Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4324–4332
Marseille, 20-25 June 2022

Modeling Noise in Paraphrase Detection

Teemu Vahtola, Eetu Sjöblom, Jörg Tiedemann, Mathias Creutz
Department of Digital Humanities
Faculty of Arts
University of Helsinki
Finland
{teemu.vahtola, eetu.sjoblom, jorg.tiedemann, mathias.creutz}@helsinki.fi

Abstract

Noisy labels in training data present a challenging issue in classification tasks, misleading a model towards incorrect decisions during training. In this paper, we propose the use of a linear noise model to augment pre-trained language models to account for label noise in fine-tuning. We test our approach in a paraphrase detection task with various levels of noise and five different languages. Our experiments demonstrate the effectiveness of the additional noise model in making the training procedures more robust and stable. Furthermore, we show that this model can be applied without further knowledge about annotation confidence and reliability of individual training examples and we analyse our results in light of data selection and sampling strategies.

Keywords: paraphrase detection, label noise

1. Introduction

Large pre-trained language models have obtained great results in many NLP tasks in recent years by utilising the dominant pre-training and fine-tuning paradigm. The language models' success on a specific downstream task is to a great extent dependent on the quality and quantity of task-specific training data. A considerable amount of annotated, high-quality training data is often necessary even for fine-tuning a large pre-trained language model in order for it to perform well on a certain downstream task. Consequently, different methods are utilised to obtain sufficient amounts of training data in a cost-effective way leading to different levels of noise.

In this paper, we study the use of an explicit noise model to address label noise – incorrectly classified training samples that mislead the model during training. The noise model can be added to a neural architecture to make the training procedures more robust and predictable when noise enters the model. We apply the task of paraphrase detection as an example to demonstrate the use of an additional noise layer that is attached to a language model during fine-tuning.

Our contributions are: (1) An implementation of a noise model that extends pre-trained language models to increase robustness when fine-tuning paraphrase detection with noisy training data. (2) A study that shows the effect of noise modeling with various degrees of noise across five different languages. (3) Experiments that demonstrate the effectiveness of noise models without explicit information about annotation quality and label confidence.

2. Training with Label Noise

Online platforms are often utilized for the acquisition of large quantities of annotations from non-experts. The quality of such crowdsourced human-made annotations has recently been questioned, for example, in machine translation evaluation (see, e.g., Graham et al. (2017); Toral et al. (2018); Läubli et al. (2020)). Annotators seem to assign higher scores to translations that more accurately conform the sentence structure of the source sentence than to those that differ in phrasing and structure, even if the latter would be more natural (Freitag et al., 2020). Paraphrase detection struggles with a similar problem. As classifying paraphrases is about assessing degrees of similarity between sequences, many instances can be difficult to classify, and the defining line between paraphrastic and non-paraphrastic sequences can shift between annotators. Aulamo et al. (2019) have studied inter-annotator agreement in the paraphrase domain. In a four-scale classification task of identifying potential paraphrases in the Opusparcus corpus (Creutz, 2018), they found inter-annotator agreement to be as low as 59.9%. Relaxing the task into binary classification increased inter-annotator agreement to 83.1% Furthermore, they hypothesize that a deeper expertise in the topic results in more coherent categorization of paraphrases.

Paraphrase annotation and machine translation are not the only examples where the complexity of NLP tasks causes annotation noise. A similar picture can be seen in data sets for sentiment and emotion classification (Strapparava and Mihalcea, 2010) and natural language inference (Gururangan et al., 2018), to name a few. In general, some level of noise is unavoidable even with extensive annotator training (Bayerl and Paul, 2011) and needs to be addressed in one way or another. Sometimes, automation and heuristics can help to facilitate (semi-)automatic annotation (Mohammad, 2012; Araque et al., 2019) but that comes with additional caveats.

Erroneous annotations in the training data typically appear as label noise, that is, annotations that do not correspond to the true class of the training sample. The

misclassified instances may thus lead the model to incorrect decisions during training. For instance, label noise can be a negative instance incorrectly labeled as positive in a binary classification task (false positives).

