Are Latent Sentence Vectors Cross-Linguistically Invariant?

Michael Hahn
mhahn2@stanford.edu

Abstract

Previous work [Bowman et al., 2016] has shown that variational autoencoders (VAEs) can create distributed representations of natural language that capture different linguistic levels such as syntax, semantics, and style in a holistic manner. I investigate to what extent VAEs, when trained on different languages, result in comparable representations. To this end, I train VAEs for English and French, and then train a transformation between the resulting latent spaces on the task of machine translation. An analysis of the resulting mapping from French to English sentences shows that the latent representations represent the presence of words, phrases, and the general topic. However, I do not find evidence that they also encode syntax and semantics in a cross-linguistically invariant manner.

1 Introduction

Recently, there has been growing interest in better understanding the distributed representations that neural network models create for language, and what kinds of linguistic properties they encode (or don't). A lot of attention has gone towards understanding word embeddings [Association for Computational Linguistics, 2016]. Recently, Shi et al. [2016] investigated how much syntactic knowledge is learnt by neural MT, trying to decode syntax trees from the latent representations of a Seq2Seq MT system. There also has been recent work on understanding *multilingual* word embeddings and some results indicating that word embeddings encode information in a cross-linguistically invariant manner (e.g., Smith et al., 2017).

I propose to study to what extent distributed representations of *sentences* encode information similarly across different languages. For this, I propose to train monolingual sentence representations and then learn to translate sentences by transforming these representations. More precisely, I encode a sentence into a vector, apply a transformation, and decode a sentence from the resulting vector. Encoder and decoder are trained in a monolingual fashion; only the intermediate transformation is trained on the translation task.

If latent sentence representations encode aspects like syntax and semantics in a similar way across languages, the transformation should be able to utilize these commonalities and syntax and semantics should be translated successfully. On the other hand, if such abstract linguistic structures are not encoded in a cross-linguistically invariant way in the latent vectors, it is expected that only properties closer to the surface, such as words and the general topic, will be preserved by the learned translation model.

As a candidate for promising sentence representations, I will use latent vectors created by Variational Autoencoders (VAE). Bowman et al. [2016] have shown that variational autoencoders (VAEs) can create distributed representations of natural language that capture different linguistic levels such as syntax, semantics, and style in a holistic manner superior to vanilla autoencoders.

More specifically, I expect VAEs to be an interesting candidate for the following reasons:

1. The latent representations of VAEs are dense. If training of the autoencoder has worked well (which can be challenging), every point in the latent space should decode to a reasonably fluent sentence. Points that are close by decode to similar sentences [Bowman et al., 2016]. This might encourage cross-linguistically invariant representations, since it ideally should prevent the occurrence regions in latent space that are never used in training and so do not decode to anything sensible.
2. The latent representations encode probability in their geometry. Ideally, the probability mass of a set in latent space (according to the prior, see below), corresponds to the probability mass of the sentences decoded from it. This might be beneficial since probability of expressing some meaning should depend on the domain much more than on the language. Thus, if VAEs are built for several languages on the same domain, one might expect the geometric arrangement of sentences to be similar.
3. The prior (see below) forces latent representations to be spread out. This has been argued to guard against overfitting in NLP [Miao et al., 2016], as representations supported only by few datapoints would be washed out. This suggests that the representations might be forced to make more generalizations and thus to fit more to general linguistic structure than to the strings of the given language in particular.

2 Related Work

Variational autoencoders were introduced by Kingma and Welling [2014]. Variational sequence autoencoders for NLP have been described by Bowman et al. [2016] and Hu et al. [2017].

3 Approach

3.1 Variational Autoencoders

In a first step, I build variational autoencoders for two languages. A variational autoencoder can be viewed as an autoencoder with a specific form of stochasticity in the latent layer. Let X be the set of inputs, for instance, the set of sentences. A variational autoencoder consists of an encoder $\phi : X \rightarrow \mathbb{R}^d$, a decoder ψ mapping each vector $z \in \mathbb{R}^d$ to a probability distribution over X (e.g., a language model), and a special stochastic latent layer in between. Given an input $x \in X$, let $v = \phi(x)$ be the vector created by the encoder. Then two MLPs compute vectors $f_1(v), f_2(v)$ of some dimensionality d_h. Then a hidden representation z is drawn according to

$$z \sim \mathcal{N}(f_1(v); f_2(v)) \tag{1}$$

That is, we draw d_h independent Gaussians, where the i-th variable has mean $f_1(v)_i$ and variance $\sigma^2 = f_2(v)_i$. The vector z is then used as the input to the decoder. For the purpose of this project, the mean value $E[z] = f_1(v)$ serves as the latent representation of the sentence x.

