Diffusion Kernels on Graphs and Applications to Unigram Models

Bruno Jedynak

Johns Hopkins University

February 16, 2007

Outline

- Unigram Models
- How to use diffusion principle to build a Unigram Model
- Heat equation when the space variable belongs $\boldsymbol{R}^2$
- Heat equation over a graph
- Application: Unigram models
- Another construction: Normalized Diffusion
- More Unigram models
- Experiments (joint work with Damianos Karakos)

Unigram Models

Closed vocabulary V, $\#V = K \approx 10^5$
Training set of words $x_1, \ldots, x_n$. $n(x)$ is the number of times word x has been seen in the training set.
Want to build a probability mass function π over the words of V
Such a distribution is called Unigram model.

Unigram Models

Empirical distribution.

$$\pi_0(x) = \frac{n(x)}{n} = \frac{1}{n}\sum_{i=1}^{n}\delta(x = x_i)$$

Add-β

$$\pi_{add-\beta}(x) = \frac{n(x)+\beta}{n+\beta K} = (1-\lambda)\frac{n(x)}{n} + \lambda\frac{1}{K}$$

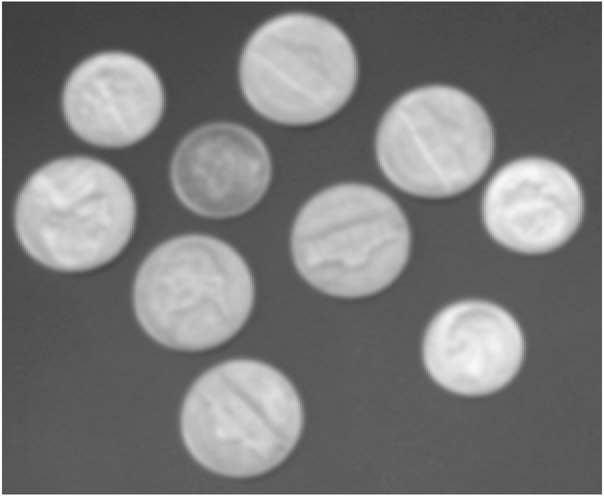
$\lambda = (\beta K)^{-1}(n + \beta K)$

Good-Turing

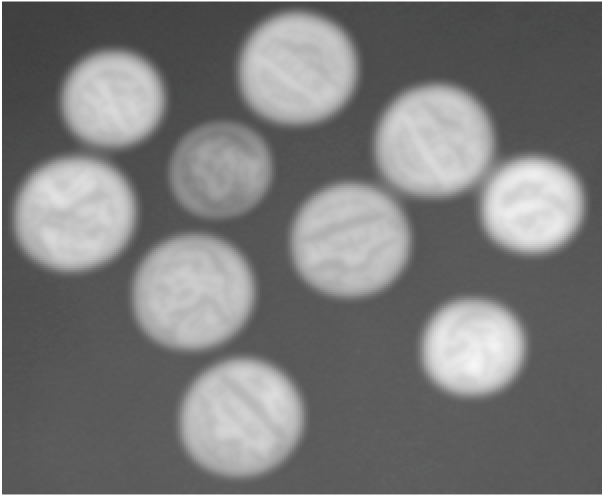
$$\begin{aligned} p_{GT}(x) &= \frac{n(x)+1}{n}\frac{r_{n(x)+1}}{r_{n(x)}} \text{ if } n(x) < M \\ &= \alpha\frac{n(x)}{n} \text{ otherwise} \end{aligned}$$

r_j is the number of words observed j times, $M = 5-10$, α is a normalizing constant.

Example of diffusion



Example of diffusion



How to use diffusion to build unigram models ?

Idea: Build a graph.
Vertices = V,
Define the edges ... start at $\pi_0(x) = n^{-1}nx)$ and diffuse and stop ...

