Fast, Effective and Self-Supervised: Transforming Masked Language Models into Universal Lexical and Sentence Encoders

Fangyu Liu, Ivan Vulić, Anna Korhonen, Nigel Collier
Language Technology Lab, TAL, University of Cambridge
{fl399, iv250, alk23, nhc30}cam.ac.uk

Abstract

Pretrained Masked Language Models (MLMs) have revolutionised NLP in recent years. However, previous work has indicated that off-the-shelf MLMs are not effective as universal lexical or sentence encoders without further task-specific fine-tuning on NLI, sentence similarity, or paraphrasing tasks using annotated task data. In this work, we demonstrate that it is possible to turn MLMs into effective universal lexical and sentence encoders even without any additional data and without any supervision. We propose an extremely simple, fast and effective contrastive learning technique, termed Mirror-BERT, which converts MLMs (e.g., BERT and RoBERTa) into such encoders in less than a minute without any additional external knowledge. Mirror-BERT relies on fully identical or slightly modified string pairs as positive (i.e., synonymous) fine-tuning examples, and aims to maximise their similarity during *"identity fine-tuning"*. We report huge gains over off-the-shelf MLMs with Mirror-BERT in both lexical-level and sentence-level tasks, across different domains and different languages. Notably, in the standard sentence semantic similarity (STS) tasks, our self-supervised Mirror-BERT model even matches the performance of the task-tuned Sentence-BERT models from prior work. Finally, we delve deeper into the inner workings of MLMs, and suggest some evidence on why this simple approach can yield effective univeral lexical and sentence encoders.

1 Introduction

Transfer learning with pretrained Masked Language Models (MLMs) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) has been extremely successful in NLP, offering unmatched performance in a large number of NLP tasks (Wang et al., 2019). Despite the wealth of semantic knowledge stored in the MLMs (Rogers et al., 2020), they do not produce high-quality lexical and sentence
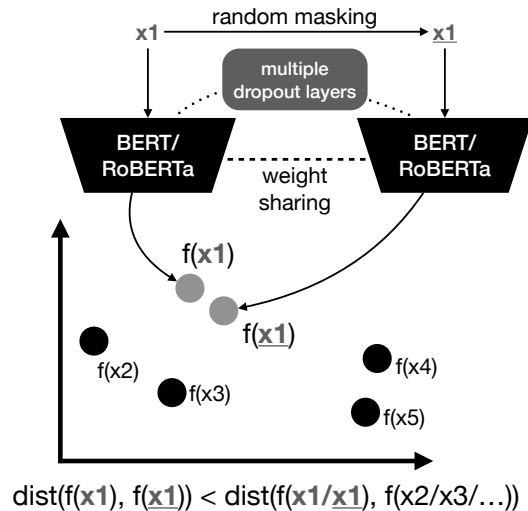

Figure 1: Illustration of the main concepts behind the proposed self-supervised Mirror-BERT method. The same text sequence can be observed from two additional "views" by: 1) performing random masking in the input space, and/or 2) by applying dropout (inside the BERT/RoBERTa MLM) in the feature space, yielding identity-based (i.e., "mirrored") positive examples for fine-tuning. A contrastive learning objective is then applied to encourage such "mirrored" positive pairs to obtain more similar representations in the embedding space relatively to negative pairs.

embeddings when used off-the-shelf, without further task-specific fine-tuning (Feng et al., 2020). In fact, previous work has shown that their performance is sometimes even below static word embeddings and specialised sentence encoders (Cer et al., 2018) in lexical and sentence-level semantic similarity tasks (Reimers and Gurevych, 2019; Vulić et al., 2020b; Litschko et al., 2021).

