HAL
open science

Distant supervision and noisy label learning for low resource named entity recognition: A study on Hausa and Yorùbá

David Ifeoluwa Adelani, Michael A Hedderich, Dawei Zhu, Esther van den Berg, Dietrich Klakow

To cite this version:

David Ifeoluwa Adelani, Michael A Hedderich, Dawei Zhu, Esther van den Berg, Dietrich Klakow. Distant supervision and noisy label learning for low resource named entity recognition: A study on Hausa and Yorùbá. ICLR Workshops (AfricaNLP & PML4DC 2020), Apr 2020, Addis Ababa, Ethiopia. hal-03359111

HAL Id: hal-03359111

https://hal.science/hal-03359111v1

Submitted on 29 Sep 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DISTANT SUPERVISION AND NOISY LABEL LEARNING FOR LOW RESOURCE NAMED ENTITY RECOGNITION: A STUDY ON HAUSA AND YORÙBÁ

David I. Adelani[1],* Michael A. Hedderich[1],* Dawei Zhu[1],* Esther van den Berg[2], Dietrich Klakow[1]
[1]Spoken Language Systems (LSV), Saarland University, Saarland Informatics Campus, Germany
[2]Institute for Computational Linguistics, Heidelberg University, Germany
{didelani,mhedderich,dzhu,dietrich.klakow}@lsv.uni-saarland.de

ABSTRACT

The lack of labeled training data has limited the development of natural language processing tools, such as named entity recognition, for many languages spoken in developing countries. Techniques such as distant and weak supervision can be used to create labeled data in a (semi-) automatic way. Additionally, to alleviate some of the negative effects of the errors in automatic annotation, noise-handling methods can be integrated. Pretrained word embeddings are another key component of most neural named entity classifiers. With the advent of more complex contextual word embeddings, an interesting trade-off between model size and performance arises. While these techniques have been shown to work well in high-resource settings, we want to study how they perform in low-resource scenarios. In this work, we perform named entity recognition for Hausa and Yorùbá, two languages that are widely spoken in several developing countries. We evaluate different embedding approaches and show that distant supervision can be successfully leveraged in a realistic low-resource scenario where it can more than double a classifier's performance.

1 INTRODUCTION

Named Entity Recognition (NER) is a classification task that identifies words in a text that refer to entities (such as dates, person, organization and location names). It is a core task of natural language processing and a component for many downstream applications like search engines, knowledge graphs and personal assistants. For high-resource languages like English, this is a well-studied problem with complex state-of-the-art systems reaching close to or above 90% F1-score on the standard datasets CoNLL03 (Baevski et al., 2019) and Ontonotes (Akbik et al., 2018). In recent years, research has been extended to a larger pool of languages including those of developing countries (Christianson et al., 2018; Zhang et al., 2018; Mayhew et al., 2019; Cao et al., 2019). Often, for these languages (like Hausa and Yorùbá studied here), there exists a large population with access to digital devices and internet (and therefore digital text), but natural language processing (NLP) tools do not support them.

One key reason is the absence of labeled training data required to train these systems. While manually labeled, gold-standard data is often only available in small quantities, it tends to be much easier to obtain large amounts of unlabeled text. Distant and weak supervision methods can then be used to create labeled data in a (semi-) automatic way. Using context (Mahajan et al., 2018; Wang et al., 2019), external knowledge and resources (Li et al., 2017; Pan et al., 2017), expert rules (Ratner et al., 2016; 2019) or self-training (Chen et al., 2018; Paul et al., 2019), a corpus or dataset can be labeled quickly and cheaply. Additionally, a variety of noise-handling methods have been proposed to circumvent the negative effects that errors in this automatic annotation might have on the performance of a machine learning classifier.

In this work, we study two methods of distant supervision for NER: Automatic annotation rules and matching of lists of entities from an external knowledge source. While distant supervision has

*Equal contribution.

been successfully used for high resource languages, it is not straight forward that these also work in low-resource settings where the amount of available external information might be much lower. The knowledge graph of Wikidata e.g. contains 4 million person names in English while only 32 thousand such names are available in Yorùbá, many of which are Western names.

Orthogonally to distant supervision, the pre-training of word embeddings is a key component for training many neural NLP models. A vector representation for words is built in an unsupervised fashion, i.e. on unlabeled text. Standard embedding techniques include Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2016). In the last two years, contextualized word embeddings have been proposed (Peters et al., 2018; Radford et al., 2019; Devlin et al., 2019). At the cost of having a much larger model size, these vector representations take the context of words into account and have been shown to outperform other embeddings in many tasks. In this study, we evaluate both types of representations.

