Как научить нейронную сеть генерировать стихи

Жариков Илья Николаевич

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

26 октября 2017 г.

Содержание

Языковые модели

Определение

Языковые модели определяют вероятность появления последовательности слов $w_1, \ldots, w_n$ в данном языке.

$$\mathbb{P}(w_1, \ldots, w_n) = \prod_{i=1}^{n} \mathbb{P}(w_i \mid w_1, \ldots, w_{i-1}).$$

Самый простой способ построения:

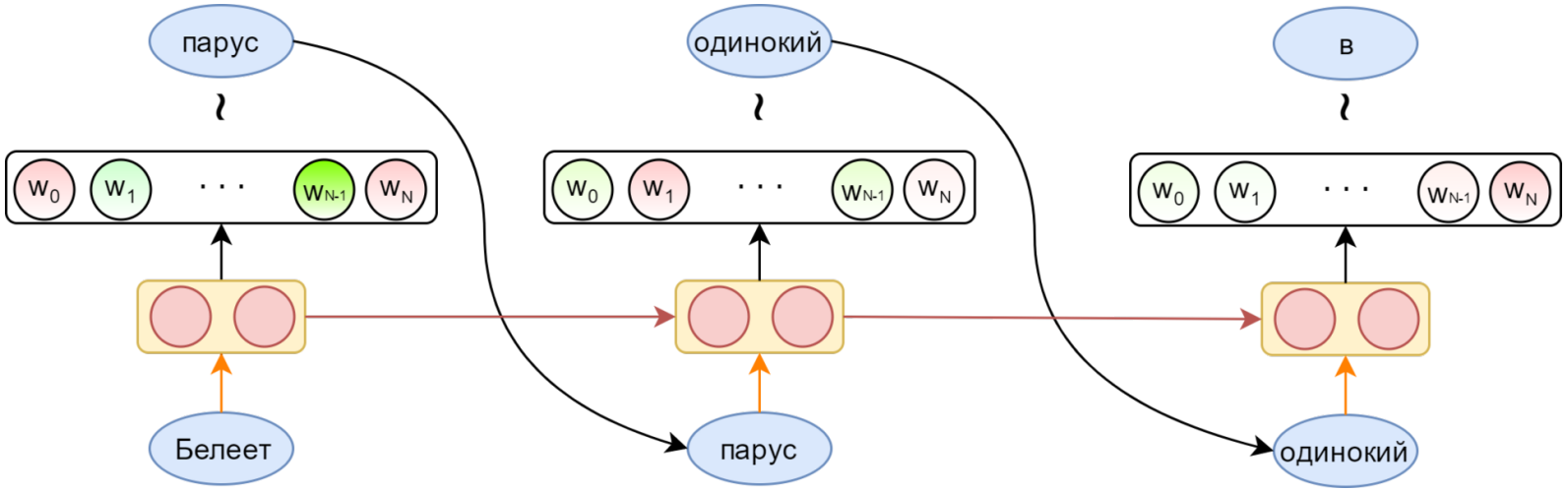
$$\mathbb{P}(w_i, \mid w_1, \ldots, w_{i-1}) \approx \mathbb{P}(w_i \mid w_{i-N}, \ldots, w_{i-1})$$

Модель:
- \+ просто реализуется
- − весьма тяжелая
- − не слишком вариативная

Статистика:
- − недостаточная

Нейросетевые языковые модели

Выход: распределения вероятностей для слов из словаря.



Рыжей стрелкой показано отображение слова в embedding.

Выходной слой: полносвязный слой размера словаря
+ функция активации **softmax**.

- размер очень большой $\Rightarrow$ Иерархический softmax; Noise contrastive estimation.