In order to address this gap, recent work has trained dual-encoder networks on labelled external resources to convert MLMs into universal language encoders. Most notably, Sentence-BERT (Reimers and Gurevych, 2019) further trains BERT and RoBERTa on Natural Language Inference (NLI) (Bowman et al., 2015; Williams et al., 2018) and

sentence similarity data (Cer et al., 2017) to obtain high-quality universal sentence embeddings. Recently, SapBERT (Liu et al., 2020) self-aligns phrasal representations of the same meaning using synonyms extracted from the UMLS (Bodenreider, 2004), a large biomedical knowledge base, obtaining lexical embeddings in the biomedical domain that reach state-of-the-art (SotA) performance in biomedical entity linking tasks. However, both Sentence-BERT and SapBERT require annotated (i.e., human-labelled) data as external knowledge: it is used to instruct the model to produce similar representations for text sequences (e.g., words, phrases, sentences) of similar/identical meanings.

In this paper, we fully dispose of any human supervision, demonstrating that transforming MLMs into universal language encoders can be achieved without external data. We propose a fine-tuning framework termed *Mirror-BERT*, which simply relies on duplicating and slightly augmenting the existing text input (or their representations) for the conversion, and demonstrate that it is possible to learn universal lexical and sentence encoders with such "mirrored" input data through self-supervision (see Fig. 1). The proposed Mirror-BERT framework is also extremely efficient: the whole MLM transformation can be completed in less than one minute on two 2080Ti GPUs. This finding further confirms a general hypothesis from prior work (Liu et al., 2020; Ben-Zaken et al., 2020; Glavaš and Vulić, 2020) that fine-tuning exposes the wealth of (semantic) knowledge stored in the MLMs for a particular application. In this case in particular, we demonstrate that the Mirror-BERT procedure can rewire the MLMs to serve as universal language encoders even without any external supervision.

We further show that data augmentation in both input space and feature space are key to the success of Mirror-BERT, and they provide a synergistic effect. Even feeding fully identical strings to the MLMs still substantially improves performance in semantic similarity tasks.

Contributions. 1) We propose a completely self-supervised approach that can quickly transform pretrained MLMs into capable universal lexical and sentence encoders, greatly outperforming off-the-shelf MLMs in similarity tasks across different languages and domains. **2)** We investigate the rationales behind why Mirror-BERT works at all, pointing out the impact of data augmentation in the input space as well as in the feature space.

2 Mirror-BERT: Methodology

Mirror-BERT consists of three main parts, described in what follows. First, we create positive pairs by duplicating the input text (§2.1). We then further process the positive pairs by simple data augmentation operating either on the input text or on the feature map inside the model (§2.2). Finally, we apply standard contrastive learning to encourage the base model to cluster the texts belonging to the same class (i.e., positives) while pushing away the negatives (§2.3).

2.1 Training Data through Self-Duplication

The key to the success of dual-network representation learning (Henderson et al., 2019; Reimers and Gurevych, 2019; Humeau et al., 2020; Liu et al., 2020, *inter alia*) is the construction of positive and negative pairs. While negative pairs can be easily obtained from randomly sampled texts, positive pairs usually need to be manually annotated. In practice, they are extracted from labelled task data (e.g., NLI) or knowledge bases that store relations such as synonymy or hypernymy (e.g., PPDB (Pavlick et al., 2015), BabelNet (Ehrmann et al., 2014), WordNet (Fellbaum, 1998), UMLS). Mirror-BERT, however, does not rely on any external data to construct the positive examples.

In a nutshell, given a set of non-duplicated strings $\mathcal{X}$, we assign individual labels to each string and build a dataset $\mathcal{D} = \{(x_i, y_i)|x_i \in \mathcal{X}, y_i \in \{1, \ldots, |\mathcal{X}|\}\}$. We then create self-duplicated training data $\mathcal{D}'$ simply by repeating every element in $\mathcal{D}$. In other words, let $\mathcal{X} = \{x_1, x_2, \ldots\}$. We then have $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots\}$ and $\mathcal{D}' = \{(x_1, y_1), (\overline{x}_1, \overline{y}_1), (x_2, y_2), (\overline{x}_2, \overline{y}_2), \ldots\}$ where $x_1 = \overline{x}_1, y_1 = \overline{y}_1, x_2 = \overline{x}_2, y_2 = \overline{y}_2, \ldots$. In §2.2, we introduce data augmentation techniques (in both input space and feature space) applied on $\mathcal{D}'$. Each positive pair $(x_i, \overline{x}_i)$ yields two different points/vectors in the encoder's representation space (see again Fig. 1), and the distance between these points should be minimised.