Heat equation in $\boldsymbol{R}^2$

$x = (x_1, x_2) \in \boldsymbol{R}^2, y = (y_1, y_2) \in \boldsymbol{R}^2, t \geq 0, \alpha > 0$
$K_t(x, y)$ is the temperature at time t at x when starting at time $t = 0$ with all the heat concentrated at y. It is called a diffusion kernel.

$$\text{for all } x, \ \text{for all } t \geq 0, \frac{\partial}{\partial t} K_t(x, y) = \alpha \triangle K_t(x, y)$$

$\triangle$ stands for Laplacian.

$$\triangle K_t(x, y) = \frac{\partial^2}{\partial^2 x_1} K_t((x_1, x_2), y) + \frac{\partial^2}{\partial^2 x_2} K_t((x_1, x_2), y)$$

Heat equation in R^2

Without restricting the domain, the solution is given by

$$K_t(x,y) = \frac{1}{4\pi\alpha t} \exp\left(-\frac{1}{4\alpha t}\left((x_1-y_1)^2 + (x_2-y_2)^2\right)\right)$$

$K_t(x,y)$ is the density of a

$$N(y, 2\alpha t\ Id)$$

If now the temperature at time 0 is given by $g(x)$ then the solution of the heat equation is the convolution

$$\int\int K_t(x,y)g(y)dy$$

Discretization of the Laplacian

$x = (x_1, x_2) \in \boldsymbol{R}^2$, $f : \boldsymbol{R}^2 \mapsto \boldsymbol{R}$

$$\begin{aligned}
\frac{\partial}{\partial x_1} f(x_1, x_2) &\approx \frac{1}{h}\left(f(x_1 + \frac{h}{2}, x_2) - f(x_1 - \frac{h}{2}, x_2)\right) \\
\frac{\partial^2}{\partial^2 x_1} f(x_1, x_2) &\approx \frac{1}{h}\left(\frac{\partial}{\partial x_1} f(x_1 + \frac{h}{2}, x_2) - \frac{\partial}{\partial x_1} f(x_1 - \frac{h}{2}, x_2)\right) \\
&\approx \frac{1}{h}\left(\frac{1}{h}\left(f(x_1 + h, x_2) - f(x_1, x_2)\right) - \right. \\
&\qquad \left. \frac{1}{h}\left(f(x_1, x_2) - f(x_1 - h, x_2)\right)\right) \\
&\approx \frac{1}{h^2}\left(f(x_1 + h, x_2) + f(x_1 - h, x_2) - 2f((x_1, x_2)\right) \\
\frac{\partial^2}{\partial^2 x_2} f(x_1, x_2) &\approx \frac{1}{h^2}\left(f(x_1, x_2 + h) + f(x_1, x_2 - h) - 2f((x_1, x_2)\right)
\end{aligned}$$

Discretization of the Laplacian

$$
\begin{aligned}
\triangle f(x_1, x_2) &= \frac{\partial^2}{\partial^2 x_1} f(x_1, x_2) + \frac{\partial^2}{\partial^2 x_2} f(x_1, x_2) \\
&= \frac{1}{h^2} \left(f(x_1 + h, x_2) + f(x_1 - h, x_2) - 2f((x_1, x_2)) + \right. \\
&\quad \frac{1}{h^2} \left(f(x_1, x_2 + h) + f(x_1, x_2 - h) - 2f((x_1, x_2)) \right.
\end{aligned}
$$

Define $\mathcal{V}(x) = \{(x_1 + h, x_2), (x_1 - h, x_2), (x_1, x_2 - h), (x_1, x_2 + h)\}$ and $d(x) = \#\mathcal{V}(x)$ then

$$
\triangle f(x) = \frac{1}{h^2} \left(\left(\sum_{y \in \mathcal{V}(x)} f(y) \right) - d(x) f(x) \right)
$$

Heat equation over a graph

$G(V,E)$ a non oriented graph.
$V = \{x_1, \ldots, x_n\}$ is the finite set of vertices.
$E \subset V \times V$ is the set of edges. If $(x,y) \in E$, we denote $x \sim y$. Assume no edge from a vertex to itself.
The degree of $x \in V$ is $d(x) = \sum_{y \in V} \delta(x \sim y)$
$f : V \mapsto \boldsymbol{R}$ can be seen as a function or as a vector $(f(x_1), \ldots, f(x_n))^T$
$H : V \times V \mapsto \boldsymbol{R}$ can be seen as a function or as a $n \times n$ matrix.

