Исследование устойчивости к орфографическим ошибкам в задачах классификации текстов и распознавания именованных сущностей

Лялин Владислав, МФТИ, 374гр
Научный руководитель: к.ф.м.н. Бурцев Михаил

Актуальность

- В мире существует множество содержащих опечатки текстов

- Существующие решения по проверке орфографии не идеальны
  (для русского языка исправляется ~85% ошибок)

- Устойчивость существующих моделей к шуму слабо исследована

- Нет общего метода, позволяющего оценить устойчивость модели к шуму

Основные результаты

- Разработан метод исследования устойчивости моделей (классификации/распознавания сущностей/…) к шуму в текстах

- Исследована устойчивость к шуму существующих моделей классификации текстов и предложены расширения, устойчивые к шуму

- Исследована устойчивость к шуму существующих моделей распознавания именованных сущностей и предложены расширения, устойчивые к шуму

План

- Моделирование орфографических ошибок
- Векторные представления слов
- Задача классификации
- Задача распознавания именованных сущностей

Шум

- Назовём текстовым шумом опечатки и орфографические ошибки
- Для определения величины зашумлённости слова берется расстояние Дамерау-Левенштейна (с перестановкой букв) от шумной словоформы до исходной
- Под исходной словоформой понимается грамматически и орфографически корректная словоформа в данном контексте
- Под шумной словоформой понимается, соответственно, любая отличающаяся от исходной словоформы

Моделирование шума

- Вероятность шума в реальных текста по некоторым оценкам достигает 10%
- Модель шума выбрана по аналогии с моделями, используемыми в литературе по моделям исправления орфографии
- Для исследуемых языков не существует открытых корпусов текстов с исправленными опечатками

Моделирование шума

Виды шума:

- удаление текущего символа с некоторой вероятностью B(1,p)
- добавление произвольного символа U{1,|A|} после текущего с некоторой вероятностью B(1,p)
- замена текущего символа на произвольный U{1,|A|} с некоторой вероятностью B(1,p)

B(1,p) - биномиальное распределение,

U{1,|A|} - равномерное распределение,

|A| - длина алфавита

План

- Моделирование орфографических ошибок (шума)
- Векторные представления слов
- Задача классификации
- Задача распознавания именованных сущностей

Векторные представления слов

- Слова не могут быть восприняты компьютером, как человеком. Требуются числовые представления.

- Простые представления слов на основе нумерации по словарю являются недостаточными, так как не учитывают многих важных аспектов, например, семантики. Требуются более сложные векторные представления. Примером может служить модель Word2Vec.

Преимущества нейросетевого представления слов

- Легко обучать
- Можно предсказывать дискретные аспекты контекста, такие как POS-тэги.
  Это также возможно делать методами подсчёта вероятностей, но высокая разреженность быстро становится проблемой
- Можно использовать дополнительную контекстную информацию

Векторные представления слов

- Word2Vec - работает с фиксированным написанием слова, не учитывают контекст

- fastText - работает с n-граммным представлением слова в дополнение к фиксированному, не учитывает контекст

- ELMo - работает со свёрточным представлением слова, учитывает контекст

- RoVe - работает с BME представлением слова, учитывает контекст

RoVe: Robust Word Vectors

$$B(w) = c_1 || \ldots || c_{n_b}$$
$$E(w) = c_{k-n_e} || \ldots || c_k$$
$$M(w) = \sum c_i$$

$$RoVe(t) = enc(BME(t), C_{left}, C_{right})$$
$$BME(w) = B(w) || M(w) || E(w)$$

- || - конкатенация векторов

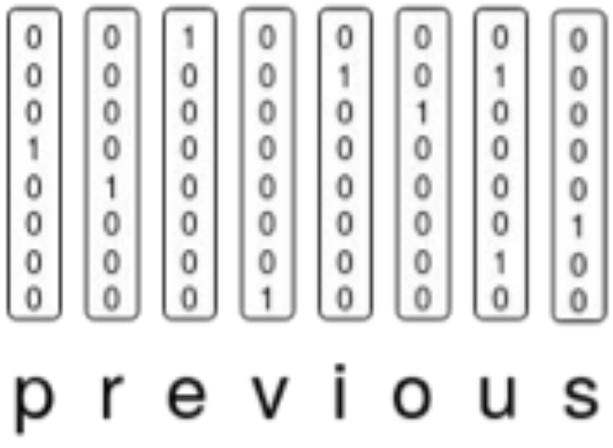
- c1..ck - вектора символов (в слове w), состоящие из нулей и одной 1 по номеру символа в алфавите