The key questions we are interested in this paper are: How do NER models perform for Hausa and Yorùbá, two languages from developing countries? Are distant-supervision techniques relying on external information also useful in low-resource settings? How do simple and contextual word embeddings trade-off in model size and performance?

2 Background & Methods

2.1 Languages

Hausa language is the second most spoken indigenous language in Africa with over 40 million native speakers (Eberhard et al., 2019), and one of the three major languages in Nigeria, along with Igbo and Yorùbá. The language is native to the Northern part of Nigeria and the southern part of Niger, and it is widely spoken in West and Central Africa as a trade language in eight other countries: Benin, Ghana, Cameroon, Togo, Côte d'Ivoire, Chad, Burkina Faso, and Sudan. Hausa has several dialects but the one regarded as standard Hausa is the *Kananci* spoken in the ancient city of Kano in Nigeria. Kananci is the dialect popularly used in many local (e.g VON news[1]) and international news media such as BBC, VOA, DW and Radio France Internationale. Hausa is a tone language but the tones are often ignored in writings, the language is written in a modified Latin alphabet. Despite the popularity of Hausa as an important regional language in Africa and it's popularity in news media, it has very little or no labelled data for common NLP tasks such as text classification, named entity recognition and question answering.

Yorùbá language is the third most spoken indigenous language in Africa after Swahilli and Hausa with over 35 million native speakers (Eberhard et al., 2019). The language is native to the South-western part of Nigeria and the Southern part of Benin, and it is also spoken in other countries like Republic of Togo, Ghana, Côte d'Ivoire, Sierra Leone, Cuba and Brazil. Yorùbá has several dialects but the written language has been standardized by the 1974 Joint Consultative Committee on Education (Asahiah et al., 2017), it has 25 letters without the Latin characters (c, q, v, x and z) and with additional characters (ẹ, gb, ṣ , ọ). Yorùbá is a tone language and the tones are represented as diacritics in written text, there are three tones in Yorùbá namely low (\), mid ("−") and high (/). The mid tone is usually ignored in writings. Often time articles written online including news articles[2] like BBC and VON ignore diacritics. Ignoring diacritics makes it difficult to identify or pronounce words except they are in a context. For example, *owó* (money), *ọ́*wọ̀ (broom), *òw*ò (business), *ọ̀*wọ̀ (honour), *ọ́*wọ́ (hand), and *ọ̀*wọ́ (group) will be mapped to *owo* without diacritics. Similar to the Hausa language, there are few or no labelled datasets for NLP tasks.

2.2 Datasets & Embeddings

The Hausa data used in this paper is part of the LORELEI[3] language pack. It consists of Broad Operational Language Translation (BOLT) data gathered from news sites, forums, weblogs, Wikipedia articles and twitter messages. We use a split of 10k training and 1k test instances. Due to the Hausa

[1] https://www.von.gov.ng/hausa/

[2] https://www.von.gov.ng/yoruba/, and https://www.bbc.com/yoruba

[3] https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents

data not being publicly available at the time of writing, we could only perform a limited set of experiments on it.

The Yorùbá NER data used in this work is the annotated corpus of Global Voices news articles[4] recently released by Alabi et al. (2020). The dataset consists of 1,101 sentences (26k tokens) divided into 709 training sentences, 113 validation sentences and 279 test sentences based on 65%/10%/25% split ratio. The named entities in the dataset are personal names (PER), organization (ORG), location (LOC) and date & time (DATE). All other tokens are assigned a tag of "O".

For the Yorùbá NER training, we make use of Yorùbá FastText embeddings (Alabi et al., 2020) and multilingual-BERT[5] that was trained on 104 languages including Yorùbá. Instead of the original FastText embeddings Bojanowski et al. (2016), we chose FastText embeddings trained on a multi-domain and high-quality dataset (Alabi et al., 2020) because it gave better word similarity scores.

