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

Michael A. Hedderich Dietrich Klakow
{mhedderich,dietrich.klakow}@lsv.uni-saarland.de
Spoken Language Systems (LSV), Saarland Informatics Campus,
Saarland University, Saarbrücken, Germany

Aim

In low-resource settings, labeled datasets are usually small. Raw data can be labeled cheaply with crowd-sourcing or automatic techniques, but these labels tend to contain many errors. This makes training difficult. We present a method to successfully leverage this additional, cheap, noisy data.

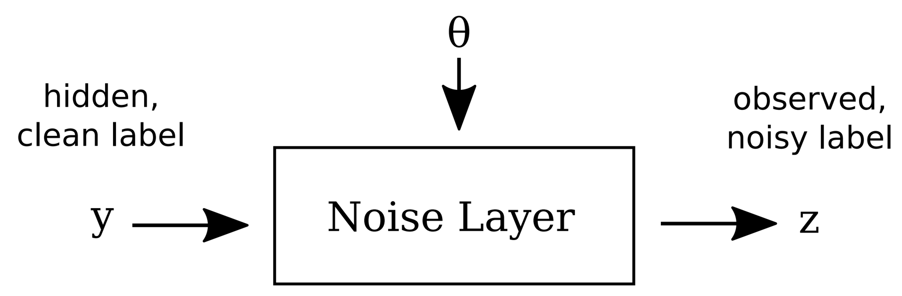
Setting

- Small, clean dataset $(x, y) \in C$.
- Large, cheaply obtained, noisy dataset $(x, z) \in N$.
- Multi-class classification:

$$p(y = i|x; w) = \frac{\exp(u_i^T h(x))}{\sum_{j=1}^{k} \exp(u_j^T h(x))}$$

Noise Model:

- Noise Channel by Goldberger and Ben-Reuven (2017).

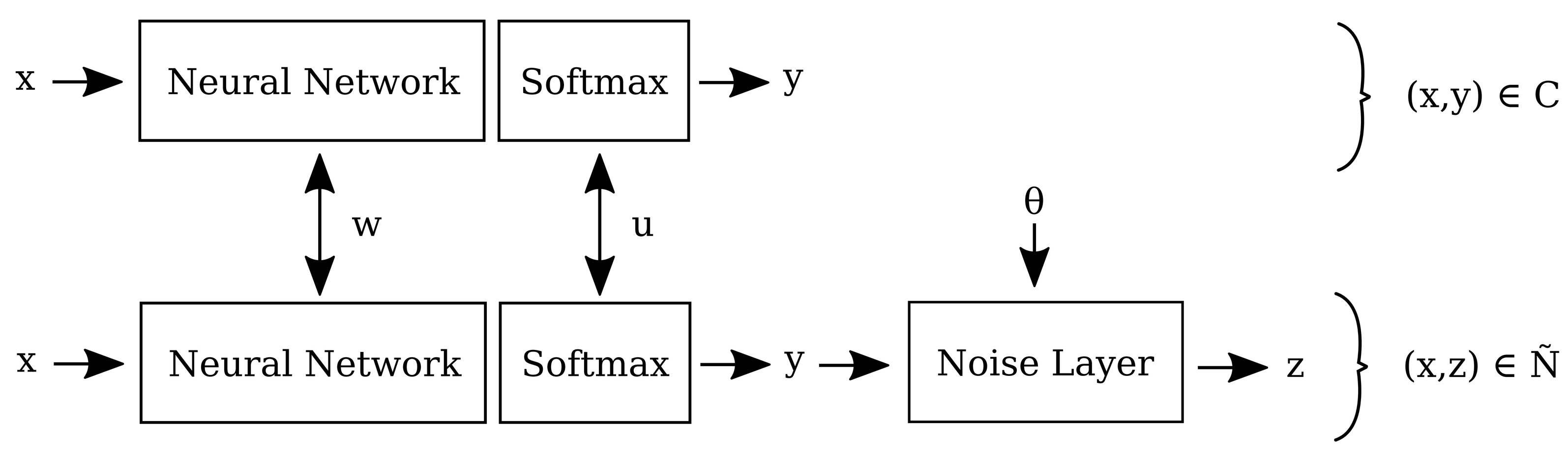


$$\theta(i, j) = p(z = j|y = i) = \frac{\exp(b_{ij})}{\sum_{l=1}^{k} \exp(b_{il})}$$

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

Proposed Model Architecture

- Base-model trained on C.
- Model with noise layer trained on N.
- Trained alternatingly (epoch-wise) with shared weights.