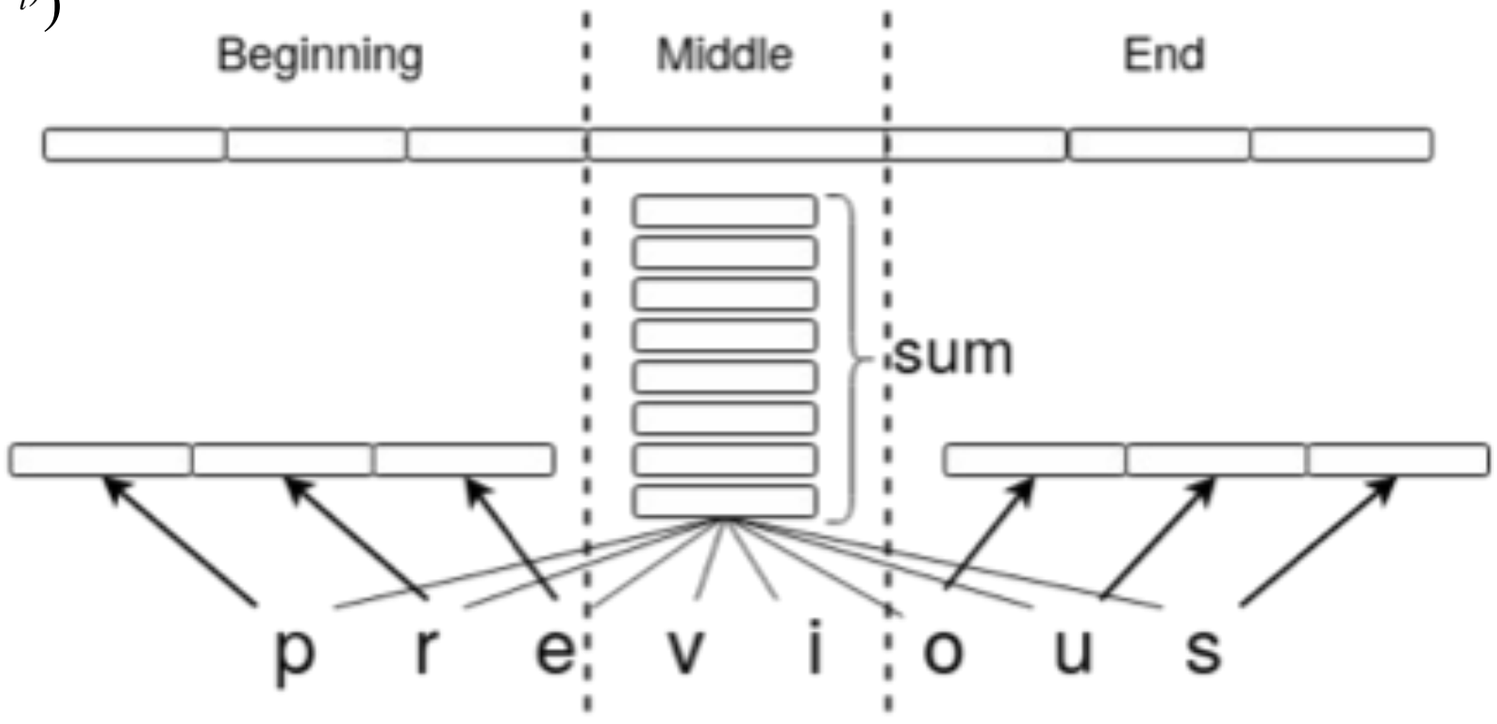
- nb - длина префикса, ne - длина суффикса, выбранные по предварительным экспериментам

- enc - некоторая функция, например, RNN, Cleft и Cright - левый и правый контексты

RoVe: Robust Word Vectors

$$L(x) = log(\sum_{i \in C} e^{-s(x,w_i)}) + log(\sum_{i \notin C} e^{s(x,w_i)})$$





$$B(w) = c_1 || \ldots || c_{n_b}$$
$$E(w) = c_{k-n_e} || \ldots || c_k$$
$$M(w) = \sum c_i$$

$$RoVe(t) = enc(BME(t), C_{left}, C_{right})$$
$$BME(w) = B(w) || M(w) || E(w)$$

План

- Моделирование орфографических ошибок (шума)
- Векторные представления слов
- Задача классификации
- Задача распознавания именованных сущностей

Задача классификации текстов

- На вход модели подается текст, на выходе ожидается метка класса

- Метрика: F1

Корпуса, на которых производилось тестирование:

- Twitter US Airline Sentiment
  14485 твитов, 3 класса

- Movie Review
  50000 отзывов, 2 класса

Постановка экспериментов

Задачи экспериментов:

- Сравнить устойчивость к шуму существующих архитектур для классификации текстов и их расширений

- Показать, что используемый искусственный шум близок к натуральному

Список экспериментов:

- Обучающая и тестовая выборки берутся без изменения.

- Обучающая и тестовая выборки: выполняется проверка орфографии и накладывается искусственный шум.

- Для обучающей выборки выполняется проверка орфографии и накладывается искусственный шум. Тестовая выборка берется без изменений.

Проверяемые модели

- CharCNN - модель, основанная на векторных представлениях символов и сверточной сети для создания скрытых представлений

- CharCNN-WordRNN - модель, основанная на векторных представлениях символов и сверточной сети для создания векторных представлений слов. Векторные представления слов обрабатываются рекуррентной нейронной сетью

- FastText - векторные представления слов порождаются моделью fastText. Векторные представления слов обрабатываются рекуррентной нейронной сетью GRU

- RoVe - векторные представления слов порождаются моделью RoVe, обрабатываются GRU

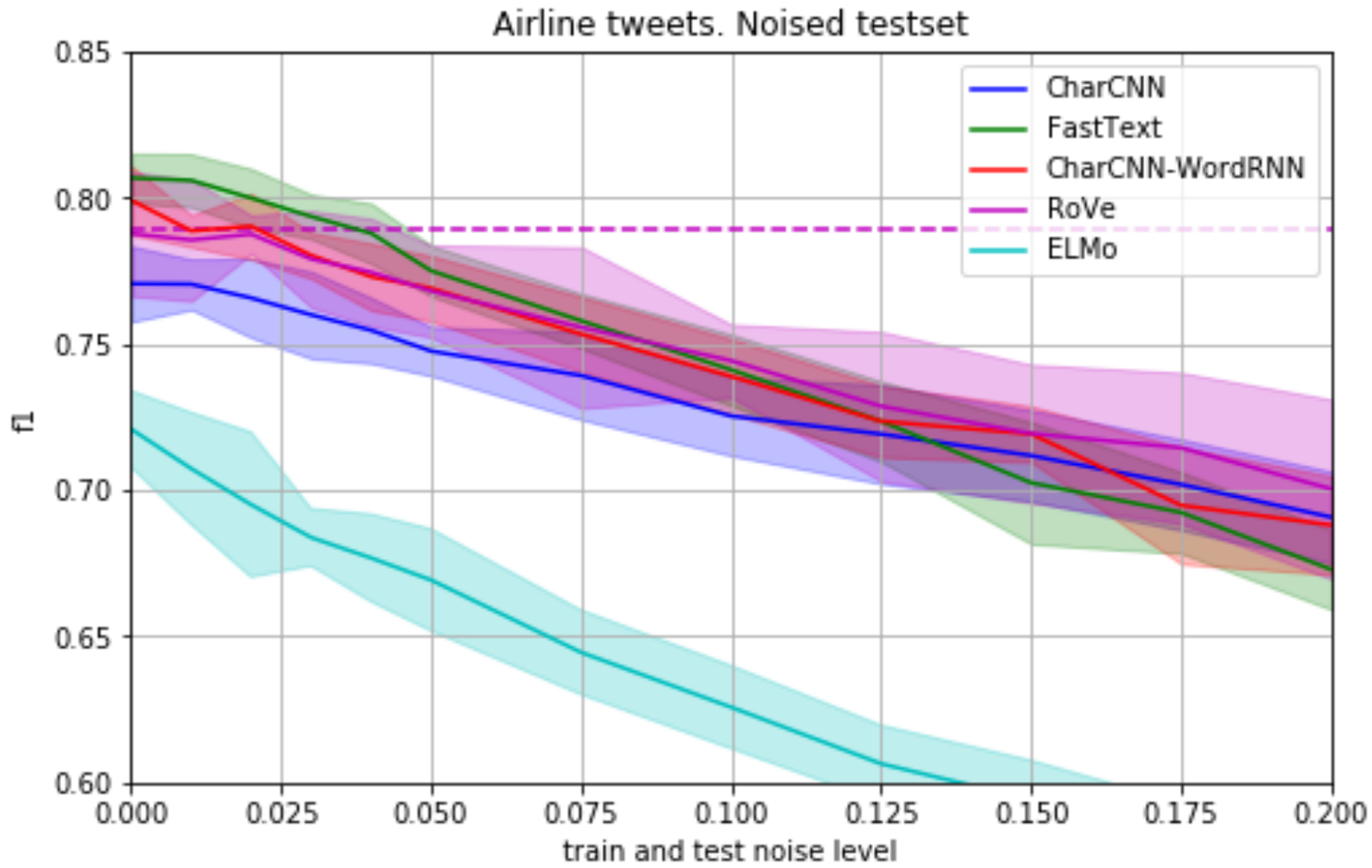
- ELMo - векторные представления слов порождаются моделью ELMo, обрабатываются GRU