Оценивание языковой модели

Используется стандартная функция потерь (cross entropy):

$$H(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_i y_i \log_2 \hat{y}_i = -\log_2 \hat{y}_k, \text{ если } y_i = \delta_{ik} \, \forall i.$$

Кросс-энтропийные потери целого предложения:

$$H(w_1, \ldots, w_n) = -\frac{1}{n} \sum_k \log_2 \mathbb{P}(w_k \mid w_1, \ldots, w_{k-1}).$$

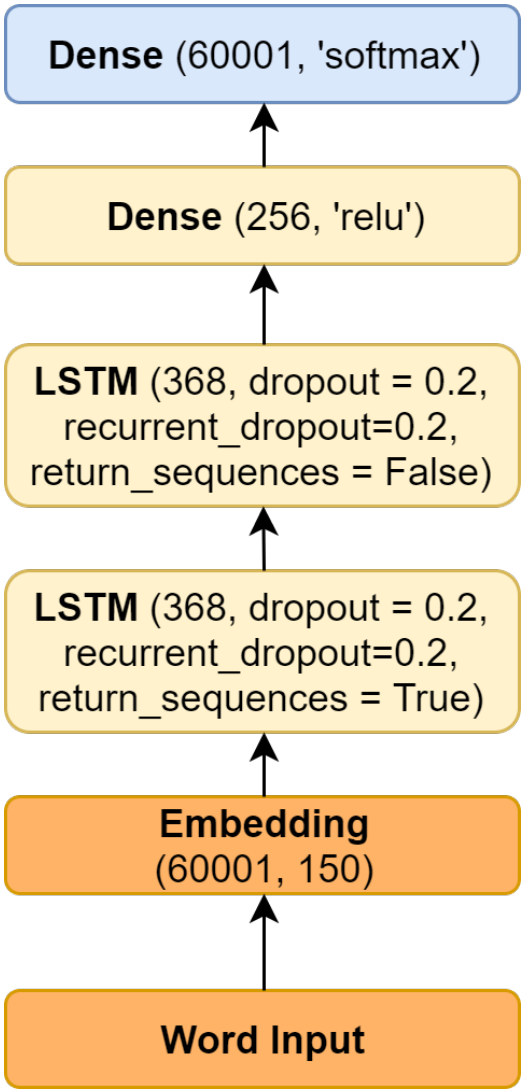
Перплексия (perplexity):

$$PP(w_1, \ldots, w_n) = 2^{H(w_1,\ldots,w_n)} = 2^{-\frac{1}{n}\sum_k \log_2 \mathbf{P}(w_k|w_1,\ldots,w_{k-1})}.$$

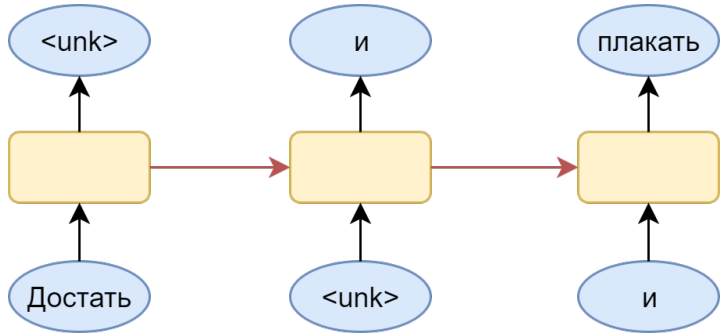
Перплексия — уровень неоднозначности генерации слова.

Plan

Построение архитектуры сети

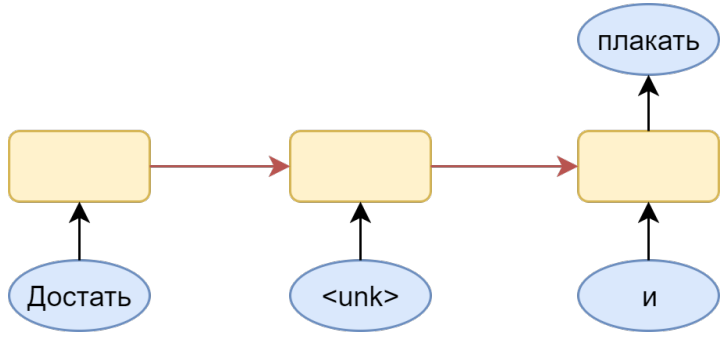


Many-to-Many:



- частое предсказание <unk> токена

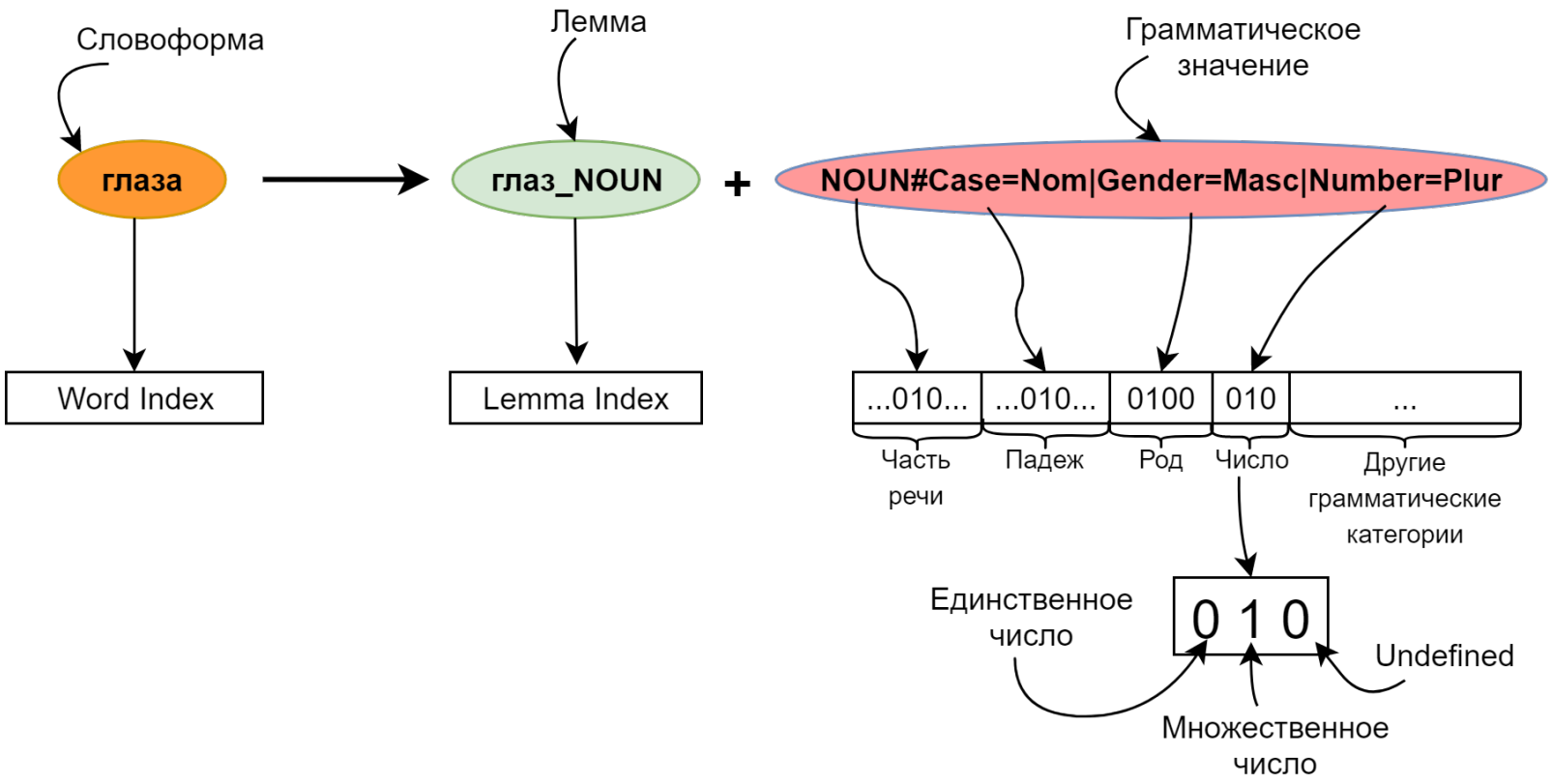
Many-to-One:



- приходится нарезать всевозможные цепочки слов из обучающей выборки

Доработка входного слоя

Нужно **сократить** размерность входного слоя:



Замечание. Грамматическое значение (битовый вектор) лучше предварительно пропустить через один или два полносвязных слоя.