Training Objective The training objective combines the log-likelihood loss for the decoder with a term encouraging the Gaussians in the hidden layer to be close to some specified prior P over $\mathbb{R}^{d_h}$:

$$L(x) = -\left(\mathrm{E}_z \log P_{Decoder}(x|z)\right) + KL(\mathcal{N}(f_1(v); f_2(v)), P) \tag{2}$$

where KL denotes the Kullback-Leibler divergence.

During training, the gradient of the first term is estimated by drawing a sample z, while the gradient of the second term can be computed directly when P is of a suitable form, such as a Gaussian [Kingma and Welling, 2014].

In my experiments, I implemented two priors P. In one set of experiments, I set $P := \mathcal{N}(0, 1)$, so that representations would naturally be concentrated around 0. This version is standard [Kingma and Welling, 2014].

In a second version, I normalize the mean $f_1(v)$, sample a point $z' \sim \mathcal{N}(\frac{f_1(v)}{\|f_1(v)\|}; f_2(v))$, and normalize z' again to obtain z. Expressed differently, I constrained the hidden representations to lie on the $h_z - 1$-dimensional sphere. The latent variable z is distributed according to the projection of the normal distribution, centered at a point on the sphere, on the sphere. For the prior P, I took the uniform distribution on the sphere. I do not know an analytic expression for the KL-divergence in this case, but for small σ^2, it should be approximated decently by the KL-divergence of the $d_h - 1$-dimensional Gaussian with a uniform distribution in a large cube in $\mathbb{R}^{d_h-1}$, which I used instead for training. For this second version, which as far as I know has not been reported in the literature, training might be easier since the prior encourages representations to be spread evenly over the sphere instead of driving them towards a single point like a Gaussian prior.

3.2 Cross-Lingual Mapping

The VAEs built in the first step define latent spaces for the two language. Crucially, the VAEs make it possible to both encode any sentence into the latent space, and to decode word sequences from any point in latent space.

In the second step, the goal is to align the two latent spaces as well as possible so that points representing semantically similar sentences are aligned. More precisely, I train a mapping $f : R^{d_h} \rightarrow R^{d_h}$ mapping the latent space of one language (the *source* language) to the latent space of the other language (the *target* language). To obtain a good map between the two latent spaces, I use the following procedure. A pair of synonymous source language and target language sentence are drawn from a bilingual corpus. The source language sentence is encoded into a latent vector z, which is then transformed into a vector $f(z)$ in the target-language latent space. Gicen $f(z)$, the target-language decoder then defines a distribution over target-language word sequences. Using backpropagation and stochastic gradient descent, I then maximize the likelihood of the gold target sentences given the respective source language sentenves. Crucially, only the parameters of the map f are updated. The encoders and decoders remain fixed.

4 Experiments

To investigate the cross-linguistic invariance of VAE latent spaces, I built variational autoencoders for English and French, and a transfer mapping f for latent representations from French to English. I used the 10^9 English-French dataset from WMT 2014, which consists of about 20 million pairs of English and French sentences. To hold the domain of monolingual and bilingual training constant, I split this set into two equal-sized sets A and B of sentence pairs, using A for the autoencoders and B for the transfer mapping. I split each of A and B into a 70 % training, 20 % development, and 10 % test set. In contrast to using the original WMT 2014 test data (note that the data that I used was intended entirely for training in WMT 2014), this has the advantage that training, devevelopment, and test set are all from the same domain.

4.1 Variational Autoencoders

I built autoencoders for English and French separately on the respective sentences from partition A. Encoders and decoders consisted of GRUs with 1028 dimensions. I set $d_h = 256$. Optimization used Adam with a learning rate of 0.001.

As previously noted by Bowman et al. [2016], training variational autoencoders for sequences is challenging, as they can get stuck in a configuration where the KL-divergence term becomes zero and the decoder performs language modeling without consideration of the input. I applied what they called *word dropout*, that is, when feeding the previous word into the decoder during training, randomly replaced words with an OOV symbol. I used a keep probability of 0.8. Also, I weighted the KL-term with a parameter λ that I increased during training from 0 to 1.

As training still proved challenging, I focused on the standard normally-distributed variable model and leave experiments with spherical latent variables to future research.

Since the loss function of a variational autoencoder on a sentence is a lower bound on the likelihood of the sentence Kingma and Welling [2014], variational autoencoders can be evaluated by dividing the loss – crucially including both the reconstruction loss and the KL divergence term – by sequence length and exponentiating the result, obtaining an upper bound on perplexity. Resulting perplexities on the test set are shown in Table I. For comparison, Bowman et al. [2016] reported a perplexity of

English	118
French	127
Bowman et al. [2016] (Penn Treebank)	119

Table 1: Perplexities of monolingual variational autoencoders (lower is better). While the datasets are different, comparison with Bowman et al. [2016] suggests that my autoencoders trained decently.