2.3 Distant and Weak Supervision

In this work, we rely on two sources of distant supervision chosen for its ease of application:

Rules allow to apply the knowledge of domain experts without the manual effort of labeling each instance. They are especially suited for entities that follow specific patterns, like time phrases in text (see also Strötgen et al. (2018)). We use them for the DATE entity. In Yoruba, date expressions are written with the keywords of "*ọjọ́*" (day), "*oṣù*" (month), and "*ọdún*" (year). Similarly, time expressions are written with keywords such as "*wákàtí*" (hour), "*ìṣẹ́jú* (minute) and "*ìṣẹ́jú-aaya* (seconds). Relative date and time expressions are also written with keywords "*ọ̀dún*" (in the year), "*lọ́ṣù*" (in the month), "*lọ́sẹ̀*" (in the week), "*lọ́jọ́*" (in the day). An example of a date expression is:

"8th of December, 2018" in Yorùbá translates to "*ọjọ́ 8 oṣù Ọpẹ̀, ọdún 2018*"

Lists of Entities can be obtained from a variety of sources like gazetteers, dictionaries, phone books, census data and Wikipedia categories (Ratinov & Roth, 2009). In recent years, knowledge bases like Freebase and Wikidata have become another option to retrieve entity lists in a structured way. An entity list is created by extracting all names of that type from a knowledge source (e.g. all person names from Wikidata). If a word or token from the unlabeled text matches an entry in an entity list, it is assigned the corresponding label. Experts can add heuristics to this automatic labeling that improve the matching (Dembowski et al., 2017). These include e.g. normalizing the grammatical form of words or filtering common false positives.

Another popular method for low-resource NER is the use of cross-lingual information (Rahimi et al., 2019). Alternatives to distant supervision are crowd-sourcing (Rehbein & Ruppenhofer, 2017) and non-expert annotations (Mayhew & Roth, 2018).

2.4 Learning With Noisy Labels

The labels obtained through distant and weak supervision methods tend to contain a high amount of errors. In the Food101N dataset (Lee et al., 2018) around 20% of the automatically obtained labels are incorrect while for Clothing1M (Xiao et al., 2015) the noise rate is more than 60%. Learning with this additional, noisily labeled data can result in lower classification performance compared to just training on a small set of clean labels (cf. e.g. Fang & Cohn (2016)). A variety of techniques have been proposed to handle label noise like modelling the underlying noise process (Lange et al., 2019) and filtering noisy instances (Yang et al., 2018; Nguyen et al., 2020). Frenay & Verleysen (2014) gives an in-depth introduction into this field and Algan & Ulusoy (2019) survey more recent approaches, focusing on the vision domain.

In this work, we experiment with three noise handling techniques. The approach by Bekker & Goldberger (2016) estimates a noise channel using the EM algorithm. It treats all labels as possibly noisy and does not distinguish between a clean and a noisy part of the data. In contrast, the method by Hedderich & Klakow (2018) leverages the existence of a small set of gold standard labels, something that - in our experience - is often available even in low resource settings. Having such a small set of clean labels is beneficial both for the main model itself as well as for the noise handling technique.

[4] https://yo.globalvoices.org/

[5] https://github.com/google-research/bert/blob/master/multilingual.md

(a) Performance of the Bi-LSTM on Hausa (PER only)

Data	Precision	Recall	F1
100% Clean	79	45	57
50% Clean + 50% Distant	59	38	42
Noise Channel	57	47	51

(b) Distant Supervision Quality (on test)

Method	P	R	F1
Overall	41	41	41
PER	36	16	22
LOC	73	54	62
ORG	17	29	22
DATE	37	66	48

(c) Distant Supervision: Coverage of Wikidata for person names (on train)

Method	P	R	F1
Wikidata	50	25	33
+ Yorùbá Names	47	25	33
+ Nigerian Names	55	39	46

Both approaches model the relationship between clean and noisy labels using a confusion matrix. This allows adapting the noisy to the clean label distribution during training. For a setting with 5 labels, it only requires $5^2 = 25$ additional parameters to estimate which could be beneficial when only few training data is available. The technique by Veit et al. (2017) (adapted to NER by Hedderich & Klakow (2018)) learns a more complex neural network to clean the noisy labels before training with them. It also takes the features into account when cleaning the noise and it might, therefore, be able to model more complex noise processes. All three techniques can be easily added to the existing standard neural network architectures for NER.