Результаты на исходных данных

Модель	Movie Review	Twitter Sentiment
CharCNN	0.74	0.77
FastText	**0.84**	0.76
CharCNN-WordRNN	0.80	**0.81**
RoVe	0.79	0.80

Результаты для Airline Twitter Sentiment (искусственный шум в тестовой выборке)

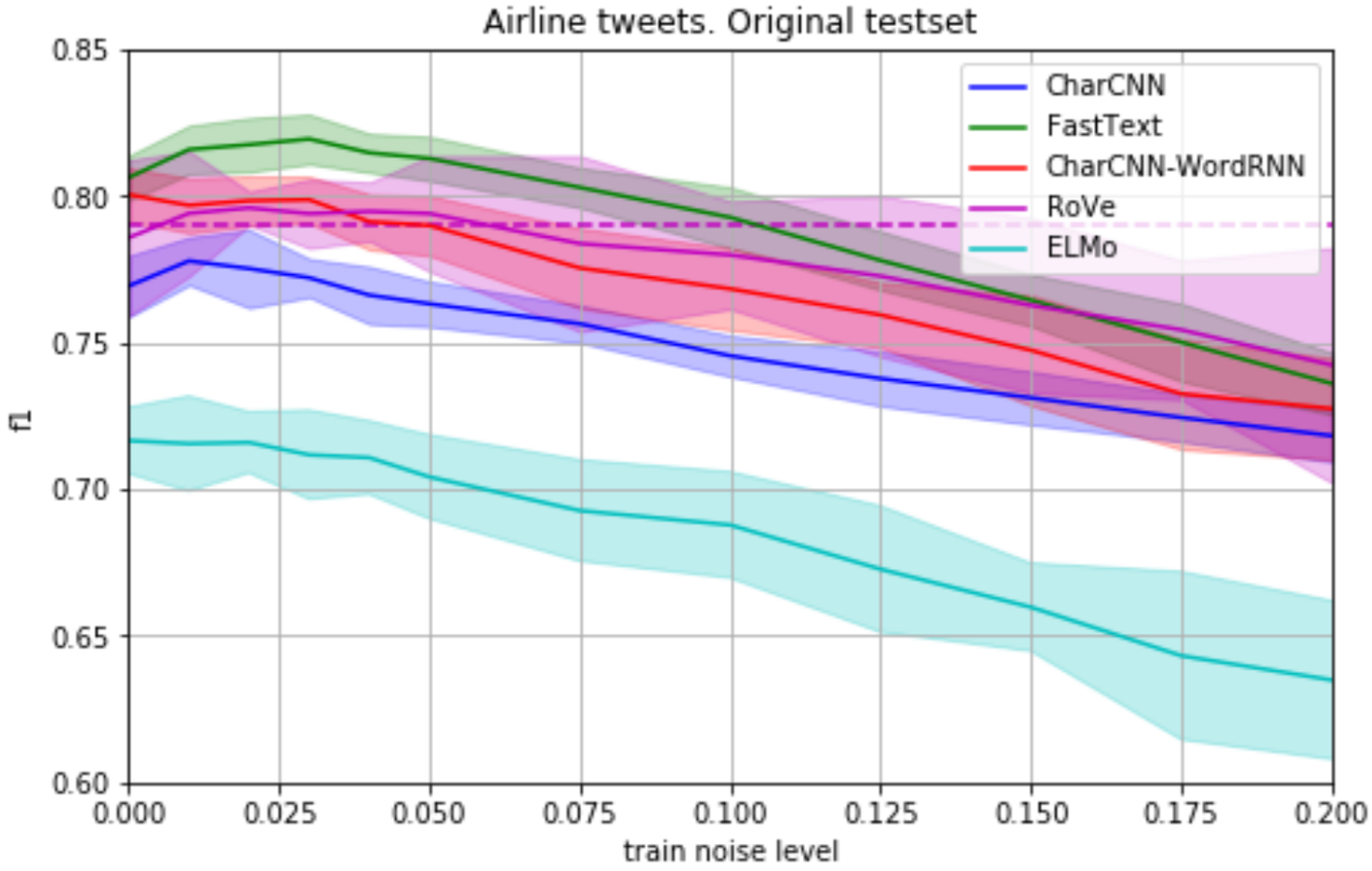
- модель FastText имеет более высокое качество на низких уровнях шума
- модель RoVe имеет более высокое качество на высоких уровнях шума



