Interval Semi-Supervised LDA: Classifying Needles in a Haystack

Svetlana Bodrunova Sergei Koltsov Olessia Koltsova
Sergey Nikolenko Anastasia Shimorina

Laboratory for Internet Studies,
National Research University Higher School of Economics, St. Petersburg

November 27, 2013

Outline

1. Topic modeling
   - Problem setting and motivation
   - Latent Dirichlet Allocation
2. Semi-Supervised LDA
   - Choosing the issues
   - SLDA and ISLDA

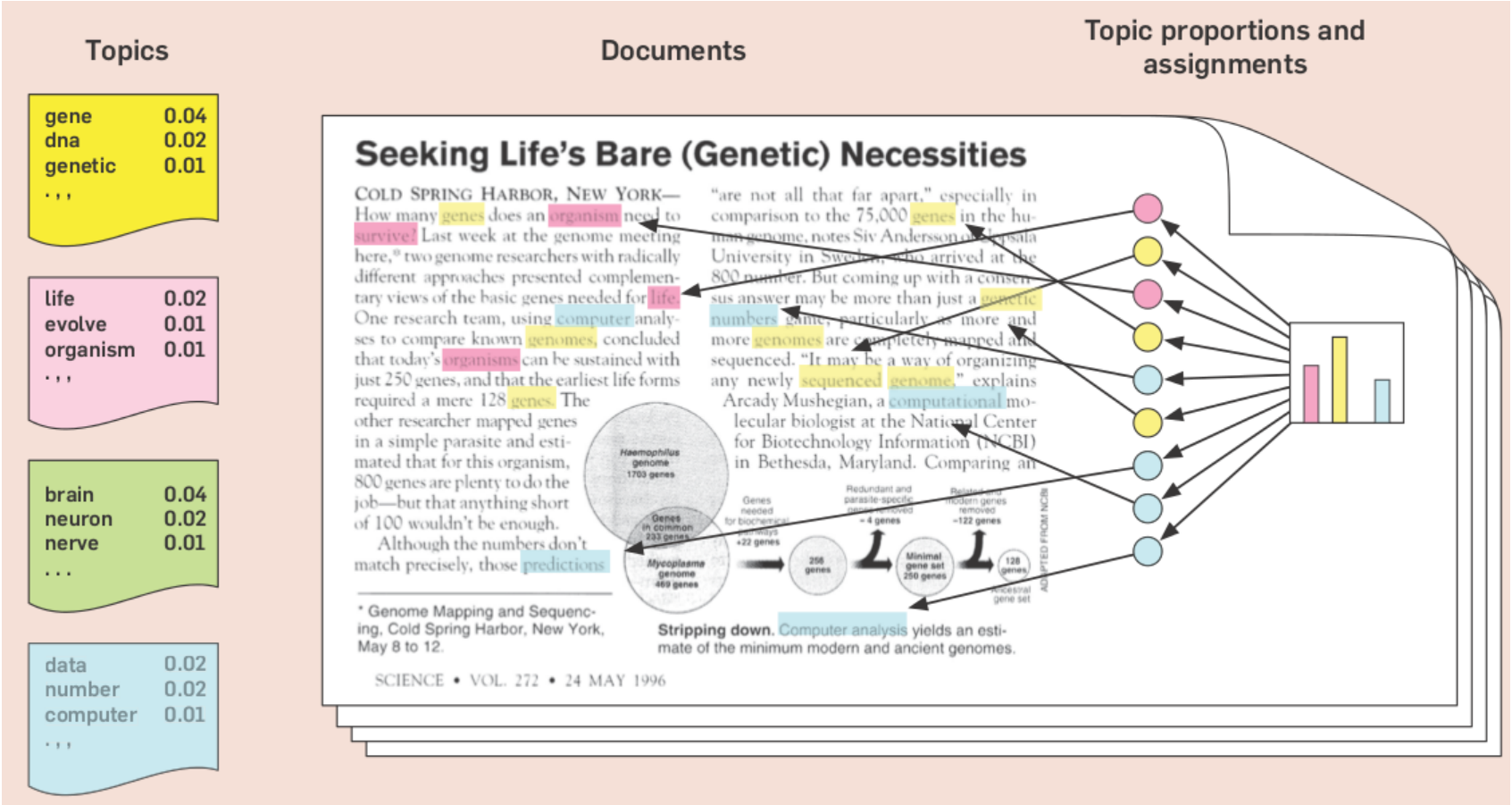
Topic modeling

- Suppose that you want to study a large text corpus.
- You want to identify specific topics that are discussed in this dataset and then either study the topics that are interesting for you or just look at their general distribution, do topical information retrieval etc.
- However, you do not know the topics in advance.