2.2 Data Augmentation

We hypothesise that, applying certain corruption techniques to (i) parts of texts or (ii) their representations, or even (iii) doing both in combination does little change to their semantics. We present two ways to corrupt text to be fed into an MLM, as illustrated in Fig. 1. First, we can directly erase or mask parts of the input text. Second, we can erase

x1: Economist Paul Krugman mainly works on trade models.

x1: Econ [MASK] Paul Krugman mainly works on trade models.

Figure 2: An example of input data augmentation via random masking.

(i.e., dropout) parts of their feature maps. Both techniques are rather simple and intuitive intuitive: (i) even when masking parts of an input sentence, humans can usually reconstruct its semantics;[1] (ii) dropping a small subset of neurons or representation dimensions, the representations of a neural network will not drift too much.

Input Augmentation: Random Masking. The idea of random masking of a small portion of input text is inspired by random cropping in visual representation learning (Hendrycks et al., 2020). In particular, starting from the mirrored pairs (x_i, y_i) and $(\overline{x}_i, \overline{y}_i)$, we randomly replace a consecutive string of length k with [MASK] in one of the x_i-s. The example (Fig. 2) illustrates the random masking procedure with $k = 5$.[2]

Feature Augmentation: Dropout. The random masking technique operating directly on text input can be applied only with sentence/phrase-level input; word-level tasks involve only short strings, usually represented as a single token under BERT's tokeniser. However, data augmentation in the feature space based on dropout, as introduced below, can be applied to any input text.

Dropout (Srivastava et al., 2014) randomly drops neurons from a neural network during training with a certain probability p. In practice, it results in the erasure of each element with a probability of p. It has mostly been interpreted as implicitly bagging a large number of neural networks which share parameters at test time (Bouthillier et al., 2016). However, here we take advantage of the dropout layers in BERT/RoBERTa to create augmented views of the input text. Assuming no random masking, given a pair of identical strings x_i and $\overline{x}_i$, their representations in the embedding space slightly differ due to the existence of multiple dropout layers in

v1 == v1
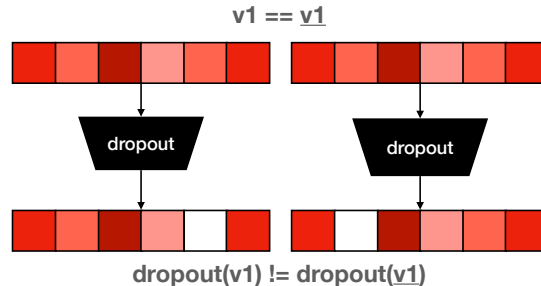

dropout(v1) != dropout(v1)

Figure 3: As the same vector goes through the same dropout layer separately, the outcomes are independent. As a consequence, two identical strings fed to BERT/RoBERTa two times will yield different representations in the MLM embedding space.

the BERT/RoBERTa architecture (Fig. 3). The two data points with distinct locations in the embedding space can be seen as two augmented views of the same text sequence. The dropout augmentations are naturally a part of the BERT/RoBERTa network. That is, no further actions need to be taken to implement them.[3]

It is also possible to combine data augmentation via random masking and dropout. We also evaluate this combined data augmentation variant.

2.3 Contrastive Learning

Let $f(\cdot)$ denote the encoder model. Fine-tuning the encoder then operates on the data constructed in §2.2. Given a batch of data $\mathcal{D}'_b$, we leverage the popular InfoNCE loss (Oord et al., 2018) to cluster/attract the positive pairs together and push away the negative pairs in the embedding space:

$$\mathcal{L}_b = -\sum_{i=1}^{|\mathcal{D}_b|} \log \frac{\exp(\cos(f(x_i), f(\overline{x}_i))/\tau)}{\displaystyle\sum_{x_j \in \mathcal{N}_i} \exp(\cos(f(x_i), f(x_j))/\tau)}.$$
(1)

