大域的な類似度を考慮した未知語分散表現

五十川真生 † 梶原智之 ‡ 荒瀬由紀 †
大阪大学大学院情報科学研究科 † , 大阪大学データビリティフロンティア機構 ‡

研究背景

研究背景

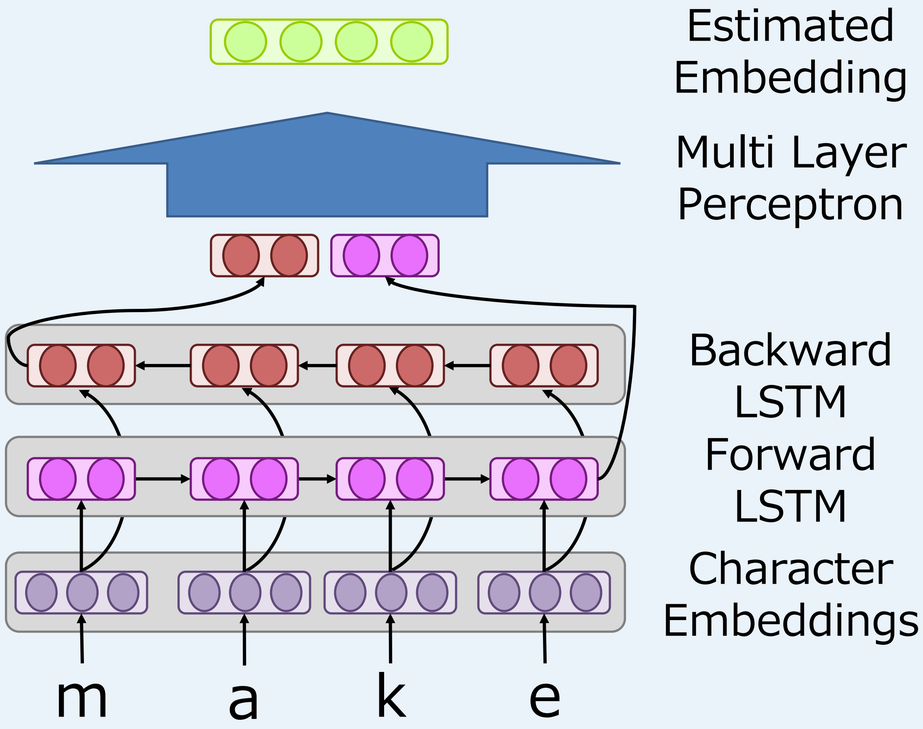
自然言語処理タスクにおいて
単語分散表現が有用

問題点

未知語の分散表現の生成が困難

既存研究（Mimic） *1

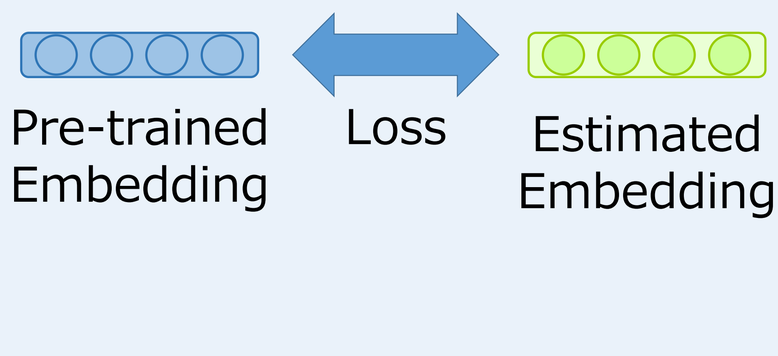
文字ベクトルから単語ベクトルを生成



*1 Pinter et al, Mimicking Word Embeddings using Subword RNNs, EMNLP, 2017

既存研究（Mimic）

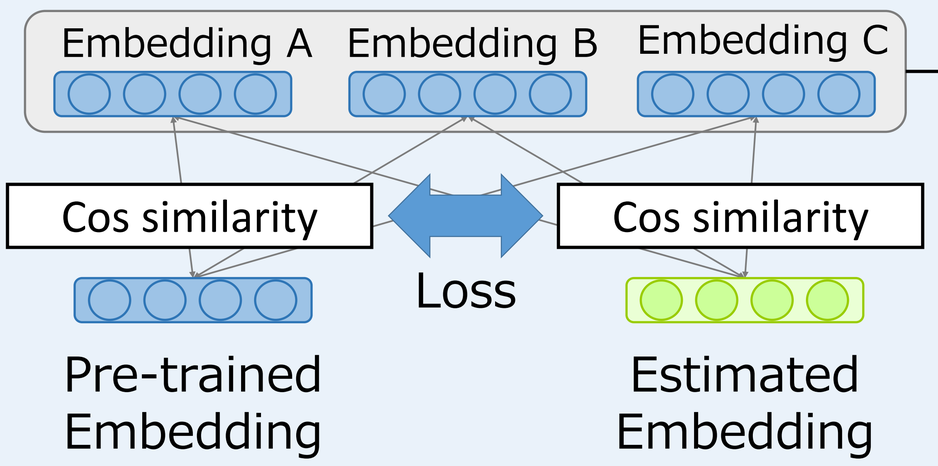
分散表現の要素間の
平均二乗誤差を最小化



局所的な学習手法

提案手法

類似度の平均二乗誤差を最小化



GloVeから選択

- 半分は類似度が高い順に選択
- 残り半分はランダム

大域的な学習手法

評価実験

実験設定

- GloVeによって生成した
  頻度上位10万語の分散表現を
  学習データ(内1000件は開発データ)に使用
- 類似度を比較する語は100件
- Mimic+提案手法は
  それぞれのLossを平均して学習
- **単語間類似度**及び**SentEval**で評価

単語間類似度

	単語のペア数	GloVe	Mimic	提案手法	Mimic+提案手法
Stanford Rare Word	2,034	0.440	0.314	0.296	**0.329**
Mixed	11,766	0.420	0.161	0.222	**0.262**

Mixedはhttps://github.com/mfaruqui/eval-word-vectors
のデータセットを合わせたもの

SentEval

GloVeの語彙を頻度上位5万語にし、語彙外の語の分散表現を各手法で生成

task	学習データ数	テストデータ数	未知語の割合	GloVe (全語彙)	GloVe (5万語)	Mimic	提案手法	Mimic+提案手法
MR	11k	11k	7.51%	0.806	0.765	**0.820**	**0.776**	**0.823**
SUBJ	10k	10k	7.10%	0.923	0.910	**0.930**	**0.915**	**0.929**
SST	67k	1.8k	7.65%	0.851	**0.824**	0.814	0.824	0.818
TREC	6k	0.5k	4.19%	0.872	**0.900**	0.802	0.894	0.808
SICK-E	4.5k	4.9k	1.36%	0.862	0.855	**0.860**	**0.858**	**0.859**
MRPC	4.1k	1.7k	5.56%	0.747	0.813	0.795	0.812	**0.826**
STS 2014	0	3.7k	6.67%	0.700	0.543	**0.672**	**0.649**	**0.669**
STS B	5.7k	1.4k	6.76%	0.756	0.727	**0.733**	0.722	**0.733**
SICK-R	4.5k	4.9k	1.36%	0.883	0.878	**0.881**	**0.879**	**0.881**

実験結果

単語間類似度

Mimicより提案手法が**相関高**

SentEval

Mimicも提案手法も
GloVeより**性能高**

今後の予定

文字以外の情報を利用して
分散表現を生成する