最先端NLP2020

“Kernelized Bayesian Softmax for Text Generation”

(Miao+, NeurIPS 2019)

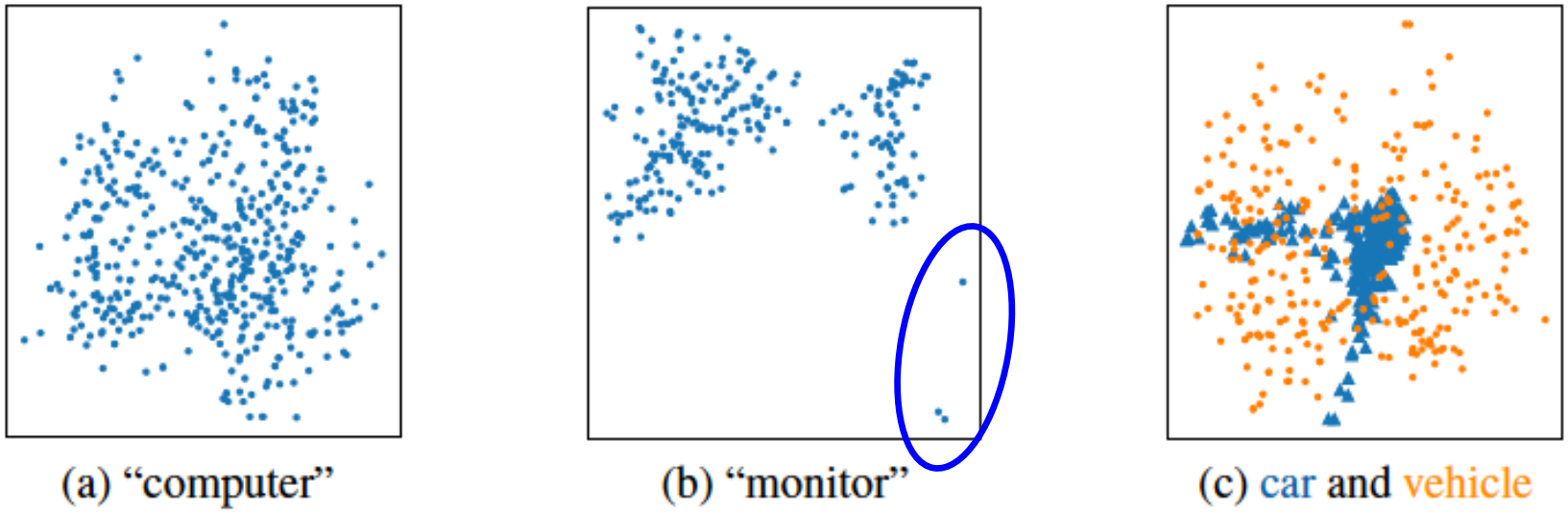
持橋大地
統計数理研究所 数理・推論研究系
daichi@ism.ac.jp

2020-9-25(金)

どんな論文?

- ニューラルネットで最後に単語を出力するSoftmaxレイヤを改良
- 単語ベクトルが一つに決まっていると、複数の意味がある場合に問題 → 混合モデルにしてSoftmaxを計算
  - カーネル関数による内積を用いることで、柔軟なモデル化
- カーネル関数のパラメータも自動学習
  → 意味によって分散が異なる (一般的な単語には大きな分散)
- Seq2seq, Transformerと接続することで高い生成性能

