Character Based Language Models Through Variational Sentence and Word Embeddings

Zaccary Alperstein and Kevin Dsouza

Character based model

Pros :

- Character based language models have a much smaller input space thus requiring less memory
- They do not have problems with out of vocabulary words, or lexemes
- No softmax bottleneck
- Languages with rich morphology like German, Finnish, Turkish and Russian are modelled better by character level language models as they can have extremely large vocabularies

  Ex. ‘***Unabhaengigkeitserklaerungen***’ : independence declarations

Cons :

- exponential explosion in possible character combinations
- Long term dependencies become **really** long

Wordpiece model (Google's NMT)

Word: Jet makers feud over seat width with big orders at stake

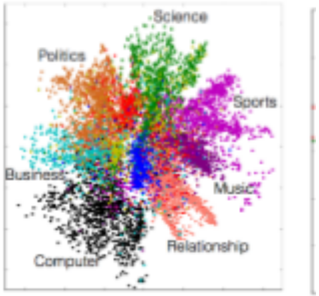
wordpieces: _J et _makers _fe ud _over _seat _width _with _big _orders _at _stake

- Create a bigger vocabulary by making wordpieces and handle OOV words by breaking them into letters with special tokens attached
- Wordpiece created by language model by iteratively combine N-grams and re-training
    - Clearly heuristic as it is intractable to explore most N-grams
    - Also try training mixed character/word model

Still looking for a general approach to language modelling with an infinitely large dictionary

Variational Autoencoders as Language Models

- Text VAEs recently achieved SOTA on a few language modelling datasets (yelp, yahoo)
- VAEs allow us to build our models with well defined latent variables



(a) Yahoo

Citation: Improved Variational Autoencoders for Text Modeling using Dilated Convolutions/ Semi-supervised Lerning with Deep Generative Models

Literature survey

1. Bahuleyan, Hareesh, et al. "Variational Attention for Sequence-to-Sequence Models." *arXiv preprint arXiv: 1712.08207* (2017).
2. Sønderby, Casper Kaae, et al. "Ladder variational autoencoders." *Advances in neural information processing systems*. 2016.
3. Vaswani, Ashish, et al. "Attention is all you need." *Advances in Neural Information Processing Systems*. 2017.
4. Bowman, Samuel R., et al. "Generating sentences from a continuous space." *arXiv preprint arXiv:1511.06349* (2015).
5. Kim, Yoon, et al. "Character-Aware Neural Language Models." *AAAI*. 2016.
6. Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." *arXiv preprint arXiv:1609.08144* (2016).
7. Hwang, Kyuyeon, and Wonyong Sung. "Character-level language modeling with hierarchical recurrent neural networks." *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017.
8. Goyal, Prasoon, et al. "Nonparametric variational auto-encoders for hierarchical representation learning." *arXiv preprint arXiv:1703.07027* (2017).

Attention is all you need

- Purely attention based models have recently gained SOTA in sequence translation

