Contextual Distributed Models for Sequences

aka Latent parameterizations, Family Learning, ...

Benjamin Piwowarski

LIP6, Université Pierre et Marie Curie / CNRS

9th of November, 2015

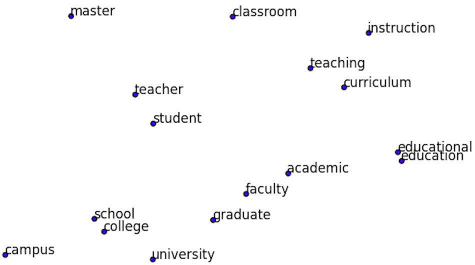
Introduction

Importance of context

- Non generative process
  - *User*, *Location* in Personalized Web Search
- Generative process
  - *Speaker* in spoken language understanding
  - *Topic* and *style* in document generation
  - *User* in web browsing
  - *User* and *Activity* in music listening
  - ...

Distributed Representations

- Allow to capture geometrically dependencies between entities (users, objects, etc.)
- Used successfully in image, text, etc.
- Able to take into account various sources of information



This talk

- Taking into account context
    - Learn $p(x|\mathcal{D})$ on a dataset $\mathcal{D}$
    - Learn $p(x|\mathcal{D}_c)$ for a given dataset $\mathcal{D}_c << \mathcal{D}$, taking into account $p(x|\mathcal{D})$
- ...using distributed representations
- Goal: learn a representation that can be reused
    - Predicting search results for a user in a given context
    - Predicting the satisfaction of a user
    - Predicting the next recommendation to a user

Taking into account Context

Overview

Different ways to take into account parameters

- Explicit factors
- Families of parameters
- Prior on parameters

Probabilistic models

Latent variables describing the context

Latent Dirichlet Allocation (text)

Word generation is governed by the themes the document deals with:

$$\begin{aligned} \theta &\sim \mathit{Dirichlet}\left(\alpha\right) \\ z &\sim \mathit{Multinomial}(\theta) \\ w &\sim \mathit{Multinomial}(z) \end{aligned}$$

Problem of estimating (the maximum likelihood) of

$$p\left(\theta | d_1 \ldots d_n\right)$$

"Latent dirichlet allocation," *Ng, Blei, and Jordan (2003)*

Family Discovery

- Mixture of experts
  1. Learn a mixture of experts over the dataset

$$p(x) = \sum_i p(x|\mathcal{M}_i)\, p(\mathcal{M}_i)$$

  2. Estimate $p(\mathcal{M}_i|\mathcal{D}_c)$
- Learn families of models as a subset S of the parameter space
  1. Single Model
  2. Separate Models
  3. Affine subspace (determined by PCA)
  4. Affine patch
  5. Coupled Map (auto-encoder)

and project a new solution onto the family space

"Family Discovery.," *Omohundro (1995)*

Parameterizing probabilistic models

- Global idea:
  - Parametric model
  - Modify some parameters using context-specific parameters

Parameterizing probabilistic models

Gesture recognition with HMMs

The gesture is parameterized with θ

$$\begin{aligned} \hat{\mu}_j &= W_j\theta + \mu \\ p(\mathbf{x}_t | q_t = j, \theta) &= \mathcal{N}\left(\mathbf{x}_t; \hat{\mu}_j(\theta), \Sigma_j\right) \end{aligned}$$

where $\mathbf{x}_t$ is a 6-dimensional vector (position of each hand); θ is re-estimated for each example

"Parametric hidden Markov models for gesture recognition," *Wilson and Bobick (1999)*

Extension of Wilson and Bobick to parameterization of variance $\Sigma_j(\theta)$ in handwriting recognition

"Handling signal variability with contextual markovian models," *Radenen and Artieres (2014)*

Regularization

- Learning a model on a dataset $\mathcal{D}$

$$\theta^* = \operatorname*{argmin}_{\theta} R(\theta; \mathcal{D})$$

- Contextualization : the parameters are used as a "prior" for the new optimization

$$\theta_c = \operatorname{argmin} R(\theta; \mathcal{D}) + \Delta(\theta, \theta^*)$$