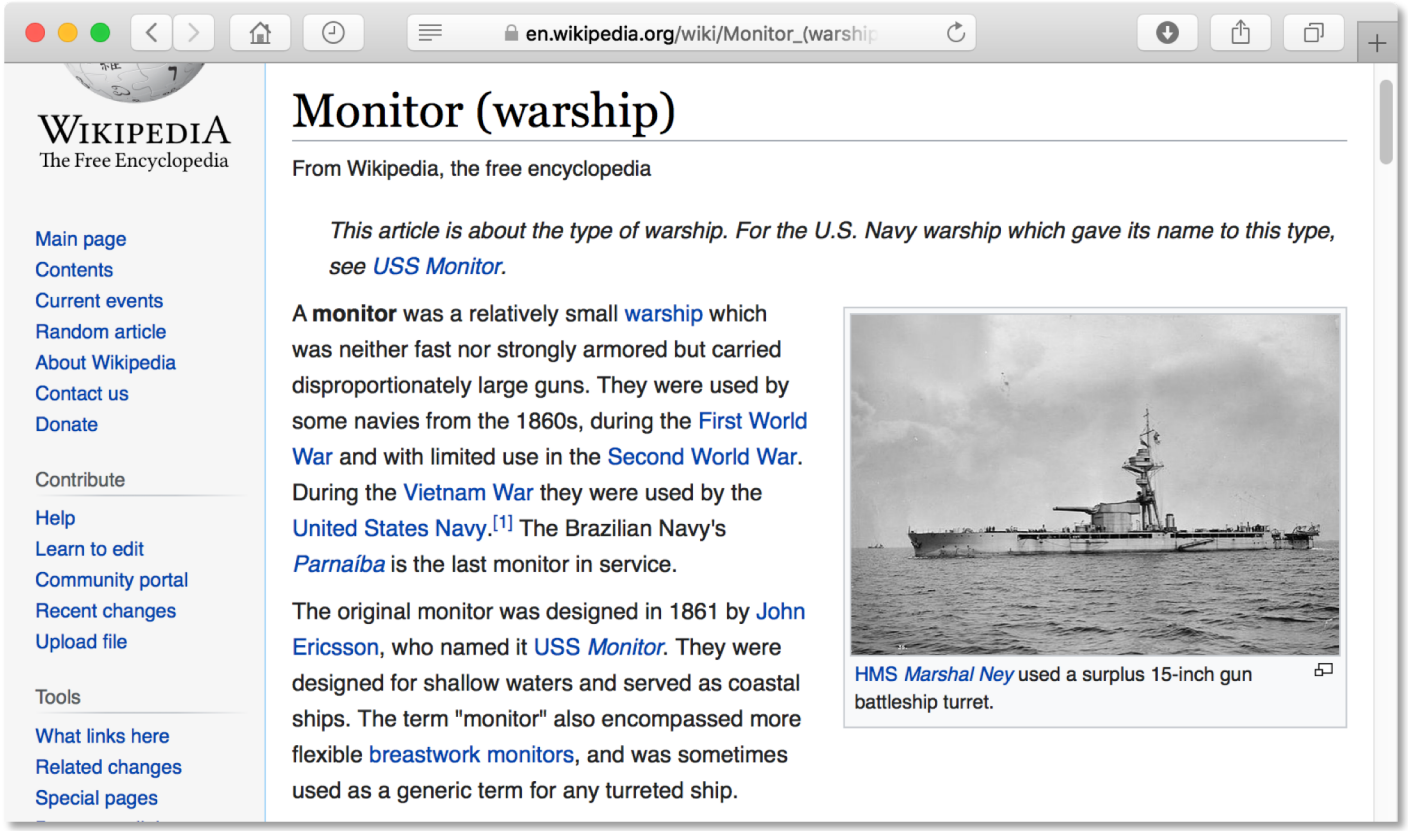
背景



- 単語の意味は必ずしも一意ではない
  - 上は、BERTによる単語のEmbeddingの例
- 3つの性質
  (1) 単語の意味は複数のクラスタに分かれる →図(b)
  (2) 各クラスタの分散はそれぞれ異なる →図(a),(c)
  (3) 外れ値が存在する →図(b)

