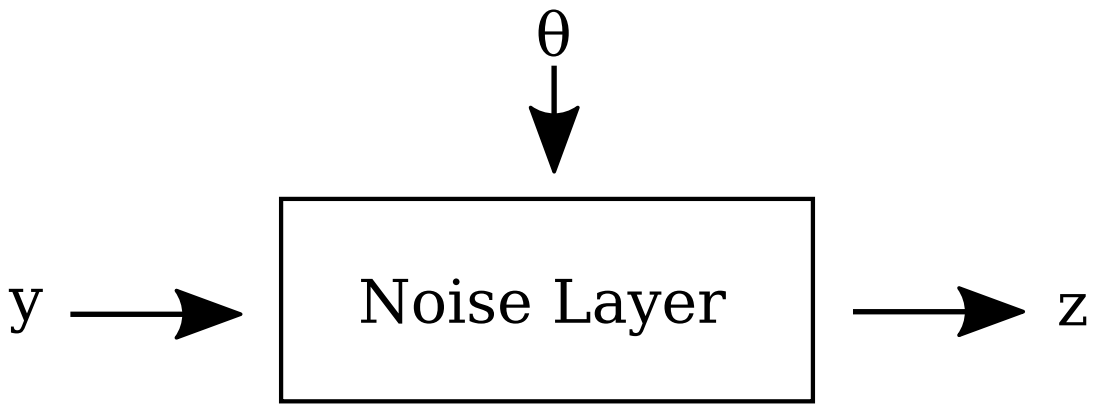
- Models trained alternatingly, each for one epoch
- Shared weights

Model Structure



- Controlling amount of noisy data compared to clean
- $\tilde{N}$ new random subset of N in each epoch

Initialization of θ



- Produce noisy labels z for $(x, y) \in C$
- Initialize θ: $b_{ij} = \log\left(\frac{\sum_{t=1}^{|C|} 1_{\{y_t=i\}} 1_{\{z_t=j\}}}{\sum_{t=1}^{|C|} 1_{\{y_t=i\}}}\right)$

Experiments

- Named Entity Recognition task
- Small subset of English CoNLL03 NER dataset as C
- All CoNLL03 data as raw data

Tjong Kim Sang and De Meulder: *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition*. Conference on Natural Language Learning (2003).

Experiments

- Named Entity Recognition task
- Small subset of English CoNLL03 NER dataset as C
- All CoNLL03 data as raw data
- Many synthetic noise models have strong assumptions (e.g. uniform or permutation)
- Automatic labeling as source of noisy data N

Tjong Kim Sang and De Meulder: *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition*. Conference on Natural Language Learning (2003).

Automatic Annotation of Named Entities

LOC	**PER**	**ORG**
Washington	Henry	United Nations
Melbourne	Elisabeth	ACL
Saarbrücken	Victoria	WHO
...	...	...

- E.g. from gazetteers, Wikipedia, census data

Dembowski et al.: *Language Independent Named Entity Recognition using Distant Supervision*. Language and Technology Conference (2017).

Automatic Annotation of Named Entities

LOC	**PER**	**ORG**
Washington	Henry	United Nations
Melbourne	Elisabeth	ACL
Saarbrücken	Victoria	WHO
...	...	...

Melbourne, state capital of Victoria

Dembowski et al.: *Language Independent Named Entity Recognition using Distant Supervision*. Language and Technology Conference (2017).

Automatic Annotation of Named Entities

LOC	**PER**	**ORG**
Washington	Henry	United Nations
Melbourne	Elisabeth	ACL
Saarbrücken	Victoria	WHO
...	...	...

Melbourne, state capital of Victoria

LOC

Dembowski et al.: *Language Independent Named Entity Recognition using Distant Supervision*. Language and Technology Conference (2017).

Automatic Annotation of Named Entities

LOC	**PER**	**ORG**
Washington	Henry	United Nations
Melbourne	Elisabeth	ACL
Saarbrücken	Victoria	WHO
...	...	...

Melbourne, state capital of Victoria

LOC 0 0 0

Dembowski et al.: *Language Independent Named Entity Recognition using Distant Supervision*. Language and Technology Conference (2017).

Automatic Annotation of Named Entities

LOC	**PER**	**ORG**
Washington	Henry	United Nations
Melbourne	Elisabeth	ACL
Saarbrücken	Victoria	WHO
...	...	...

Melbourne, state capital of Victoria

LOC 0 0 0 PER

Dembowski et al.: *Language Independent Named Entity Recognition using Distant Supervision*. Language and Technology Conference (2017).

