A Simple Geometric Method for Cross-Lingual Linguistic Transformations with Pre-trained Autoencoders

Maarten De Raedt
Chatlayer.ai by Sinch

Fréderic Godin
Chatlayer.ai by Sinch

Pieter Buteneers
Sinch

Chris Develder
Ghent University - Imec

Thomas Demeester
Ghent University - Imec

Abstract

Powerful sentence encoders trained for multiple languages are on the rise. These systems are capable of embedding a wide range of linguistic properties into vector representations. While explicit probing tasks can be used to verify the presence of specific linguistic properties, it is unclear whether the vector representations can be manipulated to indirectly steer such properties. For efficient learning, we investigate the use of a geometric mapping in embedding space to transform linguistic properties, without any tuning of the pre-trained sentence encoder or decoder. We validate our approach on three linguistic properties using a pre-trained multilingual autoencoder and analyze the results in both monolingual and cross-lingual settings.

1 Introduction

Recently, the design of sentence encoders, monolingual (Kiros et al., 2015; Conneau et al., 2017) and multilingual (Artetxe and Schwenk, 2019; Feng et al., 2020) has enjoyed a lot of attention. Many works have used probing tasks to investigate the presence of specific linguistic properties in sentence representations (Adi et al., 2016; Conneau et al., 2018; Conneau and Kiela, 2018; Ravishankar et al., 2019; Hewitt and Manning, 2019; Chi et al., 2020). However, it remains unclear to what extent these linguistic properties can be actually steered by manipulating the representations. By analogy to the definition of style-transfer from Li et al. (2018), we refer to modifying a particular linguistic property in a given text (e.g., a sentence's tense) while preserving all of the property-independent content as *linguistic property transfer*.

Training dedicated models to transfer linguistic properties requires substantial computational effort and a lot of training data. Adding the ability to transform a new property may require an entire retraining of the text encoder and decoder. This is especially challenging for low-resource languages or when reusing or building transfer models for more than one language.

Assuming that pre-trained autoencoders capture the linguistic properties of interest, we investigate (i) whether they can be used without further tuning to efficiently transfer the properties, and (ii) whether this extends to the cross-lingual setting, when based on a multilingual pre-trained autoencoder. Our starting point is a pre-trained sentence encoder, with a corresponding decoder trained on an autoencoder objective. We show how a geometric transformation of pre-trained multilingual sentence embeddings can be efficiently learned on CPU for transferring specific linguistic properties. We also experiment with cross-lingual linguistic property transfer, using a language-agnostic pre-trained encoder.

In summary, this paper presents a set of preliminary experiments on linguistic property transfer, and shows that there may be value in further research on manipulating distributed representations to efficiently tackle language generation tasks.

2 Related work

Linguistic properties usually denote the grammatical behavior of linguistic units in sentences. This contrasts with styles which concern semantic aspects of sentences such as sentiment and gender. Nevertheless, transferring linguistic properties can be situated in the broader style transfer setting.

Style transfer systems can be categorized into (i) methods that learn *disentangled* representations, in which the content is explicitly separated from the style, making the style aspect controllable and interpretable (Hu et al., 2017; Shen et al., 2017; Zhao et al., 2018; Fu et al., 2018; Logeswaran et al., 2018; John et al., 2019) and (ii) methods that learn *entangled* representations in which the content and style are not explicitly separated (Mueller et al., 2017; Dai et al., 2019; Liu et al., 2020; Wang et al.,

