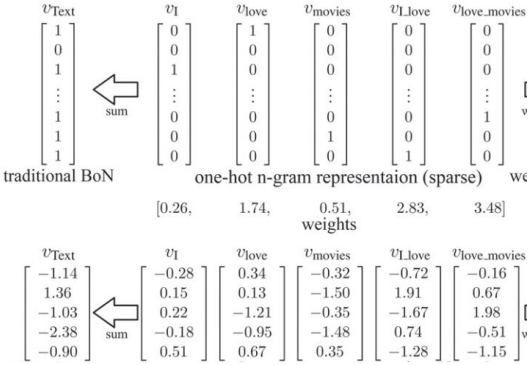
Analysis of Bag-of-n-grams Representation Properties

Team 4
Qi Huang, Zhanghao Chen, Zijie Lu, Ye Yuan

Motivation

- The simple and straightforward bag-of-n-grams model outperforms more sophisticated models such as RNN and CNN
- This suggests bag-of-n-grams representation may contain most, if not all information
- **How much information is contained? What kinds of information are contained?**



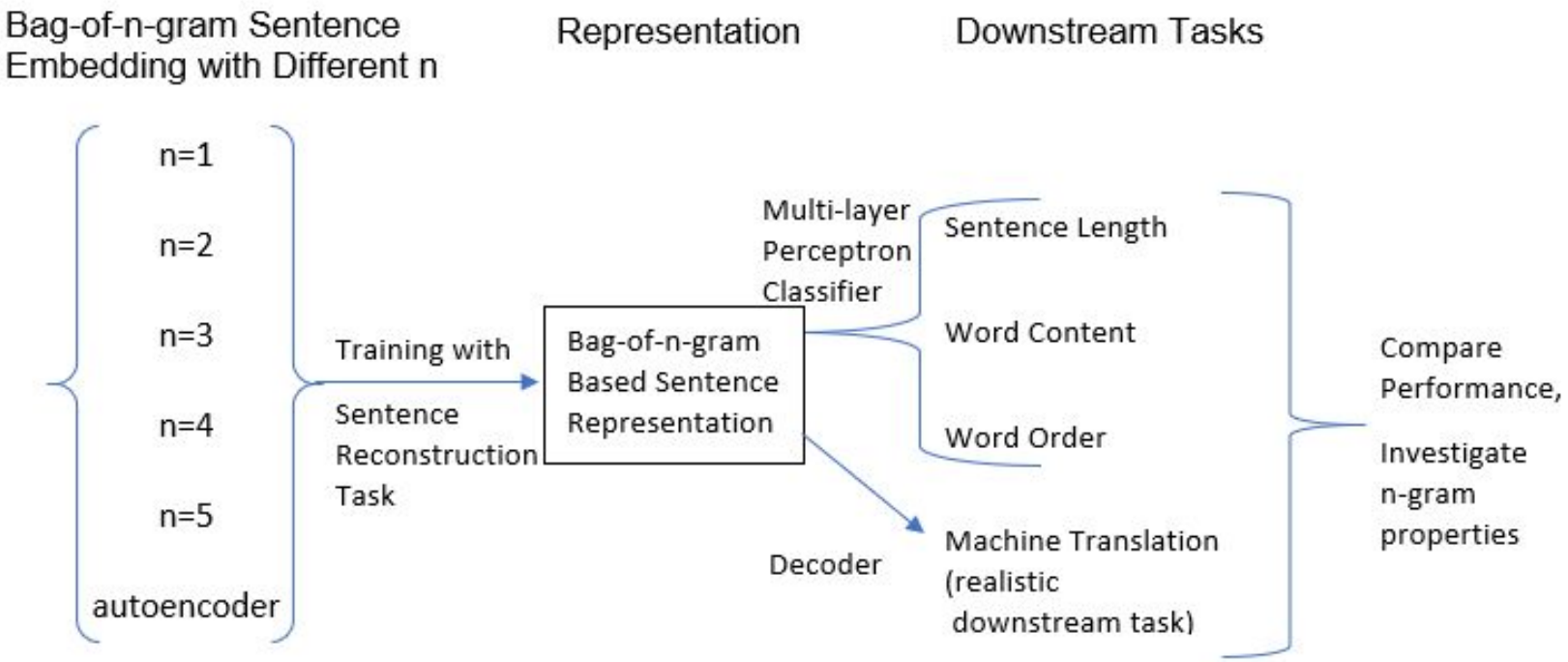
Picture adopted from (Li et al, 2017)

Hypothesis

We expect that...