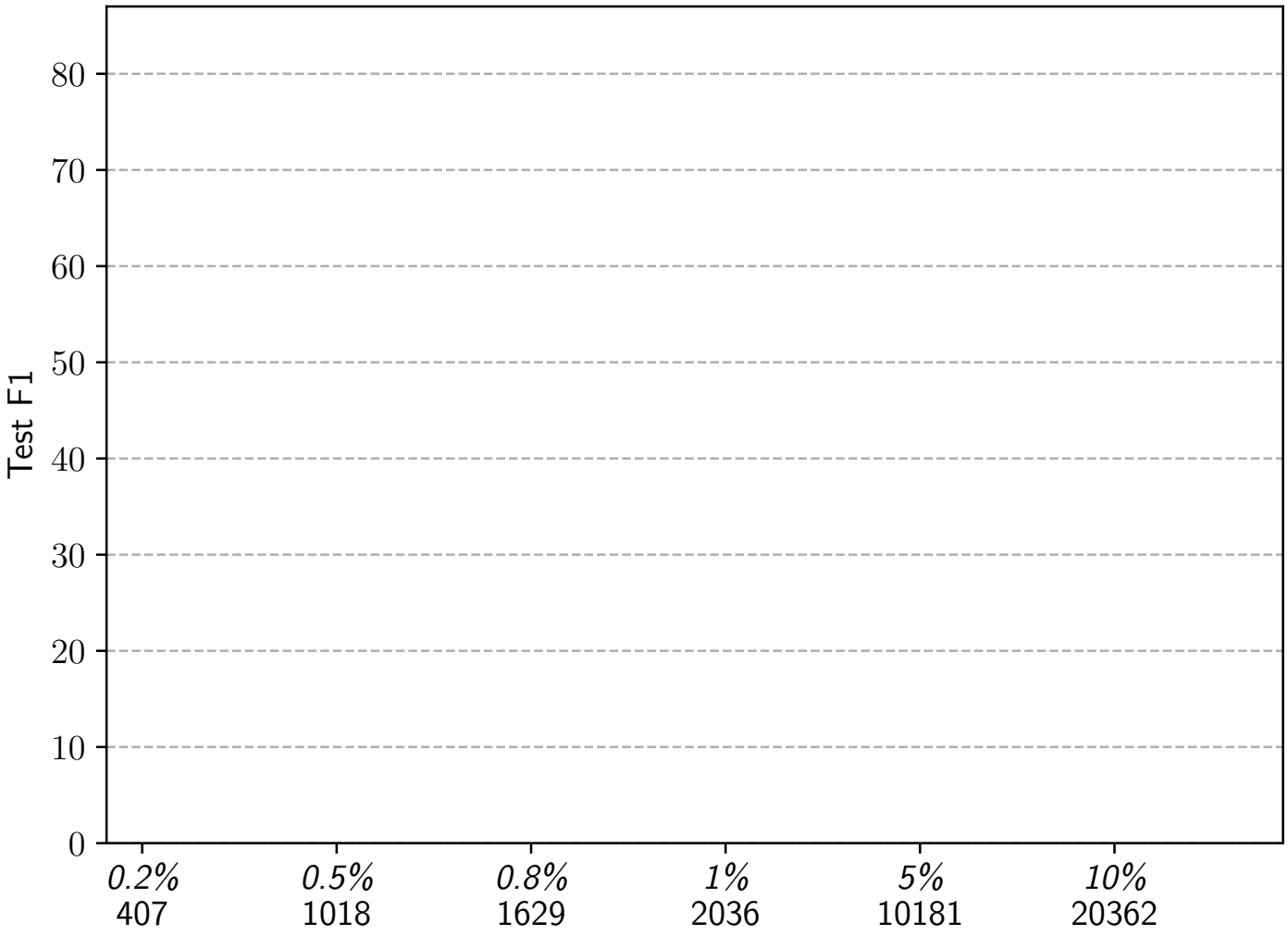
Automatic Annotation of Named Entities

LOC	**PER**	**ORG**
Washington	Henry	United Nations
Melbourne	Elisabeth	ACL
Saarbrücken	Victoria	WHO
...	...	...