- Thus, you need to somehow extract what topics are discussed and find which topics are relevant for a specific document.
- Moreover, you want to do it in a completely unsupervised way because you do not know anything except the text corpus itself (as opposed to text categorization where you usually know the categories in advance).
- This is precisely the problem that *topic modeling* solves.

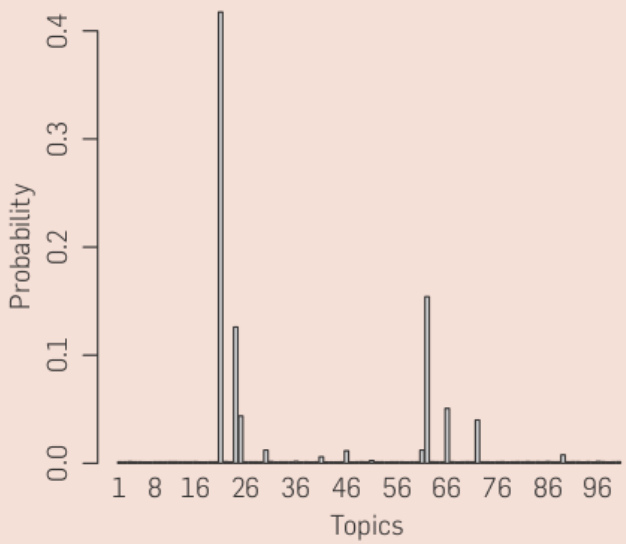
LDA

- Latent Dirichlet Allocation, LDA: the modern model of choice for topic modeling.
- In naive approaches to text categorization, one document belongs to one topic (category).
- In LDA, we (quite reasonably) assume that a document contains several topics:
  - a topic is a (multinomial) distribution on words (in the bag-of-words model);
  - a document is a (multinomial) distribution on topics.

Pictures from [Blei, 2012]



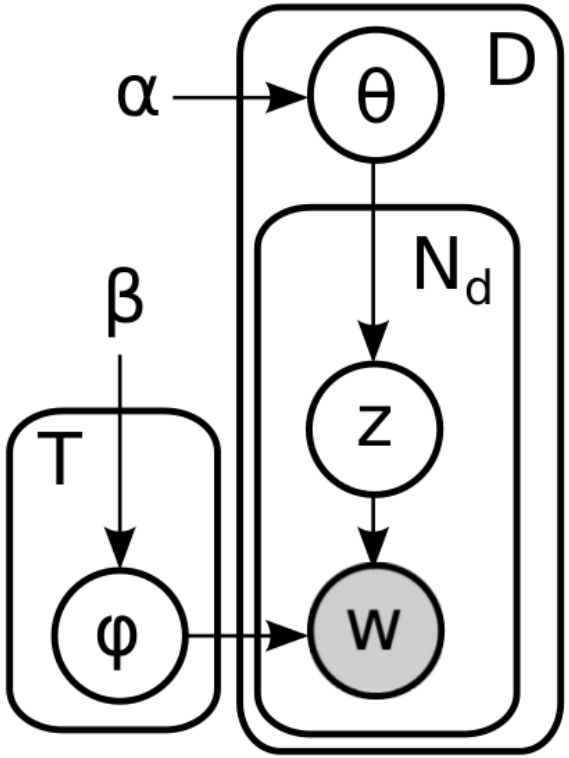
Pictures from [Blei, 2012]



