YANDEX.ALGORITHM GENERAL CONVERSATION CHALLENGE

Антюхов Денис

ЗАДАЧА СОРЕВНОВАНИЯ

- **Цель:** подбор реплик, подходящих по смыслу для данного момента разговора и способных заинтересовать пользователя в продолжении беседы

- **Задача:** построить модель, ранжирующую реплики по степени их уместности в диалоге

ДАННЫЕ

- **Структура:** короткие эпизоды диалогов (2-4 реплики)
  Каждый эпизод состоит из двух частей – контекста и набора финальных реплик - кандидатов:

  контекст_2: Персонаж А говорит реплику
  контекст_1: Персонаж В отвечает на нее
  контекст_0: Персонаж А произносит вторую реплику
  кандидат_1: Персонаж В отвечает на вторую реплику
  кандидат_6: Персонаж В отвечает на вторую реплику

- **Объем:** 17178 уникальных контекстов, для каждого из которых ~6 реплик-кандидатов

- **Разметка:** реплики-кандидаты размечены асессорами по степени релевантности следующим образом:
  0: реплика не имеет никакого смысла в данном контексте
  1: реплика уместна, но не интересна (тривиальна)
  2: реплика уместна и интересна (нетривиальна, мотивирует продолжать разговор)

  Для каждой метки был известен confidence weight - степень согласованности асессоров

МЕТРИКА

- Для каждого эпизода - вернуть реплики кандидаты в порядке убывания скора
- Метрика соревнования - Normalized Discounted Cumulative Gain (nDCG)

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} \qquad \mathrm{nDCG_p} = \frac{DCG_p}{IDCG_p}$$

СТРУКТУРА РЕШЕНИЯ

- **Unsupervised признаки:**
  - 70 шт
  TF-IDF, avg-w2v, WMD, Levenshtein, etc etc