- Precision: 53% | Recall: 27% | F1: 36%
- Quickly annotate large corpora

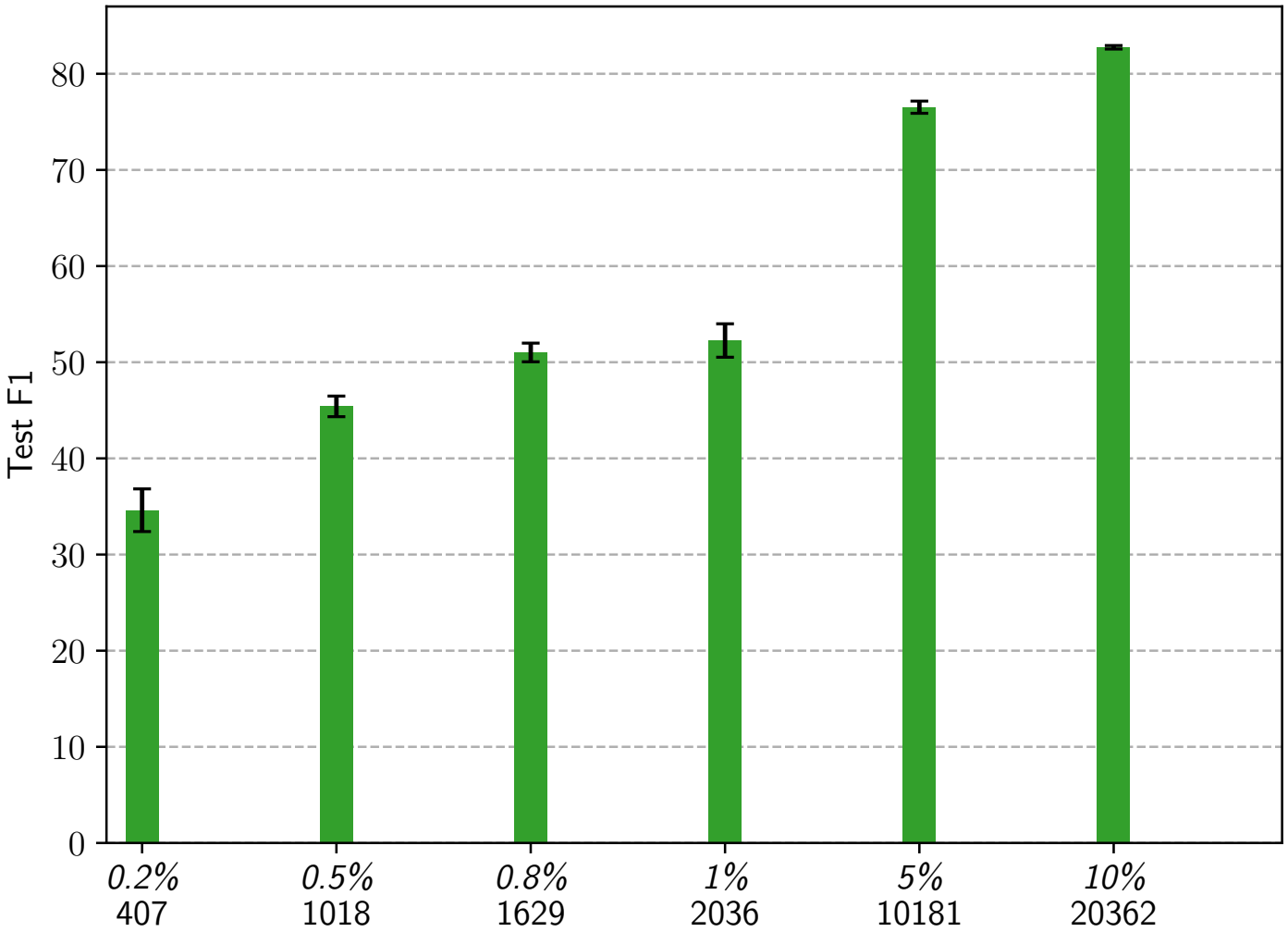
Dembowski et al.: *Language Independent Named Entity Recognition using Distant Supervision*. Language and Technology Conference (2017).

Performance

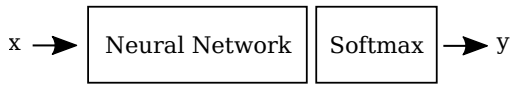


Sizes of clean and subsampled, noisy data $|C| = |\tilde{N}|$
(as *percentage of full corpus* and as absolute size).

