УДК 004.89 **Информационные технологии**

Батурин Михаил Михайлович, *студент-магистр, Калужский филиал ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана» (национальный исследовательский университет)*

Белов Юрий Сергеевич, *к.ф. -м.н., доцент, Калужский филиал ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана» (национальный исследовательский университет)*

ИСПОЛЬЗОВАНИЕ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ, ДОЛГОЙ КРАТКОСРОЧНОЙ ПАМЯТИ И ОЦЕНОК ВНИМАНИЯ ДЛЯ РАЗЛИЧЕНИЯ АВТОРСТВА ТЕКСТА

Аннотация: Различение авторства — это задача определения того, принадлежат ли два произведения одному и тому же лицу. В этой работе предлагается структура «Согласен - Не согласен» (A2D), новый метод решения задачи распознавания авторства. Это двухэтапная структура, основанная на глубоком обучении, состоящая из сети «Согласен» и «Не согласен». На первом этапе она определяет атрибуты авторства своей сетью «Согласен». Впоследствии, через сеть «Не согласен», фреймворк пытается определить авторство нового набора данных (совершенно не связанного с обучающим набором данных).

Ключевые слова: Различение авторства, глубокое обучение, машинное обучение, сверточные нейронные сети (CNN), долгая краткосрочная память (LSTM), обработка естественного языка (NLP).

Annotation: Authorship discrimination is the task of determining whether two works belong to the same person. This paper proposes an Agree-Disagree (A2D) framework, a new method for solving the problem of authorship recognition. It is a two-stage structure based on deep learning, consisting of a network of Agree and

Disagree. In the first step, it defines authorship attributes with its Agree network. Subsequently, through the Disagree network, the framework tries to determine the authorship of the new dataset (completely unrelated to the training dataset).

Keywords: Authorship discrimination, deep learning, machine learning, convolutional neural networks (CNN), long short term memory (LSTM), natural language processing (NLP).

Введение

A2D не зависит от предшествующих знаний, специфичных для набора данных [3], и может учиться только на атрибутах авторства тренировочного набора данных, чтобы определить, принадлежат ли два разных произведения одному и тому же автору. Фреймворк A2D может успешно выявлять, какой автор стоит за псевдонимом, к примеру, при помощи него можно раскрыть псевдонимы известного американского писателя Вашингтона Ирвинга. Не редко, критичным недостатком процесса разделения авторства является то, что у нас нет никаких предварительных знаний об авторах [1]. Система «Согласен - Не согласен» способна устранить данное ограничение.

Обзор платформы A2D

Структура A2D представляет собой комбинацию двух идентичных сетей: «Согласен» (A) и «Не согласен» (D). Как показано на рисунке 1, сеть A обучается по набору данных идентифицированных авторов (т.е. набору данных, содержащему помеченные фрагменты, созданные многими авторами). Одномерный свёрточный слой сети A обучается обнаруживать характеристики авторства из представленного набора данных. Таким образом, мы формируем свёрточный слой этой сети, чтобы различать авторские характеристики.