- **Supervised признаки:**
  - 7 шт
  - 10-Fold OOF predictions

- **Модель ансамблирования:**
  - 1 шт
  - LightGBM регрессия

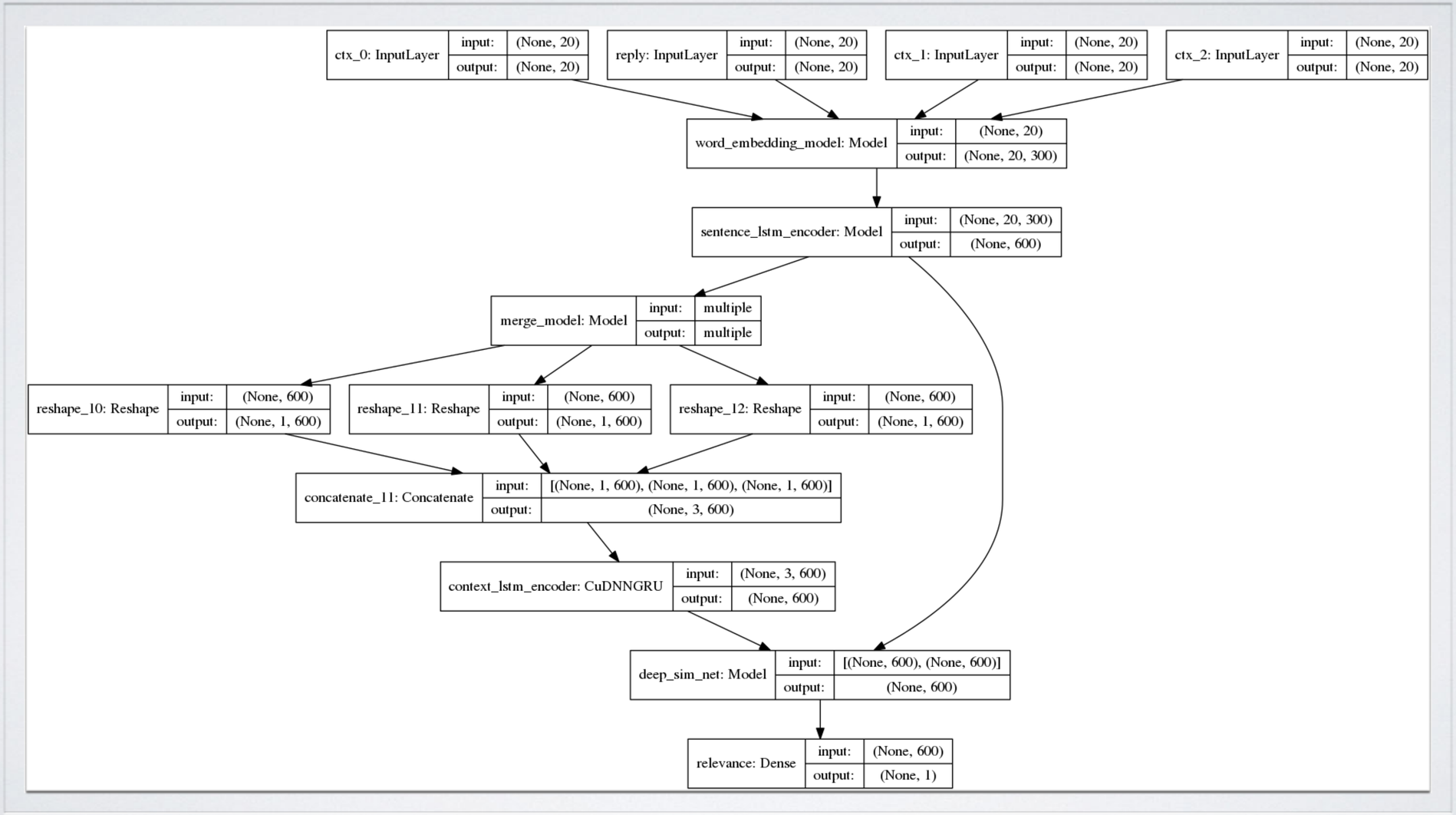
ПРОСТЫЕ ПРИЗНАКИ

- WMD-relax, AVG-W2V, TF-IDF
- Jaccard, Levenshtein distance
- ROUGE-1,3,5
- Длины реплик, наличие знаков пунктуации и пр.

НЕЙРОСЕТЕВАЯ МОДЕЛЬ

- Shared LSTM-энкодер уровня предложения
- GRU-энкодер уровня контекста
- MLP-голова для вычисления релевантности