Performance

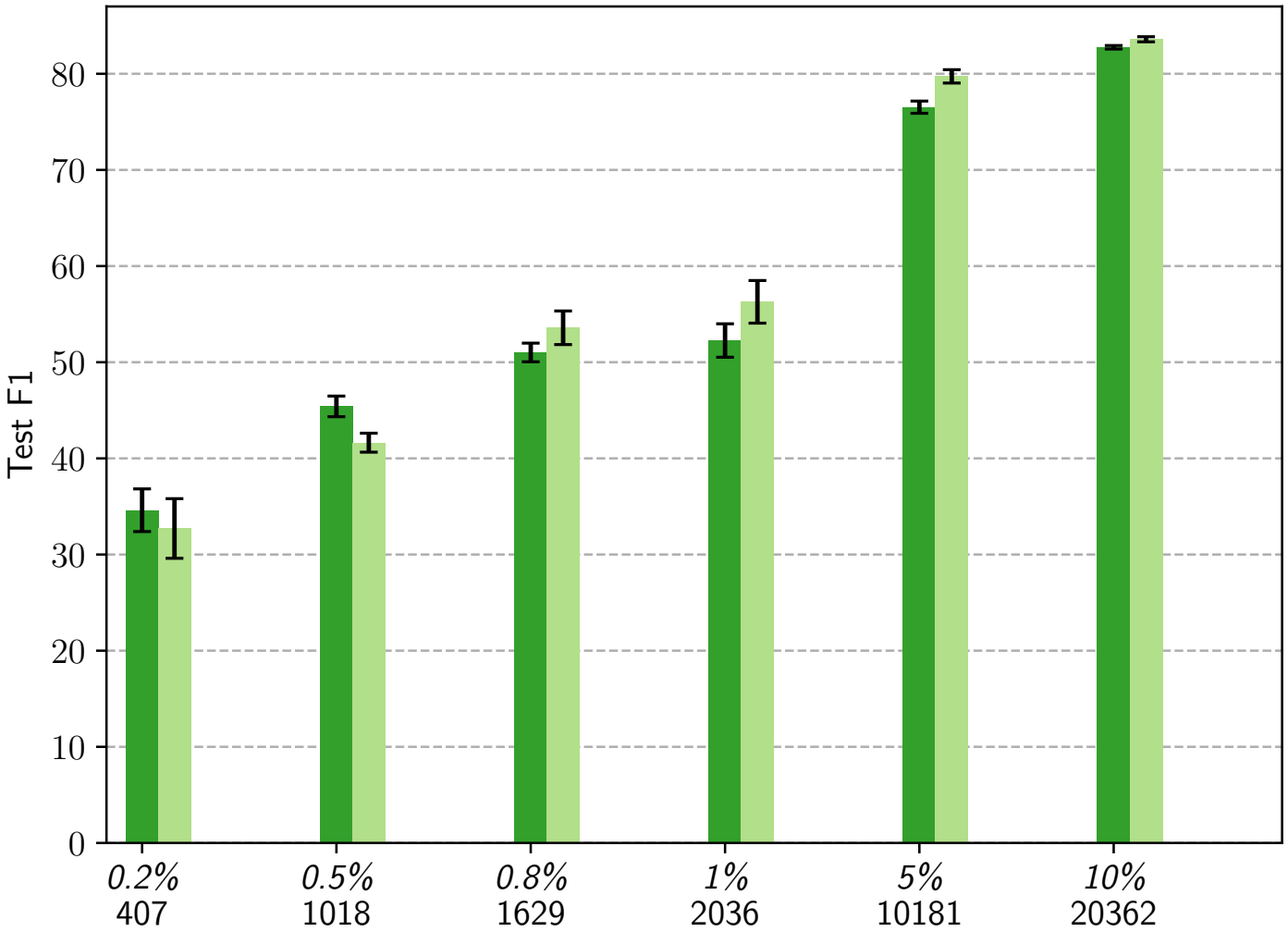


base-model



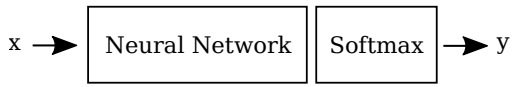
Sizes of clean and subsampled, noisy data $|C| = |\tilde{N}|$
(as *percentage of full corpus* and as absolute size).