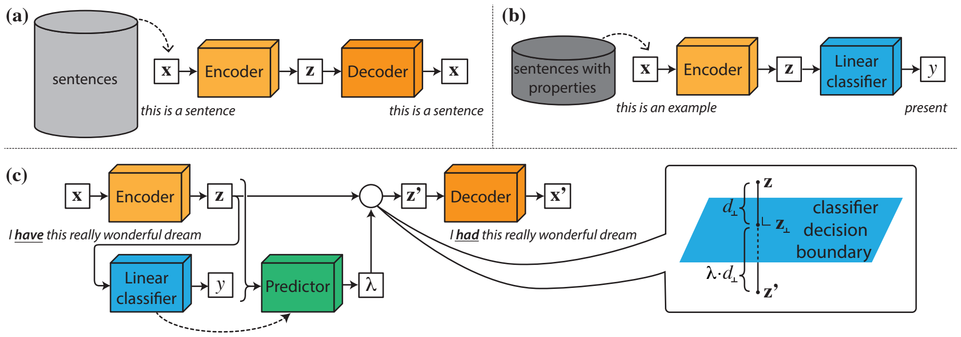

Figure 1: (a) Pretrained autoencoder (encoder ENC, decoder DEC). (b) linguistic property classifier $\mathcal{C}$. (c) Geometric transformation of the sentence representation to shift $\mathbf{z}$ according to λ beyond the decision boundary of $\mathcal{C}$, the shifted encoding $\mathbf{z}'$ is then given as input to the decoder resulting in the sentence $\mathbf{x}'$ with the transferred property.

2019; Duan et al., 2020). Our approach falls under the entangled methods because encoder-decoder systems trained on an autoencoding objective yield representations in which there is no explicit separation between content and style. Conceptually, our method is most similar to Duan et al. (2020), but differs in (i) that it can use any existing pre-trained autoencoder as opposed to training an autoencoder from scratch on a variational objective, (ii) that a simple geometric transformation is applied on the representations instead of training a computational heavy neural transformation network, and (iii) that it generalizes to the cross-lingual setting.

3 Linguistic Property Transfer

Our system consists of three components: (1) a pre–trained multilingual autoencoder, (2) linear classifiers for the targeted linguistic properties and (3) a component that geometrically transforms sentence embeddings to transfer the selected properties in the dense sentence representation space. These components are presented schematically in Fig. 1.

We start from a pre-trained autoencoder (Fig. 1a) that consists of an encoder (ENC : $\mathcal{X} \rightarrow \mathbb{R}^n$) which maps sentences ($\mathbf{x} \in \mathcal{X}$) to vectors ($\mathbf{z} \in \mathbb{R}^n$), and a decoder (DEC : $\mathbb{R}^n \rightarrow \mathcal{X}$) that maps the vectors $\mathbf{z}$ back to the corresponding sentences.

The second component (Fig. 1b) is a linear classifier $\mathcal{C} : \mathbb{R}^n \rightarrow \mathcal{Y}$ that takes as input a sentence encoding $\mathbf{z}$ and outputs a linguistic property label. We will limit our experiments to binary properties, i.e., $\mathcal{Y} = \{0, 1\}$.

Finally, the last component (Fig. 1c), performs a geometric transformation. It allows flipping the value of the selected linguistic property by projecting the original encoding $\mathbf{z}$ into the opposite half-space with respect to the property classifier, over an estimated distance λ. This leads to the *transferred* encoding $\mathbf{z}'$, designed to be decoded into a sentence $\mathbf{x}'$ close to the original sentence, but with the transformed target property.

The three components shown in Fig. 1 are further described below.

3.1 Pretrained Autoencoder

For the pre-trained autoencoder shown in Fig. 1a, we use Language Agnostic Sentence Representations (LASER) (Artetxe and Schwenk, 2019). LASER encodes sentences of 93 languages into a single vector space, such that semantically similar sentences in different languages have similar vectors. For our experiments, we leave the LASER encoder unchanged and train separate decoders for English and Dutch, by optimizing the likelihood $p(\mathbf{x}|\mathbf{z})$, with $\mathbf{z} = \text{ENC}(\mathbf{x})$. The decoder consists of a single-layer 1024-dimensional hidden state LSTM (Hochreiter and Schmidhuber, 1997).

3.2 Linear Property Classifier

Our approach assumes that both labels of the considered property are linearly separable in $\mathbf{z}$ space. A linear classifier $\mathcal{C}$ is trained on examples of the linguistic property. With the coefficients $\mathbf{w} \in \mathbb{R}^n$ and bias $b \in \mathbb{R}$, its decision boundary is characterized by the affine hyperplane

$$\mathcal{H} = \{\mathbf{z} \in \mathbb{R}^n : \mathbf{z} \cdot \mathbf{w} + b = 0\}. \qquad (1)$$

Logistic regression was used for the results presented in this work.

