Coping with Noisy Training Data Labels in Paraphrase Detection

Teemu Vahtola and **Mathias Creutz** and **Eetu Sjöblom** and **Sami Itkonen**
{teemu.vahtola, mathias.creutz, eetu.sjoblom, sami.itkonen}@helsinki.fi
Department of Digital Humanities
Faculty of Arts
University of Helsinki

Abstract

We present new state-of-the-art benchmarks for paraphrase detection on all six languages in the Opusparcus sentential paraphrase corpus: English, Finnish, French, German, Russian, and Swedish. We reach these baselines by fine-tuning BERT. The best results are achieved on smaller and cleaner subsets of the training sets than was observed in previous research. Additionally, we study a translation-based approach that is competitive for the languages with more limited and noisier training data.

1 Introduction

Paraphrase detection is a task where a model is trained to recognize from a collection of phrases, whether two phrases carry approximately the same meaning. Paraphrase detection provides a natural way to study how well models operate on semantic abstractions by recognizing similar meanings from syntactically and lexically different sequence pairs. Models that are capable of operating on such a semantic level are useful in a number of Natural Language Processing tasks, such as Information Retrieval or Machine Translation, by providing semantically coherent alternative queries or translation hypotheses for a sequence of words.

In this paper, we study paraphrase detection trained on noisy data in six European languages. We train paraphrase detection models by fine-tuning BERT (Devlin et al., 2018) and explore the effect of noisy labels in the training data. We reach new paraphrase detection benchmarks on all six languages in the Opusparcus (Creutz, 2018) sentential paraphrase corpus, a collection of paraphrase pairs extracted from the OpenSubtitles2016 (Lison and Tiedemann, 2016) movie and television subtitle collection. The new benchmarks are obtained with considerably smaller and cleaner subsets of the training set than was observed in previous research. By exploring different proportions of noisy labels in the training data, we identify a level at which fine-tuned BERT models still appear to be robust to the noisy labels. We also present a competitive translation-based approach that is especially useful for languages with more limited and noisier paraphrase data in our experimental setup.

Opusparcus contains training, development and test sets in English (en), Finnish (fi), French (fr), German (de), Russian (ru), and Swedish (sv). For each language, the development and test sets consist of a few thousand sentence pairs that have been annotated manually. The annotators have indicated on a four-grade scale to which extent two proposed sentences mean the same thing. The training sets are considerably larger, comprising millions of sentence pairs, but lack manual annotations. However, the training data have been ordered in such a way that the sentence pairs that are most likely to be true paraphrases are placed first and the pairs of sentences that are least likely to carry the same meaning are placed last. This ranking is produced automatically, according to metrics based on point-wise mutual information. Consequently, when training a paraphrase detection model on these training sets, one needs to decide how much of the available training data to use, weighing the benefits of a smaller and cleaner set versus a larger but noisier set.

Paraphrase detection based on Opusparcus has previously been studied by Sjöblom et al. (2018), who train two different neural network models to represent the phrases and apply a cosine distance-based metric to detect the paraphrases. The authors observe that the networks are rather robust to noise and they can benefit from more data, albeit increase in data size results in increased amounts of noise within the training data. Whereas Sjöblom et al. (2018) use neural networks trained from scratch, we approach paraphrase detection by fine-tuning language-specific BERT models for sequence classification.

Further previous explorations of Opusparcus include the fine-tuning of pre-trained BERT models

Proceedings of the 2021 EMNLP Workshop W-NUT: The Seventh Workshop on Noisy User-generated Text, pages 291–296

for the purpose of word similarity assessment in English and Finnish (Garí Soler and Apidianaki, 2020) as well as German text summarization (Paraschiv and Cercel, 2020). Pragst et al. (2020) compare different classifiers in different paraphrasing-related tasks on different corpora, including Opusparcus. Fabre et al. (2021) investigate automatic paraphrase generation on Opusparcus as well as other datasets. None of these previous works, however, establish new benchmarks for paraphrase detection on the Opusparcus test sets.

An alternative to using large neural language models, such as BERT, is to leverage the capability of machine translation systems instead. Wieting et al. (2017) and Iyyer et al. (2018) create paraphrases using massive backtranslation; from a parallel corpus of two languages (bitext), paraphrases in the first language are obtained by automatically translating the corresponding sentence in the second language back to the first language. In our proposed translation-based approach for paraphrase detection, we do not need bitexts. We simply use existing machine translation models to translate our paraphrase candidates in one language (for instance, English) to another language (such as Spanish) and see whether the two source sentences (in English) have common translations in the target language (Spanish), in which case they are considered to be paraphrases. However, we need to take into consideration that machine translation is not perfect and produces noisy results. We discuss our methods in detail in Section 2. The results of our experiments are presented in Section 3, and we conclude in Section 4.