Regularization: example

Truncated gradient

Adapting Deep RankNet for Personalized Search

$$w \leftarrow w - \eta T_1 \left(\frac{\partial C}{\partial w}, a, \tau \right)$$

where C is the cost, a the the output of the neuron, and τ a threshold.

$$T_1 (v, a, \theta) = \begin{cases} \max (0, v - a) & v \in [0, \tau] \\ \min (0, v + a) & v \in [-\tau, 0] \\ v & \text{otherwise} \end{cases}$$

Adapting deep RankNet for personalized search, Song, Wang, and He (2014)

Representation Learning and Sequences

Representation Learning

- Projecting points from the original point to a manifold
- Hypotheses

  "The Manifold Tangent Classifier.," *Rifai, Dauphin, Vincent, Bengio, and Muller (2011)*

  - *The semi-supervised learning hypothesis*
    Learning $p(x)$ can improve our classification $p(y|x)$
  - *The (unsupervised) manifold hypothesis*
    Real world data presented concentrate in the vicinity of non-linear sub-manifolds
  - *The manifold hypothesis for classification*
    Points of different classes concentrate along different sub-manifolds

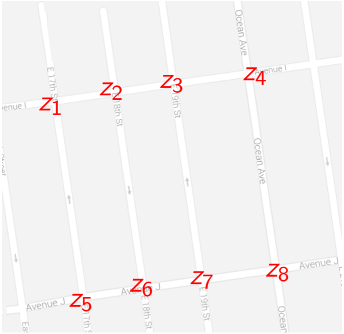
Representation Learning

- Represent complex objects (as vectors in $\mathbb{R}^n$)
    - Nodes in a graph
    - Sequences (state as a vector)
    - Context

  Examples:
    - Language Models
    - Spoken Language Understanding
    - Handwriting
    - ...

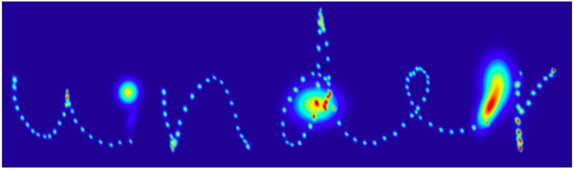
Distributed States (traffic prediction)

$$\mathcal{L}(\theta) = \frac{1}{O} \sum_{it} m_i \Delta \left(f_\theta \left(z_i^{(t)} \right), x_i^{(t)} \right) + \lambda_1 \sum_{it} \left\| z_i^{(t+1)} - h_\gamma \left(z_i^{(t)} \right) \right\| + \lambda_2 \sum_{ijt} e_{ij} \left\| z_i^{(t)} - z_j^{(t)} \right\|$$



"Car-Traffic Forecasting: A Representation Learning Approach.," *Ziat, Contardo, Baskiotis, and Denoyer (2015)*

Handwriting sequences



$$p\left(x_{t+1}|s_t\right) = \sum_{j=1}^{M} \pi_t^j \mathcal{N}\left(x_{t+1}|\mu_t^j, \sigma_t^j, \rho_t^j\right)$$

$$\hat{y}_t = \left(\hat{e}_t, \left\{\hat{\pi}_t^j, \hat{\mu}_t^j, \hat{\sigma}_t^j, \hat{\rho}_t^j\right\}\right) = b_y + \sum Ws$$