2.1. Related Work

Machine learning practitioners use varied methods to alleviate the effect of label noise to the model. To obtain high-quality data consisting only of reliable training examples, incorrectly classified examples can be identified and removed. However, the process of identifying and removing misclassified examples can be extremely time-consuming, and moreover, it is not always desired to remove training examples from the data, especially if the data is of limited size to begin with. Detailed information about approaches to training with noisy labels is presented in Frenay and Verleysen (2014). Noisy training data can be augmented with clean data for a model to learn to account for the corrupted labels in the original data. This line of work includes, for example, Li et al. (2017), who use knowledge distillation from a model trained on small and clean data to re-weight noisy labels to soft labels in a noisy training set. Veit et al. (2017) train a network with a clean sub-sample of the training data to reduce noise in the rest of the data. Annotating a sufficient number of clean training examples for training an auxiliary network can be costly, and ready-made additional clean data is rarely available. Thus, methods for learning models directly from data that includes noisy labels should be analysed further.

Augmenting machine learning models with a noise processing layer to directly transform latent labels to noisy labels to explicitly model the effect of noise has been studied comprehensively for image processing. Mnih and Hinton (2012) implement a robust loss-function that reduces the effect of noisy labels to the image classifier by treating true observations as noisy observations. Sukhbaatar et al. (2015), Jindal et al. (2016), and Patrini et al. (2017) integrate a linear noise layer on top of a base model's prediction layer to transform the output of the base model to correspond to the noisy label distribution in the data. Jindal et al. (2019) extend the noise layer experimentation from image processing to multiple text-classification tasks. These works train the base neural networks from scratch, whereas we are curious to examine to what extent a strong pre-trained language model can benefit from an additional noise layer when the noise layer is applied into the fine-tuning stage. We reproduce the model proposed by Jindal et al. (2019) to paraphrase detection, but instead of training a sequence classification model from scratch, we use BERT (Devlin et al., 2019) as our base model and exploit the noise processing layer for transforming the outputs of an additional classification head to correspond to the noisy label distribution in the training data during fine-tuning.

Jindal et al. (2019) apply the same noise processing layer to all training samples and assume that the noise is instance-independent. Another strategy is to deal with instance-dependent noise by incorporating explicit information about specific training samples into the model reflecting the label confidence or label ambiguity in the data (Haque et al., 2021; Berthon et al., 2021). We implement a strategy for augmenting our general noise model with label confidence values extracted from the training data ranking to test the impact of such information.

2.2. Paraphrase Detection with Noisy Labels

We use five languages from Opusparcus to evaluate our approach.[1] Opusparcus is a sentential paraphrase corpus consisting of data sets in six European languages. The training sets are automatically ranked based on a probabilistic score of two sequences being paraphrases. As such, the data does not contain explicit annotations but the score can be used to define cutoff points for positive paraphrase examples. A small sample of manually annotated paraphrase pairs across various ranges of scores makes it possible to control the expected noise level in the selected training data (Creutz, 2018), which is useful for our experiments below. The data is described in more detail in section 4.1.

Previous research about paraphrase detection on Opusparcus include Sjöblom et al. (2018) and Vahtola et al. (2021). However, that work does not attempt to mitigate the effect of noisy labels directly. Instead, the objective has been in finding an optimal trade-off between the number of training examples and noisy labels for a deep neural network trained from scratch (Sjöblom et al., 2018), or a proportion of noisy labels in the training data up to which a pre-trained language model (BERT in this case) is still capable of obtaining high accuracy in paraphrase detection (Vahtola et al., 2021). Vahtola et al. (2021) show that a fine-tuned BERT can perform with up to 20% of corrupted training data after which the model starts collapsing. In contrast to previous work, we attempt to alleviate the effect of noise in the data with an explicit noise model to address considerably higher proportions of noisy labels and to make it possible to handle label noise even without existing confidence values for the labels in the training data.

3. Model Description

The *baseline model* implements a binary classification layer on top of fine-tuned language models (language-specific BERT in our case) with no explicit noise modeling. In addition to that, we experiment with two noise models: (i) a *label noise model* that takes as input the output prediction from the base model's classification layer, and transforms the predictions to account for a noisy label distribution in the training set, and (ii) a

[1] Opusparcus also includes French, which we do not include in this work due to technical difficulties in implementing the noise models using FlauBERT (Le et al., 2020) as the base model.