2 Experiments

We evaluate the performance of language-specific BERT models on paraphrase detection based on differently sized subsets consisting of different proportions of noise sampled from the Opusparcus training data. We also evaluate our translation-based model on the paraphrase detection task.

2.1 General Setting

All our experiments involving BERT are conducted using the Hugging Face Transformers (Wolf et al., 2020) library. For each language, we choose a corresponding model from the collection of pre-trained transformers. The language-specific models we use are *bert-base-cased* (Devlin et al., 2018), *TurkuNLP/bert-base-finnish-cased-v1* (Virtanen et al., 2019), *flaubert/flaubert_base_cased* (Le et al., 2020), *bert-base-german-cased* (Chan et al., 2020), *DeepPavlov/rubert-base-cased* (Kuratov and Arkhipov, 2019), and *KB/bert-base-swedish-cased* (Malmsten et al., 2020) for English, Finnish, French, German, Russian, and Swedish respectively. We fine-tune the models with an additional sequence classification layer for three epochs and choose the best performing model based on results achieved on the development set. The models are trained with 1e-5 learning rate and 0.3 weight decay.

For the machine translation experiments we use pre-trained translation models from the OPUS-MT project (Tiedemann and Thottingal, 2020).

	de	en	fi	fr	ru	sv
1M	90	97	83	95	85	85
2M	87	90	80	90	85	65
5M	**75**	80	**65**	80	**70**	**60**
10M	60	75	60	75	60	45
20M	50	**70**	45	**60**	55	–
30M	–	60	–	–	–	–

Table 1: Estimates of proportions of correctly labeled paraphrases [%] in growing subsets of the Opusparcus training sets.[1] The figures in boldface show the approximate optimal data sizes and their corresponding noise levels reported by Sjöblom et al. (2018) for their recurrent neural network model.

2.2 Data selection for BERT fine-tuning

We sample our training examples similarly to Sjöblom et al. (2018), who picked a certain number of *assumed* positive examples of paraphrases from the beginning of the training sets, and sampled an equal number of assumed negative examples by randomly pairing sentences. Sjöblom et al. (2018) experimented with data sets containing between one and thirty million positive examples. The level of correct labels in these subsets of data are shown in Table 1, which has been compiled from values reported by Creutz (2018) and Sjöblom et al. (2018). The reported best models were trained on data sets containing between 60 % and 80 % of presumably correctly labeled paraphrases. Thus, the level of noise was fairly high, between 20 and 40 percent.

In the present work, we study models fine-tuned on up to five million positive examples, but we

[1] See pages 12–13 of the presentation slides at https://helda.helsinki.fi//bitstream/handle/10138/237338/creutz2018lrec_slides.pdf

also introduce smaller sets of 100 000 and 500 000 positive examples. These 100k and 500k sets are expected to contain very few incorrectly labeled paraphrases.

We use the Opusparcus development sets to systematically study how well our models perform on different sizes of training data with different proportions of label noise. The final results are reported on the Opusparcus test sets, and only for the best models, according to the development sets. The development and test sets have been annotated by hand, and are not expected to include noisy labels.

2.3 Translation-based method

In order to decide whether two sentences carry the same meaning, one could ideally investigate whether they have a common translation in another language. We were curious to find out whether such an approach could work for Opusparcus, in particular for the languages with more limited and noisier training sets available (Swedish, Finnish, and Russian).

In these experiments, we employ existing neural machine translation models to translate the Opusparcus development and test sets. No further training takes place, so the training sets are not necessary in this setup. For a pair of sentences we create n-best lists of the 1000 most likely translations into a second language. If the two source sentences have translations in common, we infer that the sentences are likely to be paraphrases.

