Few-Shot Representation Learning for Out-Of-Vocabulary Words

Ziniu Hu, Ting Chen, Kai-Wei Chang, Yizhou Sun
University of California, Los Angeles

Our Code:

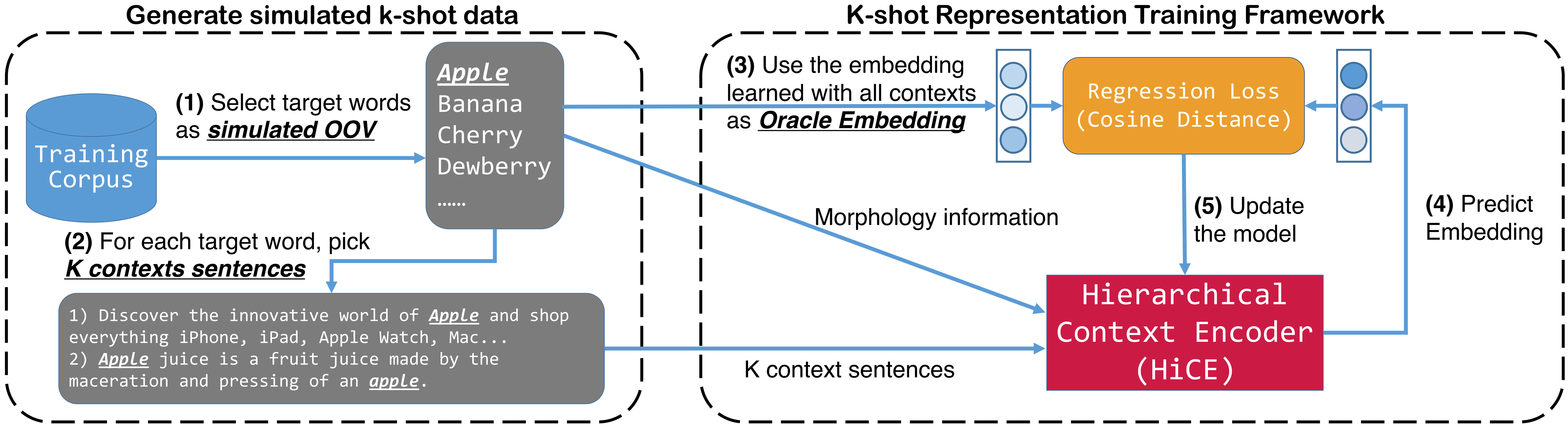
Learning representation for Out-Of-Vocabulary (OOV) words is challenging

Challenge: OOV in new corpus sometimes have insufficient contexts, and its morphology might not be informative enough.

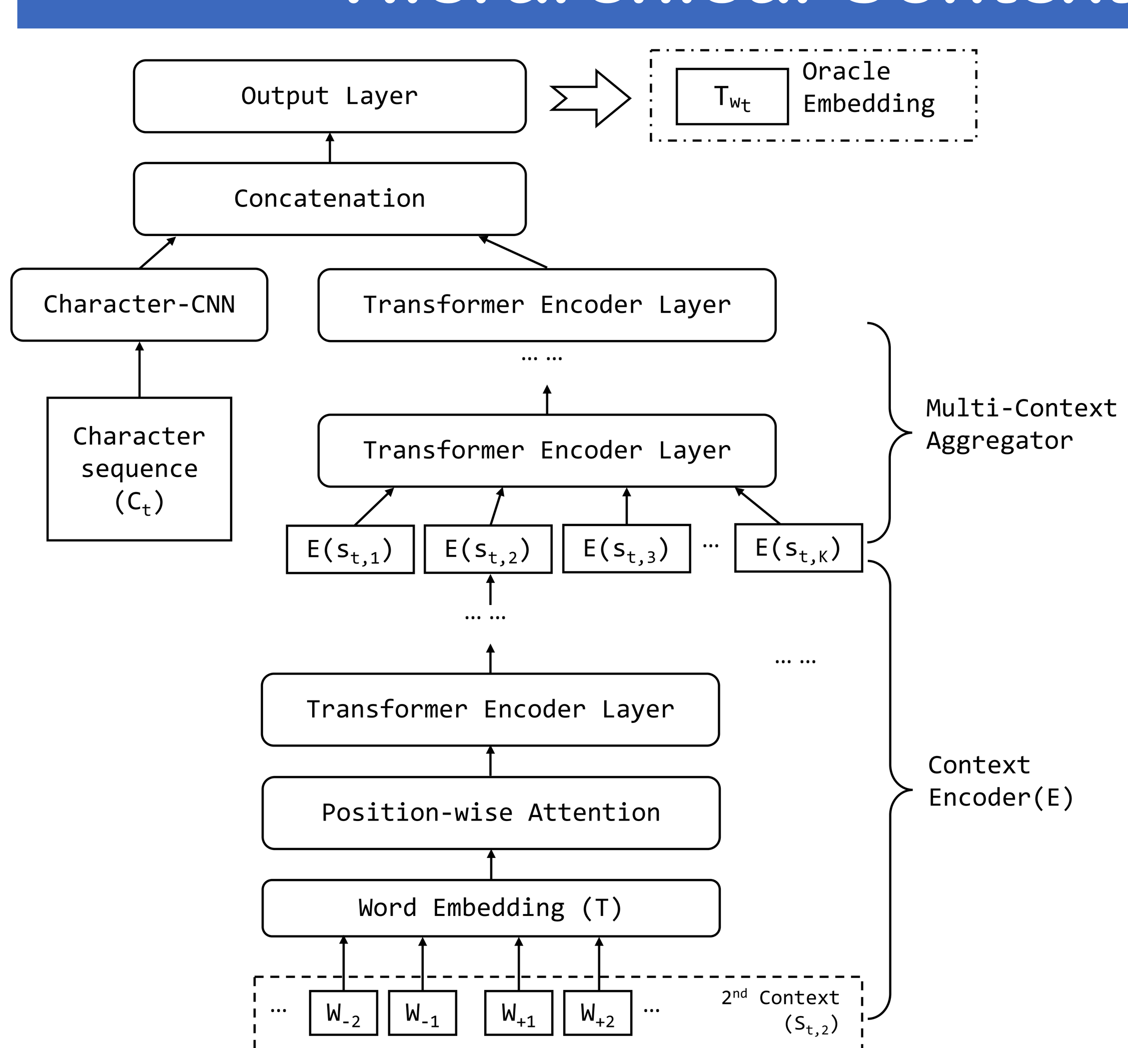
Goal: Can we learn accurate embedding for ***OOV words*** by observing their usages in ***only a few context sentences***?