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

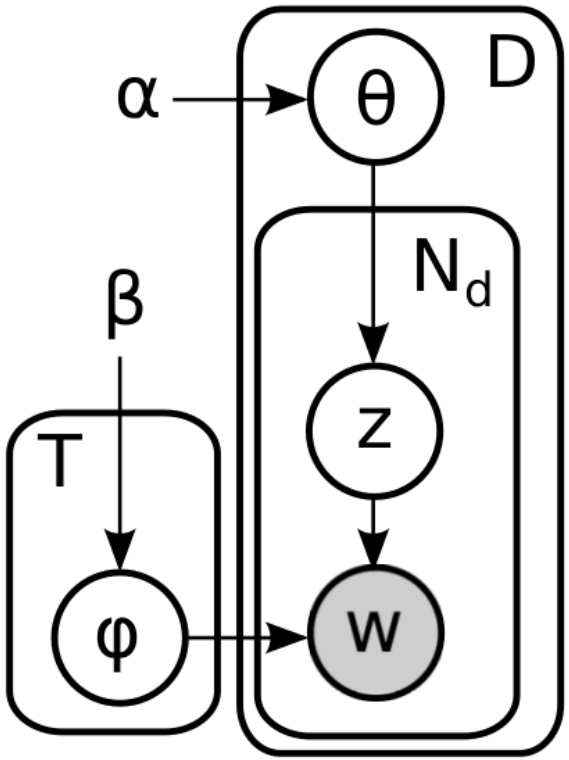
LDA

- LDA is a hierarchical probabilistic model:
  - on the first level, a mixture of topics φ with weights z;
  - on the second level, a multinomial variable θ whose realization z shows the distribution of topics in a document.
- It's called Dirichlet allocation because we assign Dirichlet priors α and β to model parameters θ and φ (conjugate priors to multinomial distributions).



LDA

- Generative model for the LDA:
  - choose document size $N \sim p(N \mid \xi)$;
  - choose distribution of topics $\theta \sim \mathrm{Dir}(\alpha)$;
  - for each of N words w_n:
    - choose topic for this word $z_n \sim \mathrm{Mult}(\theta)$;
    - choose word $w_n \sim p(w_n \mid \varphi_{z_n})$ by the corresponding multinomial distribution.



- So the underlying joint distribution of the model is

$$p(\theta, z, w, N \mid \alpha, \beta) = p(N \mid \xi)\, p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta)\, p(w_n \mid z_n, \beta).$$

LDA: inference

- There are two major approaches to inference in complex probabilistic models with a very complicated factor graph, like LDA:
  - *variational approximations* simplify the graph by approximating the underlying distribution with a simpler one, but with new parameters that are subject to optimization;
  - *Gibbs sampling* approaches the underlying distribution by sampling a subset of variables conditional on fixed values of all other variables.

LDA: inference

- Both variational approximations and Gibbs sampling are known for the LDA; we will need the collapsed Gibbs sampling:

$$p(z_w = t \mid \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) \propto q(z_w, t, \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) =$$
$$= \frac{n^{(d)}_{-w,t} + \alpha}{\sum_{t' \in T} \left(n^{(d)}_{-w,t'} + \alpha \right)} \frac{n^{(w)}_{-w,t} + \beta}{\sum_{w' \in W} \left(n^{(w')}_{-w,t} + \beta \right)},$$

  where $n^{(d)}_{-w,t}$ is the number of times topic t occurs in document d and $n^{(w)}_{-w,t}$ is the number of times word w is generated by topic t, not counting the current value z_w.
- Gibbs sampling is usually easier to extend to new modifications, and this is what we will be doing.