- Randomly subsample N to $\tilde{N}$ in each epoch to prevent noise from being too dominant.
- Initialize weights of θ using pairs of clean and noisy labels:

$$b_{ij} = \log\left(\frac{\sum_{t=1}^{|C|} 1_{\{y_t=i\}} 1_{\{z_t=j\}}}{\sum_{t=1}^{|C|} 1_{\{y_t=i\}}}\right)$$

Automatic Annotation of Named Entities

- Technique by Dembowski et al. (2017) uses external lists and gazetteers of entities (persons, organizations and locations).
- If a word appears in a list, assign corresponding entity class.
- Allows to quickly and cheaply annotate large corpora.
- On CoNLL data: precision 53%, recall 27% → noisy.

Dataset

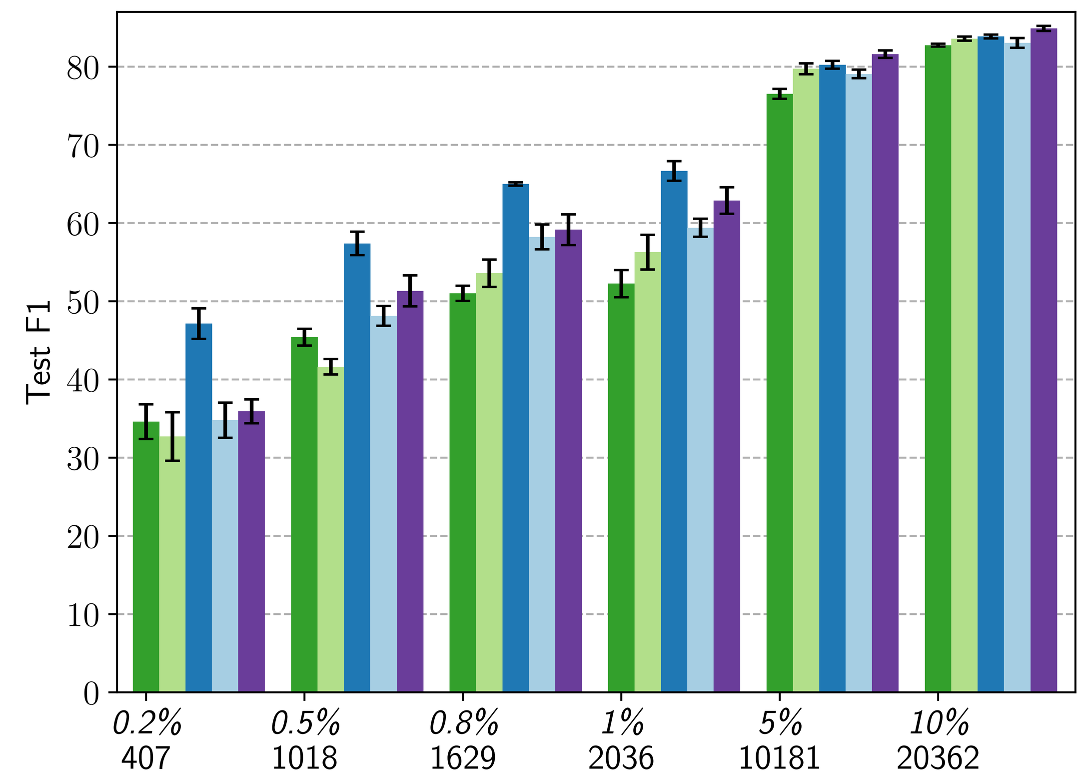
- Subsample of English CoNLL-2003 NER corpus as C.
- All of corpus as raw, unlabeled data with automatically annotated labels as N.

