Детектирование смены автора в тексте

А.Фаттахов,* Р.Кузнецова,* and В.Стрижов*

Московский Физико-Тенический Институт

E-mail: fattahov.ao@phystech.edu; rita.kuznetsova@phystech.edu; strijov@phystech.edu

Аннотация

Существует два различных подхода для задачи поиска заимствований: "внешний"и "внутренний". При "внешнем"необходимо найти документ, из которого произошло заимствование. В данной работе рассматриватеся "внутренний"подход, при котором нет доступа к внешнему корпусу, по которому ведется проверка и необходимо основываться исключительно на анализе самого текста. При этом предполагается, что авторы имеют свой уникальный стиль написания, и смена стиля может быть сщуественным сигналом к подозрению на заимствование. В работе рассматривается алгоритм на основе GAN, генерирующих предложения в определенном стиле. Результаты работы алгоритма сравниваются с двумя базовыми алгоритмами: на основе текстовых статистик и на основе рекуррентной сети LSTM с использованием векторных представлений слов. Вычислительный эксперимент проводится на выборке с соревнования PAN-2018.

Введение

С развитием сети Интернет задача поиска заимствований становится все более сложной и актуальной. Если "внешний"анализ заимствований(external plagiarism detection) подразумевает попарное сравнение подозрительного текста с определенным набором внешних текстов, то задача поиска "внутренних"заимствований(internal plagiarism detection)

состоит в анализе исключительно подозрительного текста. При этом алгоритмы анализа должны учитывать стиль письма и выявлять признаки в тексте, свойственные данному автору. В то время как многие подходы нацелены на выявление автора целого текста, специфика данной задачи заключается в работе с мульти-авторскими текстами.

В PAN2011 ставилась схожая, но более сложная задача - поиск участков в тексте, на которых происходила смена автора. В PAN2016 задачу расширили на авторскую диаризацию(тут англоязычный термин), решение победителя[1]. PAN2017 Так, данная задача сейчас, в основном, решается с помощью использования различных текстовых статистик. Например, в [2] предложен статистический подход к решению задачи, основанный на tf-idf признаках, которые характеризуют документ с различных точек зрения: словесные n-граммы(в работе рассматриватеся только n=1 и n=3), пунктуации, части речи с использованием PST Tagbank Penn Treebank. Решение с использованием подобных признаков, но для обучения модели в unsupervised манере описывается в [3]. Данный подход особенно интересен, потому что он может быть полезным в нахождении фрагментов, подозрительных на плагиат в том случае, когда нет доступа к внешним источникам данных. Кроме того, в [3] применяется комбинация лексических, синтаксических признаков, которые зависят от содержания, но не обязательно от стиля автора. Этот подход может быть очень удобен в тех случаях, когда необходимо связать одно предложение с его соседними предложениями и таким образом определить уточнение границ перехода в пределах данного документа. В статье [4] описан алгоритм перевода предложений в многомерное пространство, с последующим определением стилевой функции. Помимо статистических подходов были попытки рассматривать задачу с прменением многообразий [5] и нейронных сетей [6]

В PAN2018 ставится новая задача - нужно определить, написан текст одним автором или несколькими. Решение подразумевает использование только стилевых особенностей, игнорируя сам смысл текста.

В данной работе предлагается использовать алгоритм на основе GAN [7], дескрими-

натор которой определяет не только, реальный текст или сгенерированный, но ещё и мультиавторный или нет. Результаты работы алгоритма сравниваются с двумя базовыми алгоритмами: : на основе текстовых статистик и на основе рекуррентной сети LSTM с использованием векторных представлений слов. Работа алгоритма тестируется на данных с конкурса PAN2018. Для оценки используются Accuracy.