LDA extensions

- Numerous extensions for the LDA model have been introduced:
  - *correlated topic models* (CTM): topics are codependent;
  - *Markov topic models*: MRFs model interactions between topics in different parts of the dataset (multiple corpora);
  - *relational topic models*: a hierarchical model of a document network structure as a graph;
  - *Topics over Time*, *dynamic topic models*: documents have timestamps (news, blog posts), and we model how topics develop in time (e.g., by evolving hyperparameters α and β);
  - *DiscLDA*: each document has a categorical label, and we utilize LDA to mine topic classes related to the classification problem;
  - *Author-Topic model*: information about the author; texts from the same author will share common words;
  - a lot of work on *nonparametric* LDA variants based on Dirichlet processes (no predefined number of topics).

Outline

1. Topic modeling
   - Problem setting and motivation
   - Latent Dirichlet Allocation
2. Semi-Supervised LDA
   - Choosing the issues
   - SLDA and ISLDA

LDA in qualitative studies

- In qualitative studies, one important use for LDA is to find out what issues occur in a large text corpus with the distribution of words in topics φ.
- Then we can use the distribution of topics in documents θ to find specific documents with high affinities to interesting topics and then read only those.
- However, one usually has a specific agenda in mind.

LDA in qualitative studies

- For instance, in our case study we used a Russian LiveJournal dataset (blog posts) to study ethnic discourse: how do people talk about specific ethnic groups.
- Dataset:
  - four months of LiveJournal posts written by 2000 top bloggers;
  - 235,407 documents in total;
  - after cleaning stopwords and low frequency words, 192,614 terms in the dictionary with $\approx$ 53.5 million total instances.
- We first ran vanilla LDA and looked for topics that refer to ethnic groups.
- That is, we looked at the topics and tried to find topics with top words of ethnic origin.

LDA in qualitative studies

- E.g., here are the topics related to Ukraine and Ukranians when we run LDA with 100 topics:

Ukraine	0.043	Ukraine	0.049
Ukrainian	0.029	Ukrainian	0.017
Polish	0.012	Timoshenko	0.015
Belorussian	0.011	Yanukovich	0.015
Poland	0.011	Victor	0.012
Belarus	0.010	president	0.012

LDA in qualitative studies

- And here are the topics related to Ukraine and Ukranians when we run LDA with 400 topics:

Ukraine	0.098	Ukraine	0.054	dragon	0.026
Ukrainian	0.068	Timoshenko	0.019	Kiev	0.022
Belorussian	0.020	Yanukovich	0.018	Bali	0.012
Belarus	0.018	Ukrainian	0.016	house	0.010
Kiev	0.018	president	0.015	place	0.006
Kievan	0.012	Victor	0.013	work	0.006

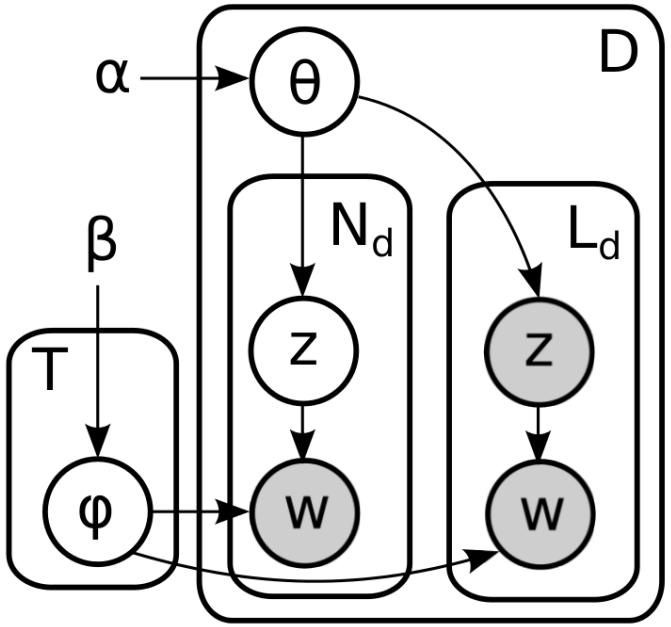
- We get one general topic, one political, and one more topic in the 400 topic case which is, frankly, trash.