Performance



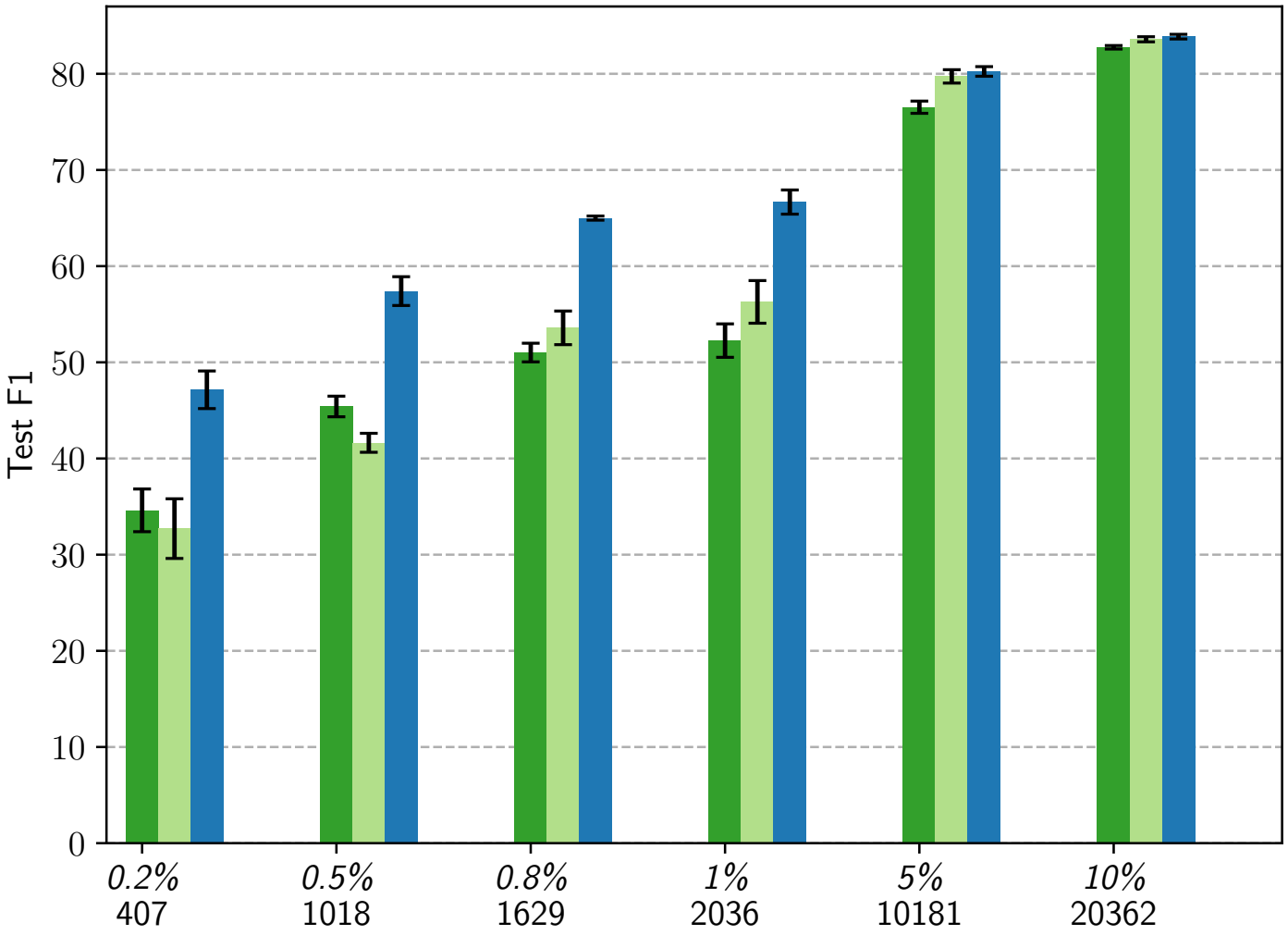
base-model

base-model + noisy data



Sizes of clean and subsampled, noisy data $|C| = |\tilde{N}|$
(as *percentage of full corpus* and as absolute size).