Доработка выходного слоя

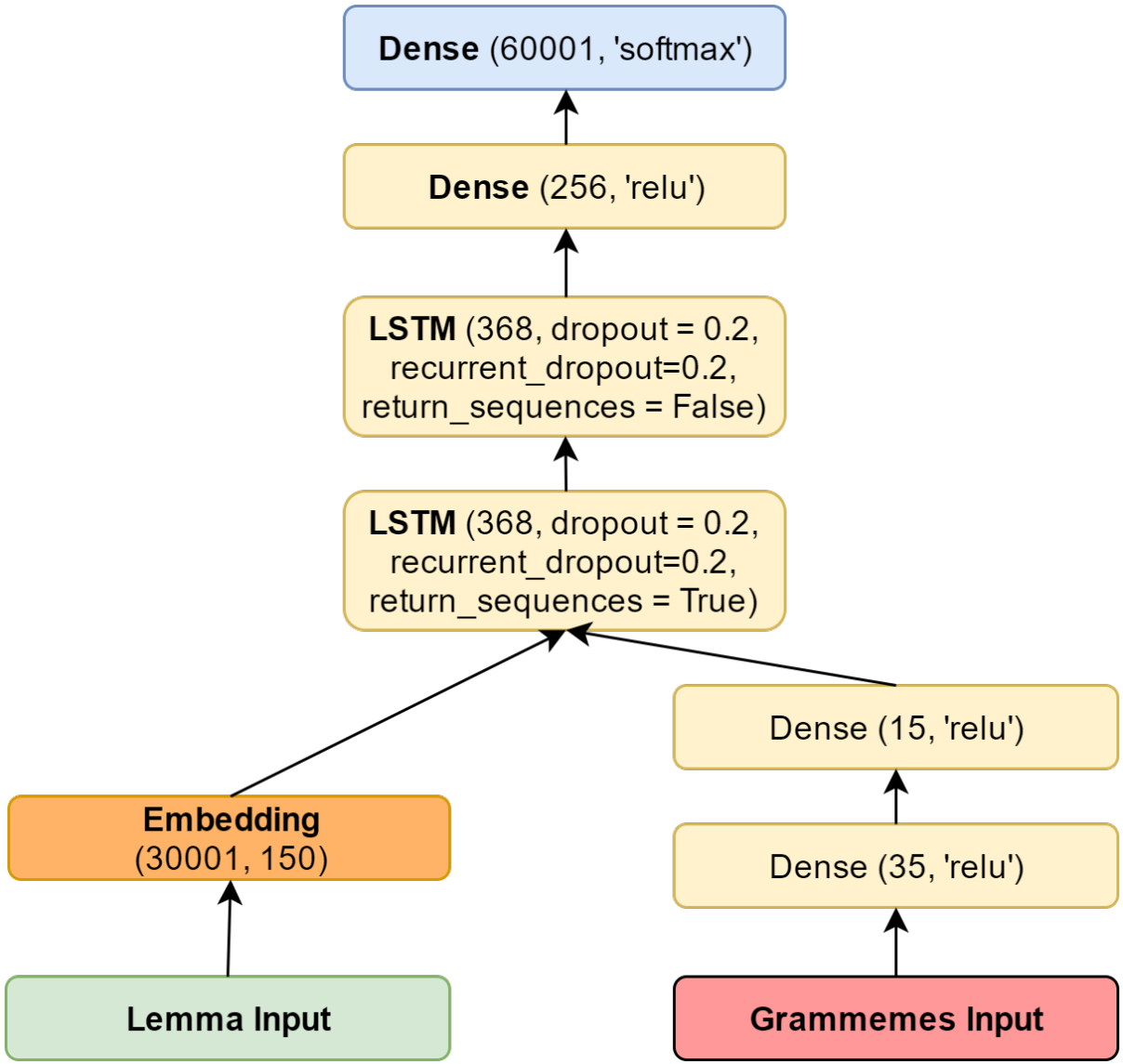
Идея: вместо индекса слова можно предсказывать всё те же лемму и грамматическое значение по отдельности.

Проблема: невозможно гарантировать наличие такого грамматического значения у данной леммы.