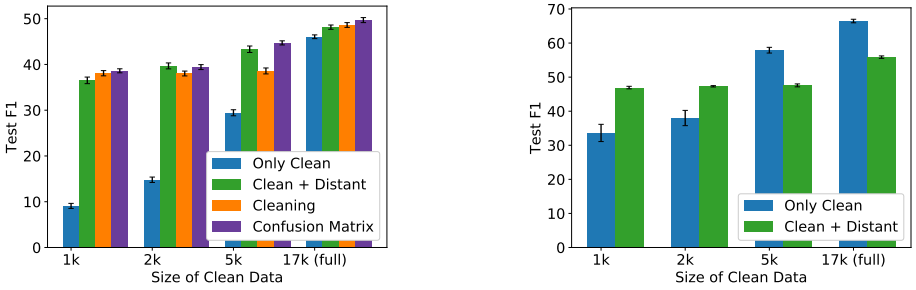
3 Models & Experimental Settings

Hausa Distant supervision on Hausa was performed using lists of person names extracted from Wikipedia data. Since we had limited access to the data, we tested a simplified binary NER-tagging setting (PERSON-tags only). As a base model, we used a Bi-LSTM model developed for Part-of-Speech tagging (Plank et al., 2016). For noise handling, we apply the **Noise Channel** model by Bekker & Goldberger (2016).

Yorùbá For Yorùbá, the entity lists were created by extracting person, location and organization entities from Wikidata in English and Yorùbá. Additionally, a list of person names in Nigeria was obtained from a Yorùbá Name website[6] (8,365 names) and list of popular Hausa, Igbo, Fulani and Yorùbá people on Wikipedia (in total 9,241 names). As manual heuristic, a minimum name length of 2 was set for extraction of PER (except for Nigerian names), LOC and ORG. The Nigerian names were set to include names with a minimum length of 3. For the DATE label, a native Yorùbá speaker wrote some annotation rules using 11 "date keywords" ("*ọjọ́*", "*ọsẹ̀*", "*osù*", "*ọdún*", "*wákàtí*" , "*ọ̀ọ́dún*", "*ọ̀ọ́dún*-un", "*ọdún*-un" "*ọ̀ọ́sẹ̀*" , "*ọ̀ọ́jọ́*", " *aago*") following these two criteria: (1) A token is a date keyword or follows a date keyword in a sequence. (2) A token is a digit. For Yorùbá, we evaluate four settings with different amounts of clean data, namely 1k, 2k, 4k and the full dataset. As distantly supervised data with noisy labels, the full dataset is used. Additionally, 19,559 words from 18 articles of the Global News Corpus (different from the articles in the training corpus) were automatically annotated.

The **Bi-LSTM** model consists of a Bi-LSTM layer followed by a linear layer to extract input features. The Bi-LSTM layer has a 300-dimensional hidden state for each direction. For the final classification, an additional linear layer is added to output predicted class distributions. For noise handling, we experiment with the **Confusion Matrix** model by Hedderich & Klakow (2018) and the **Cleaning** model by Veit et al. (2017). We repeat all the Bi-LSTM experiments 20 times and report the average F1-score (following the approach by Tjong Kim Sang & De Meulder (2003)) and the standard error.

The **BERT** model is obtained by *fine-tuning* the pre-trained BERT embeddings on NER data with an additional untrained CRF classifier. We fine-tuned all the parameters of BERT including that of the CRF end-to-end. This has been shown to give better performance than using word features extracted from BERT to train a classifier (Devlin et al., 2019). The evaluation result is obtained as an average of 5 runs, we report the F1-score and the standard error in the result section.



(a) Bi-LSTM (b) BERT

Figure 1: F1-scores and standard error for Yorùbá.

4 RESULTS

The results for Hausa are given in Table 1a. Training with a mix of 50% clean and 50% distantly-supervised data performs 15 F1-score points below using the whole 100% clean data which is to be expected due to the lower quality of the distantly-supervised labels. Using the Noise Channel closes half of this gap. Due to the limited availability of the dataset, we could unfortunately not investigate this further, but it shows already the benefits that are possible through noise-handling.

An evaluation of the distant supervision for Yorùbá is given in Table 1b. The quality of the automatically annotated labels differs between the classes. Locations perform better than person and organization names, probably due to locations being less diverse and better covered in Wikidata. With simple date rules, we obtain already a 48% F1-score. This shows the importance of leveraging the knowledge of native speakers in automatic annotations. Overall a decent annotation can be obtained by the distant supervision and it even outperforms some of the actual machine learning models in the low-resource setting. Table 1c compares using only Wikidata as data source versus adding additional, manually obtained lists of person names. While adding a list of Yorùbá names only improves recall slightly, the integration of Nigerian names helps to boost recall by 13 points.