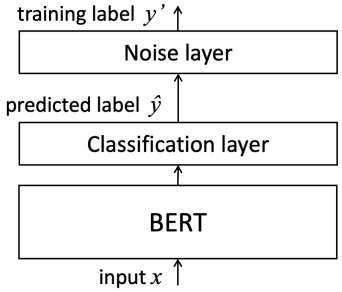

Figure 1: The input signal x, consisting of a pair of sentences, is fed into a fine-tuned BERT model augmented with a classification layer, which outputs a predicted label $\hat{y}$, indicating whether the sentences are paraphrases or not. During training, an additional noise layer transforms $\hat{y}$ into the observed training label y', which is noisy and may therefore be incorrect.

label confidence model that, in addition to transforming the latent labels to noisy labels, uses auxiliary label confidence values to further guide the transformation. The confidence values are obtained from the scores that are used in ranking the sentence pairs in Opusparcus.
With those two models, we want to contrast the ability of a noise model to cope with label noise with or without explicit knowledge about annotation confidence in comparison to a general baseline model.

3.1. Label Noise Model

Our implementation of the additional noise model layer is based on the noise absorbing layer proposed by Jindal et al. (2019). In contrast to Jindal et al. (2019), we augment a pre-trained language model (BERT) with an auxiliary linear transformation layer Q that learns the transformation from the unnormalized base model output $\hat{y}$ to the noisy label distribution y' such that

$$p(y'|x) = \textit{softmax}(Q \times \hat{y}) \tag{1}$$

The noise layer Q is initialized as an $n \times n$ identity matrix, where n corresponds to the number of discrete labels. Figure 1 illustrates the additional noise layer on top of BERT.
We train the model to minimize the cross-entropy loss as follows:

$$L(y, y') = -\sum_{i=1}^{N} y_i \log(y'_i) \tag{2}$$

As the cross-entropy is minimized over the noisy training data, the values in the noise matrix Q are updated during error backpropagation to account for a noisy label distribution. Last, we apply L_2-regularization to the noise layer independently of the rest of the model.

3.2. Label Confidence Model

To take into account a confidence value associated with each training sample, we propose a noise layer whose weight matrix Q is a linear combination of two matrices Q_1 and Q_2. Similarly to the label noise model described above, the matrices Q_1 and Q_2 are initialized to identity matrices. Given the two matrices, the noise model output is calculated as follows:

$$Q = (cQ_1 + (1 - c)Q_2) \tag{3}$$

$$p(y'|x) = \textit{softmax}(Q \times \hat{y}) \tag{4}$$

where c is a confidence value in $[0, 1]$ for each training sample x.
For training samples with high label confidence, that is, confidence values close to 1, the weight matrix Q_1 contributes more to the layer weights. Conversely, for samples with low confidence, the weight matrix Q_2 contributes more. Intuitively, during training, the matrix Q_1 can stay close to the identity matrix while Q_2 can learn more of the noise distribution so that the base model outputs for samples with high confidence are not transformed as much as the outputs for samples with low confidence.
We calculate the confidence value c for each training sample by normalizing the scores used to rank the paraphrase candidates in Opusparcus. We normalize the scores so that the highest ranking sample in our final training set has a confidence value of 1 and the last sample has a confidence value of 0.

4. Experiments

We perform systematic experimentation of the proposed noise layer augmented models to evaluate their robustness to increasing proportions of noisy labels in the training data. The noise layer is implemented on a BERT model for sequence classification using the Hugging Face transformers library (Wolf et al., 2020). The pre-trained models we use are: Devlin et al. (2019) for English, German BERT by deepset[2] for German, Virtanen et al. (2019) for Finnish, Kuratov and Arkhipov (2019) for Russian, and Malmsten et al. (2020) for Swedish.
All BERT models, with or without the added noise layers, are fine-tuned using an early stopping criterion with patience of 10. The learning rate is set to 1e-5, and the models are regularized with a weight decay of 0.1. In the training of all noise models, we additionally regularize the noise layer with a weight decay of 0.01. Because the negative instances in our data are not noisy (see Section 4.2), we train the additional noise layers using only the positive instances. For the negative instances we use the base model output both during training and inference.

4.1. Data

We use five languages – German, English, Finnish, Russian and Swedish – from the Opusparcus paraphrase corpus to evaluate the proposed noise models.