Define the Laplacian (choose $h = 1$)

$$\begin{aligned}
\triangle f(x) &= \left(\sum_{y \in \mathcal{V}(x)} f(y) \right) - d(x) f(x) \\
&= \left(\sum_{y : y \sim x} f(y) \right) - d(x) f(x) \\
&= \sum_{y \in V} \left(f(y)\delta(x \sim y) - d(y) f(y) \delta(x = y) \right) \\
&= \sum_{y \in V} \left(\delta(x \sim y) - d(y)\delta(x = y) \right) f(y) \\
&= \sum_{y \in V} H(x, y) f(y) \\
&= Hf(x)
\end{aligned}$$

Laplacian

$$H(x, y) = \delta(x \sim y) - d(y)\delta(x = y)$$

$$H = A - D$$

$A(x, y) = \delta(x \sim y)$ is the adjacency matrix of G
$D(x, y) = d(x)\delta(x = y)$ is the degree matrix. D is diagonal.

Heat Equation

$x, y \in V,\ t \geq 0.$
$K_t(x, y)$ is the temperature at x at time t when starting with a unit temperature at y at time 0.
$K_0(x, y) = \delta(x = y)$ which in matrix notation is $K_0 = Id$
We define the heat equation for a fixed $y \in V$ as:

$$\text{for each } x \in V,\ \text{for each } t \geq 0, \frac{\partial}{\partial t} K_t(x, y) = HK_t(x, y)$$

Notate $u_t(x) = K_t(x, y)$

$$\begin{aligned} \frac{\partial}{\partial t} u_t(x) &= \sum_{z \in V} H(x, z) u_t(z) \\ &= \left(\sum_{z : z \sim x} u_t(z) \right) - d(x) u_t(x) \\ &= d(x) \left(\left(\frac{1}{d(x)} \sum_{z : z \sim x} u_t(z) \right) - u_t(x) \right) \end{aligned}$$

Heat Equation

Claims:

The heat equation admits a unique solution $K_t = e^{tH}$

$$e^{tH} = Id + tH + \frac{t^2}{2!}H^2 + \frac{t^3}{3!}H^3 + \dots$$

$$e^{tH} = \lim_{k \to +\infty} (Id + \frac{t}{k}H)^k$$

Starting with a temperature $\pi(x)$, $x \in V$, the solution to the heat equation is $K_t\pi$

If for all x, $\pi(x) \geq 0$ and $\sum_{x \in V} \pi(x) = 1$ then for all $x \in V$ and all $t \geq 0$, $K_t\pi(x) \geq 0$ and $\sum_{x \in V} K_t\pi(x) = 1$

If G is connected $K_t\pi > 0$.

Markov Chain Interpretation

Recall $K_t = \lim_{k \to +\infty} (Id + \frac{t}{k} H)^k$
Fix $t > 0$, choose a large enough k,
Define a Markov Chain over V with $X_0 \sim \pi_0$

$$\begin{aligned} P(X_{n+1} = y | X_n = x) &= (Id + \frac{t}{k} H)(x, y) \\ &= \delta(x = y) + \frac{t}{k} (\delta(x \sim y) - d(x)\delta(x = y)) \\ &= \delta(x = y)(1 - \frac{t}{k} d(x)) + \frac{t}{k} \delta(x \sim y) \end{aligned}$$

