Topic Enhanced Language Model (Proposal)

Ao Liu
University of Massachusetts Amherst
College of Information and Computer Science
aoliu@cs.umass.edu

Abstract

We want to design a novel language model that is capable of capturing local semantics, which we view as a subset of global semantics, in order to help predicting the next word given some context. Traditional RNN-based language models are good at capturing the local syntactic structures of a word sequence, but difficult to remember global semantic dependencies due to the vanishing gradient problem. However, such semantics can be learn by a topic model. We can take the advantages of unsupervised topic models to enable language models having better expressiveness. The proposed model can have more capacity to model natural language and potentially benefit downstream tasks, such as producing highly contextualized word embeddings, text generation and document ontology classification and ranking.

1 Introduction

Statistical language modelling has evolved from traditional N-gram language models to neural language models over the past two decades (Bengio et al., 2003). Traditional N-gram approaches rely on Markov assumption, which usually has a high cost to learn long-distance dependency. The introduction of RNN-based language model relieves such burden by computing the conditional probability of a word given all previous words using hidden state(Mikolov et al., 2010). Despite the difference of model architectures, statistical language models factorize the probability of joint probability of a given sequence $S = (w_1, ..., w_N)$ into conditional probabilities: $P(S) = \prod_{t=1}^{N} P(w_t|w_{m:t-1})$, where $w_{m:t}$ is the sequence $w_m, w_{m+1}, ..., w_t$. Notice that when $m = t - N + 1$, the factorization is for N-gram models, whereas when $m = 1$, the conditional probability is based on all previous words.

Although RNN-based language models have higher generalizability by capturing long-distance information, they may in practice encounter overfitting issues (Srivastava et al., 2014). A general goal of language model is to capture both the syntax and the semantic coherence in the language. Syntactic structures mostly depend on local context, and semantic dependencies can have arbitrary long distance.

Probabilistic topic model is one way to learn such semantic coherence (Blei and Lafferty, 2009). A general goal of probabilistic topic models is to learn abstract topics formed by groups of words and to express documents as combinations of those topics. By doing so, such abstract topics are then embedded with semantic dependencies over the words and could potentially be utilized by language model to capture long-distance information without overfitting.

Previous approach that incorporates topic model (Dieng et al., 2016) shows that such approach can help improving the perplexity of language models. Their approach only excludes the effects of stop words and the topic model remains global. In general, a sentence only contains a subset of the global topics in a document. If we consider the global semantics for every sentence equally, local topic alternation is then ignored and may affect the precision of modelling sentences.

State-of-the-art approach of language model (Yang et al., 2017) views the task as a matrix factorization problem and introduces mixture of softmax (MoS) to allow the probability matrix to have high rank in order to increase the expressiveness of a language model. However, the learned latent states may be hardly interpretable due to the nature of deep neural networks.

2 Preliminary Experiment

To evaluate our model, we will first compare the model perplexities over language modelling datasets, such as Penn Treebank (PTB) and WikiText. After that, we would also want to test its performance on downstream tasks - text generation and document ontology classification and ranking.

3 Evaluation

Evaluation will be based on my completion of each goal set during meetings. Progress will also be evaluated on the documentation of code written and the error analysis on my implementation of the system. Specific goals will be adjusted based on feedback given in meetings.

4 Timeline

Specific dates will be adjusted as the need arises.

September 20th, 2018: Finish implementing the baseline language models if needed

September 27th, 2018: Finish analysis on baseline language models

October 11th, 2018: Finish designing a language model that can incorporate Haw-Shiuan's topic model

November 1st, 2018: Finish implementing the designed language model

November 15th, 2018: Complete some experiments and analysis on the language model

November 29th, 2018: Complete all the experiments and analysis on the language model

December 6th, 2018: Finish fist draft for submitting to NAACL

January 24th, 2019: Start working on downstream tasks

February 7th, 2019: Finish implementing the baseline models for text generation and document ontology classification and ranking if needed

February 14th, 2019: Finish analysis on baseline models

March 7th, 2019: Finish designing and implementing the text generation model

March 21th, 2019: Complete all the experiments and analysis on the text generation model

April 11th, 2019: Finish designing and implementing the document ontology classification and ranking model

April 25th, 2019: Complete all the experiments and analysis on the document ontology classification and ranking model

May 2nd, 2019: Finish the first draft of final report

May 9th, 2019: Submit the final report

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

David M Blei and John D Lafferty. 2009. Topic models. *Text mining: classification, clustering, and applications*, 10(71):34.

Adji B Dieng, Chong Wang, Jianfeng Gao, and John Paisley. 2016. Topicrnn: A recurrent neural network with long-range semantic dependency. *arXiv preprint arXiv:1611.01702*.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

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.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W Cohen. 2017. Breaking the softmax bottleneck: a high-rank rnn language model. *arXiv preprint arXiv:1711.03953*.