[2]https://www.deepset.ai/german-bert

Index	Sentence 1	Sentence 2	Ranking score
8	It was a difficult and long delivery .	The delivery was difficult and long .	77.5163
2 501	He doesn 't know what he 's doing .	He has no idea what he 's doing .	60.5163
520 103	None of your business .	That was none of your damn business .	26.9842
1 000 589	He liked it .	She liked it .	22.1814
1 300 948	What 's this all about ?	Why do you need him ?	20.0698

Table 1: Examples from the Opusparcus English training data illustrating label noise. The first column indicates the location of a sentence pair in the data, after all sentence pairs have been sorted according to a PMI-based score shown in the fourth column. The higher the score, the more likely the sentences are paraphrases. Therefore, the most likely paraphrastic sentence pairs are located in the beginning of the training set followed by less likely paraphrastic sentence pairs in a decreasing order.

The corpus consists of automatically annotated training and manually annotated evaluation and test sets. The automatic annotation of the training sets is based on a PMI weighted probability of a given sentence pair consisting of two paraphrastic sentences. The most probable paraphrase pairs are positioned in the beginning of the training data, followed by less probable pairs in a descending order. Table 1 provides examples. Albeit having differences on their syntactic and lexical level, the first two sentence pairs carry the same approximate meanings, respectively, and occur early in the training data. The third sentence pair includes a difference in tone but still carries the same approximate meaning. In the fourth example, the surface forms differ only in the choice of the subjective pronoun. The sentences in the last example are unrelated. Collecting the first 1.5 million sentence pairs from the English training set and treating the collected sentence pairs as positive examples, that is, classifying them as paraphrases, results in a subset where the probability of acquiring incorrectly classified sentence pairs is approximately 10%. Thus, the proportion of noisy training examples can be controlled by selecting *n* examples from the beginning of the file so that the use of a larger *n* will produce a larger proportion of noisy labels. The format of the data enables a natural evaluation of the model robustness on different proportions of noisy labels in the training data. For a detailed description of the construction of the corpora, we refer the reader to Creutz (2018).

To evaluate the robustness of the models to the increasing proportions of noisy labels in the training data, we sample corresponding noise buckets for each language ranging from an estimated proportion of noisy labels between 5% and 40%. These buckets are generated by increasing the data sizes for each language independently so that the estimated proportion of noisy labels increases in increments of 5 percentage points. As a result, we produce eight training data subsets for each language that differ in size but match in their estimated noise distributions.

4.2. Data Sampling for Fine-Tuning

Each respective training data subset consists of an equivalent number (n) of positive and negative examples. The positive examples are always collected starting from the beginning of all available training data, meaning they comprise the n most likely paraphrases in the data. We use three different sampling strategies to obtain an equivalent number of negative examples to the positive examples, illustrated in Figure 2.

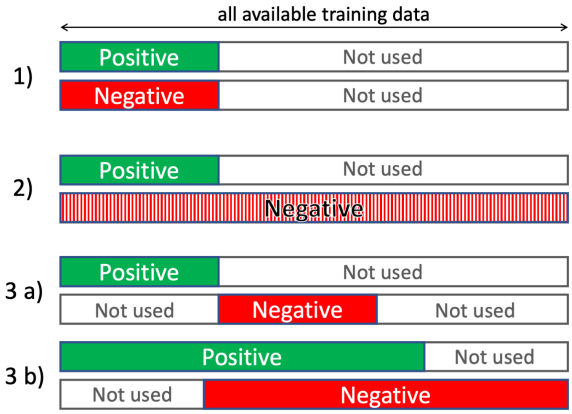

Figure 2: Our approaches for sampling negative training examples (sentence pairs that are not paraphrases).

First, we produce negative examples from the same subset as the positive ones by pairing sentences randomly within the set.

Second, we match the number of positive examples by randomly sampling a corresponding number of negative examples from all available training data.

Third, we sample the negative examples similarly to Sjöblom et al. (2018). We obtain a matching number of examples following the positive example subset from the training data and randomly shuffle the pairs (3a in Figure 2). If the sum of the positive and the required negative examples exceeds the number of total examples in the training data, we sample the missing pairs from the end of the subset used as positive examples and shuffle the pairs (3b).

As Sjöblom et al. (2018) note, these methods may in theory introduce false negatives to the training data, but the probability for this is expected to be so low that

it does not affect the models' capability to generalise well to unseen data. Note, since we expect the negative examples to be (predominantly) clean of noisy labels, the reported proportions of noisy labels correspond to the proportion of noisy labels in the positive examples, exclusively.