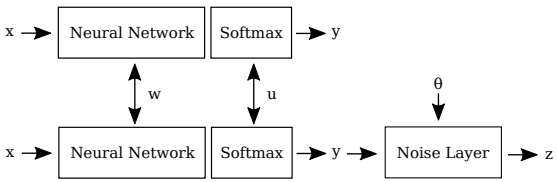
Performance



base-model

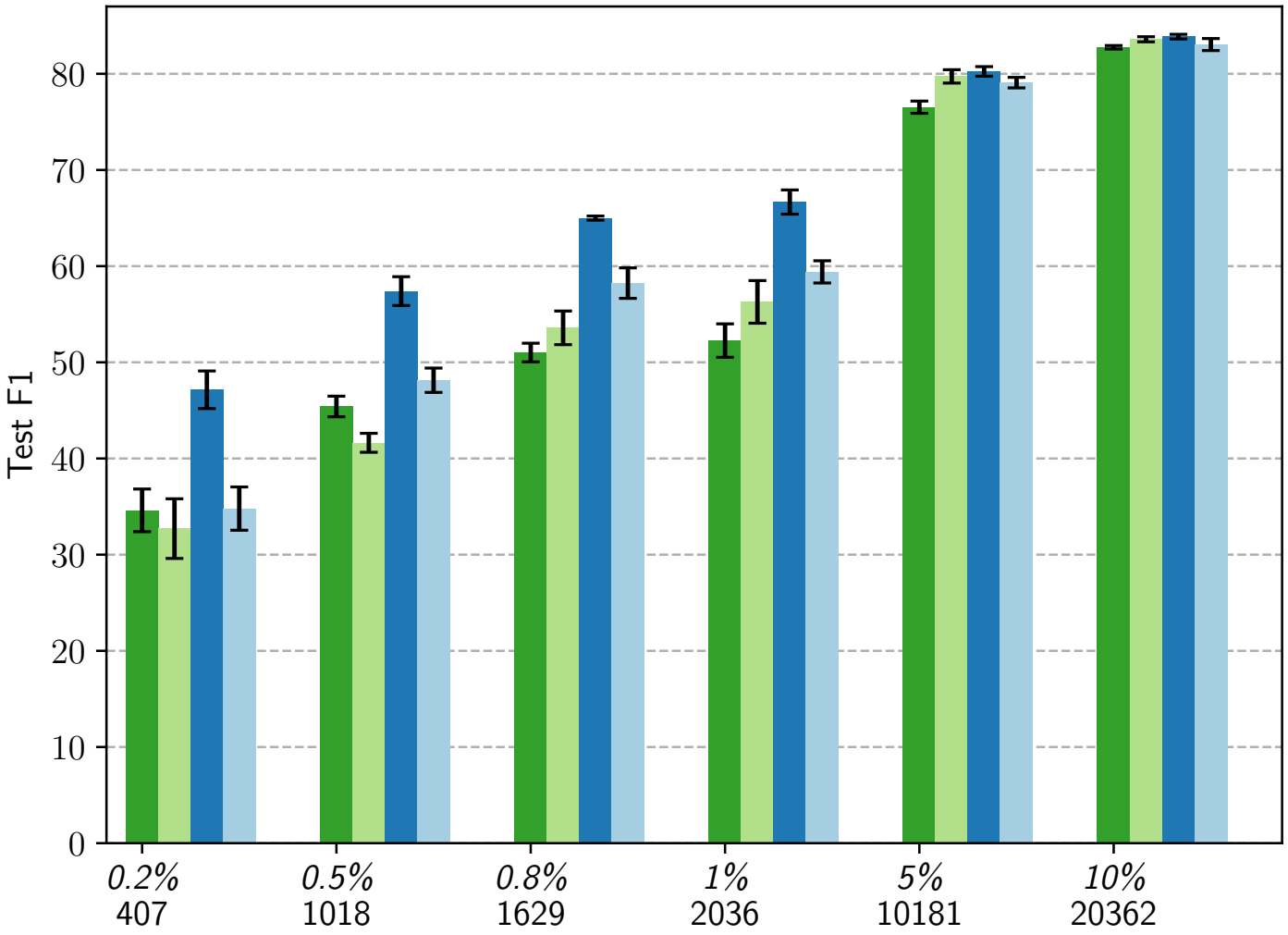
base-model + noisy data

noise-model



Sizes of clean and subsampled, noisy data $|C| = |\tilde{N}|$
(as *percentage of full corpus* and as absolute size).