3.3 Geometric Transformation

The idea behind the geometric transformation is the following: a perpendicular projection from $\mathbf{z}$ onto the decision plane $\mathcal{H}$ would make the classifier $\mathcal{C}$ most uncertain about the considered attribute, with minimal changes (in Euclidean sense) to the original vector. When removing the property information from the corresponding sentence with the opposite label, we assume it gets projected onto the same position of $\mathcal{H}$. As a result, the proposed geometric transformation comes down to shifting $\mathbf{z}$ in the direction perpendicular to $\mathcal{H}$, and beyond it, into the region where $\mathcal{C}$ would predict the opposite label of the property. The transformed representation $\mathbf{z}'$ is then decoded by DEC. The intuitive approach of simply mirroring $\mathbf{z}$ over the decision plane appears sub-optimal (see Section 4.4). The distance into the opposite half space is therefore predicted based on the input (see Section 3.4).

The geometric shift of $\mathbf{z}$ in the direction of $\mathcal{H}$ can be derived with basic geometry, for which what follows is a brief sketch. By construction, $\mathbf{w}$ is perpendicular to the plane described by $\mathbf{z} \cdot \mathbf{w} = 0$, which in turn is parallel to $\mathcal{H}$, given Eq. (1), such that $\mathbf{w} \perp \mathcal{H}$. With that, the perpendicular projection $\mathbf{z}_\perp$ of $\mathbf{z}$ onto $\mathcal{H}$ can be written as

$$\mathbf{z}_\perp = \mathbf{z} + \beta\mathbf{w}, \quad \text{with} \quad \beta = -\frac{\mathbf{z} \cdot \mathbf{w} + b}{||\mathbf{w}||^2},$$

after substituting $\mathbf{z}_\perp \in \mathcal{H}$ into Eq. (1).
We finally express the transformation of $\mathbf{z}$ onto $\mathbf{z}'$ beyond $\mathcal{H}$ as

$$\mathbf{z}' = \mathbf{z}_\perp + \lambda(\mathbf{z}_\perp - \mathbf{z}) \qquad (2)$$

where the parameter $\lambda \geq 0$ represents the distance of $\mathbf{z}'$ from $\mathcal{H}$, relative to the distance $||\mathbf{z}_\perp - \mathbf{z}||$ on the original side of the decision plane (indicated as $d_\perp$ in Fig. 1).

3.4 Projection Distance Predictor

As mentioned above, we propose estimating the most suitable value of λ, corresponding to how far on the other side of the decision plane $\mathbf{z}$ needs to be projected to get optimal transfer results. To that end, we use a contextual multi-armed bandit (CMAB) (Auer, 2002), a simple and efficient form of reinforcement learning with a single state, which in our setting is the sentence representation $\mathbf{z}$. For a new input $\mathbf{z}$, the bandit needs to select the value of λ that best allows transferring the property with Eq. (2), while preserving the content of the associated sentence $\mathbf{x}$.

The bandit method allows using a non-differentiable reward, but other choices of algorithm are possible. Our model's goal is to preserve the content of the original sentence $\mathbf{x}$ while changing its property y to y'. Hence, our CMAB reward consists of (i) a linguistic property reward r_{prop} and (ii) a content-preserving reward r_{content}. To compute r_{prop}, we pass the decoded transformed sentence $\mathbf{x}' = \text{DEC}(\mathbf{z}')$ back into the encoder and use the predicted likelihood of the corresponding linear property classifier for target y' as the reward:

$$r_{\text{prop}}(\mathbf{x}', y') = \begin{cases} \sigma(\text{ENC}(\mathbf{x}') \cdot \mathbf{w} + b) & y' = 1 \\ 1 - \sigma(\text{ENC}(\mathbf{x}') \cdot \mathbf{w} + b) & y' = 0 \end{cases}$$

with $\sigma(.)$ the logistic function. For r_{content}, we directly optimize the BLEU-score (Papineni et al., 2002) between the original sentence $\mathbf{x}$ and the transferred sentence $\mathbf{x}'$. Intuitively, this leads to the minimum number of changes that are required to transfer y to y' and thus encourages the content preservation between $\mathbf{x}$ and $\mathbf{x}'$:

$$r_{\text{content}}(\mathbf{x}, \mathbf{x}') = \text{BLEU}(\mathbf{x}, \mathbf{x}')$$

For the final reward $r(\mathbf{x}, \mathbf{x}', y')$, the harmonic mean of $r_{\text{prop}}(\mathbf{x}', y')$ and $r_{\text{content}}(\mathbf{x}, \mathbf{x}', y')$ appeared a suitable choice, encouraging the model to jointly ensure the correct target property (high r_{prop}) as well as preserve the sentence content (high r_{content}).