Based on an evaluation of the results obtained using the different sampling strategies for the negative samples, we did not observe a clear difference justifying the selection of one strategy over the other. Therefore, we report results that are averaged over the three negative sampling strategies.

4.3. Results from the Label Noise Model

Figure 3 presents the averaged results of the fine-tuned models on Opusparcus development sets over all negative sampling strategies.

The figures represent a comparison of three models: a fine-tuned BERT without an additional noise processing layer (BERT Basic), a fine-tuned BERT with the proposed label noise model (BERT Noise), and a fine-tuned BERT augmented with the proposed label confidence model (BERT Confidence).

Overall, BERT Noise obtains higher accuracy scores compared to BERT Basic except for German and for data selections with very small noise levels in English. The difference between both models increases with growing noise levels as expected. All noise models certainly drop in performance when noise is added but, especially for Finnish and Swedish, the performance stays high. In general, one would not know the noise level of the training data and the result is, hence, very encouraging in terms of training robustness. German seems to be an outlier in our experiments but even there the noise model does at least not hurt the performance except in the cleanest data sets with noise levels below 10%.

4.4. Results from the Label Confidence Model

Next, we evaluate the impact of the confidence information on the noise modeling performance. We train BERT Confidence using the same hyperparameters as before, introducing auxiliary confidence values to the model. The results show that our model is not able to pick up additional benefits from those values and the model behaves very similarly to BERT Noise. Alike BERT Noise, BERT Confidence outperforms BERT Basic in Finnish and Swedish on all noise levels, and in the other languages when trained on the noisier subsets. Similar behaviour between the two noise models indicates that the label noise model can effectively be utilised even when we do not have further knowledge about the noise level in the data and the reliability of individual labels.

5. Discussion

The results suggest that the additional noise processing layer is most beneficial in scenarios where the training data is especially noisy, and utilising the noise model does not affect performance in a deteriorating manner in other test conditions either. In most situations, reducing noise by selecting training data only from a certain distribution of the data is not possible, or information about the quality of the data is not known *a priori*. In situations where the training data is assumed or known to be noisy, utilising the noise model is recommended, as the results show that this simple addition makes the model considerably more robust to noise.

Across the five languages, the English results are the best in absolute terms, with accuracies over 90%. Figure 3 (bottom right) may provide an explanation, as it can be seen that English has more training data than any other language for each noise level. German, Finnish and Swedish all reach accuracies in the high 80's. If the amount of training data were the decisive factor, we would expect German to outperform Finnish, which should outperform Swedish by a slight margin. However, this is not what happens. Furthermore, the Russian results are the worst ones, with accuracies well below 80%, although the amount of training data is comparable to that of the other languages, except for noise levels 5 and 10%.

5.1. Effects of Negative Data Sampling and Unbalanced Classes

One of our observations is that all models, across languages and noise levels, predict the positive (paraphrase) class significantly more often than the negative (non-paraphrase) class. For instance, at the 10% noise level, BERT Basic predicts the positive class as follows: 78% (Russian), 77% (English), 76% (German), 69% (Finnish) and 64% (Swedish). This is not unexpected, due to the way the negative non-paraphrase training samples were created. Because the samples were created by randomly pairing sentences from the training set, we would expect there to be very few non-paraphrase pairs which are close in surface form, that is, the hardest cases for the model to recognize as non-paraphrases. This gives the model a signal that for a pair of sentences to be non-paraphrastic, the surface forms should probably have very little in common, and other pairs are likely to be paraphrases.

Furthermore, increasing the noise level of the training set exacerbates this problem. As the label noise increases, the models increasingly overpredict the paraphrase class. This results from the fact that our noise is asymmetric: we only have noisy samples where the assumed class is a paraphrase but the sample in reality is a non-paraphrase, further strengthening the signal that many different types of phrase pairs are paraphrastic. However, this problem is clearly worse for BERT Basic than for BERT Noise. For instance, when BERT Basic predicts 84% positives for Russian at the 40% noise level, BERT Noise only predicts 66% positives.

