Feature-Dependent Confusion Matrices for Low-Resource NER Labeling with Noisy Labels

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

Abstract

In low-resource settings, the performance of supervised labeling models can be improved with automatically annotated or distantly supervised data, which is cheap to create but often noisy. Previous works have shown that significant improvements can be reached by injecting information about the confusion between clean and noisy labels in this additional training data into the classifier training. However, for noise estimation, these approaches either do not take the input features (in our case word embeddings) into account, or they need to learn the noise modeling from scratch which can be difficult in a low-resource setting. We propose to cluster the training data using the input features and then compute different confusion matrices for each cluster. To the best of our knowledge, our approach is the first to leverage feature-dependent noise modeling with pre-initialized confusion matrices. We evaluate on low-resource named entity recognition settings in several languages, showing that our methods improve upon other confusion-matrix based methods by up to 9%.

1 Introduction

Most languages, even with millions of speakers, have not been the center for natural language processing and are counted as low-resource for tasks like named entity recognition (NER). Similarly, even for high-resource languages, there exists only few labeled data for most entity types beyond person, location and organization. Distantly- or weakly-supervised approaches have been proposed to solve this issue, e.g., by using lists of entities for labeling raw text (Ratinov and Roth, 2009; Dembowski et al., 2017). This allows obtaining large amounts of training data quickly and cheaply. Unfortunately, these labels often contain errors and learning with this noisily-labeled data is difficult and can even reduce overall performance (see, e.g. Fang and Cohn (2016)).

A variety of ideas have been proposed to overcome the issues of noisy training data. One popular approach is to estimate the relation between noisy and clean, gold-standard labels and use this noise model to improve the training procedure. However, most of these approaches only assume a dependency between the labels and do not take the features into account when modeling the label noise. This may disregard important information. The global confusion matrix (Hedderich and Klakow, 2018) is a simple model which assumes that the errors in the noisy labels just depend on the clean labels.

Our contributions are as follows: We propose to cluster the input words with the help of additional, unlabeled data. Based on this partition of the feature space, we obtain different confusion matrices that describe the relationship between clean and noisy labels. We evaluate our newly proposed models and related baselines in several low-resource settings across different languages with real, distantly supervised data with non-synthetic noise. The advanced modeling of the noisy labels substantially improves the performance up to 36% over methods without noise-handling and up to 9% over all other noise-handling baselines.

2 Related Work

A popular approach is modeling the relationship between noisy and clean labels, i.e., estimating $p(\hat{y}|y)$ where y is the clean and $\hat{y}$ the noisy label. For example, this can be represented as a noise or confusion matrix between the clean and the noisy labels, as explained in Section 3. Having its roots in statistics (Dawid and Skene, 1979), this or similar ideas have been recently studied in NLP (Fang and Cohn, 2016; Hedderich and

Klakow, 2018; Paul et al., 2019), image classification (Mnih and Hinton, 2012; Sukhbaatar et al., 2015; Dgani et al., 2018) and general machine learning settings (Bekker and Goldberger, 2016; Patrini et al., 2017; Hendrycks et al., 2018). All of these methods, however, do not take the features into account that are used to represent the instances during classification. In (Xiao et al., 2015) only the noise type depends on x but not the actual noise model. Goldberger and Ben-Reuven (2016) and Luo et al. (2017) use the learned feature representation h to model $p(\hat{y}|y, h(x))$ for image classification and relation extraction respectively. In the work of Veit et al. (2017), $p(y|\hat{y}, h(x))$ is estimated to clean the labels for an image classification task. The survey by Frenay and Verleysen (2014) gives a detailed overview about other techniques for learning in the presence of noisy labels.

Specific to learning noisy sequence labels in NLP, Fang and Cohn (2016) used a combination of clean and noisy data for low-resource POS tagging. Yang et al. (2018) suggested partial annotation learning to lessen the effects of incomplete annotations and reinforcement learning for filtering incorrect labels for Chinese NER. Hedderich and Klakow (2018) used a confusion matrix and proposed to leverage pairs of clean and noisy labels for its initialization, evaluating on English NER. For English NER and Chunking, Paul et al. (2019) also used a confusion matrix but learned it with an EM approach and combined it with multi-task learning. Recently, Rahimi et al. (2019) studied input from different, unreliable sources and how to combine them for NER prediction.