The machine translation models have been trained on large amounts of parallel corpora, including bitexts from the OpenSubtitles corpus. The sentence pairs extracted into Opusparcus are also based on sentence alignments across languages in OpenSubtitles. Therefore, there is a risk that Opusparcus test data might have been included in the training of the machine translation models, which would lead to unreliable results. In our current experiments, we have avoided this risk by translating the Opusparcus languages only to languages excluded from Opusparcus: Spanish (es), Dutch (nl) and Polish (pl). This guarantees that no bitext used in the training of the machine translation models has been included in any phase of the creation of Opusparcus.[2]

It is conceivable to think that for two sentences to be paraphrases it is sufficient for them to have one translation in common. In practice, this is not the case. Firstly, the concept of having the same meaning is about degrees of similarity rather than a discrete dichotomy. Secondly, as machine translation systems are not perfect and produce noisy results, the fact that we can find common translations is not a guarantee for the sentences to be paraphrases. Within the list of the 1000 most likely translations, we might find truncated translations, which do not cover the entire source sentence. Thus, two sentences that clearly mean different things can still have translations in common, typically such "simplified" or "summarized" fragments.

We have used the Opusparcus development sets, separately for each language, to find the optimal level of overlap in the translated n-best lists. Two sentences are considered to be paraphrases if they have a higher overlap in the number of shared translations than this inferred threshold value.

As an addition, which further improves performance, we use the development sets to determine for each source language to which of the target languages (Spanish, Dutch, Polish) one should translate in order to obtain the best paraphrase detection performance.

Final results of the translation-based approach are computed on the test sets for only the optimal target language and the corresponding optimal overlap threshold.

3 Results

The performance of our fine-tuned BERT models with respect to different amounts of data and noise are presented in Figure 1. The results show that while the baseline models by Sjöblom et al. (2018) benefit from adding more and noisier data, this is not the case with BERT to the same extent. We notice three types of behaviour of BERT on different languages. With Finnish and Swedish, the languages with most limited training data, BERT obtains high results with a small and clean training subset, and adding more noisy data is only detrimental to the model. With the higher-resource languages, French, German and Russian, we notice an initial increase in performance when adding more data. The models peak at 1 million positive examples and start to deteriorate when more noisy data is added. The results of the highest-resource

[2]Note that this is not a problem for the BERT models that we use for fine-tuning, even though two of them (French and German BERT) have been trained partly on OpenSubtitles data. This is because these BERT models are based on monolingual data rather than bitexts, so there are no "bridges" between potential paraphrases via another language.

	de	en	fi	fr	ru	sv
BERT	**88.9 (1)**	**93.3 (0.5)**	**85.0 (0.1)**	**82.6 (1)**	**77.8 (1)**	**87.7 (0.1)**
Translation	85.1 (pl)	89.5 (pl)	84.9 (nl)	73.4 (pl)	75.0 (es)	85.8 (nl)
Sjöblom et al. (2018)	86.7 (4)	92.1 (20)	82.5 (3.5)	77.9 (22)	70.3 (5)	82.1 (5)

Table 2: Paraphrase detection accuracies [%] on the test sets of the six languages. For BERT and the model by Sjöblom et al. (2018), the figures within brackets indicate the amount of training data used (in millions of sentence pairs assumed to be positive examples of paraphrases). For the translation based model, the brackets indicate which target language produced the best translations for parapahrase detection.

language in the Opusparcus data, English, first increase and then plateau. It seems that the model can reach high performance already on 500 000 positive examples and adding more data when fine-tuning does not result in notable amelioration of the model. However, in our experiments with English, the proportion of noise does not exceed 20 %. Common for all models is that they reach the best performance already before or at 1 million positive examples. Only a fraction of the number of training examples required to train the models of Sjöblom et al. (2018) are necessary for significantly better performance when using BERT.

The final test results are presented in Table 2. The best results in all languages are obtained by fine-tuning BERT. However, we reach fairly competitive results with the translation models especially in the languages with less and noisier training data, Finnish, Swedish and Russian, which outperform the previous results by Sjöblom et al. (2018). Translating into Polish turned out to produce the best results for the high-resource languages (German, English, French), whereas Dutch was the best target language for Finnish and Swedish. As there were no translation models available from Russian to Dutch and Polish, the only option for Russian was to translate into Spanish.