119 for their most comparable variational autoencoder on the Penn Treebank.While the datasets are different, this result shows that my VAEs trained quite well.

4.2 Assessing Cross-Language Invariance

For the transfer mapping f, I used a network with one hidden ReLU layer and a linear output layer. I trained f on the bilingual sentence pairs from the training set partition B by minimizing cross-entropy on the English gold translation, given the French source sentence.

Gradients are obtained at the softmax layers of the English decoder, and then backpropagated up to the transfer MLP modeling f, at which point they are used to update its parameters.

During training, crucially only this two-layer MLP is trained. This is important, since it ensures that the latent spaces purely reflect the monolingual VAE representations and are not affected by the bilingual information present in the translation task.

In these experiments, I limited the sequence length to 20 and cut off longer source and target sentences. This was to ensure that gradients would flow successfully through the decoder, as initial experiments suggested that longer sequence lengths were detrimental for this.

To assess to what extent the latent representations show cross-language invariance, I manually inspected a sample of 200 sentences decoded from French sentences in the development set and compared to the English gold translations. My findings are as follows:

- Only a few decoded sentences preserve the meaning in a straightforward way, such as (Here, I'm omitting the French input sentences and only show the synonymous English target translations and the decoded sentences):

Target:	Partnering Agreements Provincial/Territorial Agreements
Decoded:	The partnership with the provinces and territories .

- Decoded sentences more often share some content words or phrases with the target translation, but the meanings can be quite different:

Target:	Participation has reportedly increased with repayment rates reaching 00 % .
Decoded:	UNK is to be increased by 00 % in 0000 .
Target:	GDP growth is slower than expected .
Decoded:	The growth of the UNK and UNK 's UNK is 0.0 % .
Target:	Since an increasing number of companies , especially large producers and retailers , have and continue to green their businesses , national green building standards have been established in the United States and Canada .
Decoded:	A number of firms are engaged in R&D and R&D and development , and are the to be the most important
Target:	extended the taxation of UNK life insurance premiums to the first $ 00,000 of coverage .
Decoded:	the application of the tax paid by the applicant for a period of 00 years , the date of which
Target:	E. Additional Information For more information about Info Source , the Access to Information Act or the Privacy Act , you may contact :
Decoded:	The information is provided to the Agency for the purposes of the Act and the Act to the Act and the

- In other cases, the semantic relation between the target and the decoded sentence is a much more vague kind of similarity:

Target:	The Centre also hosts several special events each year to give Members and their families a chance to learn about their community , meet new people or just have some fun .
Decoded:	The UNK also provided a forum for the people to discuss and discuss the issues and challenges of the UNK .
Target:	The expected completion date of this project is (day , month , year) .
Decoded:	UNK , the date of which is 00 (0) (0) (0) (0) (
Target:	These supporting documents are manually checked for their accuracy by a program officer performing Section 00 verification .
Decoded:	This and other information you are required to complete the application form , please contact the Office 's Office of the
Target:	They want UNK analytical UNK who are a source of new ideas , understand the innovation process and are at ease in the world of new technology .
Decoded:	UNK and their knowledge and skills are diverse and diverse . UNK

- While decoded sentences typically share words or the general topic with the target translation, the relations between words and the statements made by the sentences are usually very different:

Target:	The net expense for the Commission 's employee severance plan for the year ended 00 March 0000 was $ 0,000 (0000 - $ 0,000) .
Decoded:	And , the employer 's pay is payable to the employee in the year of the year for which the employee
Target:	In its 0000 Trade Policy Review of Canada , the WTO concluded that Canada 's trade and investment regime remains one of the world 's most transparent .
Decoded:	EU 's Policy on Trade and Investment in 0000 is the to the UNK of the 0000 Convention on the

- There are a few recurrent patterns that erroneously show up in many decoded sentences, such as 'Committee of Ministers', or 'Council of Europe':

Target:	Article 0 Cooperation with other International Financial Institutions 0 .
Decoded:	However, the Committee of Ministers ' meeting of the Committee of Ministers ' s Committee on Budgets ' Committee on

Presumably, the translation mapping concentrates a lot of probability mass in certain regions of the hidden space. This is one of the problems that a spherical hidden variable might alleviate.