3 Global Noise Model

We assume a low-resource setting with a small set of gold standard annotated data C consisting of instances with features x and corresponding, clean labels y. Additionally, a large set of noisy instances $(x, \hat{y}) \in N$ is available. This can be obtained e.g. from weak or distant supervision. In a multi-class classification setting, we can learn the probability of a label y having a specific class given the feature x as

$$p(y = i|x) = \frac{exp(u_i^T h(x))}{\sum_{l=1}^{k} exp(u_l^T h(x))} \quad (1)$$

where k is the number of classes, h is a learned, non-linear function (in our case a neural network) and u is the softmax weights. This is our base model trained on C. Due to the errors in the labels, the clean and noisy labels have different distributions. Therefore, learning on C and N jointly can be detrimental for the performance of predicting unseen, clean instances. Nevertheless, the noisy-labeled data is still related to C and can contain useful information that we want to successfully leverage. We transform the predicted (clean) distribution of the base model to the noisy label distribution

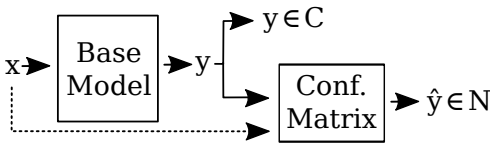


Figure 1: Visualization of the noisy labels, confusion matrix architecture. The dotted line shows the proposed new dependency.

$$p(\hat{y} = j|x) = \sum_{i=1}^{k} p(\hat{y} = j|y = i)p(y = i|x). \quad (2)$$

The relationship is modeled using a confusion matrix (also called noise or transformation matrix or noise layer) with learned weights b_{ij}:

$$p(\hat{y} = j|y = i) = \frac{\exp(b_{ij})}{\sum_{l=1}^{k} \exp(b_{il})} \quad (3)$$

The overall architecture is visualized in Figure 1. An important question is how to initialize this noise layer. As proposed by Hedderich and Klakow (2018), we apply the same distant supervision technique used to obtain N from unlabeled data on the already labeled instances in C. We thus obtain pairs of clean y and corresponding noisy labels $\hat{y}$ for the same instances and the weights of the noise layer can be initialized as

$$b_{ij} = \log\left(\frac{\sum_{t=1}^{|C|} 1_{\{y_t=i\}} 1_{\{\hat{y}_t=j\}}}{\sum_{t=1}^{|C|} 1_{\{y_t=i\}}}\right). \quad (4)$$

Following the naming by (Luo et al., 2017), we call this the global noise model.

4 Feature Dependent Noise Model

The global confusion matrix is a simple model which assumes that the errors in the noisy labels

depend on the clean labels. An approach that also takes the corresponding features x into account can model more complex relations. Veit et al. (2017) and Luo et al. (2017) use multiple layers of a neural network to model these relationships. However, in low resource settings with only small amounts of clean, supervised data, these more complex models can be difficult to learn. In contrast to that, larger amounts of unlabeled text are usually available even in low-resource settings. Therefore, we propose to use unsupervised clustering techniques to partition the feature space of the input words (and the corresponding instances) before estimating the noise matrices. To create the clusters, we use either Brown clustering (Brown et al., 1992) on the input words or k-means clustering (Lloyd, 1982) on the pretrained word embeddings after applying PCA (Pearson, 1901).

In sequence labeling tasks, the features x of an instance usually consist of the input word $\iota(x)$ and its context. Given a clustering Π over the input words $\{\iota(x) \mid (x, y) \in C \cup N\}$ consisting of clusters $\Pi_1, ..., \Pi_p$, we can group all clean and noisy instances into groups