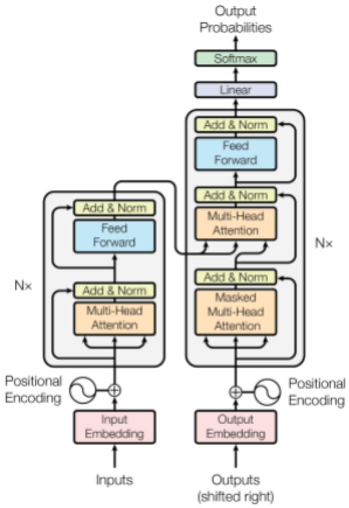
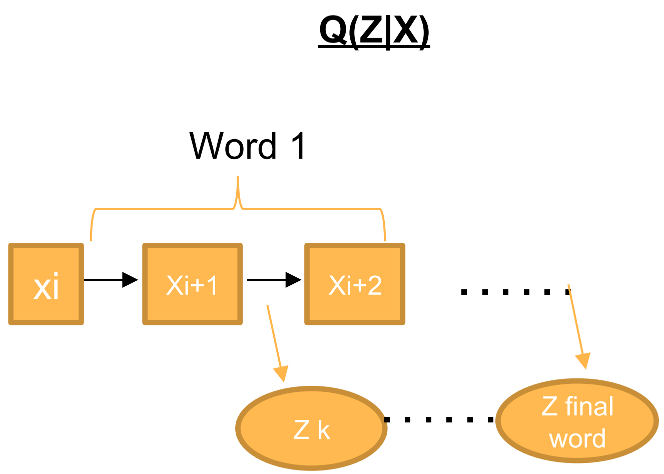
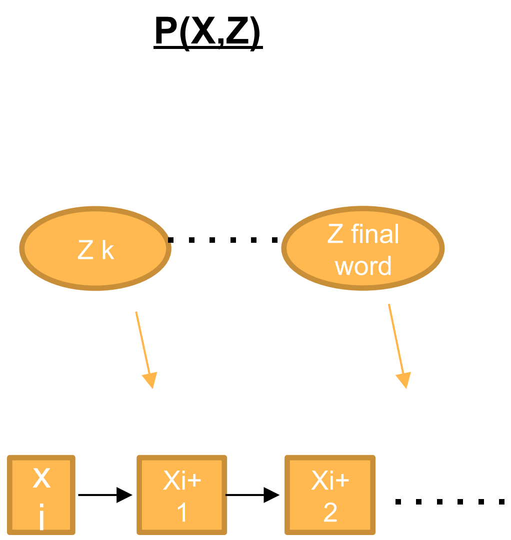


Figure 1: The Transformer - model architecture.