τ denotes a temperature parameter; $\mathcal{N}_i$ denotes all negatives of x_i, which includes all $x_j, \overline{x}_j$ where $i \neq j$ in the current data batch (i.e., $|\mathcal{N}_i| = |\mathcal{D}_b| - 2$). Intuitively, the numerator is the similarity of the self-duplicated pair (positive) and the denominator is the sum of the similarities between x_i and all other strings besides $\overline{x}_i$ (negatives). The loss encourages the positive pairs to be relatively close comparing to the negative ones.[4]

---
[3]Note that random masking is applied on only one side of the positive pair while dropout is applied on all data points.
[4]We also experimented with another state-of-the-art contrastive learning scheme proposed by Liu et al. (2020). There, hard triplet mining combined with multi-similarity loss (MS

3 Experimental Setup

Evaluation Tasks: Lexical. We evaluate on both domain-general and domain-specific tasks: word similarity and biomedical entity linking. For the word similarity task, we rely on the Multi-SimLex evaluation set (Vulić et al., 2020a), which contains human-elicited word similarity scores for multiple languages. For biomedical entity linking, we use NCBI-disease (NCBI) (Doğan et al., 2014), BC5CDR-disease (BC5-d), BC5CDR-chemical (BC5-c) (Li et al., 2016), AskAPatient (Limsopatham and Collier, 2016) and COMETA (stratified-general split, Basaldella et al. 2020) as our evaluation datasets. The first three datasets are in the scientific domain (i.e., the data have been extracted from scientific papers), while the latter two are in the social media domain (i.e., extracted from online forums discussing health-related topics). We report Spearman's rank correlation coefficients (ρ) for word similarity; accuracy @1/@5 is the evaluation measure in biomedical entity linking tasks.

Evaluation Tasks: Sentence-Level. We use SemEval 2012-2016 datasets (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (STS-b) (Cer et al., 2017), SICK-Relatedness (SICK-R) (Marelli et al., 2014) for English; STS SemEval-17 data is used for Spanish and Arabic (Cer et al., 2017), and we also evaluate on Russian using a dataset available online.[5] We again report Spearman's ρ rank correlation.

Mirror-BERT:Training Resources. For (general-domain) word-level representations, we use the top 10k[6] most frequent words in each language.[7] For biomedical name representations, we randomly sample 10k names from the UMLS. For sentence-level English tasks, we sample 10k sentences (without labels) from the training set of the STS Benchmark; for Spanish, Arabic and Russian, we sample 10k sentences from the WikiMatrix dataset (Schwenk et al., 2019).

Training Setup and Details. The hyperparameters of word-level models are tuned on SimLex-999 (Hill et al., 2015); biomedical models are tuned

model	en	et	ar	zh	ru	es	pl
fastText	<u>.434</u>	**.447**	<u>.409</u>	.428	**.435**	**.488**	.396
MONO-BERT-CLS	.105	.160	.210	.277	.177	.152	.257
MONO-BERT-mt	.267	.106	.220	.398	.202	.177	.217
+ Mirror	**.556**	.308	**.538**	**.639**	<u>.365</u>	.296	**.444**
mBERT-CLS	.062	.074	.047	.204	.063	.039	.051
mBERT-mt	.105	.094	.101	.261	.109	.095	.087
+ Mirror	.358	.134	.097	<u>.501</u>	.210	<u>.332</u>	.141

Table 1: Word similarity evaluation on Multi-SimLex (Spearman's ρ). "MONO" denotes monolingual models (a complete listing of used monolingual MLMs can be found in App. §A). "mBERT" denotes multilingual BERT. "mt" denotes using mean-pooling of representations of all output tokens while "CLS" means using the [CLS] token's representation. **Bold** and <u>underline</u> denote highest and second highest scores per column.

model	scientific language						social media language			
	NCBI		BC5-d		BC5-c		AskAPatient		COMETA	
	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
SapBERT	.920	.956	.935	.960	.965	.982	.705	.889	.659	.779
PubMedBERT	.778	.869	.890	.938	.930	.946	.425	.496	.468	.532
+ Mirror	.909	.948	.930	.962	.958	.979	.590	.750	.603	.713

Table 2: Biomedical entity linking evaluations in the scientific and social media domain.

on COMETA (zeros-shot-general split). Sentence-level models are tuned on the development set of STS-b. The τ in Eq. (1) is 0.04 for biomedical phrase-level and all sentence-level models; 0.2 for other word-level models. All models use a dropout rate (p) of 0.1. Sentence-level models use a random masking rate of $k = 5$, while we set $k = 2$ for biomedical phrase-level models, and we do not employ random masking for word-level models. All lexical models are trained for 2 epochs with a maximum token length of 25. All sentence-level models are trained for 1 epoch with a max sequence length of 50.

