FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Institute of Formal and Applied
Linguistics
Malostranské náměstí 25
118 00 Praha
Czech Republic

Review of Michal Auersperger's PhD thesis proposal

Michal Auersperger submitted PhD thesis proposal with title *Disentangled Representations for Natural Language Processing*. The idea of disentangled representation is that it should, to quote the proposal, "capture and **isolate** all potentially relevant characteristic of the data".

Proposal Content

Section 2 contains the definition of disentanglement, using the generative modeling framework, which assumes the data **x** is generated from a latent variable **z**. Disentangled representation for data **x** is then a code **c**, in which the individual generative factors are stored as independently as possible (e.g., "time of day" or "roll" of a given photograph), which is deemed desirable for several reasons. However, as mentioned in Section 2.4, it is not clear what these factors should be for example for word representation. Therefore, in NLP, disentangled representations are usually understood in a narrower sense of isolating some predefined features from the representation (i.e., sentence representation could explicitly model "colloquiality" or "sentiment").

Institute of Formal and Applied Linguistics
Malostranské nám. 2/25, 118 00 Praha 1
Czech Republic
phone: 95155 4278, fax: 257 223 293
e-mail: ufal@ufal.mff.cuni.cz

Three methods of disentanglement representation induction are described in Section 3. Two of them, GANs (generative adversarial networks) and VAEs (variational autoencoders), are instances of generative modeling methods. While GANs train only a decoder $p(x \mid z)$, VAEs train both the decoder and an encoder $q(x \mid z)$, and in such a way, that z can be sampled from $N(0, I)$, which means it should be fairly disentangled, given that the individual components of z should be independent. The third described method is based on meta-learning (learning to learn) approach, based on the assumption that retraining on changed distribution of data should be faster and more efficient for disentangled representations.

Section 3.7 discusses the applications of generative models in NLP. An important aspect of generating sentences with GANs is that gradient backpropagation through discrete variables is required. Even if several approaches to tackle the issue have been proposed (the Gumbel-softmax, a.k.a. Concrete, distribution, or REINFORCE-like algorithms), VAEs seem to attract more attention and research.

Most importantly, Section 4 presents three experiments the author plan to execute:

- **Learn BERT-like masked language model with disentanglement using VAE-like approach.**
  To evaluate the disentanglement, the idea is that disentangled representations should speed up transfer learning. The author believes that POS tagging would be a suitable transfer task, given the language modeling objective.
- **Learn disentangled representation by classification.**
  The goal is again to employ the VAE-like approach of the encoder generating a distribution over the latent variable, but with a POS tag decoder.
- **Meta-learning from sentences with gradually changing distribution.**
  Using an encoder/decoder framework, the encoder and the decoder will be alternately trained (with the other being fixed) on sentences with changing distribution. Two specific approaches for distribution shift are proposed – complexity (approximated by length) and topic (generated by a topic model like LDA).

Strengths

Better representations are of course desirable in NLP. Even if hugely successful methods have appeared recently, they are usually extremely data-hungry and performance on out-of-domain data deteriorates

dramatically. Disentangled representation would probably bring improvements on these fronts.

I like that the goal is "blue-sky science" enough, while promising possible improvements.

Note that a paper performing an experiment similar to the second one proposed in Section 4 won the Best paper award on EMNLP 2019 (Li and Eisner, 2019).

Weaknesses

Even though the author is in his third year, the proposal does not contain any original results. That is not necessarily a problem, but I would expect it would still take several years to collect enough material for the thesis.

The quite generic notion of disentanglement might make it difficult to design experiments and also to evaluate them. Further, incremental improvements might not be enough, and novel ideas might be required (that is not by itself bad, it is science after all). Finally, the experiments can be easily very computationally demanding (as the first and third proposed experiments), to show promising results compared to recent techniques like BERT or GPT-3.