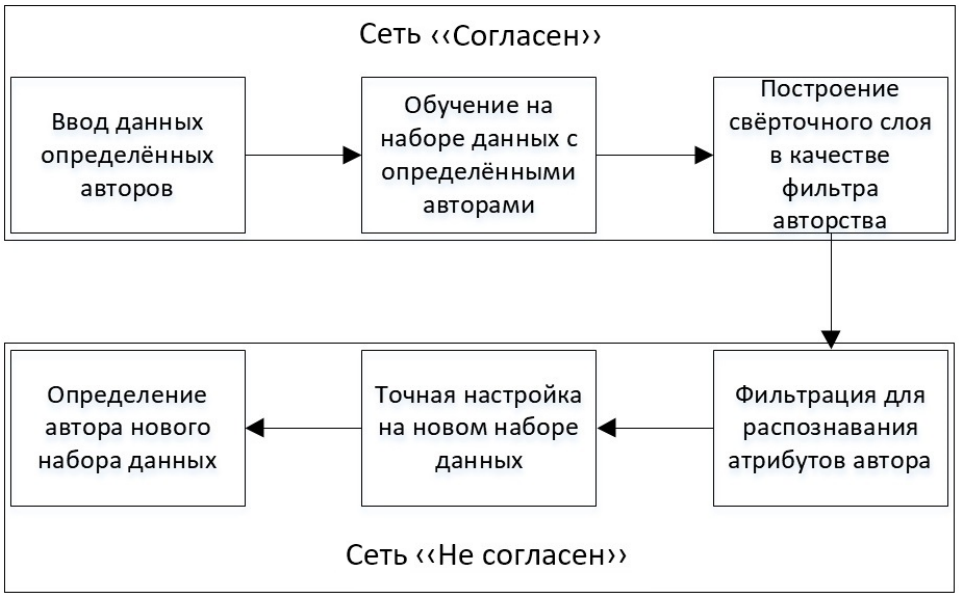

Рисунок 1. Рабочий процесс фреймворка A2D

После того, как сеть «Согласен», как упомянуто выше, обучена, наша структура A2D готова взять на себя задачу различения авторов, то есть по двум (литературным) произведениям она решит, написаны ли они разными авторами. Теперь тот же свёрточный слой сети «Согласен» будет использоваться в сети «Не согласен», работая как «фильтр авторства» для сети «Не согласен». Нам необходимо создать объединенный набор данных, используя два заданных литературных произведения. Затем эта сеть «Не согласен» настраивается на основе этого комбинированного набора данных без обновления параметров свёрточного слоя. Таким образом, свёрточный слой предоставляет данные уровню внимания сети «Не согласен», на основании которого эта сеть принимает решение об авторстве двух литературных частей объединенного набора данных.

Сеть «Согласен»

Для данной пары работа-автор *(x, y)* наша задача в сети «Согласен» (рисунок 2) состоит в том, чтобы предсказать y^A, где y^A является автором данной записи *x* в сети согласования.

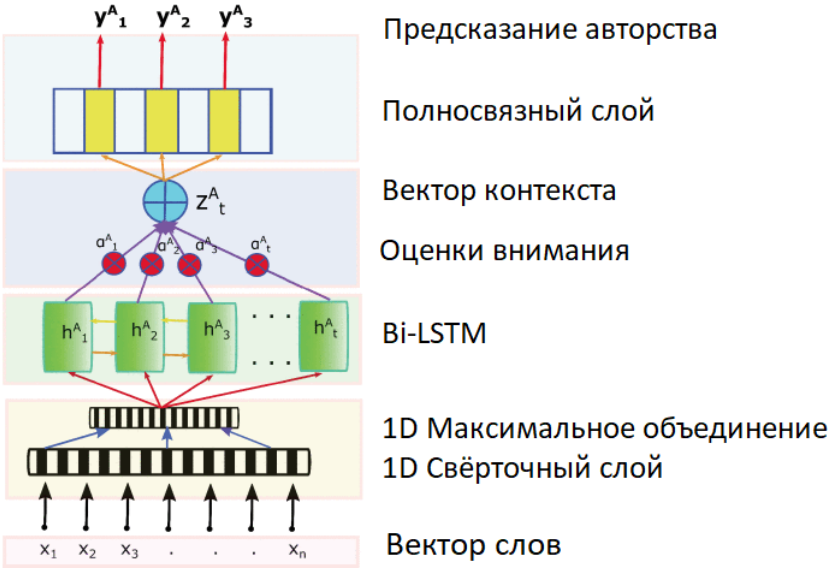

Рисунок 2. Строение сети «Согласен»

Одномерная свёрточная нейронная сеть может извлекать признаки и классифицировать тексты после преобразования слов в корпусе предложений в векторы [5]. Пусть x представлен максимальной длиной n слов. Следовательно,

$$x_{1:n} = x_1 \oplus x_2 \oplus x_3 \oplus \dots \oplus x_n \quad (1)$$

Фильтр размерности w^A применяется к перечню из h слов для создания нового признака c_i^A,

$$c_i^A = fc(w^A * x_{i:i+h-1} + b_C^A) \quad (2)$$

Здесь b_C^A - выражение предвзятости, а fc - нелинейная функция. Этот фильтр применяется к каждому возможному перечню слов в предложении для создания вектора признаков,

$$c^A = [c_1^A, c_2^A, \dots] \quad (3)$$

Операция объединения применяется к карте параметров и принимает максимальное значение параметров, соответствующее конкретному фильтру. Локальные векторы признаков, извлеченные свёрточными слоями, должны быть объединены для получения глобального вектора признаков с фиксированным размером, независимым от длины предложения [4]. Подход с максимальным объединением в k раз вынуждает сеть захватывать наиболее полезные локальные параметры, создаваемые свёрточными слоями.

$$c_{m_i}^A = \max(c_{i:i+k-1}^A) \quad (4)$$

Мы используем двунаправленный LSTM для кодирования данных из стадии максимального объединения. Определяем скрытые состояния как прямого, так и обратного направлений $\vec{h}_t^A$ и $\overleftarrow{h}_t^A$ для заданного временного шага t. Затем мы объединяем их, чтобы получить h_t^A.

