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

https://www.aclweb.org/anthology/P19-1402.pdf

목차

- 개념 정리
- 서론
- 모델
- Experiments

개념 정리

- Few-Shot Learning
    - 간단 정의 : 아주 적은 데이터로도 데이터의 특징을 식별하도록 하는 것. 즉, 적은 데이터로 새로운 기술을 배우거나, 새로운 환경에 빠르게 적응할 수 있도록 설계하는 것.

개념 정리

- Few-Shot Learning
    - 기존 deep learning의 경우,
      수백장 사진을 통해 강아지라는
      class의 특성을 학습 →
      사람이 **한 장의** 강아지의 사진을 통
      강아지라는 class의 concept를
      학습 →
      network 학습시에도 class 별로
      몇 장(K-Shot)의 이미지를 보여주
      network를 학습시킴.
        - K-Shot : 각 class 별
          Data의 갯수(K)

[example : 4-Shot]

개념 정리

- OOV word
  - Out Of Vocabulary word → <unk>
  - Vocabulary : 기계가 알고 있는 단어들의 집합
  - example
    - i go to school　i go to home

      ↓ integer encoding　↓ integer encoding

    - 1 3 2 0　　1 3 2 0

index	word
0	<unk>
1	i
2	to
3	go

개념 정리

- word embedding
  - 단어를 dense vector의 형태로 표현하는 방법

- word embedding vector
  - word embedding 결과로 나온 vector

- ex)

word embedding vector

go $\xrightarrow{\text{word embedding}}$ [-4.3, -1.4, 2.5, 2.5]

서론

- 문제점
  - in real-word scenarios, OOV words that do not appear in training corpus emerge frequently.

    → infer embeddings for OOV words that are not observed in the training corpus (D_T) based on a new testing corpus (D_N)
    - D_N is usually much smaller than D_T
    - the OOV words might only occur for a few times in D_N
    - thus, it is difficult to directly learn their embedding from D_N

서론

- 해결방법
  - Few-Shot Regression Framework →
    to infer embeddings for OOV words that are not observed in the training corpus (D_T) based on a new testing corpus (D_N)

  - HiCE(anttention-based **Hi**erarchical **C**ontext **E**ncoder) →
    to leverage both sentence examples and morphological information

  - MAML(Model Agnostic Meta-Learning) →
    to assist the fast and robust adaptation of a pre-trained HiCE model
    (새로운 도메인이나 Downstream task에 적용하기 위해)

Model : Few-Shot Regression Framework

- 목표
  - infer embeddings for OOV words that are not observed in the training corpus (D_T) based on a new testing corpus (D_N).

- training objective

$$\hat{\theta} = \arg\max_{\theta} \sum_{w_t} \sum_{\mathbf{S}_t^K \sim \mathbf{S}_t} \cos\left(F_\theta(\mathbf{S}_t^K, C_t), T_{w_t}\right), \tag{1}$$

  - $\{w_t\}_{t=1}^{N}$: N words as the target words.
  - $S_t^K \sim S_t$: the K sentences containing target word w_t are randomly sampled from all the sentences containing w_t.
  - $F_\theta(\cdot)$: word embedding learning algorithm.
  - T_{w_t} : target word's embedding as oracle embedding.

Model : HiCE

- 장점
  1. analyze the complex semantics of context.
  2. aggregate multiple pieces of context information for comprehensive embedding prediction.
  3. incorporate morphological features.