Our Approach: 1) Design a model with enough capacity to capture semantics and relations in context sentences.
2) Train the model by simulating an OOV scenario, such that the model learns to generalize well for OOV on a new corpus.

Our Solution: Formulate as K-shot Regression Problem

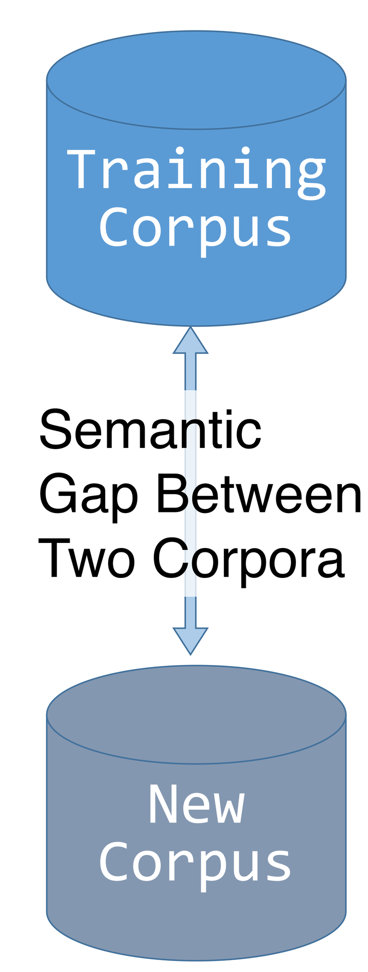


Hierarchical Context Encoder



- Lower **Context Encoder** encodes a single context sentence into vector E(s)
- Upper **Multi-Context Aggregator** combines multiple context vectors E(s) with different attention
- The morphological information is incorporated by a Char-CNN.

Robust Domain Adaptation with MAML



- **Goal:** Learn to adapt to new domain with a few examples
- **Solution:** Model Agnostic Meta-Learning (MAML) [1]
- **Procedure**:
  1) Run a few steps on training corpus (D_T) to obtain a promising initialization.
  2) Update the model on new corpus (D_N) with the gradient calculated at that initialization point.

Initialization: $\theta^* = \theta - \alpha \nabla_\theta \mathcal{L}_{D_T}(\theta)$.

Meta Update: $\theta' = \theta - \beta \nabla_\theta \mathcal{L}_{D_N}(\theta^*)$
$= \theta - \beta \nabla_\theta \mathcal{L}_{D_N}(\theta - \alpha \nabla_\theta \mathcal{L}_{D_T}(\theta))$

[1] Chelsea Finn et al. Model-agnostic meta-learning for fast adaptation of deep networks. ICML' 17

Intrinsic Evaluation (Chimera Dataset)

- Chimera provides K context sentences for a OOV word and six probe words, with human label of similarity between each probe word and the OOV word.
- The task requires the estimated embedding similarity close to the human label

Spearman correlation.