$$y_t = f\left(\hat{y}_t\right)$$

"Generating Sequences With Recurrent Neural Networks," *Graves (2013)*

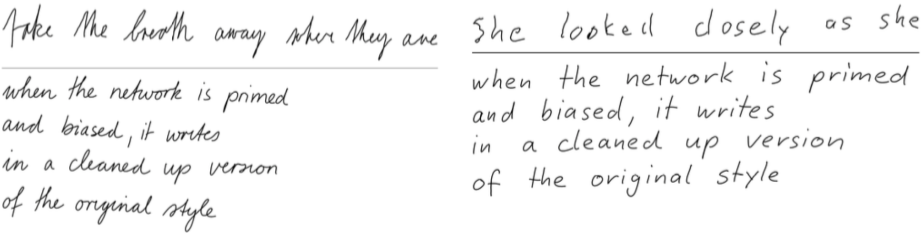
Neural Network Language Models

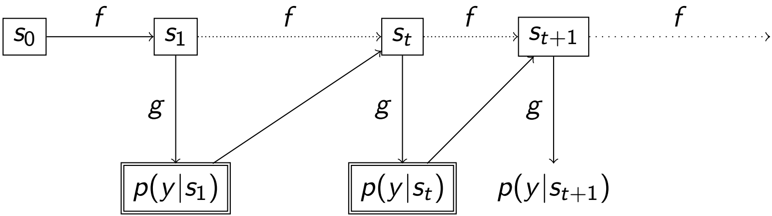
1. Lookup table : a word $w \leftrightarrow a_w \in \mathbb{R}^p$
2. Computing the state
   1. Convolution $s_k = f\left(w_{k-d} \dots w_{k-1}\right) \in \mathbb{R}^k$
   2. Recursion $s_k = f\left(s_{k-1}, w_{k-1}\right)$
3. Probability distribution over words $p\left(w|s_k\right) \propto \exp\left(\langle b_w, s_k \rangle\right)$

"Natural language processing (almost) from scratch," *Collobert et al. (2011)*

Taking context into account in distributed models

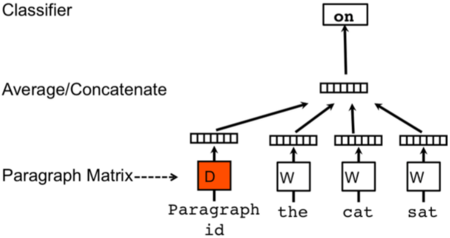
Priming





"Generating Sequences With Recurrent Neural Networks," *Graves (2013)*

Parameterized LM



Estimates

$$p\left(w_k \middle| w_{k-l}, \ldots, w_{k-1}, \theta_d\right)$$

"Distributed Representations of Sentences and Documents," *Le and Mikolov (2014)*

Problem

- Given a family of functions $\mathcal{F}$ and a context space $\mathcal{C}$, we want to find a function Φ such that

$$\begin{aligned} \Phi : \mathcal{F} \times \mathcal{C} &\rightarrow \mathcal{F} \\ f \times \theta_c &\rightarrow f_c \end{aligned}$$

- Neutral element θ_e such that $\Phi(f, \theta_e) = f$ (we want to learn a general model)
- Problems:
  - What class of functions
  - Balance between complexity and generalization

Information Retrieval

- Inputs
  - Sets of documents
    ... "I don't have a vendetta against Bambi. I really don't give a darn. It was just my personal opinion," Louise Bates Ames, associate director of the Gesell Institute of Human Development, said Wednesday.
  - A query
    Document will report judicial proceedings and opinions on contracts for surrogate motherhood.
- Goal: rank relevant documents before non relevant ones
- Measure: Mean Average Precision over queries

$$AP = \sum_{k=1}^{N} \delta\left(d_k \text{ is relevant}\right) \underbrace{\frac{\sum_{i \leq k} \delta\left(d_i \text{ is relevant}\right)}{k}}_{\text{precision at rank } k}$$

Language Models

Language Models Hypothesis (Information Retrieval)

The query should be generated by the document language model

$$p\left(q_1 \ldots q_n | \mathcal{M}_{d_1}\right) > p\left(q_1 \ldots q_n | \mathcal{M}_{d_2}\right)$$
$$\Leftrightarrow d_1 \text{ is more relevant than } d_2$$

- Unigram model

$$p\left(w_1 \ldots w_n | \mathcal{M}_d\right) \overset{H}{=} \prod_i p\left(w_i | \mathcal{M}_d\right) \overset{ML}{=} \prod_i \frac{\mathrm{tf}\left(w_i, d\right)}{\mathrm{tf}\left(d\right)}$$

Limits

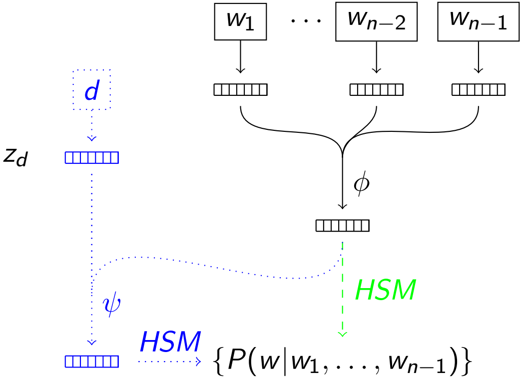
- Dependencies are not taken into account
  Extending to higher order Markov chain is not straightforward

$$p\left(w_1 \ldots w_n | \mathcal{M}_i\right) = \prod_i p\left(w_i | w_{i-1}, \ldots, w_{i-k}, \mathcal{M}\right)$$