The experimental results for Yorùbá are given in Figure 1. The setting differs from the experiments with Hausa in that there is a small clean training set and *additional*, distantly-supervised data. For the Bi-LSTM model, adding distantly-supervised labels always helps. In the low-resource settings with 1k and 2k labeled data, it more than doubles the performance. Handling the noise in the distant supervision can result in slight improvements. The noise-cleaning approach struggles somewhat while the confusion matrix architecture does give better results in the majority of the scenarios. Training on 5k labeled data with distantly supervised data and noise handling, one can obtain a performance close to using the full 17k manually labeled token.

The Bi-LSTM model has 1.50 million parameters (1.53 million for the cleaning model), while BERT has 110 million parameters. There is a clear trade-off between model size and performance. The BERT model is 70 times larger and obtains consistently better results due to its more complex, contextual embeddings pretrained on more data. Still, the F1-score also drops nearly half for the BERT model in the 1k setting compared to the full dataset. For 1k and 2k labeled data, the distant supervision helps to improve the model's performance. However, once the model trained only on clean data reaches a higher F1-score than the distant supervision technique, the model trained on clean and distantly-supervised data deteriorates. This suggests that the BERT model overfits too much on the noise in the distant supervision.

5 CONCLUSION

In this study, we analysed distant supervision techniques and label-noise handling for NER in Hausa and Yorùbá, two languages from developing countries. We showed that they can be successfully leveraged in a realistic low-resource scenario to double a classifier's performance. If model size is

[6] http://www.yorubaname.com/

not a constraint, the more complex BERT model clearly outperforms the smaller Bi-LSTM architecture. Nevertheless, there is still a large gap between the best performing model on Yorùbá with 66 F1-score and the state-of-the-art in English around 90.

We see several interesting follow-ups to these evaluations. In the future, we want to evaluate if noise handling methods can also allow the more complex BERT model to benefit from distant supervision. Regarding the model complexity, it would be interesting to experiment with more compact models like DistilBERT (Sanh et al., 2019) that reach a similar performance with a smaller model size for high-resource settings. In general, we want to investigate more in-depth the trade-offs between model complexity and trainability in low-resource scenarios.

ACKNOWLEDGMENTS

The experiments on Hausa were possible thanks to the collaboration with Florian Metze and CMU as part of the LORELEI project. Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 232722074 – SFB 1102 / Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102, the EU-funded Horizon 2020 projects ROXANNE under grant agreement No. 833635 and COMPRISE (`http://www.compriseh2020.eu/`) under grant agreement No. 3081705.

REFERENCES

Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1638–1649, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/C18-1139`.

Jesujoba O Alabi, Kwabena Amponsah-Kaakyire, David I Adelani, and Cristina España-Bonet. Massive vs. curated word embeddings for low-resourced languages. the case of Yor\ub\'a and Twi. In *LREC*, 2020.

Görkem Algan and Ilkay Ulusoy. Image classification with deep learning in the presence of noisy labels: A survey. *CoRR*, abs/1912.05170, 2019.

F. O. Asahiah, O. A. Odejobi, and E. R. Adagunodo. Restoring tone-marks in standard yoruba electronic text: Improved model. *Computer Science*, 18(3):301–315, 2017.

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. Cloze-driven pretraining of self-attention networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 5359–5368, 2019. doi: 10.18653/v1/D19-1539. URL `https://doi.org/10.18653/v1/D19-1539`.

Alan Joseph Bekker and Jacob Goldberger. Training deep neural-networks based on unreliable labels. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pp. 2682–2686, 2016. doi: 10.1109/ICASSP.2016.7472164. URL `https://doi.org/10.1109/ICASSP.2016.7472164`.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

Yixin Cao, Zikun Hu, Tat-Seng Chua, Zhiyuan Liu, and Heng Ji. Low-resource name tagging learned with weakly labeled data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 261–270, 2019. doi: 10.18653/v1/D19-1025. URL `https://doi.org/10.18653/v1/D19-1025`.

Chenhua Chen, Yue Zhang, and Yuze Gao. Learning how to self-learn: Enhancing self-training using neural reinforcement learning. In *2018 International Conference on Asian Language Processing, IALP 2018, Bandung, Indonesia, November 15-17, 2018*, pp. 25–30, 2018. doi: 10.1109/IALP.2018.8629107. URL `https://doi.org/10.1109/IALP.2018.8629107`.

