МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
"МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ)"

ФАКУЛЬТЕТ ИННОВАЦИЙ И ВЫСОКИХ ТЕХНОЛОГИЙ
КАФЕДРА АНАЛИЗА ДАННЫХ

Выпускная квалификационная работа по направлению

01.03.02 «Прикладные математика и информатика»

НА ТЕМУ:

ТЕМАТИЧЕСКАЯ СЕГМЕНТАЦИЯ РАЗГОВОРОВ КОНТАКТНОГО ЦЕНТРА

Студент _____ Батаев В.В.

Научный руководитель д.ф-м.н _____ Воронцов К.В.

Зам. зав. кафедрой д.ф-м.н, профессор _____ Бунина Е.И.

МОСКВА, 2017

Содержание

Часть 1

Тематическая сегментация текста

1.1 Постановка задачи

Сформулируем формальную постановку задачи. Пусть дано предложение на естественном языке, которое можно считать последовательностью термов: $w_1, \dots w_n$. В качестве термов могут выступать отдельные слова, n-граммы, коллокации. Для каждого терма требуется поставить метку темы, то есть некоторой обобщающей данное слово сущности в зависимости от контекста. Таким образом решается задача *sequence labelling*, то есть обучения отображения $f : W^k \to T^k$, $k \geq 0$, W, T — конечные множества. Подпоследовательности из одинаковых тем будем называть монотематичными *сегментами*.

1.2 Метрика качества

Для оценки качества сегментации будем использовать усредненную взвешенную $f1$-меру для каждой темы по всем словам из всех предложений корпуса. В таком случае метрика качества сегментации фокусируется на отдельных словах и на точном проведении границ сегментов.

Однако не всегда важно знать точные границы сегментов, а достаточно получить только порядок следования тем. В таком случае метрикой качества может быть редакторское расстояние между последовательностями смен тем.

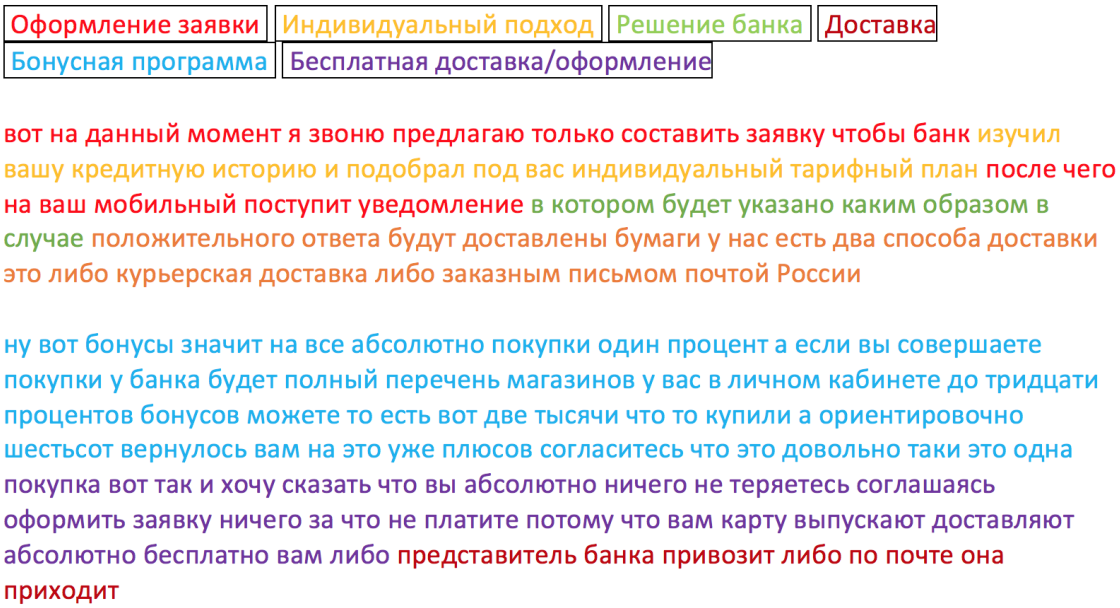

Рис. 1.1: Пример сегментации

1.3 Обзор существующих методов

При решении задачи тематической сегментации текста существует два принципиальных различных подхода. Первый основан на unsupervised learning, не требующей размеченных отсегментированных текстов, в котором основной идеей является предположение, что граница между тематическими сегментами проходит при резком смене темы локального контекста.

В работе [2] использовался алгоритм, основанный на движении скользящего окна по тексту. Для каждого положения окна строится bag-of-words вектор, сглаженный с помощью tf-idf, описывающий данное окно. По построенным векторам строилась последовательность косинусных расстояний между соседними окнами, и в ней точки максимума или точки со значениями, большими некоторого заранее определенного порога, определяются как границы сегментов. Однако же данный метод не позволяет определить темы построенных сегментов.

В работе [3] по корпусу строится вероятностная тематическая LDA модель ([1]), с помощью которой для каждого слова в некотором предложении имеется распределение по темам. Затем вектора распределений тем соседних слов сравниваются с помощью косинусной метрики и производится расстановка сегментов аналогично методу со скользящим окном.

Второй подход основан на supervised learning, использующий разметку текста на сегменты. В работе [4] предлагается решать задачу классификации для каждого слова — поставить границу сегмента после данного слова или нет, используя нейросетевую архитектуру, состоящую из рекуррентных и сверточных слоев для слов. Отметим, что в данной работе также нет прямой возможности узнать, к каким темам относятся построенные сегменты.