- Certain special types of sentences that evidently are specific for the corpus used are preserved to some extent, highlighting that the variational latent representations have adapted to the domain of the corpus:

Target:	canada.gc.ca Home > First Nations , Inuit & Aboriginal Health > Reports & Publications > Funding > Guide to Health Management Structures Institutional links
Decoded:	canada.gc.ca Home > About Health Canada > Reports & Publications > Health and Social Services > Health and Social Services

- Decoded sentences often make sense only at the beginning, such as:

Target:	August 00 , 0000 - Public Notice - Mountain View Drop Zone Training Area Project - comments required before September 00 , 0000
Decoded:	00 November 0000 , the UNK UNK the UNK for the UNK and UNK UNK and UNK
Target:	This report covers the period of June 0 , 0000 to May 00 , 0000 .
Decoded:	The report of the 0000 Report on the 0000 Report on the 0000 .
Target:	These results indicate that the percentages of green plants from UNK culture of the tested genotypes are under the control of nuclear genes and that the genes are predominantly additive .
Decoded:	These results suggest that the UNK of UNK and UNK UNK of UNK and UNK UNK and UNK UNK with UNK

- One some sentences, the decoder fails completely:

Target:	A thirteen week repeated inhalation study of ethylene UNK in rats .
Decoded:	UNK UNK UNK UNK UNK UNK UNK UNK UNK UNK UNK UNK
Target:	Anderson , Marjorie Jean – Brantford 0000-0000 Howard , Donna Louise – Mississauga 0000-0000 MacKenzie , UNK Joyce – Pembroke 0000-0000 UNK K , Louis Paul – Windsor 0000-0000 Quebec
Target:	, the UNK , UNK , UNK , UNK , UNK , UNK , UNK , UNK , UNK ,

It seems that this happens particularly when some input words are relatively rare in the context of the corpus used. Presumably, in such cases, the linear layer of the MLP has mapped the input representation into regions that have very little probability under the prior and thus do not successfully decode. The spherical approach might help again, since the prior is uniform there.

Conclusion The qualitative evaluation has shown that the transfer map can recover words, some phrases, and thematic similarity. On the other hand, we have seen that the exact meanings of decoded sentences are mostly quite different from the input sentences. In particular, no evidence is found that the mapping recovers syntactic constructions, or semantic relations between different words. This suggests that the latent representations generated by VAEs do encode surface features such as words and the topic in a way that is compatible between different languages, but not more abstract linguistic categories such as syntax and semantic relations.

Future research might get a more detailed picture of the linguistic capabilies of VAE vectors by probing them in a more targeted manner, looking at specific linguistic phenomena such as the difference between subject and object, active and passive, or past and present.

In addition to experiments with spherical latent variables suggested above, it might also be interesting to look at latent vectors generated by convolutional, instead of recurrent, autoencoders, as recently introduced by Semeniuta et al. [2017].

5 Conclusions

The goal of this project was to investigate how much of the linguistic information in the latent variables of variational autoencoders is invariant across languages. I successfully replicated a variational sequence autoencoder for sentences as described by Bowman et al. [2016] for English and French, obtaining perplexity values similar to theirs. I then investigated to what extent the latent vector representations created by these VAEs are invariant across the two languages. For this, I trained a map aligning the two language-specific latent spaces using a bilingual corpus and examined its behavior by decoding sentences from transformed latent vectors and comparing to target translations. I showed that the map can recover the presence of words, phrases, and the general topic. However, I found no evidence that the latent vectors also encode syntax and semantics in a cross-linguistically invariant manner.

References

Association for Computational Linguistics, editor. *Proceedings of of the 1st Workshop on Evaluating Vector Space Representations for NLP*. 2016. URL http://www.aclweb.org/anthology/W/W16/W16-25.pdf.

S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. Generating Sentences from a Continuous Space. In *Proceedings of CoNLL*, 2016. URL http://arxiv.org/abs/1511.06349. arXiv: 1511.06349.

Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing. Controllable Text Generation. *arXiv preprint arXiv:1703.00955*, 2017. URL https://arxiv.org/abs/1703.00955.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014. URL http://arxiv.org/abs/1312.6114. arXiv: 1312.6114.

Y. Miao, L. Yu, and P. Blunsom. Neural Variational Inference for Text Processing. In *Proceedings of ICML*, 2016. URL http://arxiv.org/abs/1511.06038.

S. Semeniuta, A. Severyn, and E. Barth. A Hybrid Convolutional Variational Autoencoder for Text Generation. *arXiv preprint arXiv:1702.02390*, 2017. URL https://arxiv.org/abs/1702.02390.

X. Shi, I. Padhi, and K. Knight. Does String-Based Neural MT Learn Source Syntax? In *Proc. of EMNLP*, 2016. URL `http://xingshi.me/data/pdf/EMNLP2016long.pdf`.

S. L. Smith, D. H. Turban, S. Hamblin, and N. Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *ICLR 2017*, 2017. URL `https://arxiv.org/abs/1702.03859`.