Based on these results, the difference between BERT Basic and BERT Noise seems to lie in that the addi-

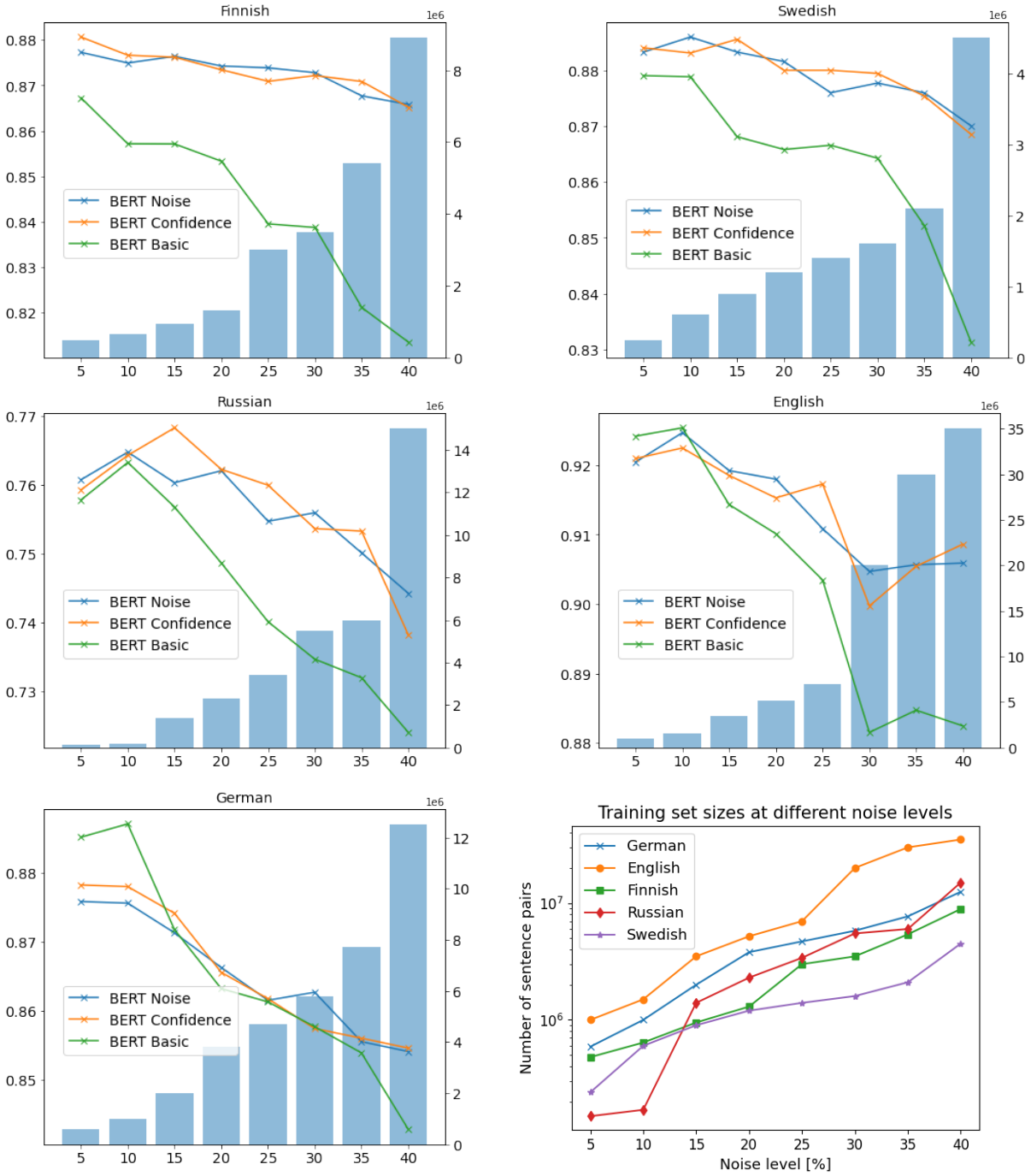

Figure 3: Paraphrase detection accuracies for models trained on data subsets with increasing noise levels. The noise levels [%] of the training sets are marked on the x axis. The bars indicate the corresponding training set sizes (labels on the secondary y axis on the right). *Bottom right:* Comparison across languages between the increasing noise level presented on the x axis, and the growing training set sizes on a logarithmic scale on the y axis.

tional noise layer alleviates the second problem resulting from the addition of asymmetric noise, confirming that the model indeed learns to correct for noise during training. However, the model cannot naturally correct for the sampling strategy used to create the non-paraphrase samples, so the problem still persists for BERT Noise to some extent.