Часть 2

Описание новых предложенных методов

Будем считать, что перед нами стоит supervised задача, поскольку качество ее решения окажется заведомо выше, нежели в unsupervised случае. Построим semi-supervised вероятностную тематическую модель, использующую априорные знания о принадлежности некоторых слов и документов к определенным темам. Выход тематической модели задает некоторое приближение для сегментации, но векторы тем слов не учитывают контекст напрямую. Для решения данной проблемы предлагаются следующие методы:

- Сглаживать вектора тем соседних слов с помощью векторов из окна фиксированной ширины
- Ввести дополнительную регуляризацию тематической модели на близость векторов тем соседних слов
- Моделировать взаимосвязи между темами слов предложения с помощью дискриминативной модели, а именно CRF

2.1 Сглаживание векторов тем соседних слов

Напомним, что результатом алгоритма тематического моделирования являются распределения $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$. Из формулы Байеса и гипотезы условной независимости слова при данном документе и теме следует:

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

Сглаживание будем производить итеративно последовательно по всем словам документа по следующей формуле:

$$g(t,d,k,i) = f(i)p(t|d,w_{k+i})$$

$$p^*(t|d,w_k) = norm_{t\in T}(\sum_{j=-h,j\neq 0}^{h} g(t,d,k,j))$$

$f(i)$ — функция затухания влияния контекста на распределение текущего слова, h — размер окна. Число итераций для документов — параметр модели.

TODO: подробное описание тематического моделирования, какие регуляризаторы используются

2.2 Условное случайное поле

В задаче тематической сегментации распределение тем при условии данного слова и документа не учитывает влияния контекста внутри предложения. Для решения данной проблемы предлагается использовать модель условного случайного поля или CRF для моделирования зависимостей между темами разных слов. Преимущество CRF в том, что такая модель является дискриминативной, а следовательно способна более точно моделировать условное распределение $p(\boldsymbol{y}|\boldsymbol{x})$ неизвестной последовательности меток $\boldsymbol{y}$ при условии обозреваемой последовательности слов $\boldsymbol{x}$.

В данной задаче будем использовать linear-chain-CRF, чей граф представляет собой линейную цепь. В таком случае условное распределение можно записать в следующем виде:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})}\prod_{t=1}^{T}\exp\{\sum_{k=1}^{K}\theta_k f_k(y_t, y_{t-1}, \boldsymbol{x}_t)\},$$

где $Z(\boldsymbol{x})$, зависящая от конкретного $\boldsymbol{x}$ функция нормализации

$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}}\prod_{t=1}^{T}\exp\{\sum_{k=1}^{K}\theta_k f_k(y_t, y_{t-1}, \boldsymbol{x}_t)\}.$$

$f_k(y, y'\boldsymbol{x}_t)$ — произвольные вещественнозначные функции, которые также называются *feature functions*. Они описывают зависимости между соседними метками тем и произвольным подмножеством входных переменных из последовательности. Подчеркиваем, что $\boldsymbol{x}_t$ может зависеть от произвольного подмножества $\{x_1, \dots x_T\}$. Таким образом условное распределение $p(\boldsymbol{y}|\boldsymbol{x})$ не накладывает каких-либо ограничений на $p(\boldsymbol{x})$.

В качестве f_k будем использовать:

- $I(x_t) = w$ — индикатора присутствия слова w в позиции t
- Вектор $p(t|w, d)$ из обученной предварительно тематической модели
- word2vec представление слова w_t
- Косинусные расстояния между word2vec представлениями словом w_t и словами некоторого окна $w_{t-h}, \dots w_{t+h}$

TODO: вывод в linear-chain CRF, обучение параметров θ_k

Часть 3

Результаты экспериментов

3.1 Входные данные

Обучающая выборка состоит из 400 диалогов контактного центра, в которых были оставлены только реплики оператора. В каждой такой реплике проставлены границы сегментов, и каждый из сегментов отнесен к некоторой теме. Число тем было выбрано равным 53 экспертом.

TODO: привести примеры тем и сегментов

3.2 Unsupervised метод с помощью векторных представлений

TODO: получение векторных представлений для окна с помощью нейросетей

3.3 Сглаживание векторов тем

Для эксперимента, описанного в 2.1, была взята функция $f(i) = \frac{1}{i^2}$. Параметры h — размер окна и $num\_iterations$ — число итераций определялись по $5 - fold$ кросс-валидации.

Результаты представлены в следующей таблице:

Table 3.1: Точность сегментации сглаживанием

Размер окна	Число итераций	weighted f1	Edit distance
3	3	0.49	
3	4	0.51	
4	3	0.48	

3.4 Метод, основанный на CRF

Признак, который является индикатором присутствия конкретного слова в текущей позиции, используется во всех моделях, поскольку является бейзлайном, который не требует какой-либо предварительной предобработки. Результаты представлены в следующей таблице:

Table 3.2: Точность сегментации CRF

Feature functions	weighted f1	Edit distance
Индикаторы слов	0.5424	
Индикаторы слов + w2v features	0.5482	
Индикаторы слов + векторы тем	0.6191	
Индикаторы слов + векторы тем + w2v features	0.6193	

Литература

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[2] Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64, March 1997.

[3] Martin Riedl and Chris Biemann. Text segmentation with topic models.

[4] Liang Wang, Sujian Li, Xinyan Xiao, and Yajuan Lyu. *Topic Segmentation of Web Documents with Automatic Cue Phrase Identification and BLSTM-CNN*, pages 177–188. Springer International Publishing, Cham, 2016.