99% доверительные интервалы

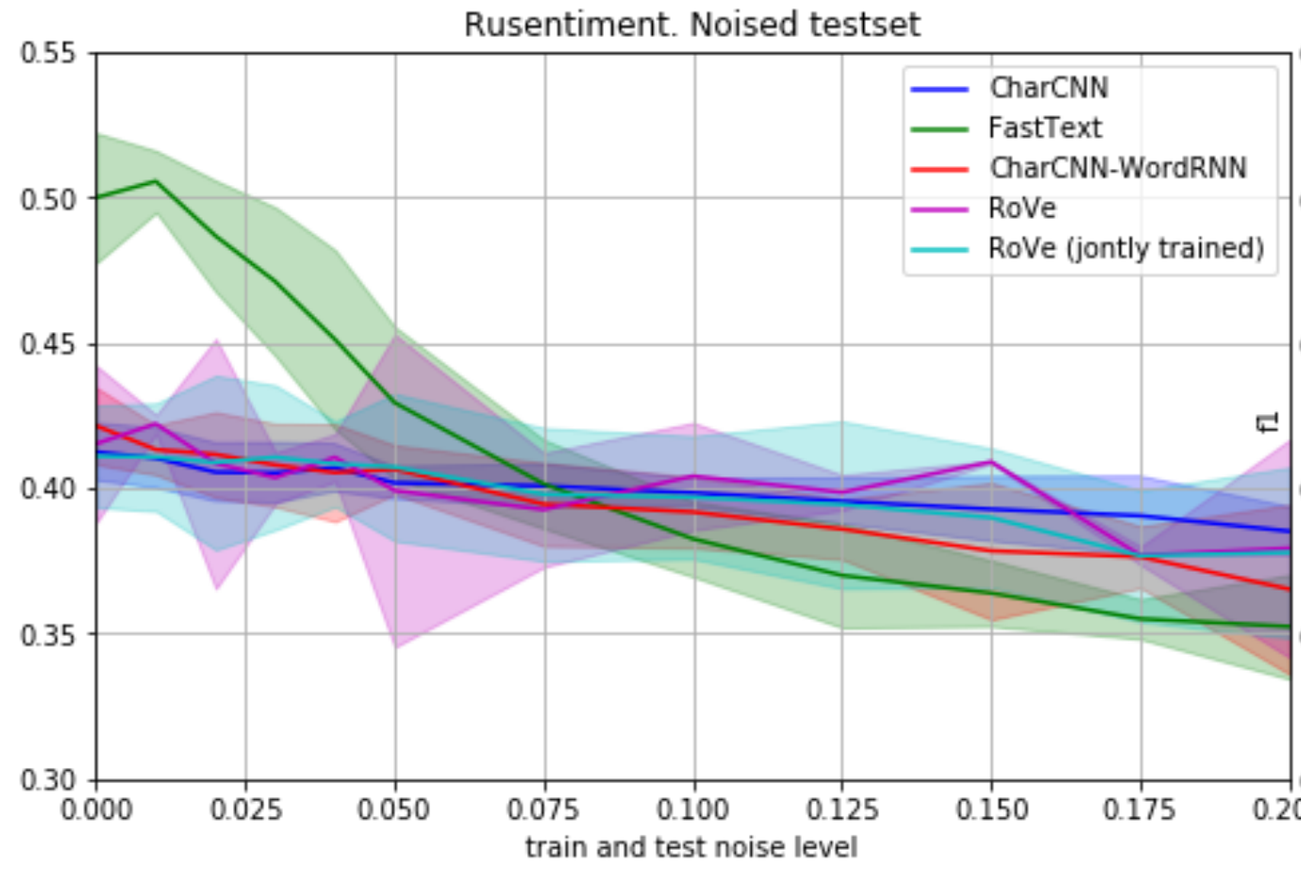
Результаты для Airline Twitter Sentiment (естественный шум в тестовой выборке)

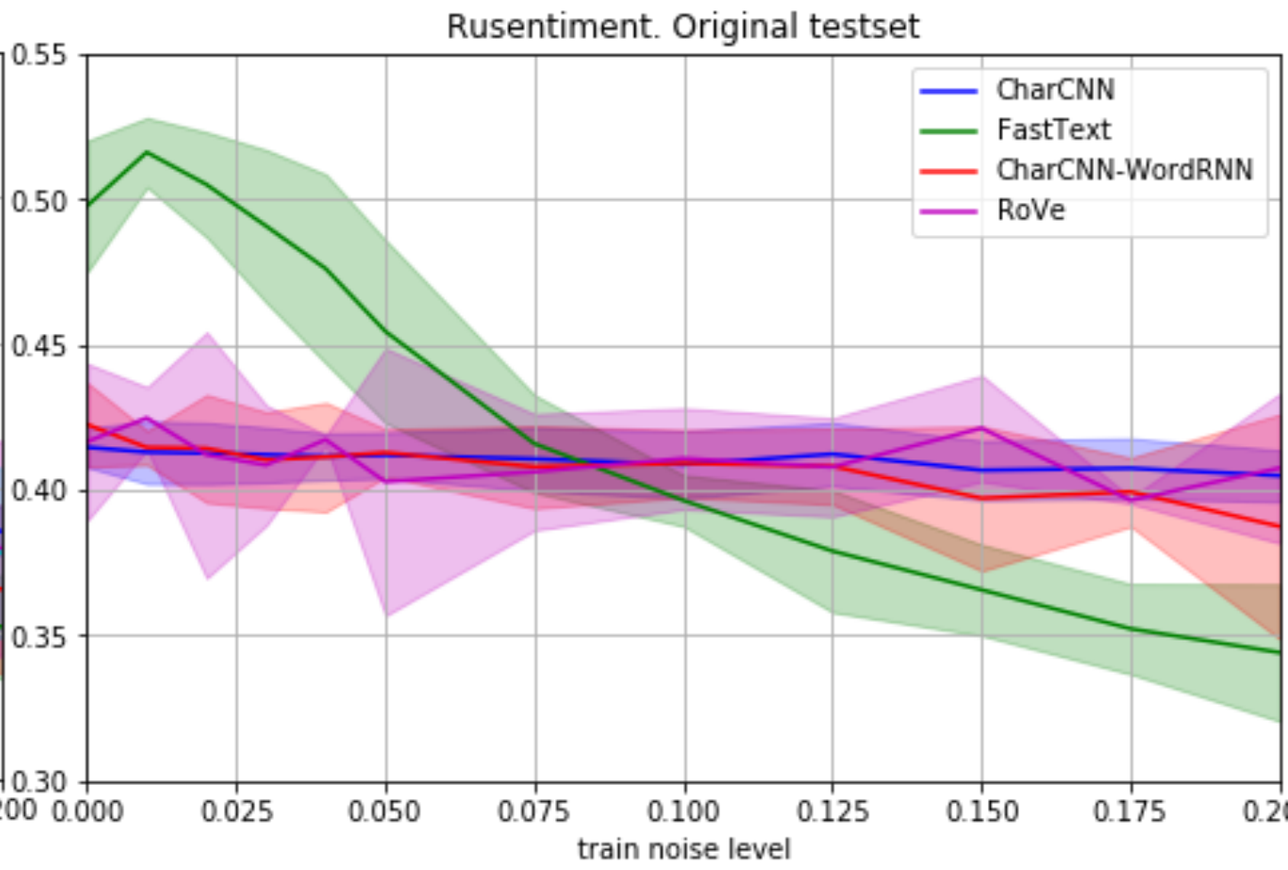
- модель FastText имеет более высокое качество на низких уровнях шума
- модель RoVe имеет сопоставимое или более высокое качество на высоких уровнях шума
- качество тестируемых систем уменьшается аналогичным предыдущему эксперименту образом