- Problem of vocabulary mismatch

Both problems can be tackled using distributed representations

Parameterized LM



Possible transformations:

- (Le and Mikolov) Component-Wise Sum
- Component-Wise Product

Procedure

- Initialization with Mikolov word2vec embeddings
- Train the general/collection language model

$$\theta^* = \operatorname*{argmax}_\theta \sum \log p_{nn}\left(w_1 \ldots w_n | \theta, z_d = e\right)$$

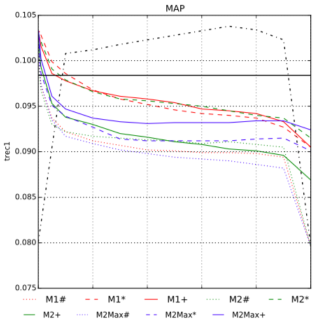
- For each document d (pre-selection with a standard IR model)
  - Compute its representation

$$z_d^* = \operatorname*{argmax}_\theta \sum \log p_{nn}\left(d_1 \ldots d_n | \theta, z_d = e\right)$$

  - Compute the score for the query

$$RSV(q, d) = \lambda p_{nn}\left(q_1 \ldots q_n | \theta, z_d\right) + (1 - \lambda)\, p_{unigram}\left(q_1 \ldots q_n | d\right)$$

Results



TREC1 (similar for others) / sum or product

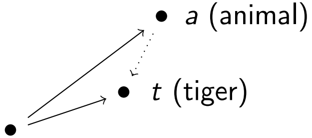
- Qualitative analysis: *underestimated* probability of document-specific terms
- Looking at alternative transformations:

Translations

- Translation Transformations

$$Id + \sum_{i=1}^{K} a_i b_i^\top$$

Example $Id + \frac{1}{\|a\|}(t - a)\, a^\top$



- Preliminary experiments have shown that translation > product/sum

Conclusion

Conclusion

- Distributed representations are useful to capture context
- Context = manipulation of the state in sequences
- Future work
  - Understand what the different transformations do
  - Recurrent models (the modified state is taken into account)
  - Other applications: user modeling (Web)

Bibliography I

- R. Collobert et al. “Natural language processing (almost) from scratch.” In: 12 (2011), pp. 2493–2537.
- A. Graves. “Generating Sequences With Recurrent Neural Networks.” In: (Aug. 2013). arXiv: `1308.0850v2 [cs.NE]`.
- Q. V. Le and T. Mikolov. “Distributed Representations of Sentences and Documents.” In: *Proceedings of the 31st International Conference on Machine Learning*. May 2014.
- A. Y. Ng, D. M. Blei, and M. I. Jordan. “Latent dirichlet allocation.” In: *Journal of Machine of Machine Learning* 3 (2003), pp. 993–1022.
- S. M. Omohundro. “Family Discovery.” In: *NIPS'14* (1995), pp. 402–408.

Bibliography II

- M. Radenen and T. Artieres. "Handling signal variability with contextual markovian models." English. In: *Pattern Recognition Letters* 35 (Jan. 2014), pp. 236–245. DOI: 10.1016/j.patrec.2013.08.015.
- S. Rifai et al. "The Manifold Tangent Classifier." In: *NIPS'14* (2011), pp. 2294–2302.
- Y. Song, H. Wang, and X. He. *Adapting deep RankNet for personalized search*. New York, New York, USA: ACM, Feb. 2014. DOI: 10.1145/2556195.2556234.
- A. D. Wilson and A. F. Bobick. "Parametric hidden Markov models for gesture recognition." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21.9 (1999), pp. 884–900. DOI: 10.1109/34.790429.

Bibliography III

A. Ziat et al. "Car-Traffic Forecasting: A Representation Learning Approach." In: *MUD@ICML* (2015), pp. 85–87.