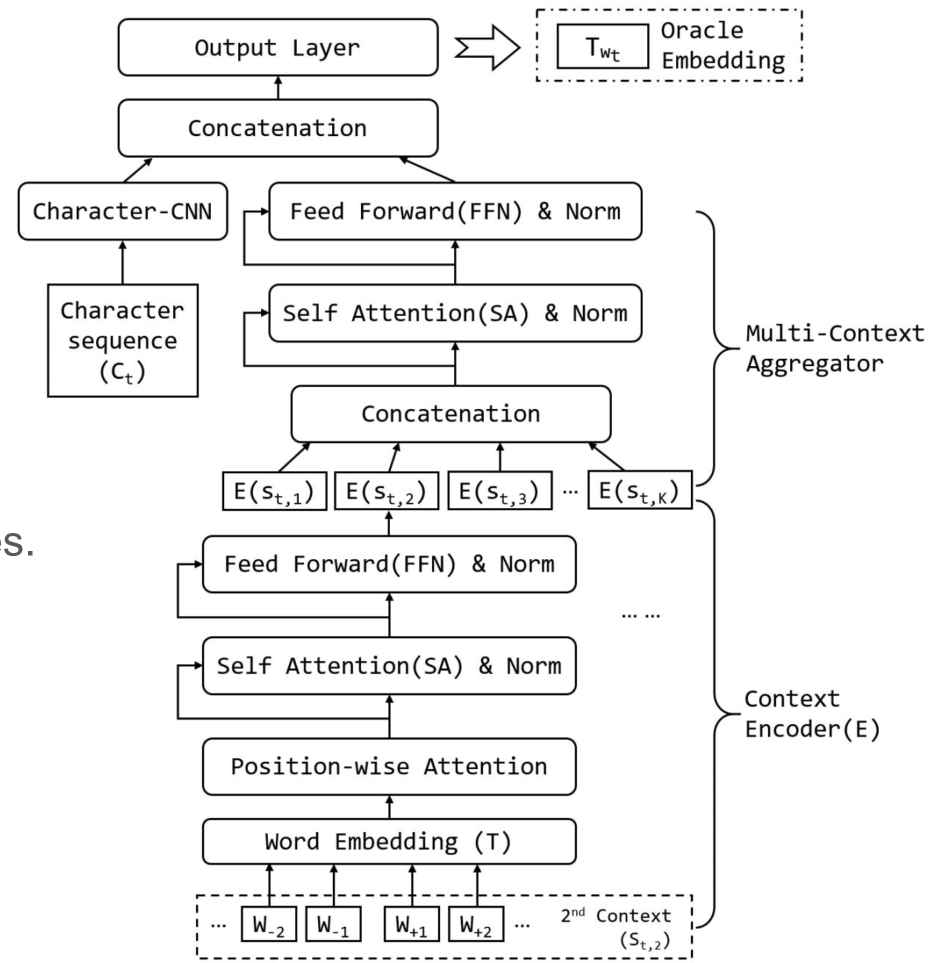

Figure 1: The proposed hierarchical context encoding architecture (HiCE) for learning embedding representation for OOV words.

Model : MAML for Fast & Robust Adaption

So far, we directly apply the learned neural regression function F_θ trained on D_T to OOV words in D_N .

문제점

This can be problematic when there exists some linguistic and semantic gap between D_T and D_N .

해결방법

Model : MAML for Fast & Robust Adaption

fine-tuning the model on D_N.

문제점

the new corpus D_N does not have enough data compared to D_T.
→ directly fine-tuning on insufficient data can be sub-optimal
and prone to overfitting.

해결방법

Model : MAML for Fast & Robust Adaption

MAML

- in each training episode, conduct gradient descent using sufficient data in D_T to learn an updated weight θ^*.

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

- then, treat θ^* as an initialized weight to optimize θ on the limited data in D_N .

$$\begin{aligned} \theta' &= \theta - \beta \nabla_\theta \mathcal{L}_{D_N}(\theta^*) \\ &= \theta - \beta \nabla_\theta \mathcal{L}_{D_N}(\theta - \alpha \nabla_\theta \mathcal{L}_{D_T}(\theta)), \end{aligned} \quad (2)$$

Experiments : Intrinsic Evaluation

- 목표
  - evaluate OOV embeddings on the chimera benchmark

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
Additive, no stop words	0.3376	0.3624	0.4080
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

Table 1: Performance on the Chimera benchmark dataset with different numbers of context sentences, which is measured by Spearman correlation. Baseline results are from the corresponding papers.

Experiments : Extrinsic Evaluation

- 목표
  - evaluate OOV embeddings on downstream benchmark

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**

Table 2: Performance on Named Entity Recognition and Part-of-Speech Tagging tasks. All methods are evaluated on test data containing OOV words. Results demonstrate that the proposed approach, HiCE + Morph + MAML, improves the downstream model by learning better representations for OOV words.

Experiments : Qualitative Evaluation

- 목표
  - illustrate how does HiCE extract and aggregate information from multiple context sentences.

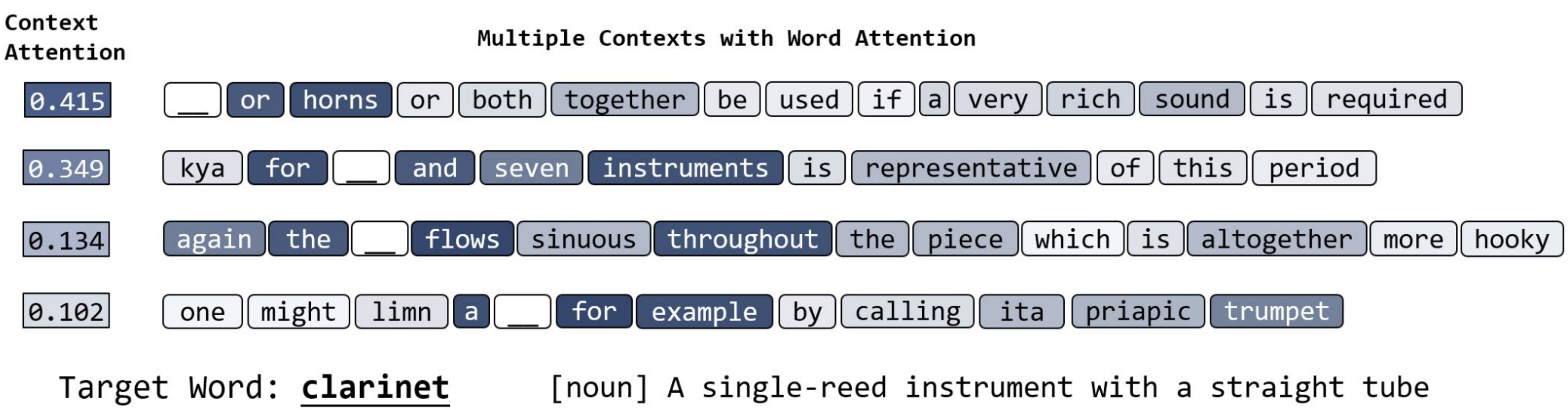

Figure 2: Visualization of attention distribution over words and contexts.

Experiments : Qualitative Evaluation

- 목표
    - illustrate how does HiCE extract and aggregate information from multiple context sentences.

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

Table 3: For each OOV in Chimera benchmark, infer its embedding using different methods, then show top-5 words with similar embedding to the inferred embedding. HiCE can find words with most similar semantics.

- FastText : find words with similar words.
- HiCE : can capture the true semantic of the OOV words.