We implement CMAB using the LinUCB with Disjoint Linear Models algorithm from Li et al. (2010), which assumes that the expected reward obtained from choosing arm λ is linear with respect to its input features (in our case, sentence encoding $\mathbf{z}$). For each discrete allowed value ('arm') for λ, LinUCB learns a separate ridge-regression model, with learnable parameters $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{b} \in \mathbb{R}^n$ (for LASER n is 1024). It predicts the reward, including an upper confidence bound (UCB), for choosing that value for λ for the given encoding $\mathbf{z}$. The hyperparameter α is used to control the wideness of the UCB, such that a larger α results in a wider UCB. Each training iteration observes a single $\mathbf{z}$ for which the arm achieving the highest potential reward (UCB) is chosen and only the parameters corresponding to its ridge-regression model are updated. Quantifying the merit of each arm for the input requires an inverse-matrix ($n \times n$) computation, 2 matrix-vector multiplications, and 2 dot products. The best arm's parameters ($\mathbf{A}$ and $\mathbf{b}$) are then updated, requiring 1 outer-vector product. During inference, the λ value of the best arm is used. The training and inference schemes are presented in Algorithms 1 and 2.

Algorithm 1: Training scheme, pseudocode adapted from Li et al. (2010)

```
input : Exploration parameter $\alpha \in \mathbb{R}_+$
        $\mathcal{A} = \{\lambda_1, ..., \lambda_k\}$
for $(\mathbf{x}_t, y_t) \in \mathcal{D}$ do
    $\mathbf{z}_t = \text{ENC}(\mathbf{x}_t)$
    for $\lambda \in \mathcal{A}$ do
        if $t = 0$ then
            $\mathbf{A}_\lambda = \mathbf{I}_{n \times n}, \mathbf{b}_\lambda = \mathbf{0}_{n \times 1}$
        $\hat{\theta}_\lambda = \mathbf{A}_\lambda^{-1}\mathbf{b}_\lambda$
        $p_{t,\lambda} = \hat{\theta}_\lambda^T \mathbf{z}_t + \alpha\sqrt{\mathbf{z}_t^T \mathbf{A}_\lambda^{-1} \mathbf{z}_t}$
    Choose $\lambda_t = \text{argmax}_{\lambda \in \mathcal{A}}\, p_{t,\lambda}$
    $\mathbf{z}'_t = \mathbf{z}_{\perp,t} + \lambda_t(\mathbf{z}_{\perp,t} - \mathbf{z}_t)$
    $\mathbf{x}'_t = \text{DEC}(\mathbf{z}'_t)$
    $\mathbf{A}_{\lambda_t} = \mathbf{A}_{\lambda_t} + \mathbf{z}_t \mathbf{z}_t^T$
    $\mathbf{b}_{\lambda_t} = \mathbf{b}_{\lambda_t} + r(\mathbf{x}_t, \mathbf{x}'_t, y_t^{'})\mathbf{z}_t$
```

Algorithm 2: Inference scheme

```
input : $\mathbf{A}_\lambda$ and $\mathbf{b}_\lambda$ for each arm
        $\lambda \in \mathcal{A} = \{\lambda_1, ..., \lambda_k\}$,
        Sentence $\mathbf{x}$ with property label $y$
$\mathbf{z} = \text{ENC}(\mathbf{x})$
for $\lambda \in \mathcal{A}$ do
    $\hat{\theta}_\lambda = \mathbf{A}_\lambda^{-1}\mathbf{b}_\lambda$
    $p_\lambda = \hat{\theta}_\lambda^T \mathbf{z}$
Choose $\lambda = \text{argmax}_{\lambda \in \mathcal{A}}\, p_\lambda$
$\mathbf{z}' = \mathbf{z}_\perp + \lambda(\mathbf{z}_\perp - \mathbf{z})$
$\mathbf{x}' = \text{DEC}(\mathbf{z}')$
return $\mathbf{x}'$
```