Problem

- So the problem is: how do we narrow our search to find topics that are specifically related to our set of issues?
- We can define an issue with a set of keywords, e.g., “Ukrainian” and “Ukraine”, “Mexico” and “Mexican” and so on.
- But the LDA model is completely unsupervised; how do we tell it to concentrate on, say, ethnic-related topics?

Semi-supervised LDA

- We construct the so-called *semi-supervised LDA* model:
  - define each issue with a set of keywords W_{sup};
  - for each $w \in W_{\text{sup}}$, fix the topic assignment z to $\tilde{z}_w$ throughout the Gibbs sampling process;
  - run regular collapsed Gibbs sampling, but with some z fixed.

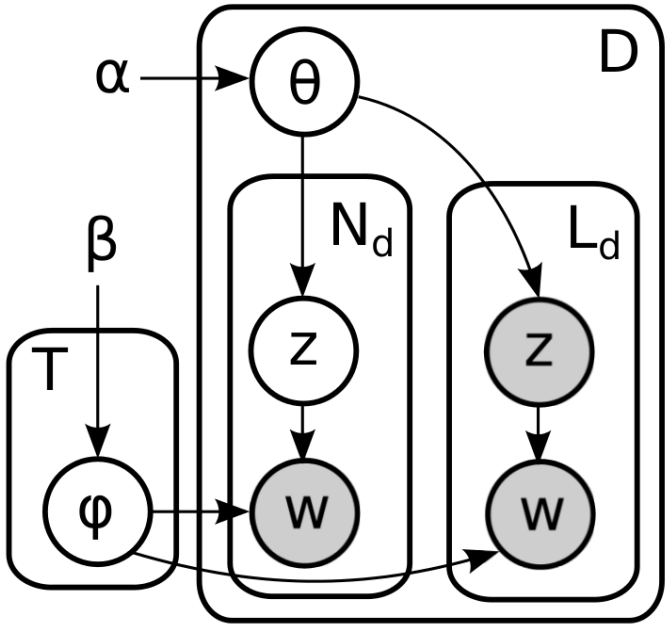


- Formally speaking, we modify the Gibbs sampler as follows:

$$p(z_w = t \mid \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) \propto \begin{cases} [t = \tilde{z}_w], & w \in W_{\text{sup}}, \\ q(z_w, t, \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) & \text{otherwise.} \end{cases}$$

Semi-supervised LDA

- E.g., we say that "Ukraine" and "Ukrainian" always belong to topic 1; then words related to Ukraine will gather around topic 1.
- As a result of this very simple modification, we can fix specific issues to specific topics.



- Formally speaking, we modify the Gibbs sampler as follows:

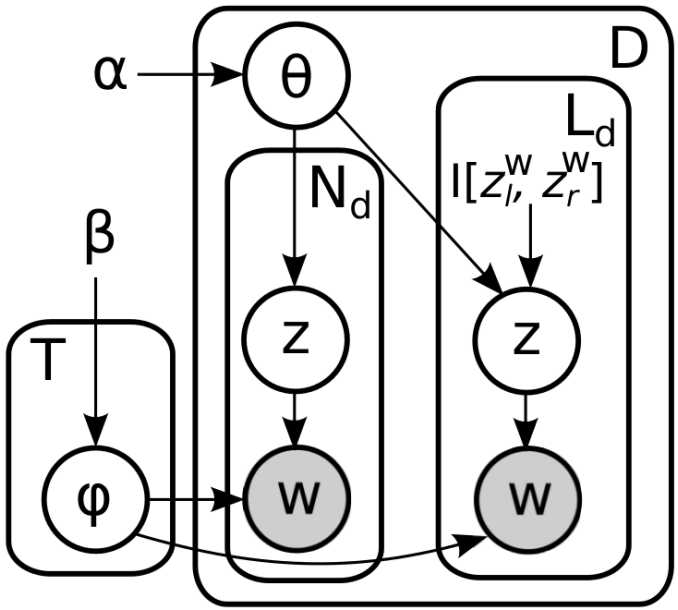
$$p(z_w = t \mid \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) \propto \begin{cases} [t = \tilde{z}_w], & w \in W_{\text{sup}}, \\ q(z_w, t, \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) & \text{otherwise.} \end{cases}$$