Methods	2-shot	4-shot	6-shot
Word2vec	0.1459	0.2457	0.2498
FastText	0.1775	0.1738	0.1294
Additive	0.3627	0.3701	0.3595
nonce2vec	0.3320	0.3668	0.3890
à la carte	0.3634	0.3844	0.3941
HiCE w/o Morph	0.3710	0.3872	0.4277
HiCE + Morph	**0.3796**	0.3916	0.4253
HiCE + Morph + Fine-tune	0.1403	0.1837	0.3145
HiCE + Morph + MAML	0.3781	**0.4053**	**0.4307**
Oracle Embedding	0.4160	0.4381	0.4427

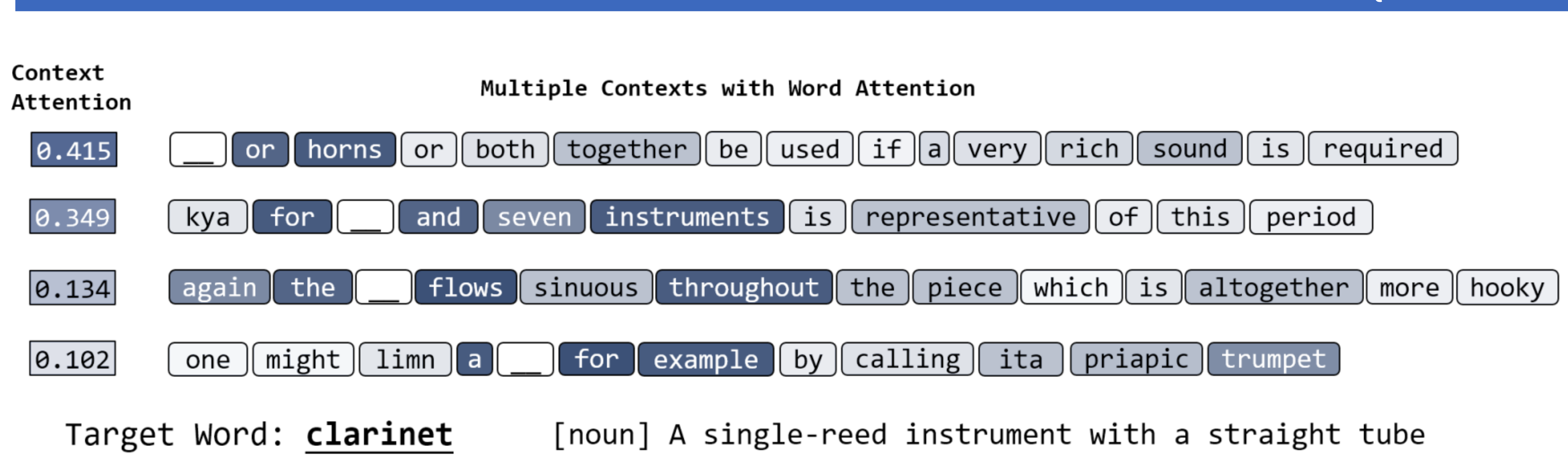
- Learn on WikiText-103 and test on Chimera Benchmark
- Morphological information is useful when context is limited, but not helpful when context is rich (4 or 6)
- MAML is helpful when context is rich

Extrinsic Evaluation (NER, POS Tagging)

Methods	Named Entity Recognition (F1-score)		POS Tagging (Acc)
	Rare-NER	Bio-NER	Twitter POS
Word2vec	0.1862	0.7205	0.7649
FastText	0.1981	0.7241	0.8116
Additive	0.2021	0.7034	0.7576
nonce2vec	0.2096	0.7289	0.7734
à la carte	0.2153	0.7423	0.7883
HiCE w/o Morph	0.2394	0.7486	0.8194
HiCE + Morph	0.2375	0.7522	0.8227
HiCE + Morph + MAML	**0.2419**	**0.7636**	**0.8286**

- Our proposed few-shot representation learning approach benefits downstream tasks.
- With MAML, the performance improves further.

Qualitative Evaluation (Attention Visualization and Case Study)



- HiCE is able to pick important words related to the target word (clarinet), such as “horns”, “instruments”, “flows”
- For sentence containing less or vague information (the forth sentence for example), HICE assigns a lower sentence attention.

OOV Word	Contexts	Methods	Top-5 similar words (via cosine similarity)
scooter	We all need vehicles like bmw c1 scooter that allow more social interaction while using them ...	Additive FastText HiCE	the, and, to, of, which cooter, pooter, footer, soter, sharpshooter cars, motorhomes, bmw, motorcoaches, microbus
cello	The instruments I am going to play in the band service are the euphonium and the cello ...	Additive FastText HiCE	the, and, to, of, in celli, cellos, ndegocello, cellini, cella piano, orchestral, clarinet, virtuoso, violin
potato	It started with a green salad followed by a mixed grill with rice chips potato ...	Additive FastText HiCE	and, cocoyam, the, lychees, sapota patatoes, potamon, potash, potw, pozzato vegetables, cocoyam, potatoes, calamansi, sweetcorn

- Additive (averaging context word embedding) makes OOV embedding near to some top frequent words as “the”, “and”, “of”.
- FastText (averaging sub-word embedding) finds words that looks similar but have a totally different meaning. For example, scooter (vehicle) vs cooter (turtle)
- Our method can capture the true semantic meaning of OOV words better.