Caitlin Christianson, Jason Duncan, and Boyan A. Onyshkevych. Overview of the DARPA LORELEI program. *Machine Translation*, 32(1-2):3–9, 2018. doi: 10.1007/s10590-017-9212-4. URL https://doi.org/10.1007/s10590-017-9212-4.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). Ethnologue: Languages of the world. twenty-second edition., 2019. URL http://www.ethnologue.com.

Meng Fang and Trevor Cohn. 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*, 2016. doi: 10.18653/v1/K16-1018.

B. Frenay and M. Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014. doi: 10.1109/TNNLS.2013.2292894.

Michael A. Hedderich and Dietrich Klakow. 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, July 2018.

Lukas Lange, Michael A. Hedderich, and Dietrich Klakow. Feature-dependent confusion matrices for low-resource ner labeling with noisy labels. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3545–3550. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1362.

Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 5447–5456, 2018. doi: 10.1109/CVPR.2018.00571.

Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *CoRR*, abs/1708.02862, 2017.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 185–201. Springer International Publishing, 2018.

Stephen Mayhew and Dan Roth. Talen: Tool for annotation of low-resource entities. In *Proceedings of ACL 2018: System Demonstrations*, 2018. URL http://aclweb.org/anthology/P18-4014.

Stephen Mayhew, Snigdha Chaturvedi, Chen-Tse Tsai, and Dan Roth. Named entity recognition with partially annotated training data. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 645–655, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1060. URL https://www.aclweb.org/anthology/K19-1060.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013. URL http://arxiv.org/abs/1301.3781.

T. Nguyen, C. K. Mummadi, T. P. N. Ngo, T. H. P. Nguyen, L. Beggel, and T. Brox. Self: learning to filter noisy labels with self-ensembling. In *International Conference on Learning Representations (ICLR)*, 2020.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1946–1958, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1178. URL https://www.aclweb.org/anthology/P17-1178.

Debjit Paul, Mittul Singh, Michael A. Hedderich, and Dietrich Klakow. Handling noisy labels for robustly learning from self-training data for low-resource sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 29–34, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-3005.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1532–1543, 2014. doi: 10.3115/v1/d14-1162. URL https://doi.org/10.3115/v1/d14-1162.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 2227–2237, 2018. doi: 10.18653/v1/n18-1202. URL https://doi.org/10.18653/v1/n18-1202.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 412–418, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2067. URL https://www.aclweb.org/anthology/P16-2067.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Afshin Rahimi, Yuan Li, and Trevor Cohn. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 151–164, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1015. URL https://www.aclweb.org/anthology/P19-1015.

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

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: rapid training data creation with weak supervision. *The VLDB Journal*, Jul 2019. ISSN 0949-877X. doi: 10.1007/s00778-019-00552-1. URL https://doi.org/10.1007/s00778-019-00552-1.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3567–3575. Curran Associates, Inc., 2016.

Ines Rehbein and Josef Ruppenhofer. Detecting annotation noise in automatically labelled data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1160–1170, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1107. URL https://www.aclweb.org/anthology/P17-1107.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

Jannik Strötgen, Michael Gertz, Graeme Hirst, and Ruihong Huang. Domain-sensitive temporal tagging. *Computational Linguistics*, 44(2), 2018.

Erik F. Tjong Kim Sang and Fien De Meulder. 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*, pp. 142–147, 2003.

Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge J. Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. doi: 10.1109/CVPR.2015.7298885.

Hao Wang, Bing Liu, Chaozhuo Li, Yan Yang, and Tianrui Li. Learning with noisy labels for sentence-level sentiment classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pp. 6285–6291, 2019. doi: 10.18653/v1/D19-1655.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2691–2699, 2015. doi: 10.1109/CVPR.2015.7298885.

Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. Distantly supervised ner with partial annotation learning and reinforcement learning. In *Proceedings of COLING 2018*, 2018. URL https://www.aclweb.org/anthology/C18-1183.

Boliang Zhang, Ying Lin, Xiaoman Pan, Di Lu, Jonathan May, Kevin Knight, and Heng Ji. Elisa-edl: A cross-lingual entity extraction, linking and localization system. In *Proceedings of NAACL-HLT 2018: Demonstrations*, 2018. doi: 10.18653/v1/N18-5009. URL http://aclweb.org/anthology/N18-5009.