- Bag-of-n-grams representation ($n \geq 2$) contains more information than bag-of-words (unigram) representation.
- The amount of information contained in bag-of-n-grams representation increases as n increases
- RNN encoder should still outperform bag-of-n-grams representation

Q: What information does bag-of-n-grams contain?
Approach



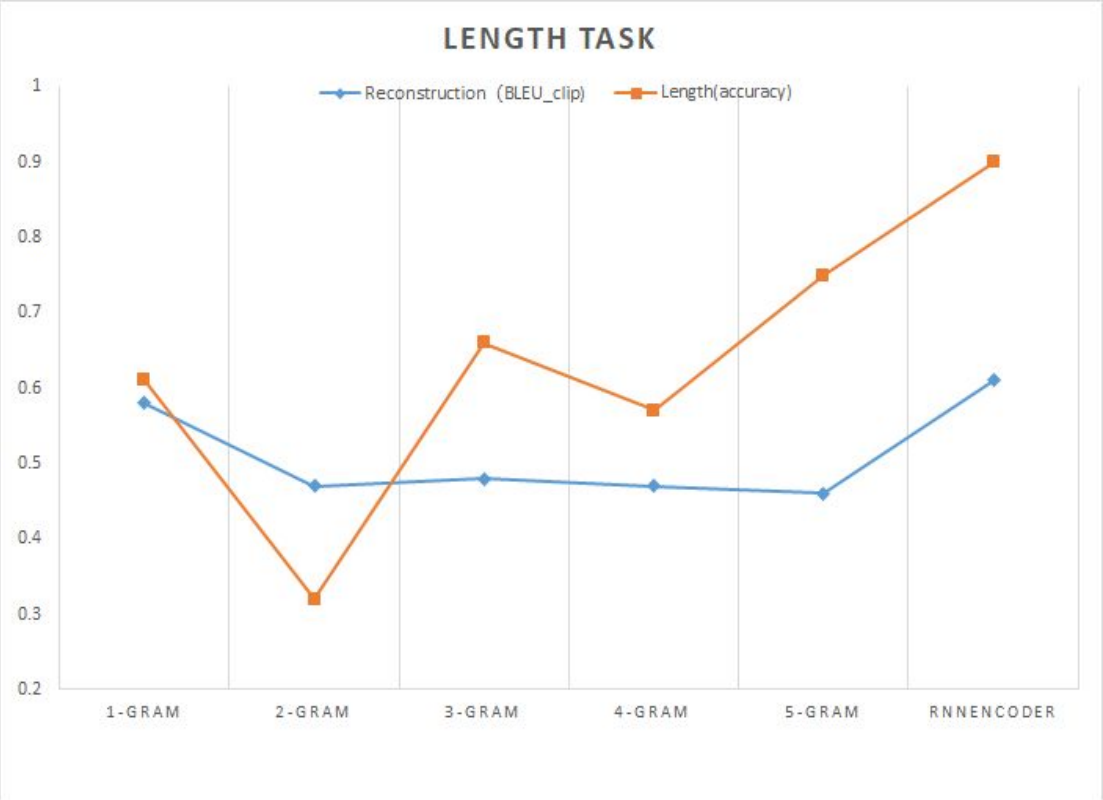
Partial Result — Sentence Reconstruction

(Metric: BLEU_clip)	Short sentences (length <= 6)	Long Sentences (length > 6)	Overall
1-gram	**0.68**	**0.45**	**0.58**
2-gram	0.56	0.35	0.47
3-gram	0.58	0.35	0.48
4-gram	0.57	0.34	0.47
5-gram	0.57	0.32	0.46
RNNEncoder	**0.69**	**0.50**	**0.61**

Dataset: Tab-delimited Bilingual Sentence Pairs
Training setting: 20000 training sentences, 5000 test sentences