4 Experiments

To investigate whether linguistic properties embedded in representations of pre-trained encoders can be transferred without finetuning, we first apply the SentEval tool from Conneau and Kiela (2018) to LASER-embeddings (Section 3.1) and identify three properties that have a strong presence. We then investigate how well our approach performs on these properties in the monolingual setting (ML), in which our CMAB model is both trained and evaluated on English sentences (**Q1**). Finally, we investigate the performance of our approach in the cross-lingual setting (CL), in which the model is trained on English but evaluated on Dutch sentences. In particular, after training on English, Dutch sentences are passed into the LASER encoder to obtain the transformed encodings $\mathbf{z}'$ which in turn are decoded by the Dutch decoder (**Q2**).

Probing Task: Accuracy (%)	
Length: 74.09	**Tense**: **89.1**
BigramShift: 68.06	CoordinateInversion: 67.82
OddManOut: 50.80	Depth: 39.2
TopConstituents: 39.2	**SubjNumber**: **90.69**
ObjNumber: **88.72**	

Table 1: Results of LASER-embeddings on the probing tasks of Conneau and Kiela (2018). In our experiments, we transfer the properties denoted in bold.

.

4.1 Linguistic Properties

Table 1 shows the results of LASER-embeddings on the probing tasks from Conneau and Kiela (2018). The high accuracies for the properties shown in bold, indicate that LASER encodes them well. In our experiments, we transfer (i) the **Tense** of the main verb which is either in the present or past, (ii) **ObjNum**, representing the number (singular or plural) of the main clause's direct object and (iii) **SubjNum**, which is the number (singular or plural) of the subject of the main clause.

4.2 Implementation and Training Data

As discussed in Section 3.1, we use LASER's encoder and train two decoders on it with around 20M English and Dutch OpenSubtitles sentences (Tiedemann, 2012; Lison et al., 2019). For each property, we train a binary logistic regression model on CPU using SentEval data, through stratified 5-fold cross-validation. We found that training the CMAB-models on SentEval led to worse results than training on OpenSubtitles. We hypothesize that this is due to a mismatch between the SentEval - and OpenSubtitles sentences on which the decoders were trained. We therefore trained, on CPU, the CMAB-models using 2500 English OpenSubtitles sentences with (noisy) property labels predicted by the SentEval classifiers. Across all experiments, we use the discrete set $\{1, 1.5, \ldots, 7\}$ as possible values for λ ('arms' of the CMAB algorithms) and set the CMAB exploration parameter α to 4.

4.3 Evaluation

We randomly selected OpenSubtitles sentences (not seen during decoder training), and for those with any of the target properties present, annotated the corresponding sentence with the flipped property. As such, 100 test-pairs $(\mathbf{x}, \mathbf{x}')$ were obtained for each property. We report human evaluation metrics: (i) the percentage of transferred sentences that have the correct property ('Label' accuracy), and (ii) the

	Property	Label (%)	All (%)	BLEU
Mono- lingual	**Tense**$_{\text{ML}}$	61	47	54.9
	ObjNum$_{\text{ML}}$	44	29	39.0
	SubjNum$_{\text{ML}}$	48	34	36.3
Cross- lingual	**Tense**$_{\text{CL}}$	51	41	49.9
	ObjNum$_{\text{CL}}$	49	43	49.0
	SubjNum$_{\text{CL}}$	56	33	32.6

Table 2: Human label accuracy ('Label') and accuracy of both label and content ('All'), and BLEU-scores of our CMAB-approach (monolingual and cross-lingual).

percentage of transferred sentences that have the correct property *and* preserve the content ('All' accuracy). We also include the BLEU-score between the transferred sentence and the gold target $\mathbf{x}'$.

4.4 Results

To answer (**Q1**), we refer to the first three rows of Table 2. Our approach switches properties in roughly half of the cases (label accuracy). However, fewer cases occur in which both the property is transferred and content is preserved. The last three rows of Table 2 display the metrics in the cross-lingual setting in which we notice similar results as in the previous setting (**Q2**). The results are encouraging, although we expect further improvements from more complex transformation approaches. Table 3 shows, for **Tense**$_{\text{ML}}$, a comparison of our CMAB approach against a baseline, that mirrors each $\mathbf{z}$ over the decision boundary i.e, $\lambda = 1$. We find that the CMAB-approach outperforms that baseline for all metrics. Moreover, Table 4 shows the distribution of the predicted arms on the test sets in the monolingual and cross-lingual settings, indicating that choosing the optimal value for λ is input-dependent. As an illustration, Table 5 lists a few examples, picked randomly from among those test items with successful label transformation and content preservation.

