TOPIC SEGMENTATION OF CONTACT CENTER CALLS TEXTS

VLAD BATAEV (BATAEV@PHYSTECH.EDU)

YANDEX SCHOOL OF DATA ANALYSIS, MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY, MOSCOW, RUSSIA

CONTACT CENTER CALLS

Speech recognition system transforms audio records of calls into text. **Outcoming** contact center calls are usually dialogs during of which operator tries to persuade the client to acquire the product.

This text could be used to make **data-driven** decisions:

- extracting most effective product features
- extracting clients' questions and objections
- quality control of the operator
- improvement of marketing scripts

First step to do this is extracting **topics** which gives an approximate understanding of text.

TOPIC EXAMPLES

- Introduction and farewell
- Product specific feature
- Operator's argument
- Client's objections

TOPIC MODELING

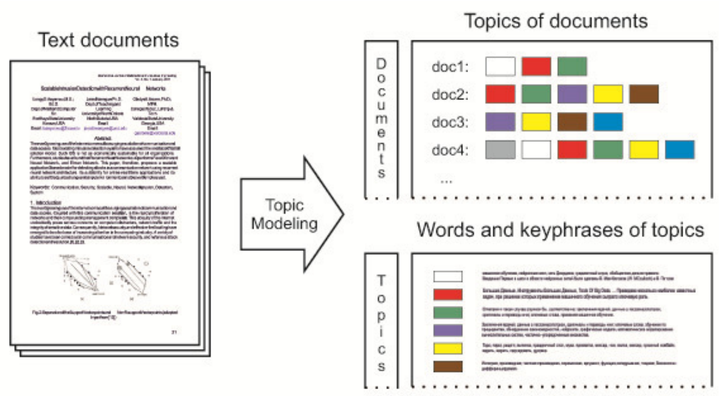
Given a collection of documents, **assume** that each observable word w in document d refers to some latent topic t:

$$p(w|d) = \sum_{t\in T} p(w|t)\, p(t|d) = \sum_{t\in T} \phi_{wt}\, \theta_{td}$$

Find:

- $\phi_{wt} \equiv p(w|t)$ — words for each topic,
- $\theta_{td} \equiv p(t|d)$ — topics for each document,

resulting in $p(w|d)$ close to $\hat{p}(w|d) \propto n_{dw}$ – frequencies of words in documents.



ADDITIVE REGULARIZATION

Regularized Likelihood:

$$\underbrace{\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td}}_{L(\Phi,\Theta)} + \underbrace{\sum_{i=1}^{n} \tau_i R_i(\Phi,\Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta}$$

Maximization is performed via modified EM-algorithm – iterative process, alternating:

E-step (Bayes' Rule for $p(t|d,w) \equiv p_{tdw}$):

$$p_{tdw} \propto \phi_{wt}\theta_{td}$$

M-step (parameters estimates):

$$\phi_{wt} \propto \left(\sum_{d\in D} n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+$$

$$\theta_{td} \propto \left(\sum_{w\in W} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+$$

Implementation:

bigARTM.org – open source library of additively regularized topic models.

SEMI-SUPERVISED REGULARIZER

For business purposes topics must be **interpreted**.

We have **prior** knowledge about some topics: we can mark a few documents with their topics and use this information in the model

$D_0 \subset D$ — set of marked documents, for each document $d \in D_0$ define $T_d \subset T$ — relevant topics related to d.

Semi-supervised regularizer:

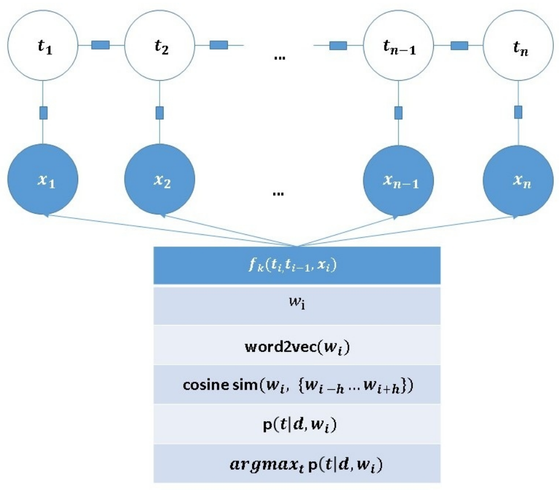
$$R(\Phi,\Theta) = \sum_{d\in D_0}\sum_{t\in T} \frac{1}{|T_d|}(-1)^{I[t\notin T_d]} \ln \theta_{td}$$

Smoothing along relevant topics and sparsing along irrelevant topics in marked documents

CONDITIONAL RANDOM FIELD

Problem: regularizer for segmentation task doesn't work quite well, otherwise topic model ignores topic shifts in a single document.