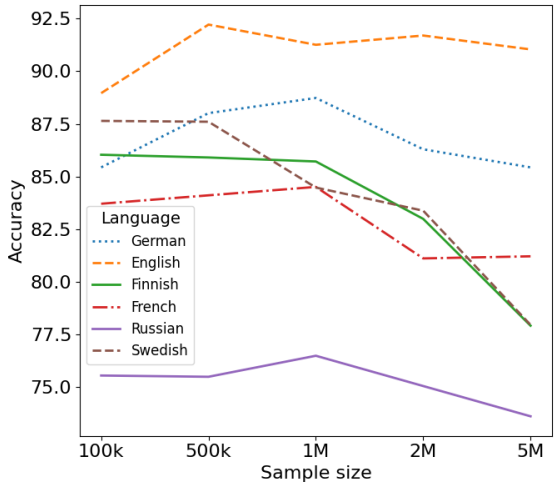

Figure 1: Paraphrase detection accuracies [%] produced by fine-tuning BERT on the Opusparcus development sets of the different languages. We explore training set sizes ranging from 100 000 to 5 million positive sentence pairs.

4 Conclusion

We have presented a new state-of-the-art benchmark based on BERT for six Opusparcus languages. Our experiments show that when fine-tuning BERT for paraphrase detection, we need only a small portion of the number of training examples that are required for training a recurrent neural network from scratch. In fact, it is detrimental to extend the amount of training data too far, as this causes the proportion of noise to go up. We find that BERT is considerably less robust to noise than the neural network models presented by Sjöblom et al. (2018). For some languages (English, Finnish and Russian) BERT appears to be robust with up to 20 % of noisy labels in the training set, whereas the experiments on French, German and Swedish indicate lower tolerance to noise in the data. Previous research on BERT (Tänzer et al., 2021) has claimed that even a considerable amount of noise in labels does not result in large performance degredations. Our findings are that we reach that point at a noise level of 20 % maximum.

In addition, we have shown that competitive results for paraphrase detection can be reached with a rather simple translation-based approach. In three languages, we were able to outperform the previous recurrent neural network-based benchmarks, and for one language (Finnish) we reached results that are almost on par with results obtained by fine-tuning BERT. This suggests that a translation-based model can be useful in situations where enough high-quality paraphrase data for fine-tuning BERT is lacking, or when an appropriate BERT model does not exist.

In future work we intend to explore means of further mitigating the effects of noisy training data. We could also study why certain target languages work better than others in the translation-based approach. Future work could also include an explicit study of the Opusparcus training set in order to identify possible other features in addition to noisy labels that affect the model performance when increasing training data. Due to high computational resources needed for training large models, we limit the number of positive examples to five million. Future work could include training a bigger English model to detect when a similar performance drop to the other languages occurs in English. Furthermore, we are interested in studying paraphrases in scenarios, where meaning is highly context-dependent.

Acknowledgments

We wish to gratefully acknowledge the Academy of Finland for their support for this research. We are thankful to *CSC — the Finnish IT Center for Science* for the generous computational resources they have provided.

References

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Betty Fabre, Tanguy Urvoy, Jonathan Chevelu, and Damien Lolive. 2021. Neural-driven search-based paraphrase generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2100–2111, Online. Association for Computational Linguistics.

Aina Garí Soler and Marianna Apidianaki. 2020. MULTISEM at SemEval-2020 task 3: Fine-tuning BERT for lexical meaning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 158–165, Barcelona (online). International Committee for Computational Linguistics.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of NAACL-HLT 2018*, pages 1875–1885, New Orleans, Louisiana.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT : des modèles de langue contextualisés pré-entraînés pour le français (FlauBERT : Unsupervised language model pre-training for French). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pages 268–278, Nancy, France. ATALA et AFCP.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden – making a swedish bert.

Andrei Paraschiv and Dumitru-Clementin Cercel. 2020. UPB at GermEval-2020 Task 3: Assessing summaries for German texts using BERTScore and Sentence-Bert. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, Zurich, Switzerland (Online).

Louisa Pragst, Wolfgang Minker, and Stefan Ultes. 2020. Comparative study of sentence embeddings for contextual paraphrasing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6841–6851, Marseille, France. European Language Resources Association.

Eetu Sjöblom, Mathias Creutz, and Mikko Aulamo. 2018. Paraphrase detection on noisy subtitles in six languages. In *Proceedings of the 2018 EMNLP Workshop W-NUT*, pages 64–73, United States. The Association for Computational Linguistics. Workshop on Noisy User-generated Text, W-NUT ; Conference date: 01-11-2018 Through 01-11-2018.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Michael Tänzer, Sebastian Ruder, and Marek Rei. 2021. Bert memorisation and pitfalls in low-resource scenarios.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish.

John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings-from back-translated bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 274–285, Copenhagen, Denmark.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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. Association for Computational Linguistics.