All models use AdamW (Loshchilov and Hutter, 2019) as the optimiser, with a learning rate of $2e - 5$, batch size of 200 (400 after duplication). If not stated otherwise, for lexical-level tasks, all models use [CLS] as the representation token; for sentence-level tasks, RoBERTa-based models use [CLS] while BERT-based models use mean-pooling performed over the last layer's output.

---

loss) is used as the learning objective. InfoNCE and triplet mining + MS loss work mostly on par, with slight gains of one variant in some tasks, and vice versa. For simplicity and brevity, we report the results only with InfoNCE.

[5] github.com/deepmipt/deepPavlovEval

[6] We tested 1k, 10k, 20k, 50k and 100k and found 10k/20k works better than the other values (see Fig. 4).

[7] https://github.com/oprogramador/most-common-words-by-language

model	STS12	STS13	STS14	STS15	STS16	STS-b	SICK-R	Avg.
Sentence-BERT	**.719**	.774	**.742**	<u>.799</u>	<u>.747</u>	<u>.774</u>	**.721**	**.754**
BERT-CLS	.215	.321	.213	.379	.442	.203	.427	.314
BERT-mt	.314	.536	.433	.582	.596	.464	.528	.493
+ Mirror	<u>.674</u>	<u>.796</u>	.713	**.814**	.743	.764	.703	.744
RoBERTa-CLS	.090	.327	.210	.338	.388	.317	.355	.289
RoBERTa-mt	.134	.126	.124	.203	.224	.129	.320	.180
+ Mirror	.648	**.819**	<u>.732</u>	.798	**.780**	**.787**	.706	<u>.753</u>

Table 3: English STS evaluations. Spearman's ρ correlation is reported.

model	Spanish	Arabic	Russian
MONO-BERT-CLS	.526	.308	.470
MONO-BERT-mt	.599	.455	.552
+ Mirror	<u>.709</u>	**.669**	<u>.673</u>
mBERT-CLS	.421	.326	.430
mBERT-mt	.610	.447	.616
+ Mirror	**.755**	<u>.594</u>	**.692**

Table 4: Spanish, Arabic and Russian STS evaluation. Spearman's ρ correlation reported.

4 Results and Discussion

4.1 Lexical-Level Tasks

Word Similarity (Tab. 1). In prior work it has been shown that SotA static word embeddings such as fastText (Mikolov et al., 2018) still typically outperform off-the-shelf MLMs on word similarity datasets (Vulić et al., 2020a). However, our results demonstrate that the Mirror-BERT procedure indeed converts the MLMs into much stronger word-level encoders. The Multi-SimLex results on 7 languages from Tab. 1[8] suggest that the + *Mirror* variant substantially improves the performance of base MLMs (both monolingual and multilingual ones), even beating fastText in 4 out of the 7 evaluation languages.

We also observe that it is essential to have a strong base MLM. While Mirror-BERT does offer substantial performance gains with all base MLMs, the improvement is more pronounced when the base model is strong (e.g., en, zh).

Biomedical Entity Linking (Tab. 2). The goal of the biomedical entity linking (BEL) task is to map a biomedical name mention to a controlled vocabulary (usually a node in a knowledge graph). While it is considered a downstream application in BioNLP, the BEL task also helps evaluate and compare the quality of biomedical name representations: it requires pairwise comparisons between the biomedical mention and all surface strings stored in the biomedical knowledge graph.