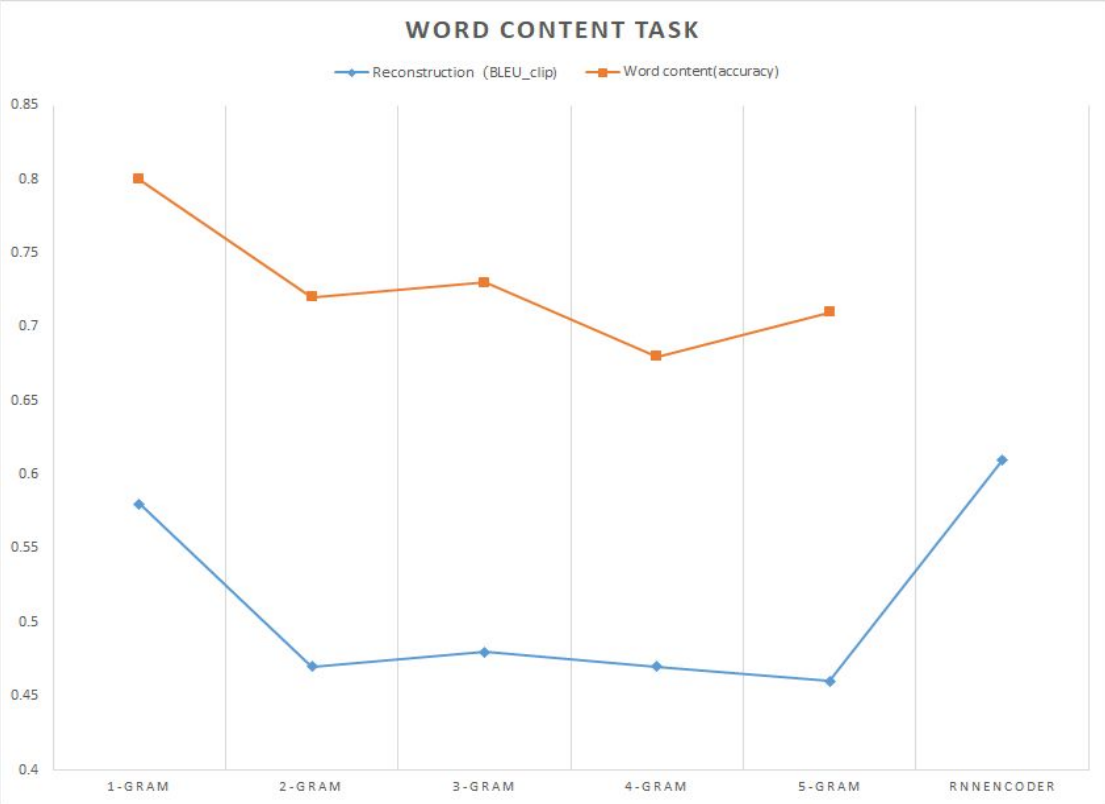
Downstream Task - Length

- Even Ns are cursed?
- Higher-order N-gram representation captures more length information?