Citation: Attention is all you need

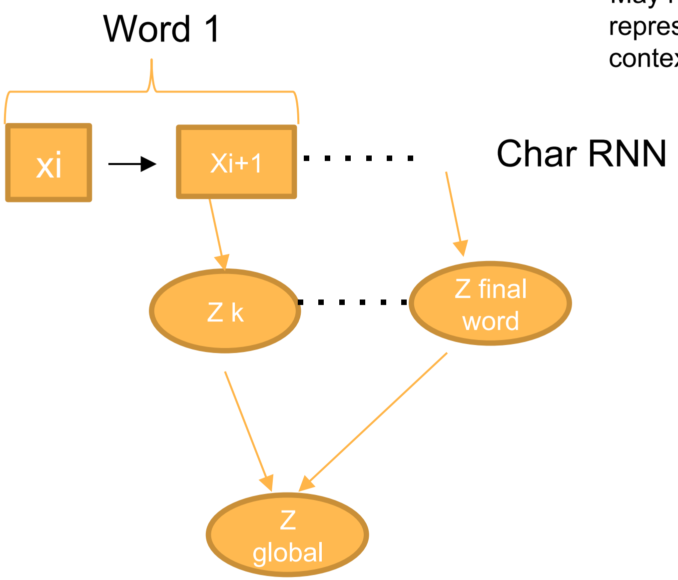
Proposed Variational Framework





Proposed Variational Framework: Hierarchial

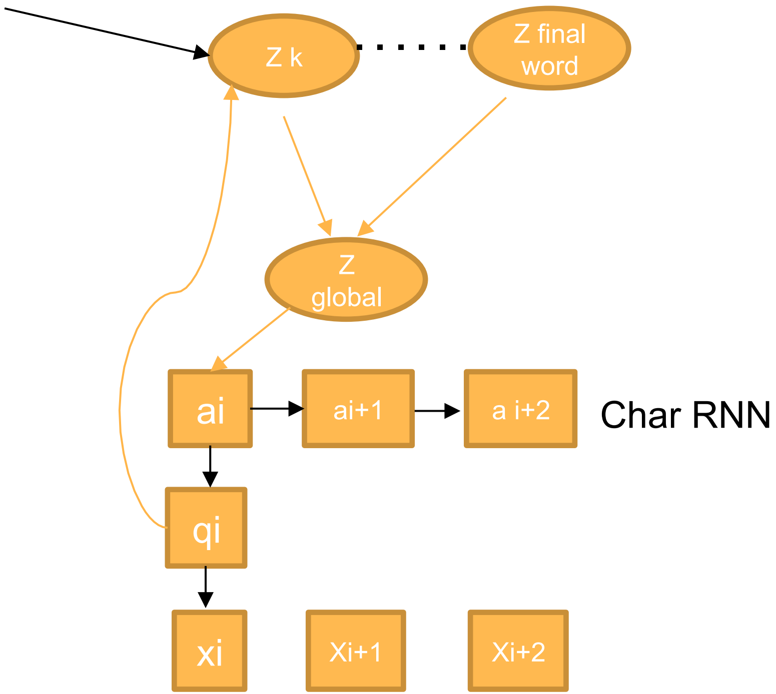
<u>Q(Z|X)</u>



Independent Gaussian(s) ?

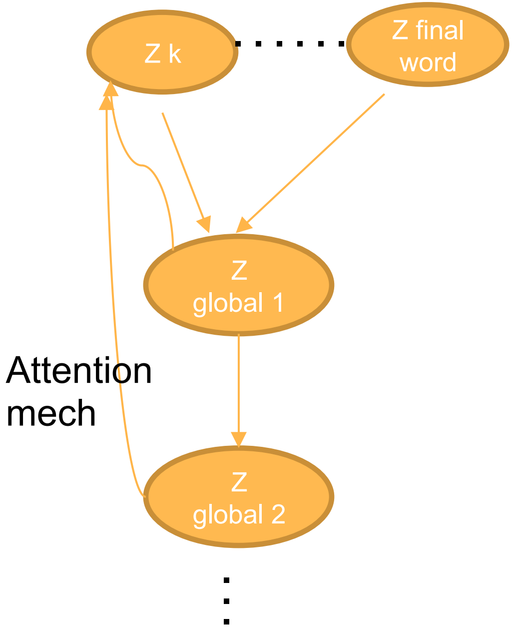
- May help word representations be context independent

<u>P(X,Z)</u>

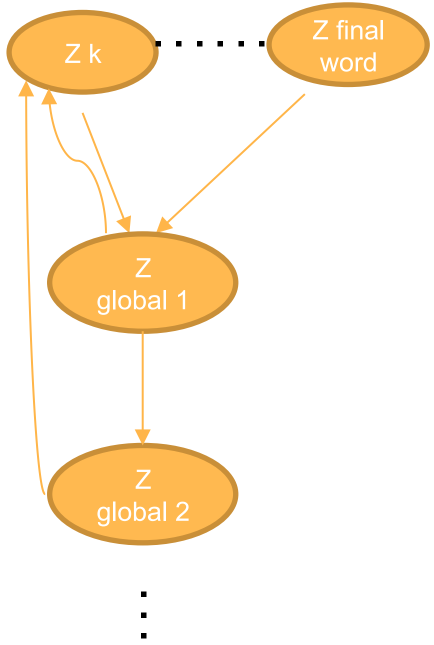


Proposed Variational Framework: Hierarchial, adding layers

Q(Z|X)



P(X,Z)



Advantages of our model

- Hierarchical posterior and prior for rich representations
- An embedding model with an infinite vocabulary
- Can be used for neural machine translation
- Possible exploration of homotopies between sentences
- Imputing missing words
- Latent variables extract meaning
  - Sentence representations
  - Word representations
  - conditioning

Evaluation / Datasets

Datasets (compare with previous work):

- Pen Tree Bank
- Yelp
- Yahoo

Evaluation (compare with other generative models like vanilla RNNs):

- ELBO
- Importance weighted estimate of NLL

Timeline

1. Algorithm writing, debugging

2. Hyper-parameter tuning on PTB

3. evaluation on other datasets, apply to translation with our pre-trained character based word encodings

Thanks

Proposed Variational Framework

Q(Z|X)

P(X,Z)