Experiments & Analysis

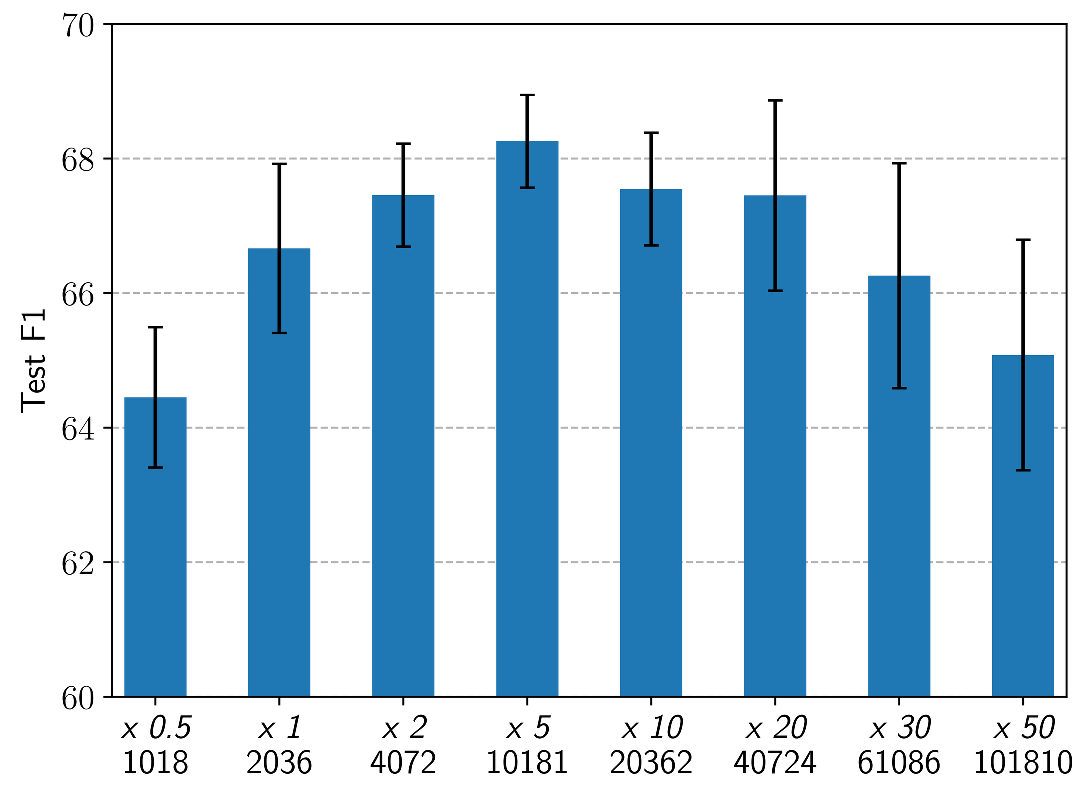
- Base-model (no noise-handling) trained only on clean data.
- Base-model trained on clean and noisy data.
- Our proposed noise-model.
- Noise-model with θ initialized using the identity matrix.
- Noise-cleaning-model based on the approach by Veit et al. (2017).

Model Comparison:

- Noisy data can hurt base-model.
- Initializing θ well is important.
- Noise-model leverages noisy data the most.



Performance for different $|C| = |\tilde{N}|$ (as *percentage of full corpus* and as absolute size).



Performance of noise-model for different $|\tilde{N}|$ (as *factor of* $|C|$ and as absolute size).

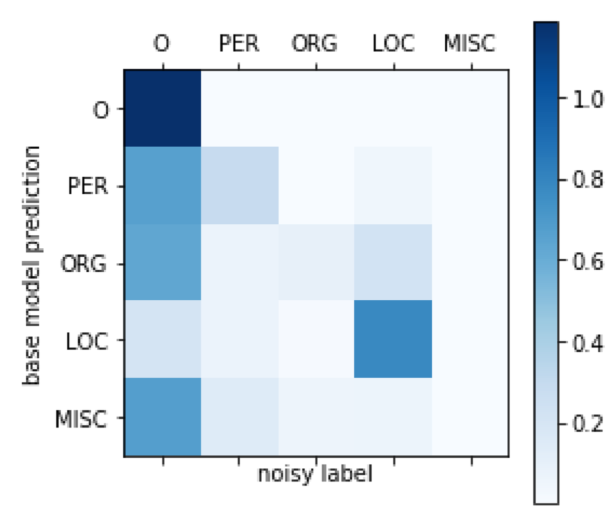
Amount of Noisy Data:

- Additional, noisy data helps as long as it is not too dominant.

Learned Weights:

Automatically annotated data N compared to learned weights.

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



- Low recall in PER and ORG reflected in high $\theta_{PER/ORG,O}$.
- High recall in LOC reflected in high $\theta_{LOC,LOC}$.

Conclusions

- Noise-model can handle the noise and leverage the additional, noisy data resulting in large performance improvements.
- Initialization of θ and subsampling $\tilde{N}$ are important factors.
- Learned noise model reflects the noise in the data.

1st Workshop on Deep Learning Approaches for Low-Resource NLP, 19 July 2018