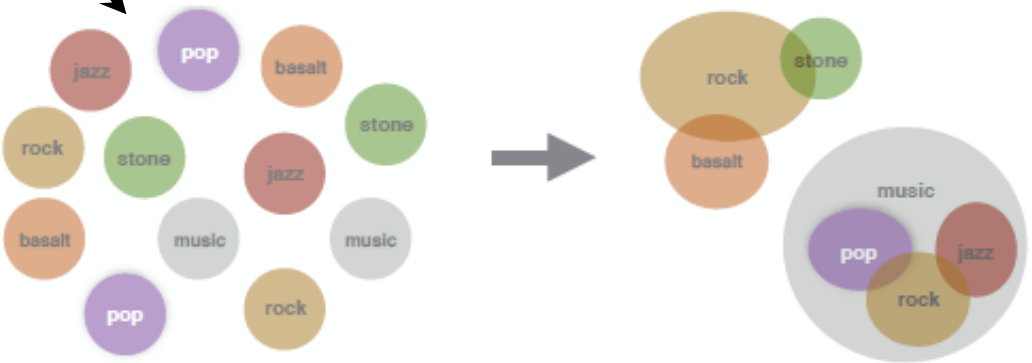
意味と外れ値

- 複数の意味を持つ単語の例: 有名なbank以外にも、“change”, “sentence”, “brother” など多数ある
- “monitor” の意味の外れ値...?



従来のアプローチ

- Word2vec, GloVe: 意味は単語に一つ
- BERT: 単語の意味はすべて動的に決まるが、大量データが必要、専用アーキテクチャ
- Vilnis and McCallum (2015): 単語の意味をガウス分布で表現、KLダイバージェンスで類似度を測る
- Athiwaratkun and Wilson (2017): 混合ガウス分布で単語の意味を表現
- Sun+ (2018): KLは数値的に不安定なため、Wasserstein距離を導入



KerBS: Kernelized Bayesian Softmax

- 単純に、意味を混合モデルとして考える
- ニューラルネットの最終層からactivation h_t が出力される時、

$$\begin{cases} p(y_t = w|h_t) & = \sum_{k=1}^{K} p(y_t = w, s_t = k|h_t) \\ p(y_t = w, s_t = k|h_t) & = \dfrac{\exp(k_\theta(h_t, \overrightarrow{w}_k))}{\sum_{w=1}^{W}\sum_{k=1}^{K}\exp(k_\theta(h_t, \overrightarrow{w}_k))} \end{cases}$$

  - 要するに、K個の意味との内積を計算して和を取っている

ガウス混合モデルとの違い?

- Vilnis and McCallum (2014), Athiwaratkun and Wilson (2017) のように埋め込み空間に多次元ガウス分布を導入するのは、一見自然なように見えるが..
- 実際の単語ベクトルは、埋め込み空間より低い次元の部分空間に存在する
- d次元の空間内のd_1次元部分空間に単語ベクトルがあった時、外れ値(d次元超立方体上に一様に分布)があると、ガウス分布で必要なノルム

$$\sum_{i=1}^{d} x_i^2$$

は、普通の点でd_1, 外れ値で$d/12$になる
→dが大きいと、外れ値にdominateされる
- 予備実験でも、混合ガウス分布は上手く行かなかった

カーネル関数

- activation hとembedding e に対し、

$$k_\theta(h, e) = |h||e|(a\exp(-\theta\cos(h, e)) - a)$$

  ここでaは正規化定数で, $a = -\dfrac{\theta}{2(e^{-\theta} + \theta - 1)}$

- $\theta \to 0$ で通常の内積に一致
- 勾配を計算すると、

$$\frac{\partial \log \mathcal{K}_\theta(h, e)}{\partial \theta} = \frac{1}{a}\frac{\partial a}{\partial \theta} - \frac{cos(h, e)exp(-\theta cos(h, e))}{\exp(-\theta cos(h, e)) - 1}$$

  - $\cos(h, e) \to 0$ のとき勾配 $\to 1/\theta$ ... 外れ値があっても勾配は一定に収まる (i.e. 外れ値に対して安定)

注：

$$k_\theta(h,e) = |h||e|(a\exp(-\theta\cos(h,e)) - a)$$
$$= |h|||e|\,a\,(\exp(-\theta\cos(h,e) - 1)$$