Model	Label (%)	All (%)	BLEU
Baseline	59	28	53.1
CMAB	**61**	**47**	**54.9**

Table 3: Comparison of the baseline ($\lambda = 1$) and the CMAB-approach for **Tense**$_{\text{ML}}$.

λ	**Tense**$_{\text{ML(CL)}}$	**SubjNum**$_{\text{ML(CL)}}$	**ObjNum**$_{\text{ML(CL)}}$
1	2.5(5)	✗	✗
1.5	23.5(31.5)	✗	✗
2	43.5(34.5)	✗	✗
2.5	14(19)	✗	13(14)
3	15(7)	3(1)	✗
3.5	✗	✗	✗
4	1.5(3)	21(✗)	✗
4.5	✗	✗(29.5)	✗
5	✗	5.5(6.5)	1.5(5.5)
5.5	✗	21(26)	✗
6	✗	24(11.5)	0.5(50)
6.5	✗	8.5(13)	22(15)
7	✗	17(12.5)	63(15.5)

Table 4: Distributions of the predicted projection distances of the CMAB for the different test sets expressed as a percentage (monolingual and cross-lingual).

	Tense (present→past)
Mono- lingual	i ask many people here . i **asked** many people here .
Cross- lingual	ik kijk naar een oude film van m ' n moeder . ik **bekeek** een oude film van mijn moeder .
	ObjNum (singular→plural)
Mono- lingual	i could tell you some story . i could tell you some **stories** .
Cross- lingual	we hebben een beter bondgenoot nodig . we hebben **betere bondgenoten** nodig .
	SubjNum (plural→singular)
Mono- lingual	families agreed to keep it quiet . **a family** agreed to keep it quiet .
Cross- lingual	monsters gaan ons opeten . **het monster gaat** ons opeten .

Table 5: Linguistic property transfer examples of the proposed system in both monolingual and cross-lingual settings

5 Conclusion and Future Work

We have introduced a simple and efficient geometric method to transfer linguistic properties which has been evaluated on three properties in both monolingual and cross-lingual settings. While there is room for improvement, our preliminary results indicate that it can allow pre-trained autoencoders to transfer linguistic properties without additional tuning, such that there is no need to train dedicated transfer systems. This potentially makes learning faster and better scalable than with existing methods. For future work, we aim at extending our method to transformer-based encoders (monolingual and cross-lingual), and will consider additional linguistic as well as more style-oriented properties.

Acknowledgments

This work was funded by the Flemish Government (VLAIO), Baekeland project-HBC.2019.2221.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Peter Auer. 2002. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

Ethan A Chi, John Hewitt, and Christopher D Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $ &!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuan-Jing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007.

Yu Duan, Jiaxin Pei, Canwen Xu, and Chenliang Li. 2020. Pre-train and plug-in: Flexible conditional text generation with variational auto-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

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

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1587–1596. JMLR. org.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. 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 1865–1874.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670.

Pierre Lison, Jörg Tiedemann, Milen Kouylekov, et al. 2019. Open subtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Dayiheng Liu, Jie Fu, Yidan Zhang, Chris Pal, and Jiancheng Lv. 2020. Revision in continuous space: Fine-grained control of text style transfer. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. In *Advances in Neural Information Processing Systems*, pages 5103–5113.

Jonas Mueller, David Gifford, and Tommi Jaakkola. 2017. Sequence to better sequence: continuous revision of combinatorial structures. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2536–2544. JMLR. org.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Vinit Ravishankar, Lilja Øvrelid, and Erik Velldal. 2019. Probing multilingual sentence representations with x-probe. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 156–168.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pages 6830–6841.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Ke Wang, Hang Hua, and Xiaojun Wan. 2019. Controllable unsupervised text attribute transfer via editing entangled latent representation. In *Advances in Neural Information Processing Systems*, pages 11034–11044.

Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. In *International Conference on Machine Learning*, pages 5902–5911.