This also construes the differences between the languages. The development sets used for testing are not balanced in terms of classes. Instead, all languages have more paraphrases than non-paraphrases in the data. English and German have the largest proportions of paraphrases in their development sets (75% and 73% respectively). Hence, it is not a bad strategy to overpredict paraphrases. As the amount of noise in the training data grows, the increasing overprediction of positives

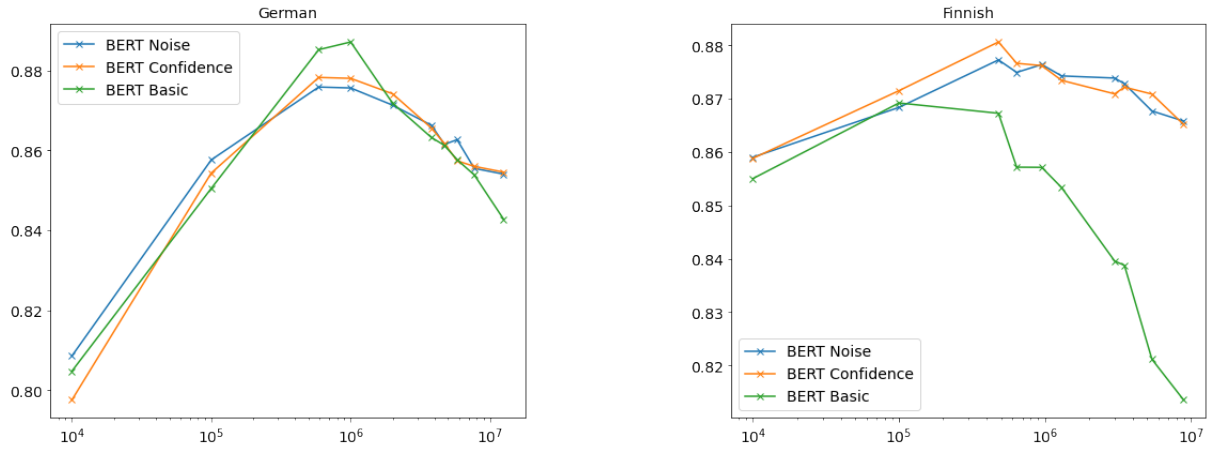

Figure 4: Paraphrase detection accuracies for all models of German and Finnish for all training set sizes after the addition of two smaller sets: 10,000 and 100,000 sentence pairs. The plots now include the same data points as in Figure 3 plus the two added ones. Note that the x axis shows the size of the training set (on a logarithmic scale) rather than the noise levels, as in the earlier figures.

by BERT Basic compensates for the overall decrease in performance, resulting in a smaller difference in performance between the BERT Basic and the BERT Noise models. This is especially true for German.

By contrast, the benefit of the BERT Noise models is the most prominent for languages where the class distribution is closer to uniform and the overprediction of positives degrades performance. This shows in the results for Finnish and Swedish, where the development sets contain no more than 61% and 57% positives, respectively. At the highest noise level, 40%, the BERT Noise models accurately predict positives in very similar proportions: 62% (Finnish) and 59% (Swedish). BERT Basic is far behind, with an overprediction of positives: 75% (Finnish) and 70% (Swedish).

5.2. Smaller Models

Another open question is the behaviour of models on even smaller training sets. To study this, we extracted subsets below the 5% noise level. Since we are unable to accurately estimate the noise level of the small sets, we decided to use the same absolute number of training examples for every language: 10,000 and 100,000 paraphrastic sentence pairs (plus an equal number of negative samples). We do not know the exact noise levels, but we expect the data to be quite clean with noise levels below 5%.

The results for the smaller data sets are similar across the languages. Figure 4 adds their performance for two languages that otherwise behave quite differently from each other. We note that in absolute terms the accuracies are lower for the smallest sizes, which is not surprising. Small amounts of training data do not outperform the use of larger, slightly noisier sets. Also, the two noise models are very similar to the BERT Basic model, which is to be expected, since the data does not contain much noise.