Solution: account for dependencies between topics inside a single document using discriminative undirected graphical model — **linear-chain-crf**



Condtional distribution of unobserved topic sequence given observed sequence of words:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})}\prod_{n=1}^{N}\Psi_n(y_n, y_{n-1}, \boldsymbol{x_n})$$

$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}}\prod_{n=1}^{N}\Psi_n(y_n, y_{n-1}, \boldsymbol{x_n})$$

$$\Psi_n(y_n, y_{n-1}, \boldsymbol{x_n}) = \exp\{\sum_{k=1}^{K}\theta_k f_k(y_n, y_{n-1}, \boldsymbol{x_n})\}$$

$f_k(y_n, y_{n-1}, \boldsymbol{x_n})$ — arbitary real valued **feature functions**.

LOOKING INTO TOPIC SEGMENTATION

Feature functions:

- topic features: $p(t|, d, w)$ and their smoothed version by fixed size context
- preliminarily trained word2vec features: word2vec embedding of context, weighted sum of word vectors from fixed size window

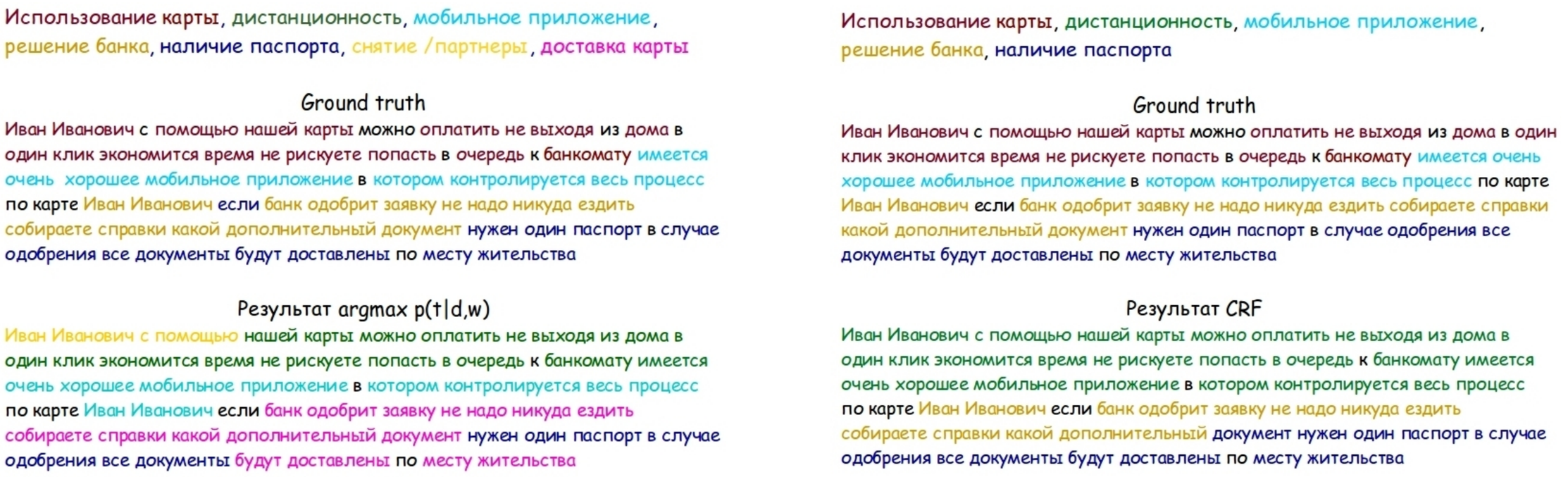


Fig. 1. CRF smoothed topics of neigboring words in contrast with $\arg\max_{t\in T} p(t|d,w)$