АРХИТЕКТУРА

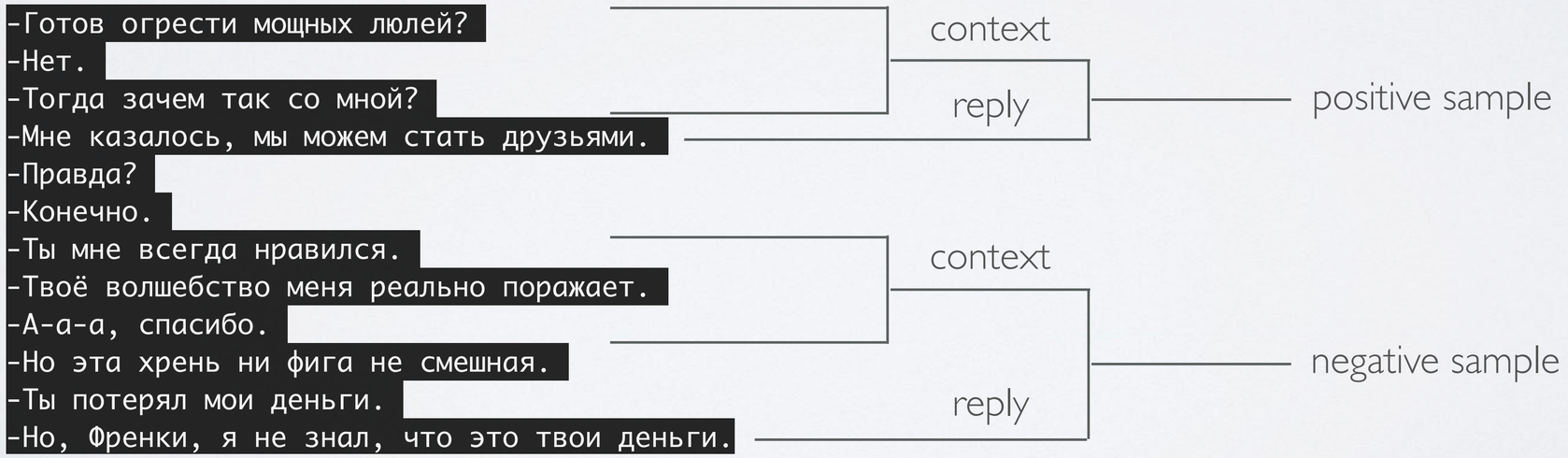


ПРОБЛЕМА

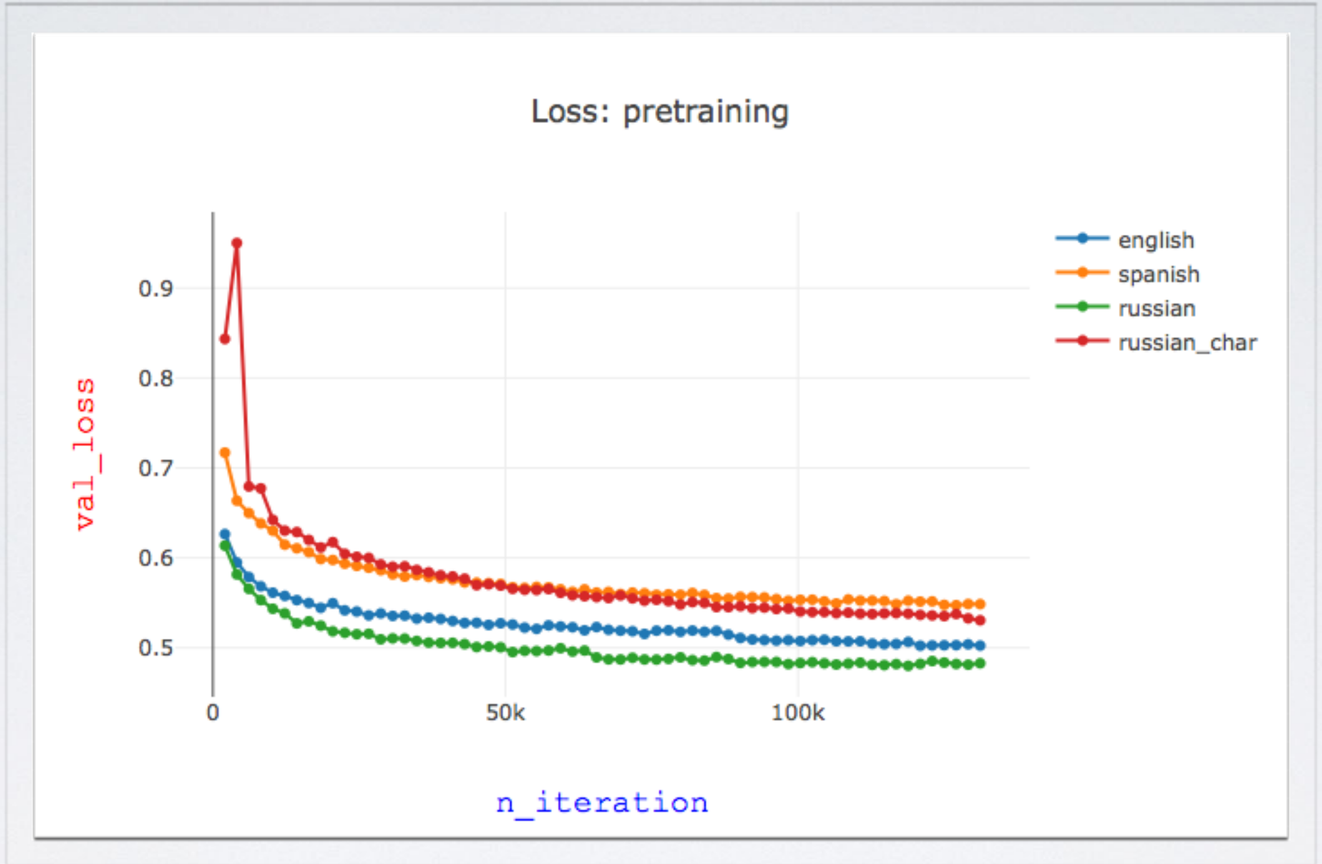
- 2-4 M обучаемых параметров в сети
- Имеющихся в train_set 100 000 сэмплов недостаточно для обучения
- Нужно больше данных!