- 正規化定数aは

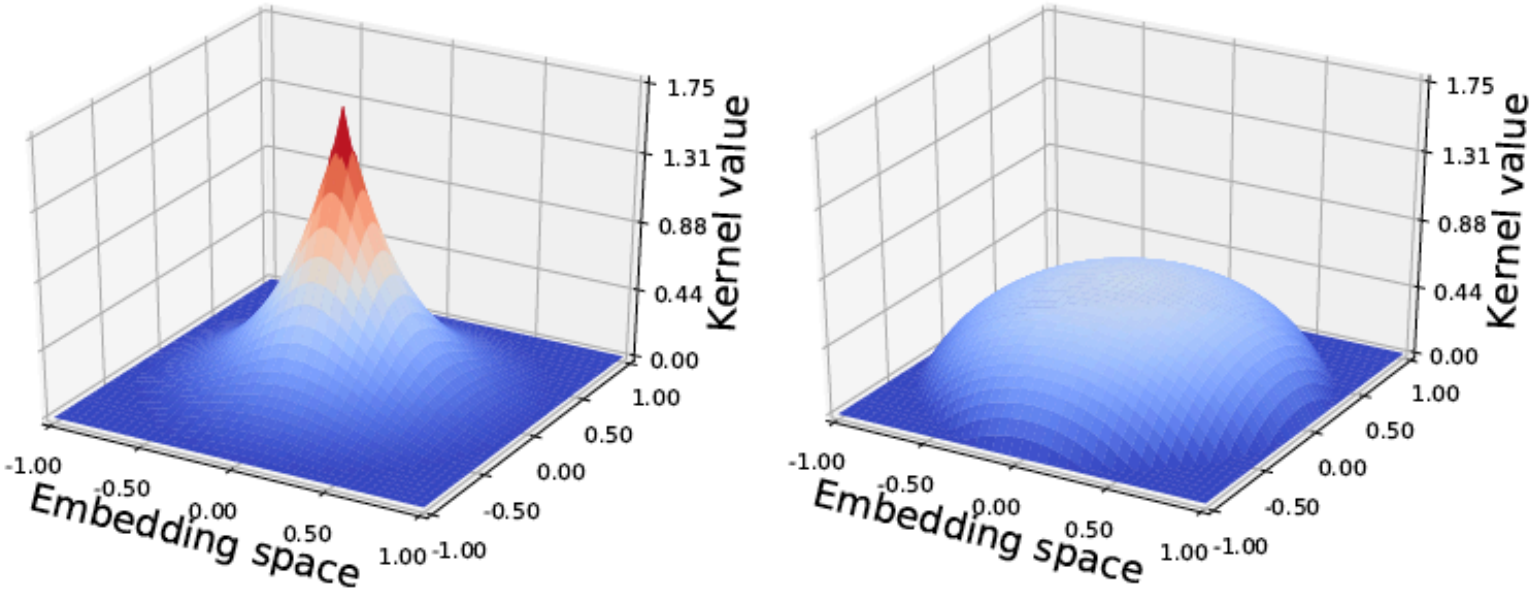
$$\int_{-1}^{1}(\exp(-\theta x) - 1)dx$$

から得られると思われるが、自分で計算するとこの値は

$$\frac{e^{\theta} - e^{-\theta}}{\theta} - 2$$

になった (バグ?)

カーネル関数のプロット



(a) $\theta = -2$ (b) $\theta = 2$

- θに応じて、ガウス分布では表せない裾の広さを持つ

学習アルゴリズム

- ランダムに合計M個の意味(sense)を単語に割り当てて、
  (1) 尤度を最大化するように、意味ベクトルと対応するカーネルパラメータθを最適化
  (2) 使われなかった意味を単語から削除して、他の単語に割り当てる
- 実験では、語彙数Vに対してM=3Vの意味を設定;
  単語あたりの最大の意味素数＝4
- 予測が当たらなかった単語 i について、
- 次の予測確率の平均P_iが閾値以下になった単語 i について、最も使われなかった意味を削除して、新しい意味に置き換える
  → 次スライド

意味素の更新

$$\log P_i \leftarrow (1-\beta)\log P_i + \log(P(y_t = i))\mathbb{1}_{i=\hat{y}_t}$$
$$U_i^j \leftarrow (1-\beta)U_i^j + \beta P(s_t = \langle i, j\rangle)\mathbb{1}_{i=\hat{y}_t}$$

- 移動平均で P_i を更新し、$P_i < \varepsilon$ となった単語iについて
  (1) U_i^j が最小の意味を削除
  (2) それを、事前分布の期待値(全体の平均) U^{new} で置き換え；対応するθ=1e-8 で初期化
- 注意：
  - この基準では、頻度が多い語(機能語)が優先して更新される
  - 予測確率が本来高い語 (isなど)・本来低い語 (abideなど)を同じ基準で扱ってよい?