Then $P(X_k = y) \approx (K_t \pi_0)(y)$

Generalized Laplacian

Define a weight function $f : E \mapsto \mathbb{R}$ stricty positive, (symmetric)
The *generalized* Laplacian is then

$$H(x,y) = f(x,y)\delta(x \sim y) - d(x)\delta(x = y)$$

with $d(x) = \sum_{y:y\sim x} f(x,y)$ then, as previously,
The heat equation admits a unique solution $K_t = e^{tH}$
Starting with a temperature $\pi(x)$, $x \in V$, the solution to the heat equation is $K_t\pi$
If for all x, $\pi(x) \geq 0$ and $\sum_{x\in V} \pi(x) = 1$ then for all $x \in V$ and all $t \geq 0$, $K_t\pi(x) \geq 0$ and $\sum_{x\in V} K_t\pi(x) = 1$
If G is connected then $K_t\pi > 0$.

Examples

- Complete graph with K vertices. $x \sim y \iff x \neq y$

$$K_t(x,y) = \frac{1}{K}(1 - e^{-Kt}) + e^{-Kt}\delta(x = y)$$

- Vertices are binary strings of length K.
  $x \sim y \iff Hamming(x,y) = 1$

$$K_t(x,y) = \frac{1}{2^K}(1 + e^{-2t})^K (\tanh(t))^{H(x,y)}$$

- Diffusion kernels are known for closed chain and certain regular trees
- Small graphs. Diagonalize H

Unigram Models from Diffusion

Choose Set of vertices $= V$. Choose the edges ...

$$\begin{aligned}\pi_t(x) &= \sum_y K_t(x,y)\pi_0(y) \\ &= \sum_y K_t(x,y)\frac{1}{n}\sum_{i=1}^n \delta(y = x_i) \\ &= \frac{1}{n}\sum_{i=1}^n K_t(x,x_i)\end{aligned}$$

Unigram Models from Diffusion. Complete Graph

Choose the complete graph over V. $x \sim y \iff x \neq y$. Start at π_0 Then

$$\begin{aligned}
\pi_t(x) &= \frac{1}{n} \sum_{i=1}^{n} K_t(x, x_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{K}(1 - e^{-Kt}) + e^{-Kt} \delta(x = x_i) \right) \\
&= \frac{1}{K}(1 - e^{-Kt}) + e^{-Kt} \frac{n(x)}{n} \\
&= \frac{n(x) + \beta}{n + \beta K}
\end{aligned}$$

Add-β estimator with

$$\beta = \frac{n}{K}(e^{Kt} - 1)$$

Unigram Models from diffusion. Data dependent graph

Define the edges as follows: $x \sim y \iff |n(x) - n(y)| \leq 1$
Computation of the kernel is difficult ...
Recall

$$K_t = \lim_{k \to +\infty} (Id + \frac{t}{k} H)^k$$

Compute $(Id + \frac{t}{3} H)^3 \pi_0$ with $t = \frac{1}{K}$ yields fast and interesting results, see later.

Normalized Diffusion

$G = (V, w)$ a weighted graph. $w : V \times V \to \boldsymbol{R}$
$w(x, y) = w(y, x)$, $w(x, y) \geq 0$ and $w(x, x) > 0$
$w(x, y)$ is interpreted as the *similarity* between x and y.
Define $d(x) = \sum_{y \in V} w(x, y)$
Define a Markov chain $X_0, X_1, \ldots$ over V with initial distribution π_0
Define a transition matrix $P(X_1 = y | X_0 = x) = T(x, y) = d^{-1}(x) w(x, y)$
Remark that T is not symmetric.