РЕШЕНИЕ:
TRANSFER LEARNING

- Предобучение на Open Subtitles (15M реплик)
- Бинарная классификация - **contrastive loss**



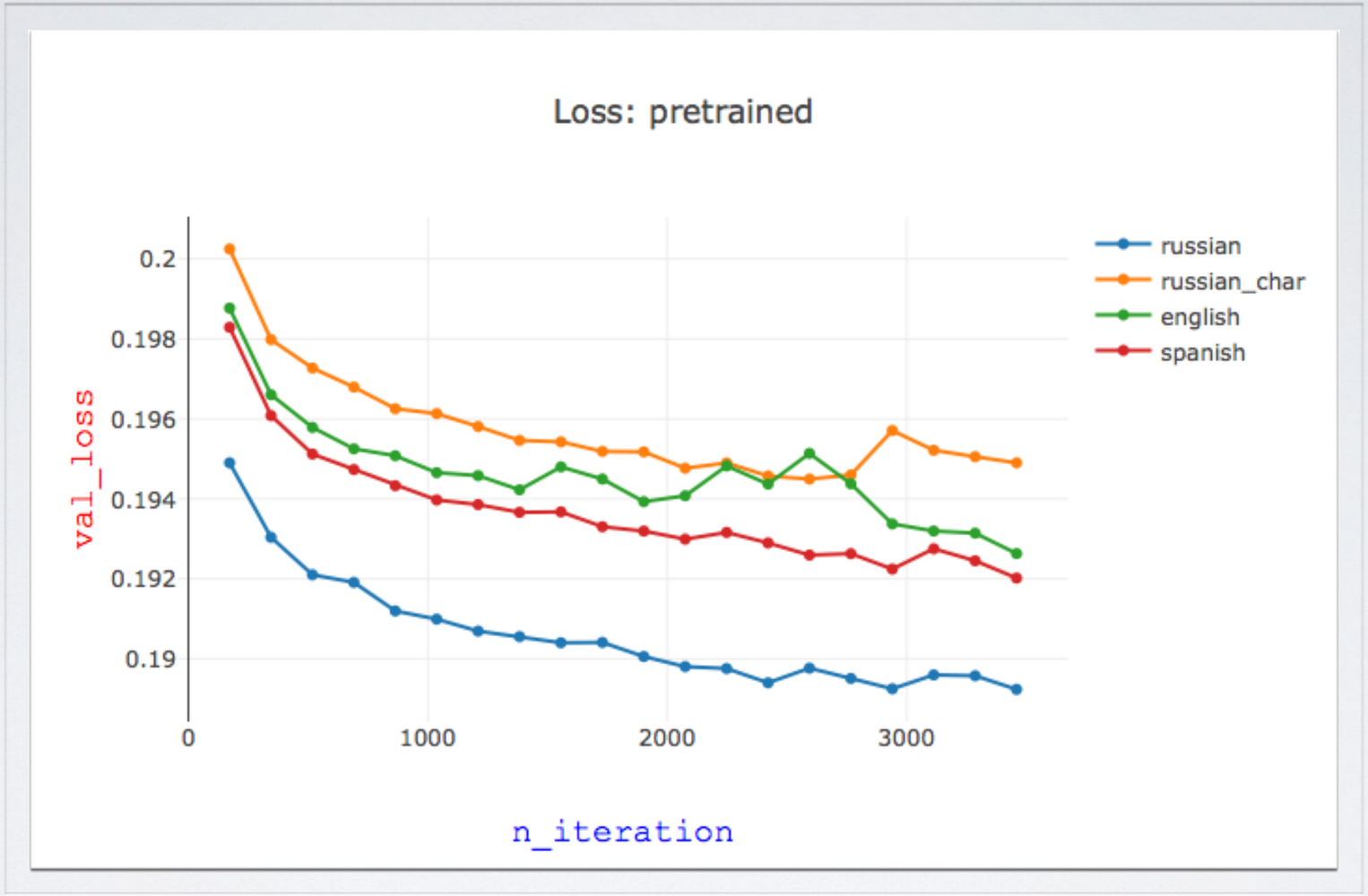
ЭТАП 1: ПРЕДОБУЧЕНИЕ



16 часов на 1080 Ti - 0.9 AUC

ЭТАП 2: ТЮНИНГ

- Замораживаем все слои кроме последних трех (MLP-голова)
- Дообучаем 16 эпох на оригинальном датасете с 10-Fold CV
- Соло-модель: **87000** на public LB


Loss: pretrained
val_loss
0.2
0.198
0.196
0.194
0.192
0.19
0
1000
2000
3000
n_iteration
russian
russian_char
english
spanish