The results summarised in Tab. 2 suggest that our + Mirror transformation achieves very strong improvements on top of the base PubMedBERT model (Gu et al., 2020). We note that PubMedBERT is a current SotA MLM in the biomedical domain. On scientific datasets, the self-supervised *PubMedBERT + Mirror* model is very close to SapBERT, which fine-tunes PubMedBERT with more than 10 million synonyms extracted from the external UMLS knowledge base.

However, in the social media domain, PubMedBERT + Mirror still cannot match the performance of the knowledge-guided SapBERT model. This result in fact reflects the nature and complexity of the task domain. For the three datasets in the scientific domain (NCBI, BC5-d, BC5-c), strings with similar surface forms tend to be associated with the same concept. On the other hand, in the social media domain, semantics of very different surface strings might be the same (e.g. *HCQ* and *Plaquenil* refer to exactly the same concept on online health forums: *Hydroxychloroquine*). This also suggests that the Mirror-BERT approach adapts PubMedBERT to a very good surface-form encoder for biomedical names, but dealing with more difficult synonymy relations (e.g. as found in the social media) does need external knowledge injection.

4.2 Sentence-Level Tasks

Similar to our lexical-level experiments in §4.1, Mirror-BERT also obtains large gains over the base MLMs in sentence similarity tasks. We break down the results into English STS (where we have a direct comparison with Sentence-BERT) and Spanish, Arabic, Russian STS.

English STS (Tab. 3). Regardless of the base

---

[8]Language codes: en (English), et (Estonian), ar (Arabic), zh (Chinese), ru (Russian), es (Spanish), pl (Polish).

model (BERT/RoBERTa), applying the + Mirror fine-tuning greatly boosts performance across all English STS datasets. Surprisingly, on average, *RoBERTa + Mirror*, fine-tuned with only 10k sentences without any external supervision, is on-par with the Sentence-BERT model, which is trained on the merged SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) datasets, containing 570k and 430k sentence pairs, respectively.

Spanish, Arabic and Russian STS (Tab. 4). In order to evaluate if the proposed data augmentation techniques, especially random masking, can generalise to other languages (and to different scripts), we also test sentence representations with Mirror-BERT in Spanish, Arabic and Russian. The results in the STS tasks again indicate very large gains for all languages, which all have different scripts, using both monolingual language-specific BERTs and multilingual BERT as base MLMs.

4.3 Further Discussion

Running Time. The Mirror-BERT procedure is extremely time-efficient. While fine-tuning on NLI (Sentence-BERT) or UMLS (SapBERT) data can take hours, Mirror-BERT with 10k positive pairs completes the conversion from MLMs to universal language encoders within a minute on two NVIDIA RTX 2080Ti GPUs. On average 15-20 seconds is needed for 1 epoch of the Mirror-BERT procedure.

Input Data Size (Fig. 4). In our main experiments in §4.1 and §4.2, we always use 10k examples for Mirror-BERT tuning. In order to assess the importance of the fine-tuning data size, we also run an analysis with other data sizes for a subset of base MLMs, and on a subset of English tasks. In particular, we evaluate the following: (i) BERT, Multi-SimLex (word-level); (ii) PubMed-BERT, COMETA (biomedical phrase-level); (iii) RoBERTa, STS12 (sentence-level). The results in Fig. 4 mainly suggest that the performance in all tasks reach their peak with around 10k and 20k examples and then gradually decrease. The word-level performance exhibits a steeper drop. We suspect that this is due to the inclusion of lower-frequency words into the fine-tuning data: embeddings of such words typically obtain less reliable embeddings (Pilehvar et al., 2018).[9]

---

[9]For word-level experiments, we used the top 100k words in English according to Wikipedia statistics. For phrase-level experiments, we randomly sampled 100k names from UMLS. For sentence-level experiments we sampled 100k sentences from SNLI and MultiNLI datasets (as the STS training set has
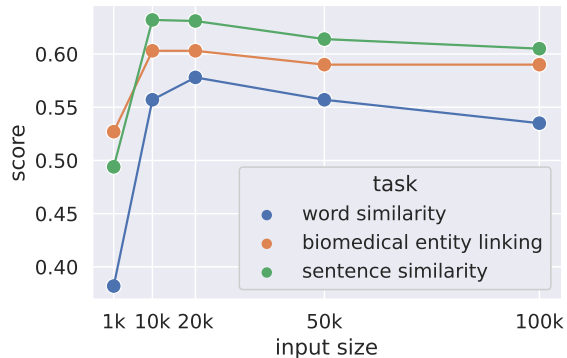

Figure 4: The impact of the number of fine-tuning "mirrored" examples (x-axis) on the task performance (y-axis). Note that the scores across tasks are not directly comparable. Word and sentence similarity tasks scores are Spearman's ρ while the BEL scores are Acc@1.

