Modeling Noise in Paraphrase Detection

Teemu Vahtola, Eetu Sjöblom, Jörg Tiedemann, Mathias Creutz

INTRODUCTION

1. Paraphrasing and paraphrase detection
2. Label noise
3. Data used in our experiments
4. Proposed models
5. Results and analysis
6. Conclusion

PARAPHRASE DETECTION

- Paraphrasing – Conveying some given meaning in a different wording.
- Paraphrase detection – Identifying whether two phrases essentially mean the same thing.

PARAPHRASE DETECTION

- ✓ It was a difficult and long delivery. – The delivery was difficult and long.
- ✓ He doesn't know what he's doing. – He has no idea what he's doing.
- ✓ None of your business. – That was none of your damn business.
- ✗ He liked it. – She liked it.
- ✗ What's this all about? – Why do you need him?

LABEL NOISE

- ✓ It was a difficult and long delivery. – The delivery was difficult and long.
- ✓ He doesn't know what he's doing. – He has no idea what he's doing.
- ✓ None of your business. – That was none of your damn business.
- ✓ He liked it. – She liked it.
- ✓ What's this all about? – Why do you need him?

DATASET

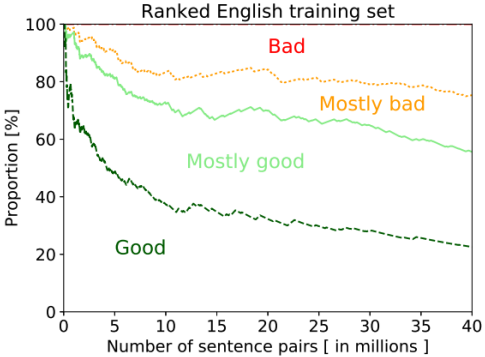
- Opusparcus (Creutz, 2018)
  - Collection of sentential paraphrases in six languages.
  - Training sets are automatically constructed and consist of millions of sentence pairs.
  - Evaluation and test sets are annotated by hand.
  - Available in the GEM benchmark (Gehrmann et al., 2021) via the Huggingface datasets library.

DATASET

- Sentence pairs are ranked based on a probabilistic score so that the most probable paraphrases occur in the beginning of the data, followed by less probable paraphrases in a descending order.

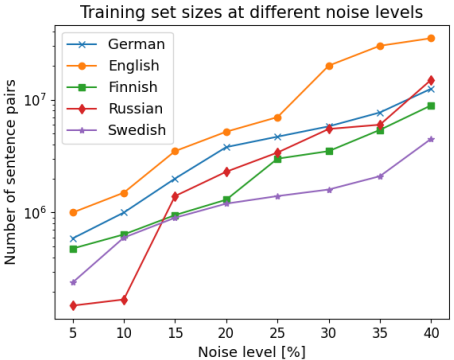
Index	Sentence 1	Sentence 2	Ranking score
8	It was a difficult and long delivery .	The delivery was difficult and long .	77.5163
2 501	He doesn ’t know what he ’s doing .	He has no idea what he ’s doing .	60.5163
520 103	None of your business .	That was none of your damn business .	26.9842
1 000 589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

TRAINING DATA



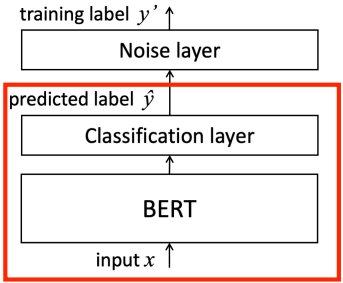
- Based on hand-labeled subset of the data, it is possible to approximate the proportion of noisy labels in a selected subset of the training data (Creutz, 2018).

TRAINING DATA



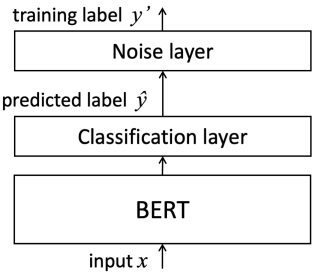
- We perform paraphrase detection for five languages: English, Finnish, German, Russian, and Swedish.
- For each language, we collect eight subsets of the training data based on a 5% increments in noisy label proportions.
- The proportional subsets are available in the GEM benchmark.
- We pair the assumed paraphrases with the same number of randomly shuffled negative examples.

BASE MODEL



- Base model – Pre-trained BERT with a sequence classification layer.
- We use language-specific BERT-base models from the Huggingface transformers library.

LABEL NOISE MODEL

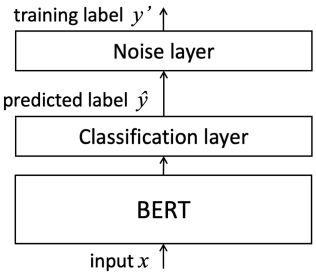


The **Label noise model** extends the base model with a linear noise layer (Jindal et al. (2019)) to transform base model outputs to noisy labels:

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

where Q is initialised as an identity matrix: $I^{2\times 2}$.