Normalized Diffusion

Recall $P(X_1 = y|X_0 = x) = T(x,y) = d^{-1}(x)w(x,y)$
$\pi_1(y) = P(X_1 = y) = \sum_{x\in V} T(x,y)\pi_0(x)$,
$\pi_k(y) = P(X_k = y) = \sum_{x\in V} T^k(x,y)\pi_0(x)$,
If G is connected, (there is a path with > 0 weights between any two vertices)

$$\lim_{k\to+\infty} \pi_k(y) = \pi(y) = \frac{d(y)}{\sum_{x\in V} d(x)}$$

Examples

Observe $x_1, \ldots, x_n$, $x_i \in V$

$$\pi_0(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x = x_i)$$

$$\begin{aligned}
\pi_1(y) &= \sum_{x \in V} T(x, y) \frac{1}{n} \sum_{i=1}^{n} \delta(x = x_i) \\
&= = \frac{1}{n} \sum_{i=1}^{n} T(x_i, y) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{w(x_i, y)}{d(x_i)}
\end{aligned}$$

Example 1

$|V| = K$, Choose $w(x, y) = \alpha\delta(x = y) + 1$, $\alpha \neq 0$
Then $d(x) = \alpha + K$

$$\begin{aligned}
\pi_1(y) &= \frac{1}{n}\sum_{i=1}^{n}\frac{w(x_i, y)}{d(x_i)} \\
&= \frac{1}{n}\frac{1}{\alpha + K}\sum_{i=1}^{n}(\alpha\delta(x_i, y) + 1) \\
&= \frac{1}{n}\frac{1}{\alpha + K}(\alpha n(y) + n) \\
&= \frac{\alpha}{\alpha + K}\frac{n(y)}{n} + \frac{K}{\alpha + K}\frac{1}{K} \\
&= \frac{n(y) + \frac{n}{\alpha}}{n + \frac{n}{\alpha}K}
\end{aligned}$$

Add-β estimator with $\beta = \alpha^{-1} n$

Example 2

$|V| = K$, Choose $w(x,y) = \delta(|n(x) - n(y)| \leq 1)$
$d(x) = r_{n(x)-1} + r_{n_x} + r_{n(x)+1}$
r_j is the number of words observed j times.

$$\begin{aligned} \pi_1(y) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\delta(|n(x_i) - n(y)| \leq 1)}{r_{n(x_i)-1} + r_{n(x_i)} + r_{n(x_i)+1}} \\ &= \frac{1}{n}\sum_{j=n(y)-1}^{n(y)+1} \frac{j r_j}{r_{j-1} + r_j + r_{j+1}} \end{aligned}$$

If $n(y) = 0$, $\pi_i(y) = \frac{1}{n}\frac{r_1}{r_0+r_1+r_2}$, $\sum_{y:n(y)=0} \pi_1(y) = \frac{1}{n}\frac{r_1}{1+\frac{r_1}{r_0}+\frac{r_2}{r_0}}$ similar to Good-Turing when r_0 is large.

Experiments (joint work with Damianos Karakos)

In our experiments, we used Sections 00-22 (consisting of $\sim 10^6$ words) of the UPenn Treebank corpus for training, and Sections 23-24 (consisting of $\sim 10^5$ words) for testing. We split the training set into 10 subsets, leading to 10 datasets of size $\sim 10^5$ tokens each. Averaged results are presented in the tables below for various choices of the training set size. We show the mean code-length, as well as the standard deviation (when available). In all cases, we chose $K = 10^5$ as the fixed size of our vocabulary.

Experiments

	mean code length	std
$\pi_\beta, \beta = 1$	11.10	0.03
π_{GT}	10.68	0.06
π_{ND}	10.69	0.06
π_{KD}	10.74	0.08

Table: Results with training set of size $\sim 10^5$.

	mean code length
$\pi_\beta, \beta = 1$	10.34
π_{GT}	10.30
π_{ND}	10.30
π_{KD}	10.31

Table: Results with training set of size $\sim 10^6$.

References:

- *diffusion Kernels on Graphs and Other discrete Spaces* Risi Imre Kondor and John Lafferty
- *A General Framework for Adaptive Regularization Based on Diffusion Processes on Graphs* Arthur D. Szalam, Mauro Maggioni and Ronald R. Coifman

Thank You