Synergy between Random Masking and Dropout (Tab. 5). We conduct our ablation studies on the English STS tasks. First, we experiment with turning off dropout, random masking, or both. With both techniques turned off, we observe large performance drops of both BERT + Mirror and RoBERTa + Mirror. Random masking appears to be the more important factor: its absence causes a larger decrease. However, the best performance is achieved when both dropout and random masking are leveraged, suggesting a synergistic effect when the two augmentation techniques are used together.

Regularisation or Augmentation? (Tab. 6). When using dropout, is it possible that we are simply observing the effect of adding/removing regularisation instead of the augmentation benefit? To answer this question, we design a more rigorous probing experiment that disentangles the effect of regularisation versus augmentation; we turn off random masking but leave the dropout on (so that the regularisation effect remains). However, instead of assigning independent dropouts to every individual string (rendering each individual string slightly different), we control the dropouts applied to a positive pair to be identical. As a result, $f(x_i) = f(\overline{x}), \forall i \in \{1, \cdots, |\mathcal{D}|\}$ in this experiment. We denote this as "controlled dropout".

In Tab. 6, we observe that, during the + *Mirror* fine-tuning, controlled dropout largely underperforms standard dropout and is even worse than not using dropout at all. As the only difference between controlled dropout and standard dropout is

---

fewer than 100k sentences).

6

model	STS12	STS13	STS14	STS15	STS16	STS-b	SICK-R	Avg.
BERT + Mirror	.674	.796	.713	.814	.743	.764	.703	.744
- dropout	.646	.770	.691	.800	.726	.745	.701	$.726_{\downarrow.018}$
- random masking	.641	.775	.684	.777	.737	.749	.658	$.717_{\downarrow.027}$
- dropout & random masking	.587	.695	.617	.688	.683	.674	.614	$.651_{\downarrow.093}$
RoBERTa + Mirror	.648	.819	.732	.798	.780	.787	.706	.753
- dropout	.619	.795	.706	.802	.777	.727	.698	$.732_{\downarrow.021}$
- random masking	.616	.786	.689	.766	.743	.756	.663	$.717_{\downarrow.036}$
- dropout & random masking	.562	.730	.643	.744	.752	.708	.638	$.682_{\downarrow.071}$

Table 5: The synergistic effect of dropout and random masking in sentence similarity tasks.

model configuration	ρ on STS12
random masking ✗; dropout ✗	.562
random masking ✗; dropout ✓	$.648_{\uparrow.086}$
random masking ✗; controlled dropout ✓	$.452_{\downarrow.110}$

Table 6: Probing the impact of dropout in a controlled experiment. English STS12 with RoBERTa + Mirror.

the augmented features for positive pairs in the latter case, this suggests that the improvement from + *Mirror* indeed stems from the data augmentation effect rather than from regularisation.

Learning from Negatives Only is Still Beneficial (Tab. 5). Interestingly, our ablation results Tab. 5 also indicate that Mirror-BERT substantially improves the base MLMs even without any positive examples, that is, even when both random masking and dropout are not used.[10] However, introducing positive examples naturally yields stronger task performance.

Learning New Knowledge or Exposing Available Knowledge? We find that running Mirror-BERT for more epochs, or with more data (see Fig. 4) does not lead to performance gains. This suggests that, instead of gaining new knowledge from the fine-tuning data, Mirror-BERT is more likely to be rewiring existing knowledge in MLMs. The experiments in the BEL task also point in the same direction: with social media data where medical terminology varies more than in the more rigorous scientific domain, the performance gap between Mirror-BERT and the UMLS-tuned Sap-BERT is still large. Motivated by these insights, in future work we also plan to experiment with a combined approach that blends Mirror-BERT and

external knowledge.