99% доверительные интервалы

Результаты для RuSentiment





План

- Моделирование орфографических ошибок (шума)
- Векторные представления слов
- Задача классификации
- Задача распознавания именованных сущностей

Задача распознавания именованных сущностей

На вход модели подается текст, на выходе для каждого слова ожидается тег. Качество подсчитывается по сущностям (то есть последовательностям тегов)

Input: Vancouver is a coastal seaport city on the mainland of British Columbia. The city's mayor is Gregor Robertson.

Location

Output: ***Vancouver*** is a coastal seaport city on the mainland of ***British Columbia***. The city's mayor is ***Gregor Robertson***.

Location

Person

Задача распознавания именованных сущностей

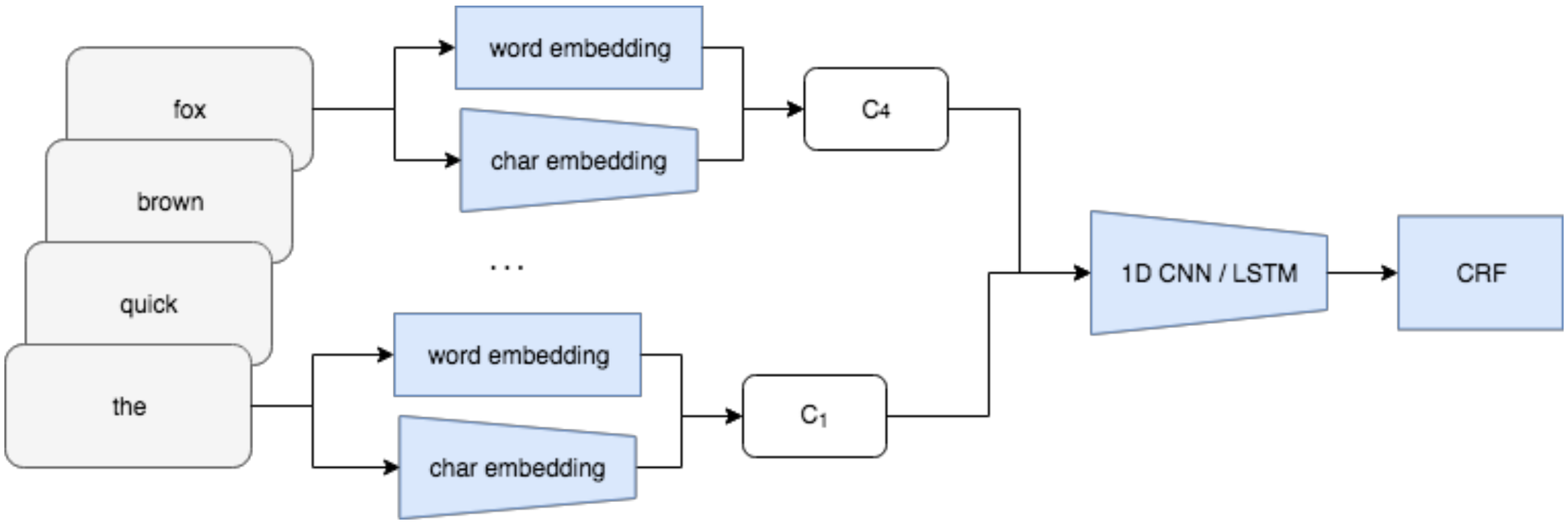
Корпуса, на которых производилось тестирование:

- Collection 5 (доразмеченный Persons-1000)
  1000 документов, размечен 5 тегами LOC, PER, ORG, MEDIA, GEOPOLIT

- CoNLL’03 (английская часть)
  946+216+241 документ, размечен тегами LOC, PER, ORG

- CAp’2017 (французский твиттер)
  3000+3645 твитов, размечен 13 тегами

Используемая архитектура



применяется для английского и русского языков, впервые применена автором для французского языка

Варианты модели

Комбинация векторных представлений слов (word embeddings) и символов (char embeddings).

Векторные представления слов:

- Word2Vec - инициализация векторами из модели Word2Vec для матрицы векторных представлений слов
- fastText - аналогично для моедли fastText
- EmbedMatrix - матрица векторных представлений выучивается в процессе обучения
- RandomEmbed - матрица векторных представлений задается случайно

Векторные представления слов на основе букв:

- nochar - без добавления признаков от побуквенного представления слова;
- CNN - сверточная сеть на уровне символов

Постановка экспериментов

Задача экспериментов:

- Проверить устойчивость к шуму существующей лучшей архитектуры для распознавания именованных и ее расширений;
- Показать, что используемый искусственный шум близок к натуральному.

Список экспериментов:

- Обучающая и тестовая выборки берутся без изменения.
- Обучающая и тестовая выборки: выполняется проверка орфографии и накладывается искусственный шум.
- Для обучающей выборки выполняется проверка орфографии и накладывается искусственный шум. Тестовая выборка берется без изменений.