Performance

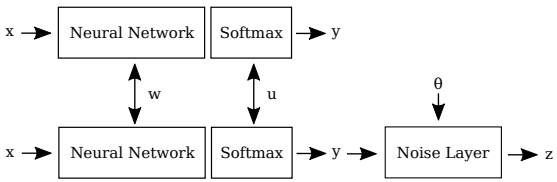


base-model

base-model + noisy data

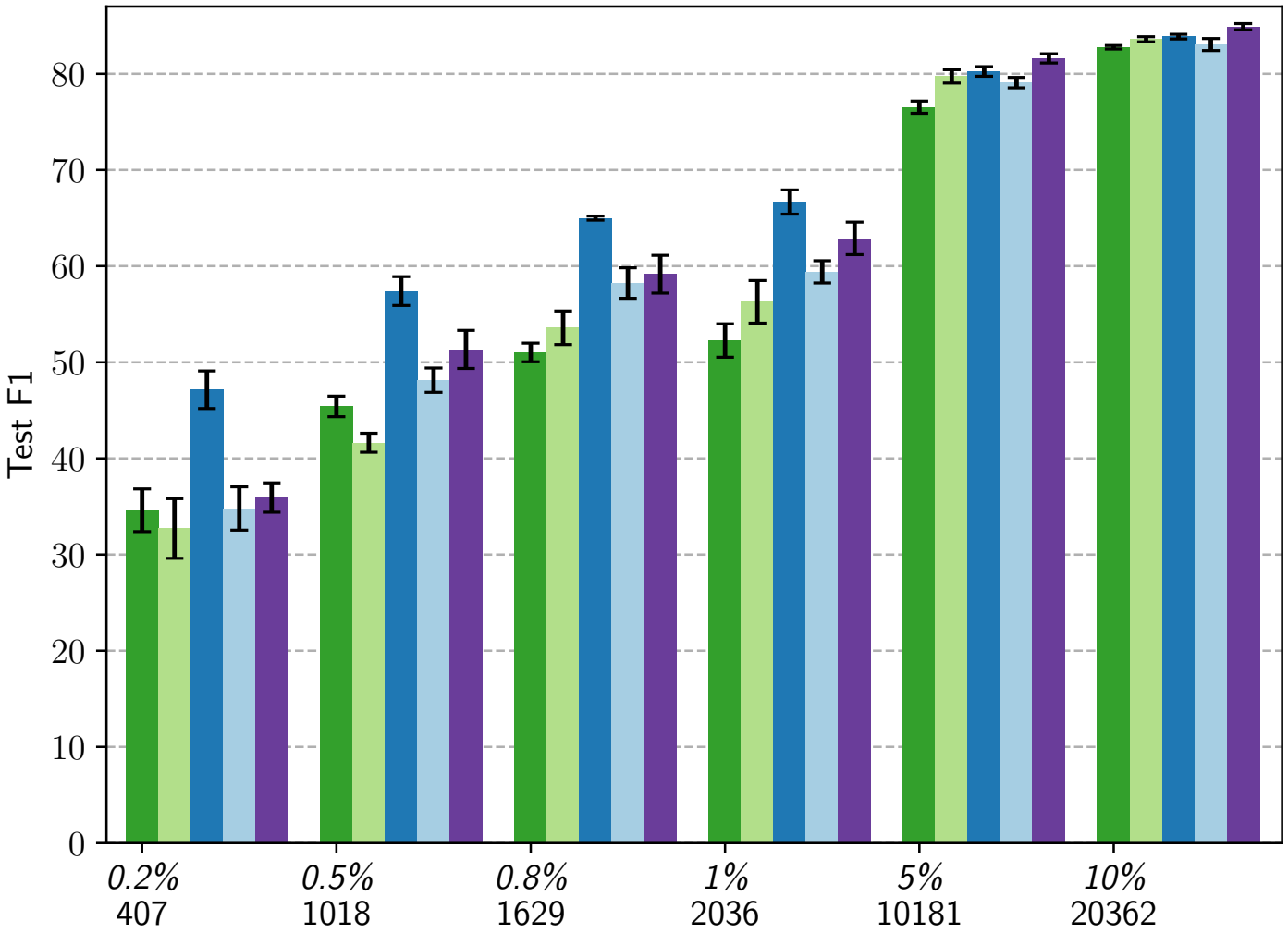
noise-model

noise-model (id-init)



Sizes of clean and subsampled, noisy data $|C| = |\tilde{N}|$
(as *percentage of full corpus* and as absolute size).