Comments Regarding the Proposed Experiments

I am a bit worried that the experiments as proposed will require a lot of computation (assuming Transformer-like architecture, note that the best Martin Popel's systems take weeks to train; similarly, BERT training requires ~weeks to train on GPUs [Google BERT is trained on TPUs, which is hard to compare, but Finnish BERT was trained for 12 days on 8 V100 GPUs, see https://arxiv.org/abs/1912.07076). Given that they will probably require non-trivial hyperparameter tuning (the latent loss strength during VAE training, the domain shift and meta-learning hyperparameters in learning to learn approach), it might take some time to execute them successfully.

In the second proposed experiment you get close to probing – the Hewitt and Liang (2019) paper is a nice read (for example, it advocates weak – linear – probes).

The EMNLP 2019 best paper Li and Eisner (2019) is based on an experiment similar to the second one proposed. The authors use pre-trained ELMo embeddings and then use a VAE-like approach to produce task-specific representation, which keeps only the information useful for the downstream task (dependency parsing is used in the paper).

Furthermore, they propose both a continuous and discrete representations. I consider the goals of the paper to be aligned with this thesis proposal, so some inspiration might be gained from it. A good thing is that using pre-trained embeddings lowers the requirements of the experiments considerably.

The disentangled representations might be useful in controlled text generation. For example, Hu et al. (2017) proposed a VAE-like encoder/decoder, where the latent variable was composed of unstructured part **z** and structured part **c**, with the goal of being able to control some aspect of the generated texts through the structured part ("sentiment" and "tense" is used in the paper). The ACL 2019 paper Vineet et al. (2019) pushed the boundaries further. This might also be an interesting direction to pursue.

Remarks to the Texts

- The references to sections are written with parentheses in Section 1; also, "section" and "Section" are both used, and a fullstop is missing at the end of Section 1.
- If the text of Section 3.3 was used in a paper, the steps from (8) to (9) should be more detailed. Similarly, more information about the reparametrization trick might also be useful.
- The prior $N(\mu, I)$ in Section 3.3.1 should probably be $N(0, I)$.
- As the end of Section 3.7, when GANs or VAEs are used in an encoder-decoder architecture, it is difficult to incorporate attention, which should probably be mentioned, given that Bahdanau et al. (2015) is cited. (I just found https://arxiv.org/abs/1712.08207 which tries to address it, nice).
- Maddison et al. (2016) and Jang et al. (2016) should also be cited when discussing Gumbel-softmax distribution in eq (12).

Random Questions

- Success of predicting individual components of a latent variable from all others seems like a natural notion of disentanglement. Do you know if someone tried to disentangle representations by using something like VAE + adversarial training on this objective (not being able to predict the components from other ones)?
- Do you have any plans about **discrete** representations?

Conclusion

I find the proposed topic to be a good one for dissertation, and the author demonstrates understanding in the methods. Therefore, I recommend the proposal to be defended. However, the lack of own results or experiments is a bit troubling (regarding the time required to finish the studies).

In Strakonice, 3rd June 2020

RNDr. Milan Straka, Ph.D.

References

Xiang Lisa Li, Jason Eisner: *Specializing Word Embeddings (for Parsing) by Information Bottleneck.* https://arxiv.org/abs/1910.00163

John Hewitt, Percy Liang: *Designing and Interpreting Probes with Control Tasks*. https://arxiv.org/abs/1909.03368

Vineet John, Lili Mou, Hareesh Bahuleyan, Olga Vechtomova: *Disentangled Representation Learning for Non-Parallel Text Style Transfer*. https://www.aclweb.org/anthology/P19-1041.pdf

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, Eric P. Xing: *Toward Controlled Generation of Text*. https://arxiv.org/abs/1703.00955

Chris J. Maddison, Andriy Mnih, Yee Whye Teh: *The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables*. https://arxiv.org/abs/1611.00712

Eric Jang, Shixiang Gu, Ben Poole: *Categorical Reparameterization with Gumbel-Softmax*. https://arxiv.org/abs/1611.01144