全体の学習アルゴリズム

Algorithm 1: Training scheme for KerBS

Input : Training corpus $\hat{Y}$, total sense num M_{sum}, wo
ratio Q, threshold ϵ;
Output : W, θ, sense allocation list L;
Initialize $W, H, \theta, U, L, step = 0$;
while *not converge* **do**
  Random select $\hat{y} \in \hat{Y}$;
  for i_t *in* T **do**
    $h_t \leftarrow f_\phi(\hat{y}_{[0:t-1]})$;
    Calculate sense probability $P(y_t = \langle i, j \rangle)$ and
    MAXIMIZE $\log(P(y_t = \hat{y}_t))$ by ADAM;
    Update $logP$ and U by Eq. (12), (13);
  end
  if $step \ \ mod \ \ Q = 0$ **then**
    for i *in* $\{1, 2, ..., V\}$ **do**
      if $logP_i < \epsilon$ **then**
        $i'_0, j'_0 \leftarrow \arg\min_{i',j'}(U^{j'}_{i'})$;
        $\theta^{j'_0}_{i'_0} \leftarrow 1e-8$; $U^{j'_0}_{i'_0} \leftarrow$ MEAN(U);
        $L[\langle i'_0, j'_0 \rangle] \leftarrow i$;
      end
    end
  end
  $step = step + 1$;
end

- 意味とカーネルパラメータはAdamで更新
- Qは意味の更新レート(公開実装では200)

実験

- 機械翻訳(MT)：独→英 196k words
  言語モデル(LM)：One-Billion-Wordコーパスから300k語
  対話生成(Dialog)：DailyDialogデータセット, 13,118対話
- 生成の結果

Table 1: Performance of KerBS on Seq2Seq.

Tasks	Metrics	Seq2Seq	Seq2Seq+ MoS [Yang et al., 2018]	SeqSeq + KerBS
MT	BLEU-4	25.91	26.45	**27.28**
LM	PPL	103.12	102.72	**102.17**
Dialog	BLEU-1	16.56	13.73	**17.85**
	Human Eval.	1.24	1.04	**1.40**

Table 2: Performance of KerBS on Transformer.

Tasks	Metrics	Transformer	Transformer + MoS [Yang et al., 2018]	Transformer + KerBS
MT	BLEU-4	29.61	28.54	**30.90**
Dialog	BLEU-1	10.61	9.81	**10.90**

実験 (2)

- 生成された例とTransformer, Mixture of Softmaxとの比較

Table 3: Examples of MT on IWSLT'16 De→En

Source	meine gebildete Mutter aber wurde Lehrerin.
Transformer	my foster mother was a teacher.
+ MoS	and my educated mother was a teacher.
+ KerBS	but my educated mother became a teacher.
Source	man erreicht niemals eine Gemeinde mit Ideen, man setzt sich mit den Einheimischen zusammen.
Transformer	you never achieve a community with ideas; you put together with local people.
+ MoS	you never get a community with ideas, you're putting together with indigenous people.
+ KerBS	you never get to a community with ideas, and you sit with the local people.

- 注意：DeepLによると、

  meine gebildete Mutter aber wurde Lehrerin.
  =「しかし、教育を受けた母は教師になった。」
  man erreicht niemals eine Gemeinde mit Ideen, man setzt sich mit den Einheimischen zusammen.
  =「郷に入りては郷に従え」

実験 (3)

- 対話応答の生成 (文脈をより考慮する必要がある)

Table 4: Examples of dialog generation on DailyDialog

Source	what do you mean ?
Seq2Seq	i mean, what s up with the and iron bars on your windows.
+ MoS	well, how can i put this? france is a terrible team.
+ KerBS	well, i mean. we always do the same thing. there s no variety in our lives.
source	now , what seems to be the trouble ?
Seq2Seq	trouble is trouble.
+ MoS	yeah. and he was.
+ KerBS	not bad. but i have a bad cold today.
source	it can t be more than fourteen days late for us to accept it .
Seq2Seq	it will just fine.
+ MoS	well, i see. have you been back to work then?
+ KerBS	maybe you re right. i think we should take it.