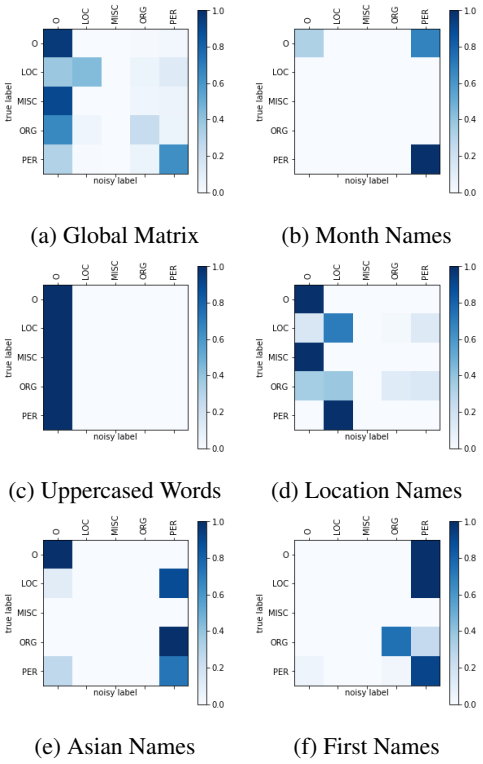
$$G_q = \{(x, y) \in C \cup N \mid \iota(x) \in \Pi_q\} \quad (5)$$

For each group, we construct an independent confusion matrix using Formulas 3 and 4. The prediction of the noisy label $\hat{y}$ (Formula 2) then becomes

$$p(\hat{y} = j|x) = \sum_{i=1}^{k} p(\hat{y} = j|y = i, G)p(y = i|x) \quad (6)$$

Since the clustering is performed on unsupervised data, in low-resource settings, the size of an actual group of instances G_q can be very small. If the number of members in a group is insufficient, the estimation of reliable noise matrices is difficult. This issue can be avoided by only using the largest groups and creating a separate group for all other instances. To make use of all the clusters, we alternatively propose to interpolate between the global and the group confusion matrix:

$$p_{\text{int}}(\hat{y} = j|y = i, G) = (1-\lambda)\cdot p(\hat{y} = j|y = i, G) + \lambda \cdot p(\hat{y} = j|y = i) \quad (7)$$



(a) Global Matrix (b) Month Names

(c) Uppercased Words (d) Location Names

(e) Asian Names (f) First Names

Figure 2: Confusion matrices used for initialization when training with the English dataset. The global matrix is given as well as five of the feature-dependent matrices obtained when using k-Means clustering for 75 clusters.

The interpolation hyperparameter λ (with $0 \leq \lambda \leq 1$) regulates the influence from the global matrix on the interpolated matrix. The selection of the largest groups and the interpolation can also be combined.

5 Experiments

We evaluate all models in five low-resource NER settings across different languages. Although the evaluation is performed for NER labeling, the proposed models are not restricted to the task of NER and can potentially be used for other tasks.

5.1 Models [1]

We follow the BiLSTM architecture from Hedderich and Klakow (2018). Only the optimizer was changed for all models to NADAM (Dozat, 2016) as this helped with convergence problems for increasing cluster numbers. The **Base** is trained only on clean data while **Base+Noise** is trained on both the clean and the noisy data without noise handling. **Global-CM** uses a global

[1]The code for all models is made available at https://github.com/uds-lsv/noise-matrix-ner

	De	En	Es	Et	Nl
Base	21.4 ± 1.0	35.9 ± 4.6	39.1 ± 1.6	36.7 ± 1.8	15.5 ± 3.0
Base+Noise	26.2 ± 0.6	50.5 ± 1.4	50.2 ± 1.0	51.5 ± 0.7	29.5 ± 2.7
Cleaning (Veit et al. 2017)	16.1 ± 4.3	52.3 ± 2.3	48.7 ± 2.3	53.8 ± 0.4	24.4 ± 5.5
Dynamic-CM (Luo et al. 2017)	32.6 ± 0.9	53.7 ± 1.8	57.6 ± 0.8	52.3 ± 0.8	36.7 ± 2.9
Global-ID-CM (H. and K. 2018)	27.1 ± 0.7	51.0 ± 1.1	50.9 ± 0.7	51.4 ± 0.6	29.9 ± 2.6
Global-CM (H. and K. 2018)	34.1 ± 1.4	52.0 ± 1.6	52.8 ± 0.6	52.3 ± 0.6	33.3 ± 2.0
Brown-CM-Freq	32.7 ± 0.7	51.3 ± 1.3	54.8 ± 1.0	53.4 ± 0.8	38.1 ± 1.7
K-Means-CM-Freq	29.7 ± 2.3	54.1 ± 2.9	52.3 ± 1.2	54.9 ± 0.8	39.8 ± 1.8
Brown-CM-IP	29.6 ± 1.1	55.5 ± 3.7	55.6 ± 1.0	52.6 ± 0.9	37.3 ± 1.5
K-Means-CM-IP	33.4 ± 1.1	53.0 ± 4.0	56.3 ± 2.1	53.3 ± 0.5	36.0 ± 1.9
Brown-CM-Freq-IP	**34.3 ± 1.4**	51.4 ± 2.3	**57.7 ± 2.4**	53.1 ± 0.9	**40.0 ± 1.3**
K-Means-CM-Freq-IP	33.1 ± 2.1	**57.6 ± 1.5**	57.2 ± 1.3	**55.2 ± 0.3**	39.7 ± 1.0