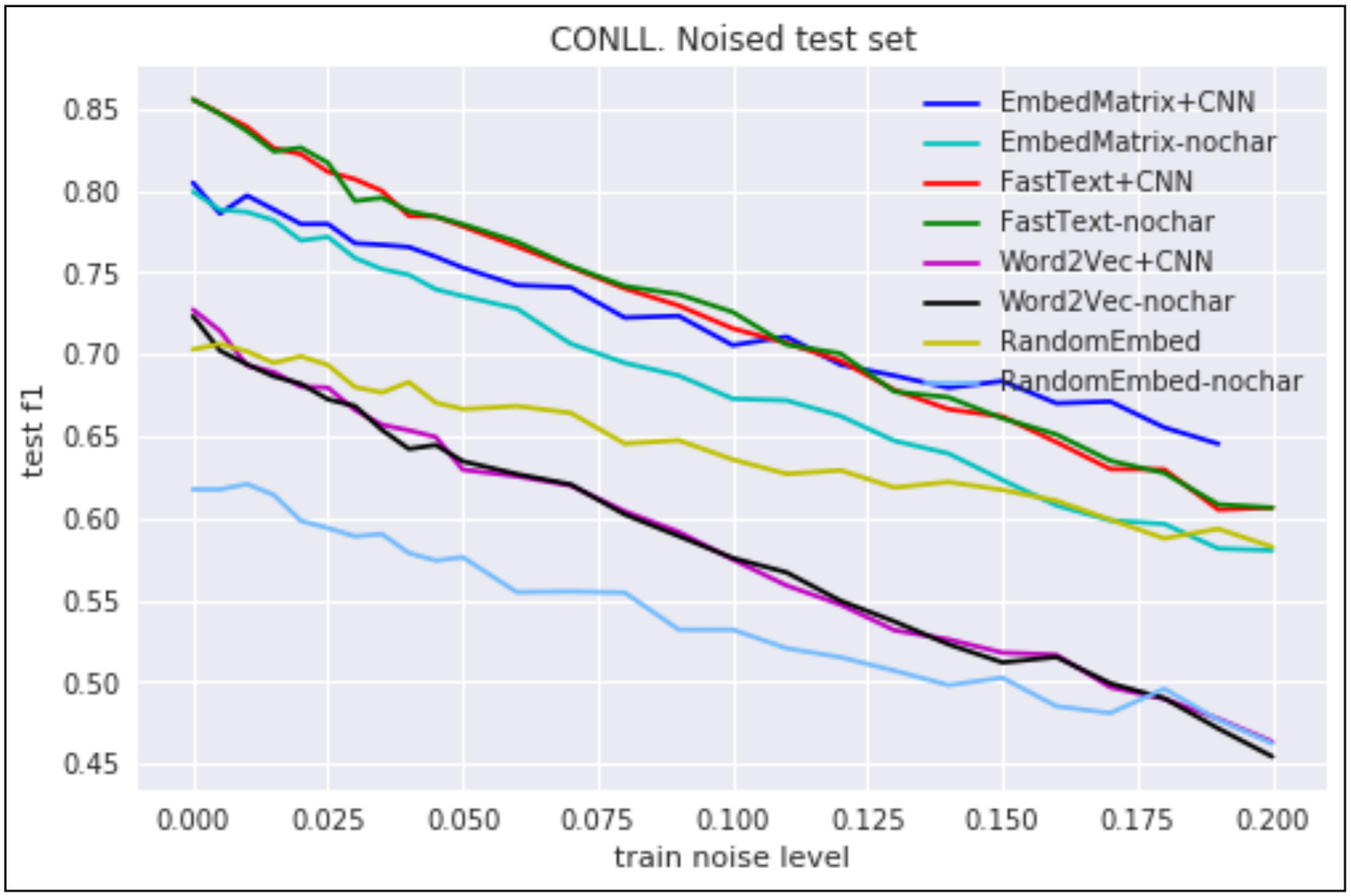
Результаты на исходных данных

Модель	CoNLL'03	Persons-1000	CAp'2017
EmbedMatrix+CNN	0.81	0.85	0.43
EmdebMatrix-nochar	0.80	0.81	0.44
RandomEmbed+CNN	0.69	0.77	0.31
RandomEmbed-nochar	0.61	0.48	0.22
FastText+CNN	0.86	0.69	0.41
FastText-nochar	0.86	0.69	0.41
Word2Vec+CNN	0.73	0.72	н/д
Word2Vec-nochar	0.72	0.72	н/д

Метрика: chunk-wise F1

Результаты для английского языка (искусственный шум в тестовой выборке)