$$\vec{h}_t^A = \overrightarrow{f_{LSTM}}(c_{m_t}^A, \vec{h}_{t-1}^A) \quad (5)$$

$$\overleftarrow{h}_t^A = \overleftarrow{f_{LSTM}}(c_{m_t}^A, \overleftarrow{h}_{t-1}^A) \quad (6)$$

$$h_t^A = concat(\vec{h}_t^A, \overleftarrow{h}_t^A) \quad (7)$$

Мы вычисляем оценку внимания (к параметрам) α_t^A по скрытым состояниям и объединяем ее с соответствующим скрытым состоянием, чтобы сгенерировать вектор контекста z_t^A.

$$\alpha_t^A = softmax(W^A h_t^A + b^A) \quad (8)$$

$$z_t^A = \sum_{t=1} \alpha_t^A h_t^A \quad (9)$$

Здесь, $$softmax(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

На последнем этапе сети «Согласен» мы применяем полносвязный слой к этому вектору контекста, чтобы предсказать результат авторства [2].

$$P(y^A|x) = f_{FC}^A(W_{FC}^A z_t^A + b_{FC}^A) \quad (10)$$

Сеть «Не согласен»

Сеть «Не согласен» (Рисунок 3) идентична сети «Согласен» за исключением свёрточного слоя.

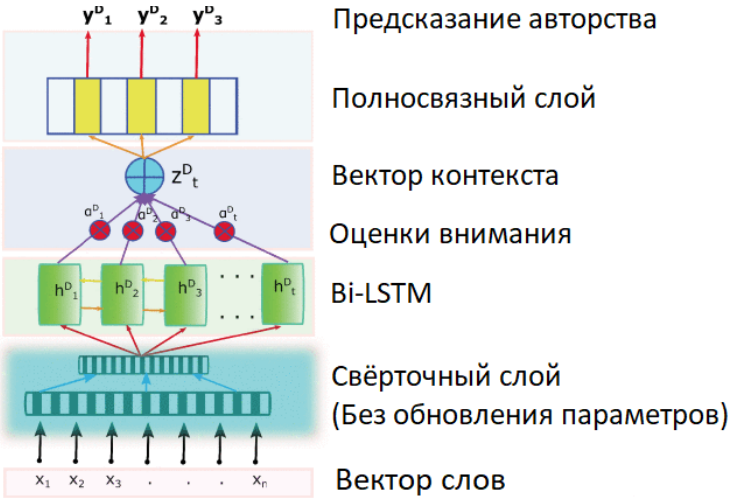

Рисунок 3. Строение сети «Не согласен»

Сеть не обновляет параметры свёрточных слоев. Она применяет те же свёрточные слои (2), (3) и (4) сети «Согласен» к векторам слов экспериментального набора данных для захвата наиболее полезных локальных параметров. Затем сеть «Не согласен» запускает Bi-LSTM для этих параметров и вычисляет оценки внимания для построения вектора контекста. На последнем этапе вычисляется вероятность авторства.

Вывод

В этой статье была представлена схема «согласен - не согласен» (A2D) для задачи распознавания авторства. Данная структура является гибкой и легко расширяемой. Система A2D является уникальной в своем роде, поскольку не зависит от каких-либо особенностей корпуса.

Библиографический список:

1. E. Stamatatos, "Authorship attribution using text distortion", Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics Long Papers, vol. 1, pp. 1138-1149, 2017, [online] Available: https://www.aclweb.org/anthology/E17-1107.
2. H. Ramnial, S. Panchoo and S. Pudaruth, "Authorship attribution using stylometry and machine learning techniques" in Intelligent Systems Technologies and Applications, Springer, pp. 113-125, 2016.
3. P. Juola and G. Mikros, "Cross-linguistic stylometric features: A preliminary investigation", Proc. Int. Conf. Stat. Anal. Textual Data, pp. 1-10, 2016.
4. S. Ruder, P. Ghaffari and J. G. Breslin, "Character-level and multi-channel convolutional neural networks for large-scale authorship attribution", arXiv: 1609.06686, 2016, [online]Available: http://arxiv.org/abs/1609.06686.
5. Y. Sari, M. Stevenson and A. Vlachos, "Topic or style? exploring the most useful features for authorship attribution", Proc. 27th Int. Conf. Comput. Linguistics, pp. 343-353, Aug. 2018, [online] Available: https://www.aclweb.org/anthology/C18-1029.