Table 1: Results of the evaluation in low-resource settings with 1% of the original labeled training data averaged over six runs. We report the F1 scores (higher is better) on the complete test set, as well as the standard error.

confusion matrix for all noisy instances to model the noise as proposed by Hedderich and Klakow (2018) and presented in Section 3. The same architecture is used for **Global-ID-CM**, but the confusion matrix is initialized with the identity matrix (instead of Formula 4) and only adapted during training.

The cluster-based models we propose in Section 4 are **Brown-CM** and **K-Means-CM**. We experimented with numbers of clusters of 5, 10, 25 and 50. The models that select only the largest groups G are marked as ***-Freq** and select either 30% or 50% of the clusters. The interpolation models have the postfix ***-IP** with $\lambda \in \{0.3, 0.5, 0.7\}$. The combination of both is named ***-Freq-IP**. As for all other hyperparameters, the choice was taken on the development set.

We implemented the **Cleaning** (Veit et al., 2017) and **Dynamic-CM** (Luo et al., 2017) models. Both were not developed for sequence labeling tasks and therefore needed to be adapted. For the Cleaning model, we followed the instructions by Hedderich and Klakow (2018). The embedding and prediction components of the **Dynamic-CM** model were replaced according to our base model. The output of the dense layer was used as input to the dynamic matrix generation. We experimented with and without their proposed trace loss.

The training for all models was performed with labels in the IO format. The predicted labels for the test data were converted and evaluated in IOB2 with the official CoNLL evaluation script. The IOB2 format would increase matrix size making the confusion matrix estimation more difficult without adding much information in practice. In preliminary experiments, this decreased performance in particular for low-resource settings.

5.2 Data

The models were tested on the four CoNLL datasets for English, German, Spanish and Dutch (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) using the standard split, and the Estonian data from Tkachenko et al. (2013) using a 10/10/80 split for dev/test/train sets. For each language, the labels of 1% of the training data (ca. 2100 instances) were used to obtain a low-resource setting. We treat this as the clean data C. The rest of the (now unlabeled) training data was used for the automatic annotation which we treat as noisily labeled data N. We applied the distant supervision method by Dembowski et al. (2017), which uses lists and gazetteer information for NER labeling. As seen in Table 2, this method reaches rather high precision but has a poor recall. The development set of the original dataset is used for model-epoch and hyperparameter selection, and the results are reported on the complete, clean test set. The words were embedded with the pretrained fastText vectors (Grave et al., 2018). The clusters were calculated on the unlabeled version of the full training data. Additionally, the Brown clusters used the language-specific documents from the Europarl corpus (Koehn, 2005).

	De	En	Es	Et	Nl
Precision	23.2	39.9	51.0	59.7	32.4
Recall	9.2	30.1	24.7	49.3	21.1
F1	13.2	34.3	33.3	54.0	25.5

Table 2: Results of the automatic labeling method proposed by Dembowski et al. (2017) on the test data.

6 Experimental Results

The results of all models are shown in Table 1. The newly proposed cluster-based models achieve the best performance across all languages and outperform all other models in particular for Dutch and English. The combination of interpolation with the global matrix and the selection of large clusters is almost always beneficial compared to the cluster-based models using only one of the methods. In general, both clustering methods achieve similar performance in combination with interpolation and selection, except for English, where Brown clustering performs worse than k-Means clustering. While the Brown clustering was trained on the relatively small Europarl corpus, k-Means clustering seems to benefit from the word embeddings trained on documents from the much larger common crawl.