We also notice that the accuracy for the smallest set is higher for Finnish (86%) than for German (80%). This could be caused by differences in the qualities of the underlying BERT models, as the amount of data used for fine-tuning is still so limited. However, another explanation may be the asymmetry between the positive and negative classes. At this level, the models do not yet overpredict the positive class as strongly as for larger training sets, which is beneficial for Finnish with its more balanced distribution of labels in the test data.

6. Conclusion

We have proposed the integration of a linear noise prediction layer into the fine-tuning step of a pre-trained language model (BERT) to implicitly account for noisy labels in the training data. We have shown that the noise model contributes to the training procedures in a way that alleviates the deteriorating effect of unknown label noise. We test the approach on a paraphrase detection task showing that the model increases robustness and stability with improved performance for four out of five languages included in our experiments.

7. Acknowledgements

This work is part of the Behind the words project, funded by the Academy of Finland. We wish to acknowledge *CSC – The Finnish IT Center for Science* for the generous computational resources they have provided.

8. Bibliographical References

Araque, O., Gatti, L., Staiano, J., and Guerini, M. (2019). Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques. *IEEE Transactions on Affective Computing*, pages 1–1.

Aulamo, M., Creutz, M., and Sjöblom, E. (2019). Annotation of subtitle paraphrases using a new web tool. In Costanza Navarretta, et al., editors, *Digital Humanities in the Nordic Countries*, number 2364 in CEUR Workshop Proceedings, pages 33–48, Germany, May. CEUR-WS.org. Digital Humanities in the Nordic Countries, DHN 2019 ; Conference date: 05-03-2019 Through 08-03-2019.

Bayerl, P. S. and Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725, December.

Berthon, A., Han, B., Niu, G., Liu, T., and Sugiyama, M. (2021). Confidence scores make instance-dependent label-noise learning possible. In *International Conference on Machine Learning*, pages 825–836. PMLR.

Creutz, M. (2018). Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Freitag, M., Grangier, D., and Caswell, I. (2020). BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online, November. Association for Computational Linguistics.

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

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2017). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June. Association for Computational Linguistics.

Haque, A., Reddi, V., and Giallanza, T. (2021). Deep learning for suicide and depression identification with unsupervised label correction. In Igor Farkaš, et al., editors, *Artificial Neural Networks and Machine Learning – ICANN 2021*, pages 436–447, Cham. Springer International Publishing.

Jindal, I., Nokleby, M. S., and Chen, X. (2016). Learning deep networks from noisy labels with dropout regularization. In Francesco Bonchi, et al., editors, *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*, pages 967–972. IEEE Computer Society.

Jindal, I., Pressel, D., Lester, B., and Nokleby, M. (2019). An effective label noise model for DNN text classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3246–3256, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Kuratov, Y. and Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for russian language.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, May. European Language Resources Association.

Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., and Li, L. (2017). Learning from noisy labels with distillation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1928–1936, Los Alamitos, CA, USA, oct. IEEE Computer Society.

Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q., and Toral, A. (2020). A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67, Mar.

Malmsten, M., Börjeson, L., and Haffenden, C. (2020). Playing with words at the national library of sweden – making a swedish bert.

Mnih, V. and Hinton, G. (2012). Learning to label aerial images from noisy data. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 203–210, Madison, WI, USA. Omnipress.

Mohammad, S. (2012). #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Patrini, G., Rozza, A., Menon, A. K., Nock, R., and Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2233–2241.

Sjöblom, E., Creutz, M., and Aulamo, M. (2018). Paraphrase detection on noisy subtitles in six languages. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 64–73, Brussels, Belgium, November. Association for Computational Linguistics.

Strapparava, C. and Mihalcea, R. (2010). Annotating and identifying emotions in text. In G. Armano, et al., editors, *Intelligent Information Access*, volume 301 of *Studies in Computational Intelligence*. Springer, Berlin, Heidelberg.

Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., and Fergus, R. (2015). Training convolutional networks with noisy labels.

Toral, A., Castilho, S., Hu, K., and Way, A. (2018). Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium, October. Association for Computational Linguistics.

Vahtola, T., Creutz, M., Sjöblom, E., and Itkonen, S. (2021). Coping with noisy training data labels in paraphrase detection. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 291–296, Online, November. Association for Computational Linguistics.

Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., and Belongie, S. (2017). Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: Bert for finnish.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

9. Language Resource References

Mathias Creutz. (2018). *Opusparcus: Open Subtitles Paraphrase Corpus for Six Languages (version 1.0)*. Kielipankki.