Постановка задачи

Поставим формально задачу классификации текста. Дана выборка $D = \{(d_k; y_k)\}$, где d_k — текстовый документ, а $y_k \in Y = \{0, 1\}$ - является ли текст мультиавторным. Выборка разбита на обучающую D_l и контрольную D_k. Принята функция ошибки S. Требуется построить отображение $g : d_i \mapsto \mathbf{d_i}$ и модель классификации $f : R^n \mapsto Y$, минимизирующую функцию ошибки S на контрольной выборке:

$$f = \frac{1}{|D_k|} \sum_{1}^{|D_k|} \arg\min_{f} S(f(\mathbf{d_i}), y_i | D_l) \tag{1}$$

Описание базового алгоритма

Алгоритм Сафина

Предлагаемый алгоритм [4] работает с частотными признаками, предоставляющими описание текста. В качестве такого признака выбран признак частоты встречаемости слов. Исходный текст подвергается предобработке: удаляются служебные символы, все буквы переводятся в нижний регистр. Также из текста удаляются стоп-слова. Текст разбивается на предложения. Затем формируется разбиение текста на сегменты s_i: если длина очередного предложения меньше минимальной длины сегмента, к этому предложению добавляется следующее за ним — процесс повторяется, пока длина сегмента s_i не превысит заданную минимальную длину. Минимальная длина сегмента является на-

страиваемым параметром алгоритма. Для каждого сегмента s_i текста строится вектор признаков. Затем строится статистика $\sigma(s_i)$, которая потом сглаживается скользящим средним:

$$\sigma'(s_i) = \frac{1}{2n+1} \sum_{\text{k = i - n}}^{i+n} \sigma(s_i) \tag{2}$$

Далее значения $\sigma'(s_i)$ исследуются на выбросы. С некоторого порога сегментов-выбросов текст классифицируется как мультиавторский.

Алгоритм на основе LSTM

В качестве альтернативного базового алгоритма использовалась рекуррентная нейронная сеть на основе векторных представлений слов. Пусть t_{ik} - слово текста. В качестве отображения $e : t_i \mapsto \mathbf{t_i}$ возьмем векторные представления слов, полученные с помощью модели word2vec [8]. Тогда $h_i = LSTM(h_{i-1}, \mathbf{t_i})$ - внутреннее состояние сети на этом слове. После двух LSTM слоев конечное состояние передается на трёхслойный перцептрон с сигмоид активацией на последнем слое из одного нейрона.

$$\begin{aligned}
\mathbf{i}_t &= \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i), \\
\mathbf{f}_t &= \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f), \\
\mathbf{o}_t &= \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o), \\
\mathbf{c}_{in} &= \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c), \\
\mathbf{c}_t &= \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \mathbf{c}_{in}, \\
\mathbf{h}_t &= \mathbf{o}_t \circ \tanh(\mathbf{c}_t).
\end{aligned} \tag{3}$$

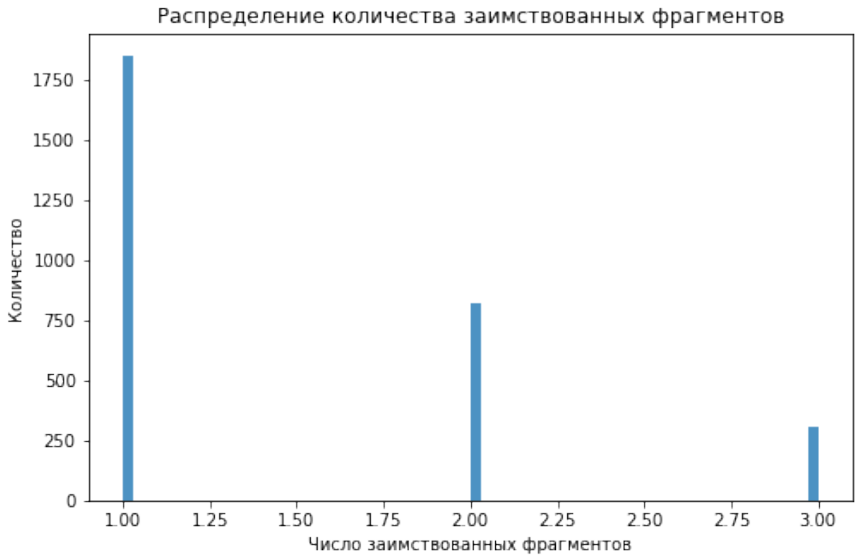
Метрика качеств

Для оценки алгоритма используется точность, которая определяется по следующей формуле:

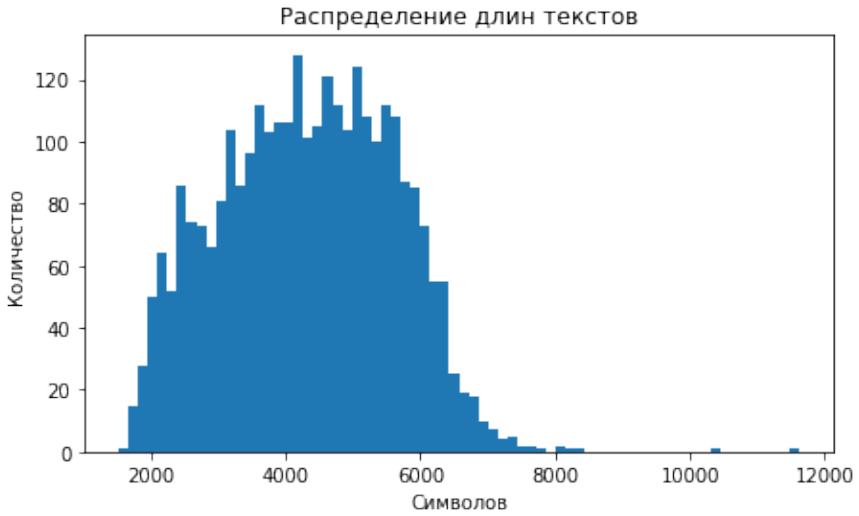
$$\mathsf{A} = \frac{1}{|D_k|} \sum_{\mathrm{k}} I(y_k = f(\mathbf{d_k})) \tag{4}$$

Описание выборки

В работе используется набор текстовых документов с соревнования PAN-2018, собранных с различных сайтов сети StackExchange network. Тематики и авторы у текстов различные. При этом у каждого текста может быть как один автор, так и несколько. Обучающая выборка состоит из 2980 текстов, при этом к каждому из них прилагается файл с разметкой. Несмотря на то, что стоит задача классификации, в разметке присутсвуют сами границы смены автора текста, что может помочь при разработке алгоритмов. Как видно из графиков, количество заимствований линейно убывает, а большинство текстов имеет длину от двух до семи тысяч.



Scheme 1: Гистограмма распределения числа заимствований



Scheme 2: Гистограмма распределения длины текстов

Качество базовых алгоритмов на контрольной выборке

Таблица 1: Качество алгоритмов

	Алгоритм на основе статистик	LSTM
Accuracy	68	64

Список литературы

(1) Kuznetsov, M.; Motrenko, A.; Kuznetsova, R.; Strijov, V. Methods for Intrinsic Plagiarism Detection and Author Diarization—Notebook for PAN at CLEF 2016. CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal. 2016.

(2) Karaś, D.; Śpiewak, M.; Sobecki, P. **2017**,

(3) Khan, J. Style Breach Detection: An Unsupervised Detection Model—Notebook for PAN at CLEF 2017. CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland. 2017.

(4) Safin, K.; Kuznetsova, R. Style Breach Detection with Neural Sentence Embeddings—Notebook for PAN at CLEF 2017. CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland. 2017.

(5) I.Molybog, V., A.Motrenko IMPROVING CLASSIFICATION QUALITY FOR THE TASK OF FINDING INTRINSIC PLAGIARISM. 2017.

(6) Kamil Safin, R. K. Style Breach Detection with Neural Sentence Embeddings. 2017.

(7) Ian J.Goodfellow, M. M. B. X. D. W. S. O. A. C. Y. B., Jean Pouget-Abadie Generative Adversarial Networks. 2014.

(8) Tomas Mikolov, G. C. J. D., Kai Chen Efficient Estimation of Word Representations in Vector Space. 2013.