Performance

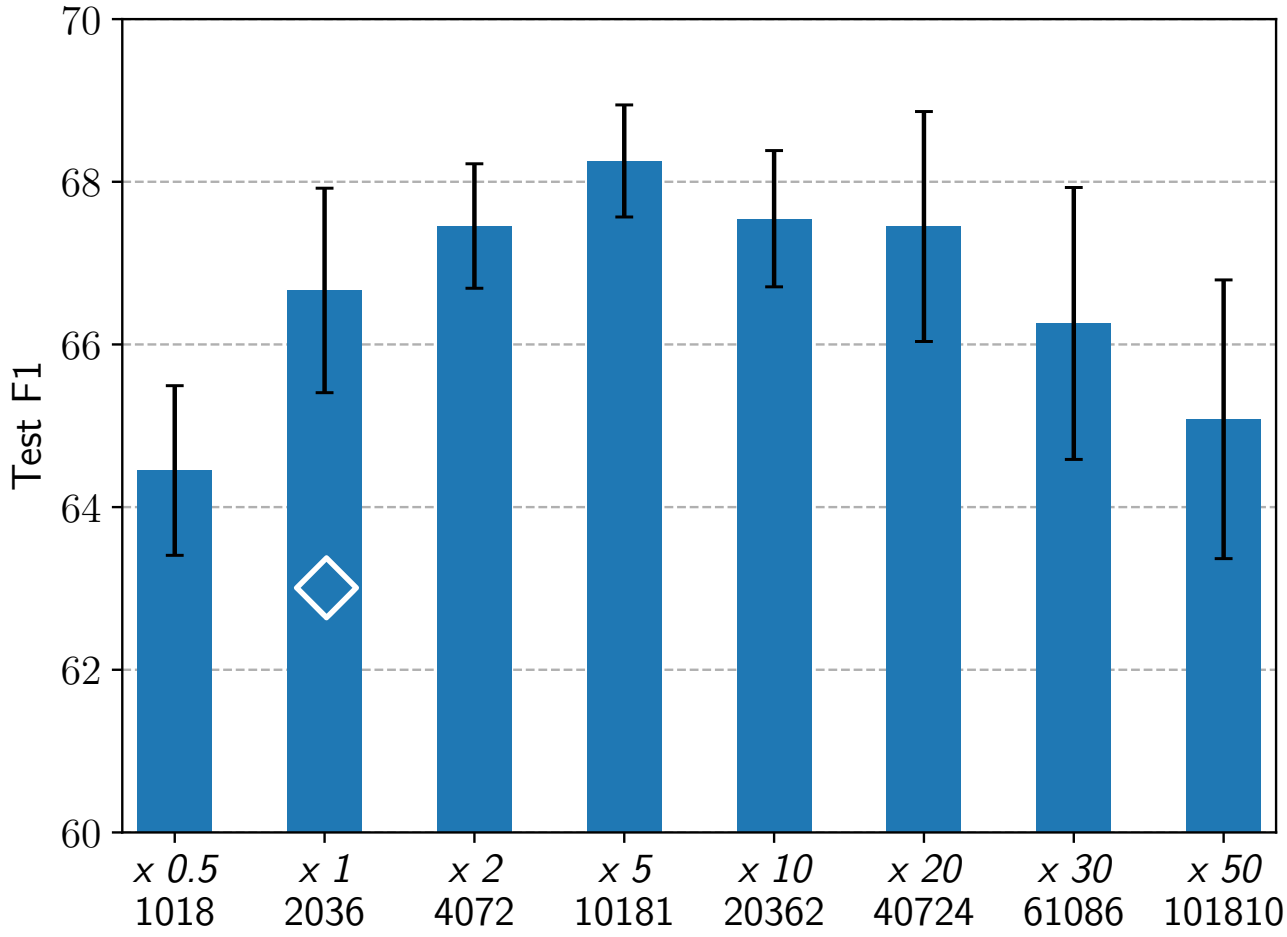


Sizes of clean and subsampled, noisy data $|C| = |\tilde{N}|$ (as *percentage of full corpus* and as absolute size).

- base-model
- base-model + noisy data
- **noise-model**
- noise-model (id-init)
- noise-cleaning-model

Veit et al.: *Learning from Noisy Large-Scale Datasets with Minimal Supervision*. CVPR (2017).

Effect of Noisy Data

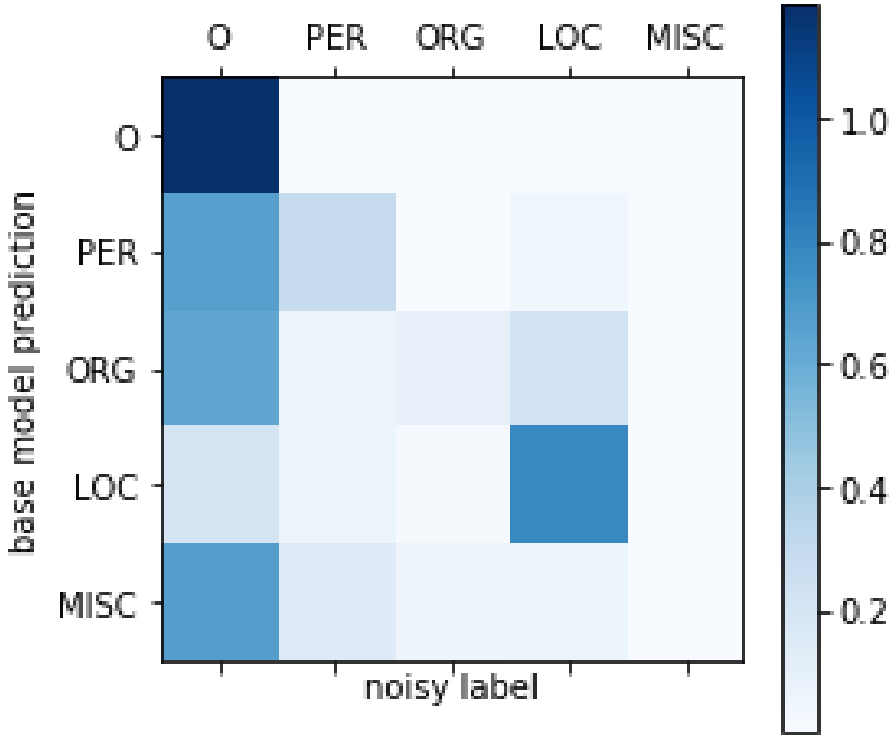


$|\tilde{N}|$ (subsampled noisy data) as *factor of* $|C|$ *(clean data)*
and as absolute number of instances

Learned Weights

Class	Recall
PER	26%
ORG	10%
LOC	65%
MISC	0%

Automatic annotation of training data



Learned noise transition weights θ

Summary

- Noise channel and layer
- Model architecture for clean and noisy data
- Automatic annotation of named entities
- Evaluation: Performance, effect of noisy data, learned weights
- Future work
  - Improve automatic annotation
  - More robustness against noise
  - Other tasks & noisy data

Summary

- Noise channel and layer
- Model architecture for clean and noisy data
- Automatic annotation of named entities
- Evaluation: Performance, effect of noisy data, learned weights
- Future work
  - Improve automatic annotation
  - More robustness against noise
  - Other tasks & noisy data