Решения:

1. Сэмплировать слово из действительно реализуемых пар;
2. Выбирать наиболее вероятное грамматическое значение среди возможных для сэмплированной леммы.

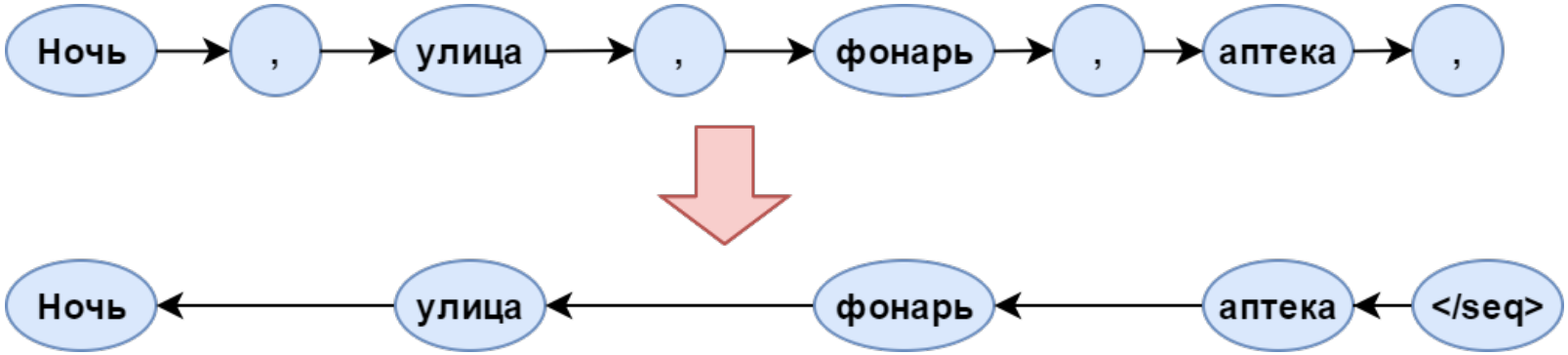
Итоговая модель



Данные для обучения

Данные: stihi.ru + морфологическая разметка.

Что было сделано с данными:



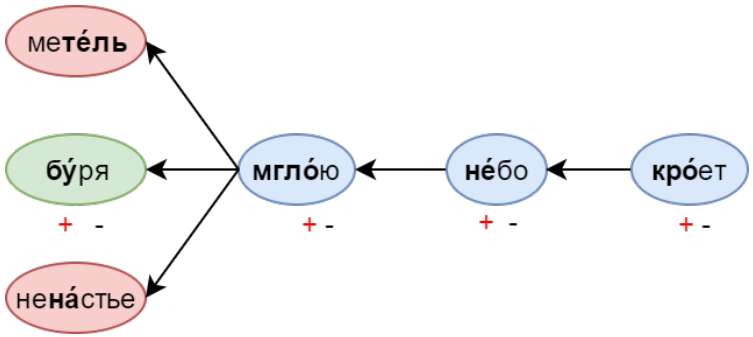
- Каждая строка рассматривалась как самостоятельная;
- При обучении ко всем строкам добавлялся завершающий символ;
- Порядок слов в строках инвертировался;
- Все знаки препинания из текстов были выброшены.

Plan

Правила-фильтры

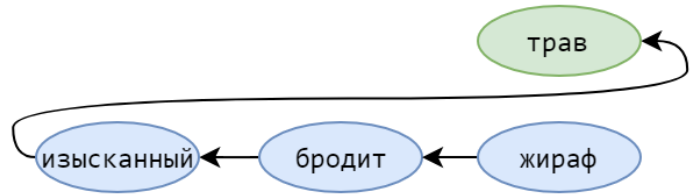
Языковая модель для построения следующего слова.
Генератор для построения стихотворений.