LABEL CONFIDENCE MODEL



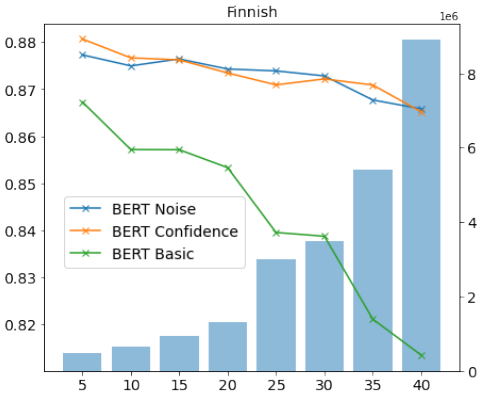
The **Label confidence model** adds auxiliary confidence values to the Label noise model to guide the transformation to noisy labels further:

$$Q = (cQ_1 + (1 - c)Q_2) \quad (2)$$

$$p(y'|x) = softmax(Q \times \hat{y}) \quad (3)$$

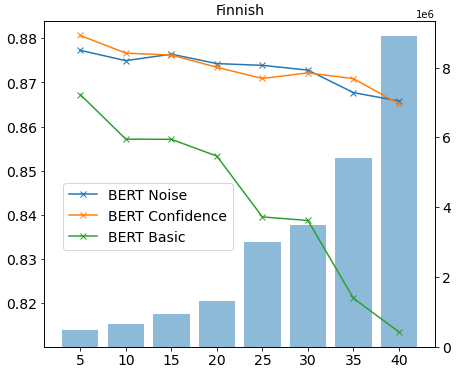
Confidence value c is normalized from the ranking scores so that $c = 1$ for the most probable sentence pair and $c = 0$ for the least probable sentence pair in the data.

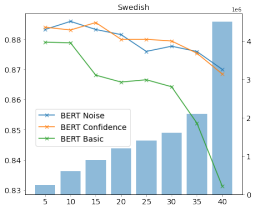
RESULTS

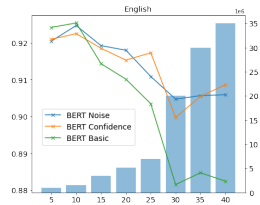


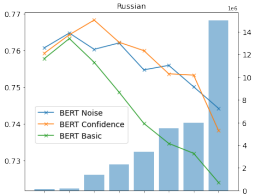
- BERT without noise layer (BERT Basic) collapses when the proportion of noisy data increases.
- The noise models can maintain higher accuracy even in considerable proportions of noisy data.
- The Label Noise Model does not require additional confidence values.

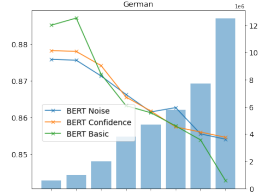
RESULTS


Finnish
1e6
BERT Noise
BERT Confidence
BERT Basic

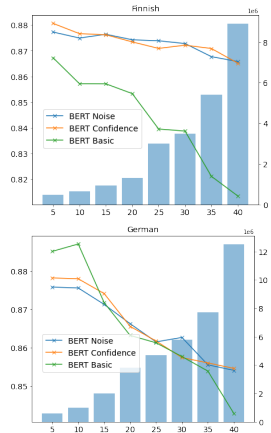

Swedish
1e6
BERT Noise
BERT Confidence
BERT Basic


English
1e6
BERT Noise
BERT Confidence
BERT Basic


Russian
1e6
BERT Noise
BERT Confidence
BERT Basic

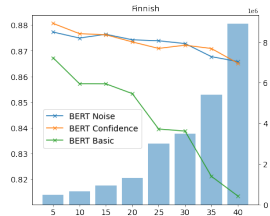

German
1e6
BERT Noise
BERT Confidence
BERT Basic

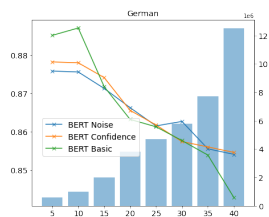
ANALYSIS



- We notice that all the models overpredict the positive class.
- Randomly shuffled negative examples have very different surface forms.
- To be a paraphrase, sentences need only a little in common.
- Adding more noisy data exacerbates this effect by adding more different types of positive examples into the training.

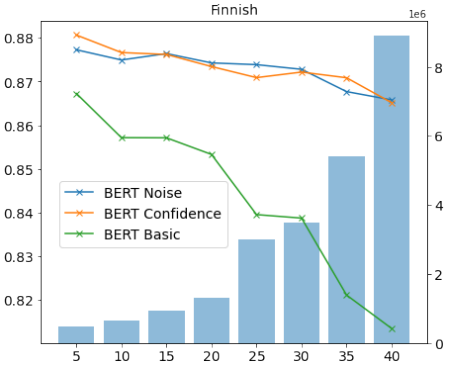
ANALYSIS





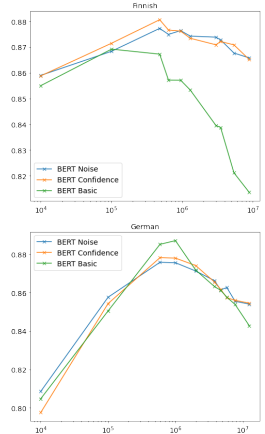
- The development sets are not balanced in terms of classes.
- For example, the German development set consists of 74% of positive classes.
- As the noise increases, the overprediction of positives compensates for the overall decrease in performance for BERT Basic, especially in German.

ANALYSIS



- The noise models alleviate the overpredicting problem.
- The Finnish development set consists of 61% of positives.
- BERT Noise predict 62% of positives for Finnish when trained with the noisiest training data subset.
- BERT Basic is far behind, overpredicting 75% of positives for Finnish trained on the same subset.

ANALYSIS



- We experiment with smaller training sets, where we expect the noisy label proportion to be below 5%.
- Training on larger, slightly noisier data outperforms training on really small but clean data.
- The small models behave similarly, because the data does not contain much noise.

CONCLUSIONS

- Integrating the noise model layer on top of a large pre-trained language model during fine-tuning alleviates the deteriorating effect of unknown label noise in the training data.
- Assessed on paraphrase detection, the model increases robustness and stability and improves results on four out of five languages included in our experiments.

Thank you for your attention!