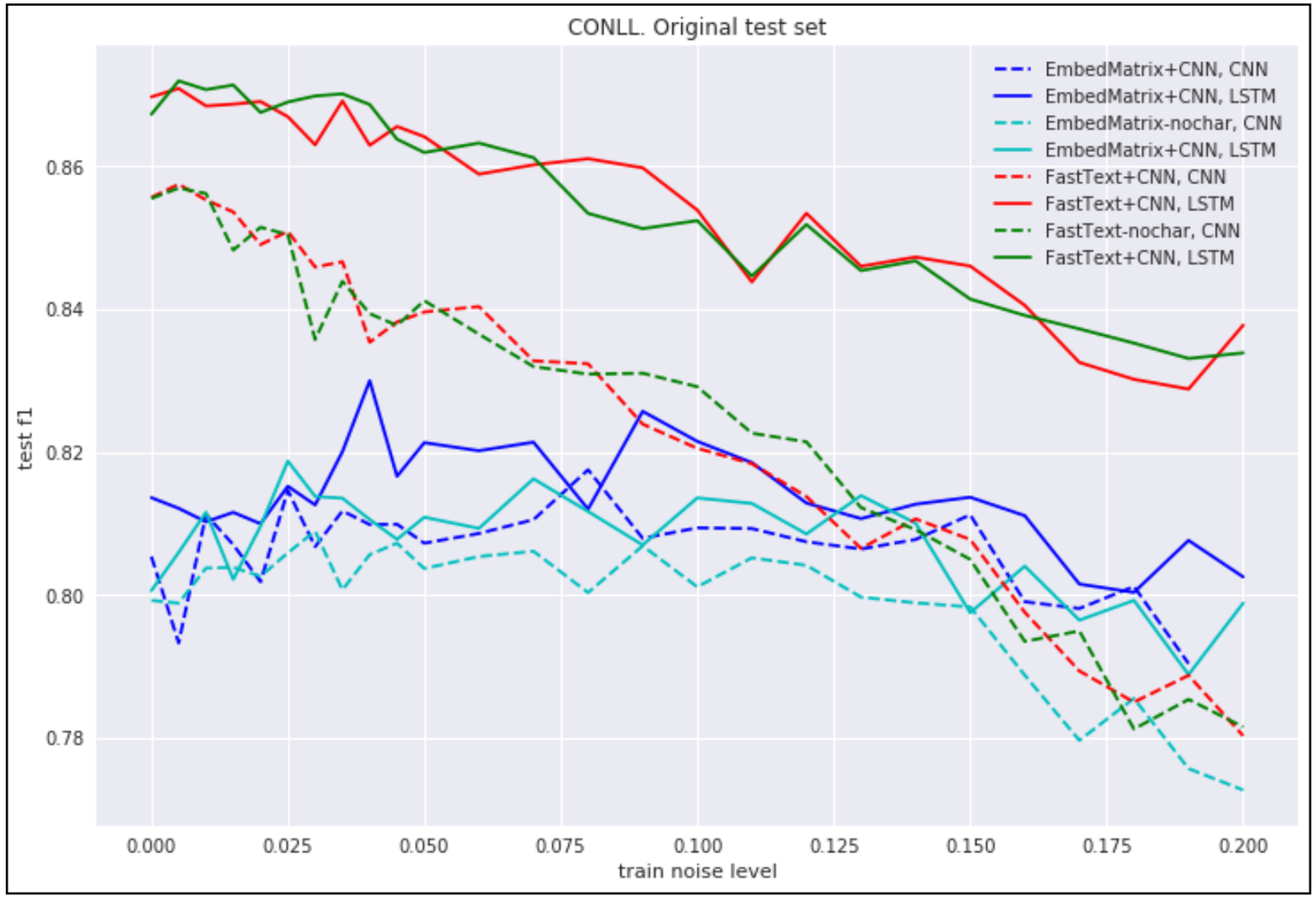
- лучше всего себя проявляют варианты модели FastText, за исключением высоких уровней шума, где лучшего всего себя показывает EmbedMatrix+CNN



Результаты для английского языка (естественный шум в тестовой выборке)

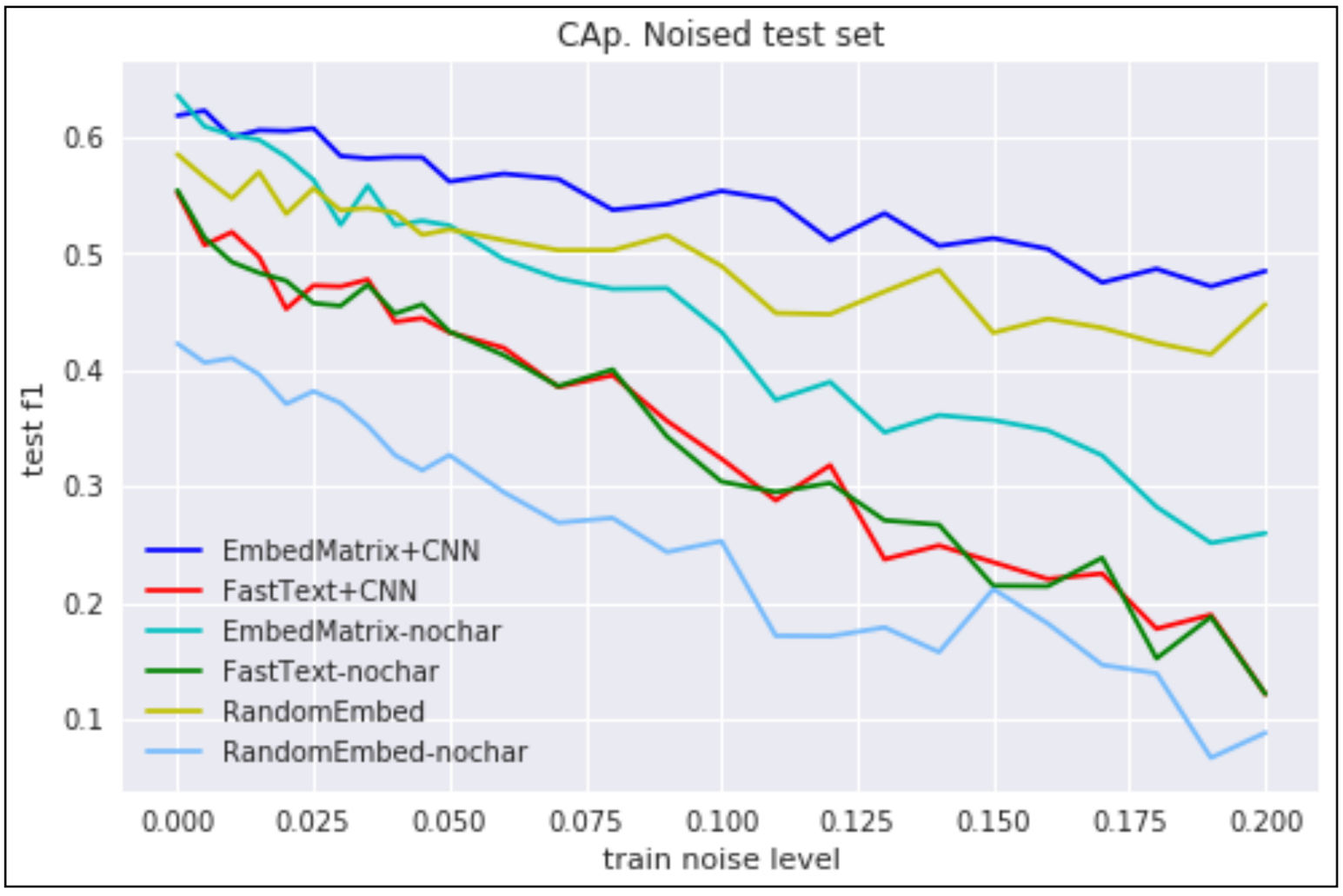
- результаты аналогичны предыдущему эксперименту для FastText и EmbedMatrix+CNN