7 Analysis

In the majority of cases, a cluster size of 10 or 25 was selected on the development set during the hyperparameter search. Increasing the number of clusters introduces smaller clusters for which it is difficult to estimate the noise matrix, due to the limited training resources. On the other hand, decreasing the number of clusters can generalize too much, resulting in loss of information on the noise distribution. For the λ parameter, a value of either 0.3 or 0.5 was chosen on the development set giving the group clusters more or equal weight compared to the global confusion matrix. This shows that the feature dependent noise matrices are important and have a positive impact on performance.

Five confusion matrices for groups and the global matrix in the English data are shown as examples in Figure 2. One can see that the noise matrix can visibly differ depending on the cluster of the input word. Some of these differences can also be directly explained by flaws in the distant supervision method. The automatic annotation did not label any locations written in all upper-case letters as locations. Therefore, the noise distribution for all upper-cased locations differs from the distribution of other location names (cf. 2d and 2c). The words April and June are used both as names for a month and as first names in English. This results in a very specific noise distribution with many temporal expressions being annotated as person entities (cf. 2b). Similar to this, first-person names and also Asian words are likely to be labeled as persons by the automatic annotation method (cf. 2f and 2e).

All of these groups show traits that are not displayed in the global matrix, allowing the cluster-based models to outperform the other systems.

8 Conclusions

We have shown that the noise models with feature-dependent confusion matrices can be used effectively in practice. These models improve low-resource named entity recognition with noisy labels beyond all other tested baselines. Further, the feature-dependent confusion matrices are task-independent and could be used for other NLP tasks, which is one possible direction of future research.

Acknowledgments

The authors would like to thank Heike Adel, Annemarie Friedrich and the anonymous reviewers for their helpful comments. This work has been partially funded by Deutsche Forschungsgemeinschaft (DFG) under grant SFB 1102: Information Density and Linguistic Encoding.

References

Alan Joseph Bekker and Jacob Goldberger. 2016. Training deep neural-networks based on unreliable labels. In *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28.

Julia Dembowski, Michael Wiegand, and Dietrich Klakow. 2017. Language independent named entity recognition using distant supervision. In *Proceedings of Language and Technology Conference*.

Y. Dgani, H. Greenspan, and J. Goldberger. 2018. Training a neural network based on unreliable human annotation of medical images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 39–42.

Timothy Dozat. 2016. Incorporating nesterov momentum into adam. In *Workshop Track of the International Conference on Learning Representations (ICLR)*.

Meng Fang and Trevor Cohn. 2016. Learning when to trust distant supervision: An application to low-resource pos tagging using cross-lingual projection. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*.

B. Frenay and M. Verleysen. 2014. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869.

Jacob Goldberger and Ehud Ben-Reuven. 2016. Training deep neural-networks using a noise adaptation layer. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Michael A. Hedderich and Dietrich Klakow. 2018. Training a neural network in a low-resource setting on automatically annotated noisy data. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*. Association for Computational Linguistics.

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. 2018. Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in Neural Information Processing Systems 31*, pages 10477–10486. Curran Associates, Inc.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, volume 5, pages 79–86.

S. P. Lloyd. 1982. Least squares quantization in PCM. In *IEEE Transactions on Information Theory*.

Bingfeng Luo, Yansong Feng, Zheng Wang, Zhanxing Zhu, Songfang Huang, Rui Yan, and Dongyan Zhao. 2017. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Volodymyr Mnih and Geoffrey Hinton. 2012. Learning to label aerial images from noisy data. In *Proceedings of the 29th International Conference on Machine Learning*, ICML'12.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1944–1952.

Debjit Paul, Mittul Singh, Michael A. Hedderich, and Dietrich Klakow. 2019. Handling noisy labels for robustly learning from self-training data for low-resource sequence labeling. *CoRR*, abs/1903.12008.

Karl Pearson. 1901. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Multilingual NER transfer for low-resource languages. *CoRR*, abs/1902.00193.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the SIGNLL Conference on Computational Natural Language Learning (CoNLL 2009)*.

Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. 2015. Training convolutional networks with noisy labels. In *Workshop Track of the International Conference on Learning Representations (ICLR)*.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alexander Tkachenko, Timo Petmanson, and Sven Laur. 2013. Named entity recognition in estonian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, BSNLP@ACL*.

Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge J. Belongie. 2017. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699.

Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. 2018. Distantly supervised ner with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 2159–2169.