Semi-supervised LDA

- But this, of course, only makes things worse.
- We had two topics about Ukraine, and maybe could get more if we increased the number of topics.
- Now we are guaranteed to have only one topic about Ukraine.
- Therefore, we need to find a way to assign issues to a *subset* of topics (would be convenient to use a contiguous interval).

Interval semi-supervised LDA

- *Interval semi-supervised LDA* model:
  - define each issue with a set of keywords W_{sup};
  - for each $w \in W_{\text{sup}}$, fix the topic assignment z to an *interval* of topics $[z_l^w, z_r^w]$;
  - this means we set topic probability to 0 outside $[z_l^w, z_r^w]$ and renormalize it inside.



- Formally speaking, we modify the Gibbs sampler as follows:

$$p(z_w = t \mid \ldots) \propto \begin{cases} [t \in [z_l^w, z_r^w]] \frac{q(z_w, t, \ldots)}{\sum_{z_l^w \le t' \le z_r^w} q(z_w, t', \ldots)}, & w \in W_{\text{sup}}, \\ q(z_w, t, \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) & \text{otherwise.} \end{cases}$$

Interval semi-supervised LDA

- This works as intended. For instance, if we apply ISLDA to the same dataset with five topics for Ukraine, we get the following:

Ukraine	0.065	Ukraine	0.062	Ukrainian	0.040	Crimea	0.046
gas	0.030	Timoshenko	0.023	Ukraine	0.036	Crimean	0.015
Europe	0.026	Ukrainian	0.022	Polish	0.021	Sevastopol	0.015
Russia	0.019	Yanukovich	0.018	Poland	0.017	Simferopol	0.008
Ukrainian	0.018	Kiev	0.015	year	0.009	Yalta	0.008
Belorussian	0.018	Victor	0.014	L'vov	0.006	source	0.007
Belarus	0.017	president	0.013	Western	0.005	Orjonikidze	0.005
European	0.015	party	0.013	cossack	0.005	sea	0.005

- Note the new topics about Ukrainian–Russian natural gas talks, Ukrainian–Polish relations, and the Crimean peninsula.
- There also have been some interesting documents uncovered in these topics that we would miss in vanilla LDA.
- Moreover, we can apply ISLDA to several different issues, setting different intervals of topics for each issue.

Sanity check: perplexity

- Of course, we have, in a way, interfered with the model, preventing it from modeling the document corpus as best it could.
- So we might lose a lot of predictive quality, which is commonly measured as *held-out perplexity*, average log probability of a held-out validation set of documents in the trained model.
- Fortunately, we have not lost much; the table shows a rather extreme case when we use almost half of all topics in the semi-supervised part.

# of topics	Perplexity, LDA		Perplexity, ISLDA	
	D_{test}	$D_{\text{test}}^{\text{key}}$	D_{test}	$D_{\text{test}}^{\text{key}}$
100	12.7483	12.7483	12.7542	12.7542
200	12.7457	12.7457	12.7485	12.7486
400	12.6171	12.6172	12.6216	12.6216

Summary

- We have presented a modification for the LDA topic model that allows to indicate specific issues (with sets of keywords) and match a specific subset of topics to them.
- We have successfully applied this modification in a qualitative case study of ethnic discourse in Russian LiveJournal.
- The predictive quality of the model does not seem to suffer much, so this modification, while useful qualitatively, does not break the quantitative probabilistic model.

Thank you!

Thank you for your attention!