- приведенные модели с использованием модуля обработки контекста LSTM показывают лучший результат для всех протестированных моделей



Результаты для французского языка (искусственный шум в тестовой выборке)

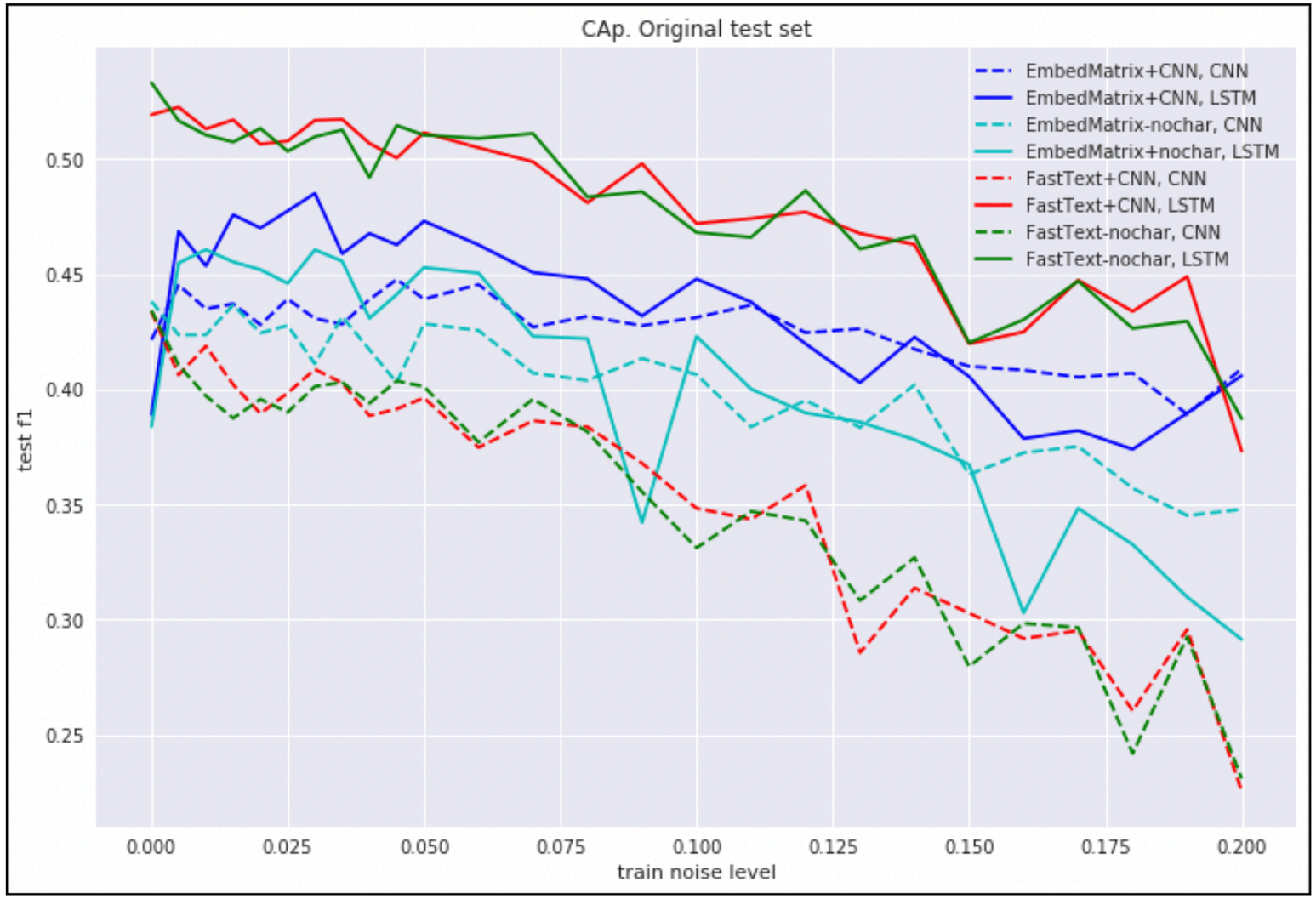
- лучше всего себя проявляют варианты модели FastText, за исключением высоких уровней шума, где лучшего всего себя показывает EmbedMatrix+CNN



Результаты для французского языка (естественный шум в тестовой выборке)

- результаты аналогичны предыдущему эксперименту для FastText и EmbedMatrix+CNN

- приведенные модели с использованием модуля обработки контекста LSTM показывают лучший результат для всех протестированных моделей



State of the art-результат на французском языке

Model	orig.	sp.-ch.
EmbedMatrix+CNN, CNN	0.42	0.63
EmbedMatrix-nochar, CNN	0.44	0.64
EmbedMatrix+CNN, LSTM	0.39	0.59
EmbedMatrix-nochar, LSTM	0.38	0.59
FastText+CNN, LSTM	0.52	0.67
FastText-nochar, LSTM	0.53	**0.69**

Предыдущее SOTA-решение имело F1 = 0.59

Публикации

- Valentin Malykh и Vladislav Lyalin. — «On Classification of Noisy Texts». — В: Записки научных семинаров ПОМИ. Серия “искусственный интеллект”. — 2018.

- Valentin Malykh и Vladislav Lyalin. — «Named Entity Recognition in Noisy Domains». — International Conference on Artificial Intelligence Applications and Innovations. — 2018.

Литература

- Sorokin, A., 2017. Spelling correction for morphologically rich language: a case study of Russian. In Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing (pp. 45-53).

- Cucerzan, S. and Brill, E., 2004. Spelling correction as an iterative process that exploits the collective knowledge of web users. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing

- Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems 2013 (pp. 3111-3119)

- Ю. В. Рубцова. Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы, 2015, №1(109), –С.72-78

- Kim Y. Convolutional neural networks for sentence classification. arXiv preprint arXiv: 1408.5882. 2014 Aug 25

- Peters E. et al. Deep contextualized word representations, Proc. of NAACL, 2018