Метрические правила:



Определяют последовательность ударных и безударных слогов в строке

Ограничения по рифме:



+ правило, запрещающее считать рифмами словоформы с одинаковой леммой

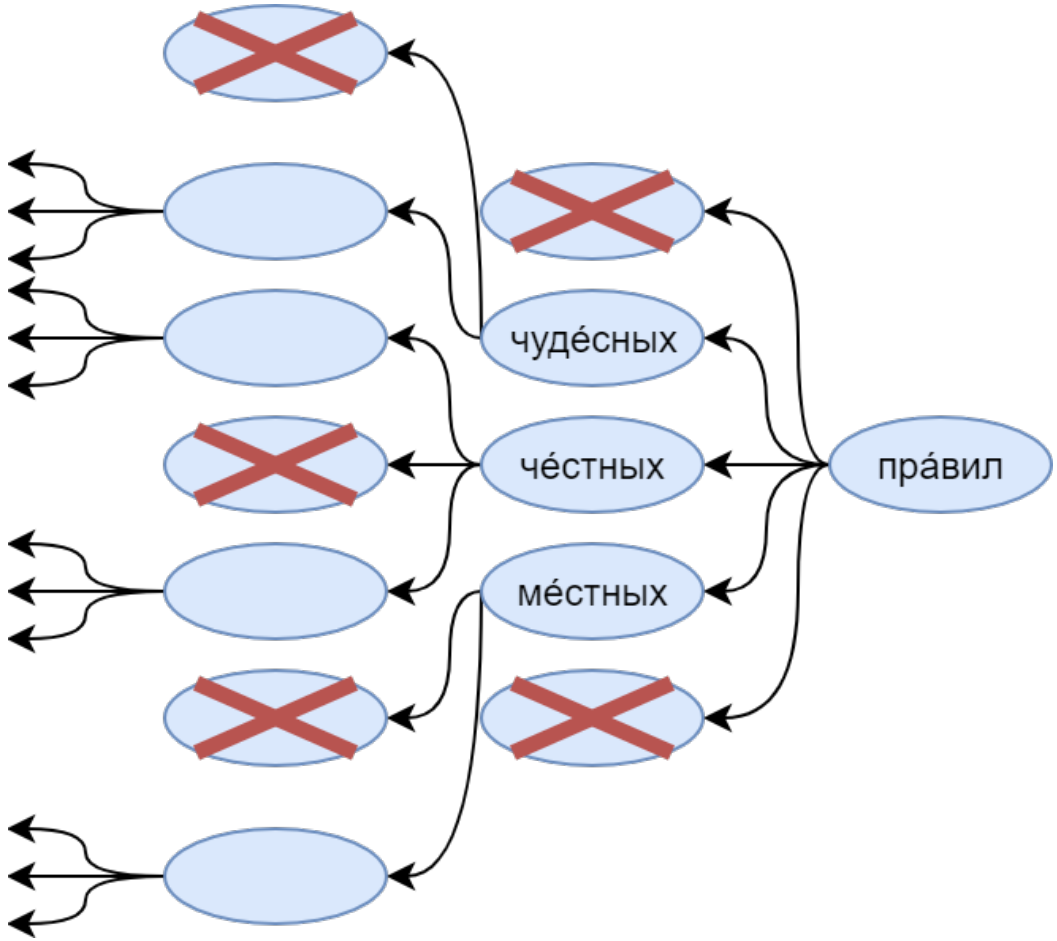
Лучевой поиск

В результате работы фильтров могло не остаться ни одного слова.

⇓

На каждом шаге сразу выбирать N путей с наивысшими вероятностями.

⇓



Примеры стихов

Так толку мне теперь грустить
Что будет это прожито
Не суждено кружить в пути
Почувствовав боль бомжика

Скучаю за твоим окном
И нежными эфирами
Люблю тебя своим теплом
Тебя стенографируя

увы стоял плачевный стул
на стуле том сидел аул
на нем сидел большой больной
сидел к живущему спиной
он видел речку и леса
где мчится стертая лиса
где водит курицу червяк
венок звонок и краковяк
сидит больной скребет усы
желает соли колбасы
желает щеток и ковров
он кисел хмур и нездоров

Не могу ответить мне поверьте
Я не знаю. Каждый день морозный
Душу гложет. У меня на сердце
Страх тревожит разум мой колхозный

Затерялся где то на аллее
Где же ты мое воспоминанье
Я люблю тебя мои родные
Сколько лжи предательства и лести
Ничего другого и не надо
За грехи свои голосовые