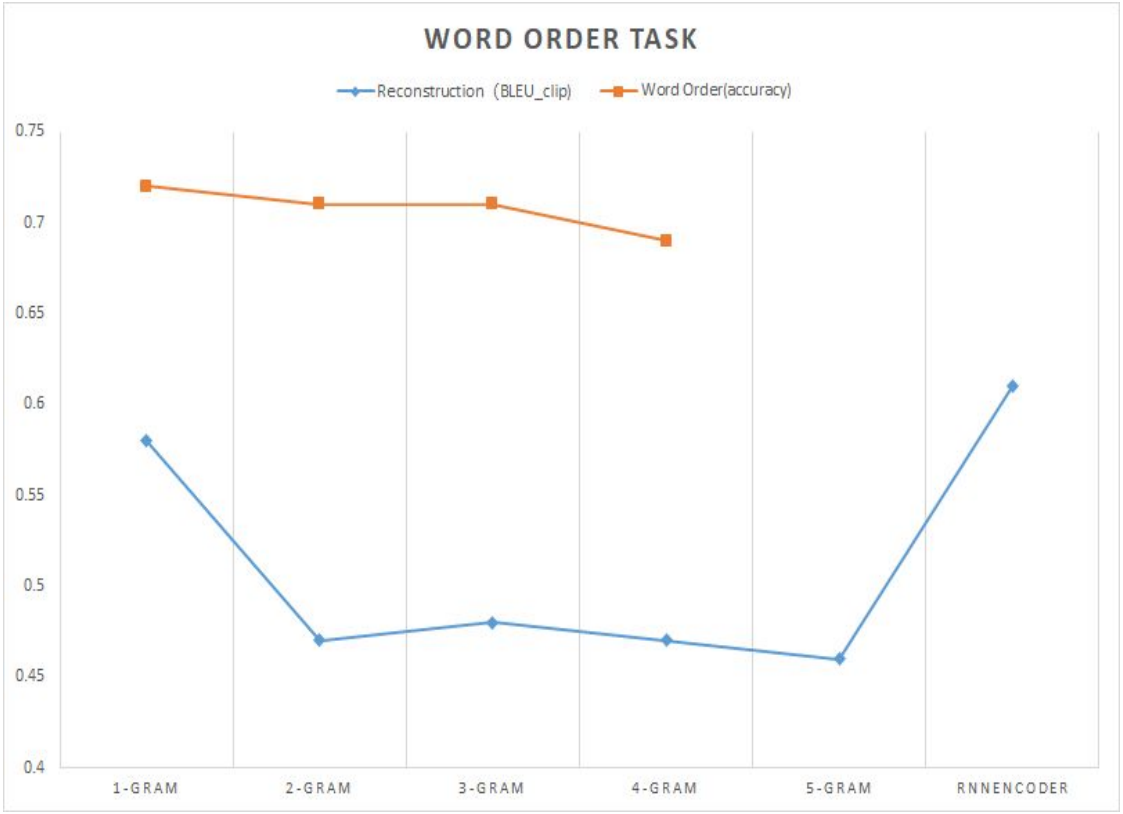
Downstream Task - Word Content

Unigram representation captured more word content information



Downstream Task - Word Order

Different N-gram representation captures similar word order information



Partial Result

- Unigram representation is surprisingly effective
  - Yet still worse than RNN encoded representation
- Higher-order N-grams does not live up to our expectation:
  - Sparsity of N-gram Dictionary
  - Insufficient Data

Bag-of-n-grams representation contain most information of a sentence?

Not really... for now

Thank you!

Questions?

Future Work

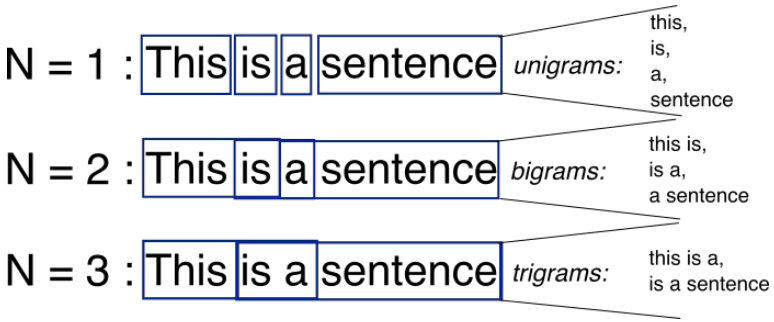
- More training, more data.
- With attention mechanism: summing N-gram representation based on N (N bags instead of one bag), train with attention, use attention weights distribution during inference to qualitatively measure the choice of n's impact on sentence reconstruction (higher the weight the more important)
- Attend to n-gram directly instead of attending to bag-of-n-gram.

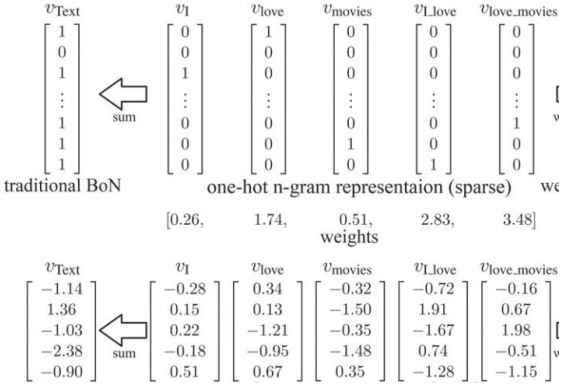
- Bag-of-N-gram is a simple model after all ... but how to measure the complexity of a neural network based model overall? (VC dimension, traditional ML algorithm?)

- N-gram can be thought of a new "language" with more severe sparsity problem..

Little bit of Background

- A big dictionary of all possible n-grams
- N-gram ➔ index ➔ one-hot vector ➔ fixed-size embedded vector
- Bag-of-n-grams representation:
  - The average sum of all embedded n-gram vector





Picture adopted from (Li et al, 2017)

Q: What information does bag-of-n-grams contain?

Analysis Approach

Step1: Learn a good bag-of-n-grams representation first!

- Use sentence reconstruction task to obtain a good bag-of-n-grams embedding

Step2: Downstream task

- Inspired by (Adi et al, 2017)
- Simple task, i.e.sentence length classification
- Adding practical task, i.e. Machine Translation
- Basic Premise: the trained embedding preserves the most information of the sentence.