実験 (4/5)

- 意味の数と単語 (ただし, 更新則が理由の可能性あり)

Table 6: Randomly selected words with different numbers of senses M after training.

Sense	1	2	3	4
word	Redwood	particular	open	they
	heal	figure	order	work
	structural	during	amazing	body
	theoretical	known	sound	power
	rotate	size	base	change

  - Specificな語は意味の数が少なく、多義語には多くの意味が自動的に割り当てられている
  - モデル全体での意味の数Mを先に指定する

パラメータθと意味の広さ

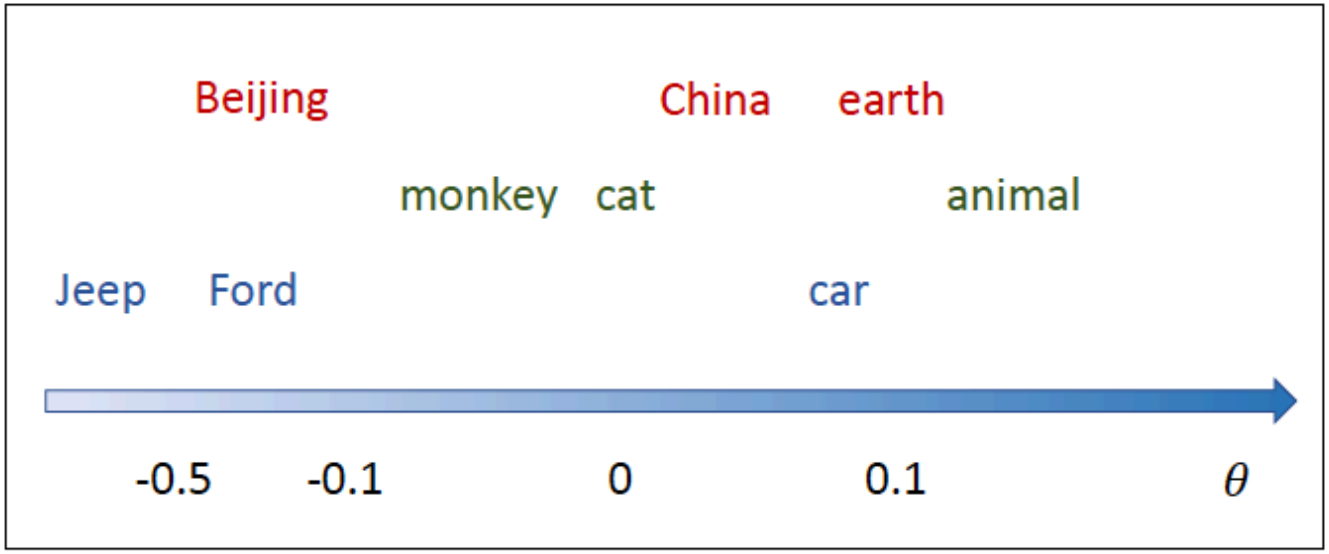


Figure 3: Words with different θ.

- θが大きいほど、カーネルで内積を取る際に遠くまで考慮
  →θは意味の「広さ」を反映
- specific→genericに並んだ上の3つの単語ペアについて、
  学習されたθがそれと正しく対応している

論文のまとめ

- 単語を出力するsoftmax関数を exp(kernel) の形で書き換え、複数の意味(sense)に対する混合モデルを考えることで、単語の多義を自動的に考慮する出力レイヤを提案
  - senseごとにカーネルのパラメータθを学習することで、意味の広さも同時に考慮
  - 計算量は、ナイーブなsoftmaxの2倍程度
- 単語ベクトルに確率分布を考えるのではなく、内積を計算する際のカーネル関数を適応的にしたところが新しい
  - ほぼ双対に近いが、より柔軟に定義できる
  - exp(cos)の形は理論的にも妥当