5 Conclusion

We have proposed Mirror-BERT, a simple, self-supervised, and highly effective approach that transforms large pretrained masked language models (MLMs) into universal lexical and sentence encoders within a minute, and without any external supervision. Mirror-BERT, based on simple unsupervised data augmentation techniques, demonstrates surprisingly strong performance in (word-level and sentence-level) semantic similarity tasks, as well as on biomedical entity linking. The large gains over base MLMs are observed for different languages with different scripts, and across diverse domains. Moreover, we analyse the main causes of the method's success, and identify that both random masking and dropout-based augmentation techniques contribute to its efficacy, yielding a synergistic effect.

Acknowledgements

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei

---

[10]Note that in this case, it always holds $x_i = \overline{x}_i$ and $f(x_i) = f(\overline{x})$. During training, this leads the numerator in Eq. (1) to be a constant. The learning collapses to the scenario where all gradients solely come from the negatives.

Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. 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 385–393, Montréal, Canada. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online. Association for Computational Linguistics.

Elad Ben-Zaken, Shauli Ravfogel, and Yoav Goldberg. 2020. BitFit: Simple parameter-efficient fine-tuningfor transformer-based masked language-model. Technical Report.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

Xavier Bouthillier, Kishore Konda, Pascal Vincent, and Roland Memisevic. 2016. Dropout as data augmentation. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015*. OpenReview.net.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Maud Ehrmann, Francesco Cecconi, Daniele Vannella, John Philip McCrae, Philipp Cimiano, and Roberto Navigli. 2014. Representing multilingual data as linked data: the case of BabelNet 2.0. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 401–408, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christiane Fellbaum. 1998. *WordNet*. MIT Press.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Goran Glavaš and Ivan Vulić. 2020. Is supervised syntactic parsing beneficial for language understanding? an empirical investigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv:2007.15779*.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

8

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019. Training neural response selection for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5392–5404, Florence, Italy. Association for Computational Linguistics.

Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2020. Augmix: A simple data processing method to improve robustness and uncertainty. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1014–1023, Berlin, Germany. Association for Computational Linguistics.

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2021. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. In *Proceedings of 43rd European Conference on Information Retrieval (ECIR 2021)*, pages 342–358.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment pre-training for biomedical entity representations. *arXiv preprint arXiv:2010.11784*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China. Association for Computational Linguistics.

Mohammad Taher Pilehvar, Dimitri Kartsaklis, Victor Prokhorov, and Nigel Collier. 2018. Card-660: Cambridge rare word dataset - a reliable benchmark for infrequent word representation models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1391–1401, Brussels, Belgium. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the ACL*, 8:842–866.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, et al. 2020a. Multi-simlex: A large-scale evaluation of multilingual and cross-lingual lexical semantic similarity. *Computational Linguistics*, pages 1–51.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020b. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

A Pretrained Encoders

A complete listing of URLs for all used pretrained encoders is provided in Tab. 7. For monolingual MLMs of each language, we made the best effort to select the most popular one (based on download counts). All models are Base models (instead of Large).

model	URL
fastText	https://fasttext.cc/docs/en/crawl-vectors.html
Sentence-BERT	https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens
SapBERT	https://huggingface.co/cambridgeltl/SapBERT-from-PubMedBERT-fulltext
BERT (English)	https://huggingface.co/bert-base-uncased
RoBERTa (English)	https://huggingface.co/roberta-base
mBERT	https://huggingface.co/bert-base-multilingual-uncased
Spanish BERT	https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased
Russian BERT	https://huggingface.co/DeepPavlov/rubert-base-cased
Chinese BERT	https://huggingface.co/bert-base-chinese
Arabic BERT	https://huggingface.co/aubmindlab/bert-base-arabertv02
Polish BERT	https://huggingface.co/dkleczek/bert-base-polish-uncased-v1
Estonian BERT	https://huggingface.co/tartuNLP/EstBERT

Table 7: A listing of HuggingFace & fastText URLs of all pretrained models used in this work.