АУГМЕНТАЦИЯ

- Перевод train/test на английский и испанский язык при помощи Google Translate
- Бесплатно без регистрации с помощью библиотеки googletrans
- Перевод всего датасета занял 12 часов
- Это позволило использовать больше предобученных word2vec моделей и добавить их в ансамбль

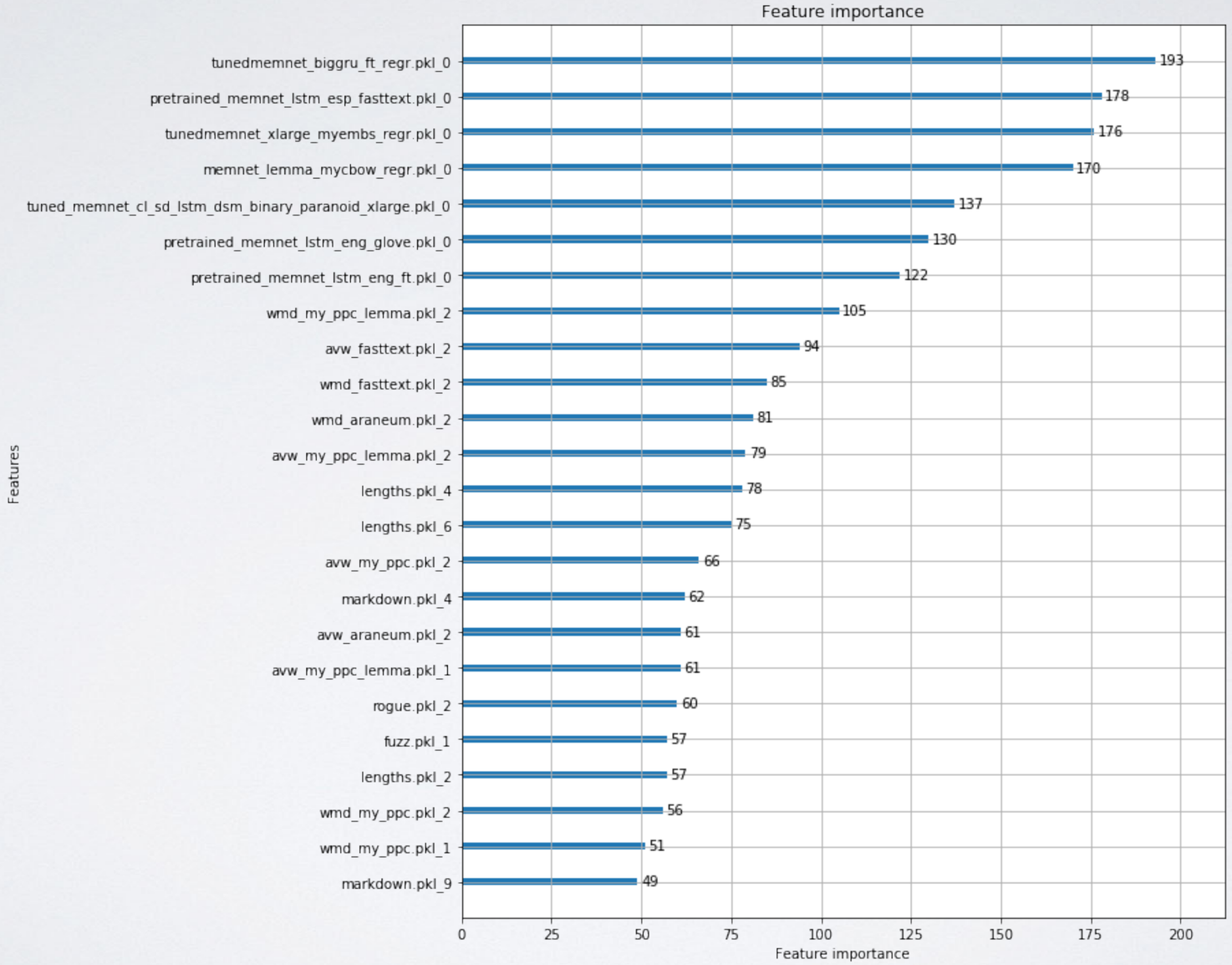
WORD EMBEDDINGS

- Русские
  - FastText
  - OPUS gensim CBOW
  - OPUS (mystemmed) CBOW

- Английские
  - FastText

- Испанские
  - FastText
  - OPUS CBOW

FEATURE IMPORTANCE



Спасибо за внимание!