ПРИМЕНЕНИЕ МНОГОЗАДАЧНОГО ОБУЧЕНИЯ ДЛЯ ОПРЕДЕЛЕНИЯ АВТОРСТВА ТЕКСТА НА ОСНОВЕ МЕХАНИЗМА КОНКУРЕНТНОГО ВНИМАНИЯ

Батурин М.М., Белов Ю.С.

ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана», филиал, Калуга, e-mail: k4dys@yandex.ru

Аннотация

В задачах определения авторства текста ключевую роль играет представление независящего от тематики произведения личного стиля автора. Таким образом, отделение содержания текста от стилистических особенностей письма автора является важной проблемой. Для решения этой проблемы зачастую используются мощные нереалистичные решения, либо вручную определённые параметры стиля текста. В этой статье предлагается применить многозадачное обучение, чтобы отделить тему текста от стиля автора. Цель предложенного подхода состоит в том, чтобы найти отдельные представления стиля и темы текста. Основной задачей является определение авторства текста, дополнительной задачей является аппроксимация темы. Применяемые для получения представлений тем модели обучаются на внешнем корпусе данных. В статье предложены механизмы конкурентного внимания и ограничения разделения-восстановления, при помощи которых двум задачам назначаются разные и конкурирующие между собой внимания, что способствует разделению темы и стиля. По результатам оценок подход, основанный на многозадачном обучении, является многообещающим, особенно при наличии набора данных с множеством пересекающихся тем. Предложенная модель разделяет тему и стиль вероятностным образом и не требует вмешательства человека.

Ключевые слова: определение авторства текста, аппроксимация темы текста, рекуррентные нейронные сети, конкурентное внимание

APPLICATION OF MULTITASK LEARNING TO DETERMINE TEXT AUTHORSHIP BASED ON MECHANISM OF COMPETITIVE ATTENTION

Baturin M.M., Belov Yu.S.

Аннотация

In the tasks of determining the authorship of a text, the presentation of the author's personal style, independent of the subject matter, plays a key role. Thus, separating the content of the text from the stylistic features of the author's writing is an important problem. To solve this problem, powerful unrealistic solutions are often used, or manually defined text style parameters. This article proposes to apply multi-task learning to separate the topic of the text from the style of the author. The goal of the proposed approach is to find separate representations of the style and theme of the text. The main task is to determine the authorship of the text, an additional task is to approximate the topic. The models used to obtain representations of topics are trained on an external data corpus. The article proposes the mechanisms of competitive attention and split-recovery constraints, by which two tasks are assigned different and competing attentions, which contributes to the separation of theme and style. Based on the results of the assessments, the multitasking learning approach is promising, especially for a dataset with many overlapping themes. The proposed model separates theme and style in a probabilistic way and does not require human intervention.

Keywords: text authorship attribution, approximation of the topic of the text, recurrent neural networks, competitive attention

1 Введение

Текст можно рассматривать как сочетание темы и стиля. Тема определяет содержание текста, а стиль отражает особый способ автора манипулировать словами. Основная идея этой статьи состоит в том, чтобы изучить отдельное представление темы и пред- ставление стиля для данного текста. В этом исследовании предлагается многозадачный подход для совместной оптимизации основной задачи – атрибуции авторства и вспомогательной задачи – аппроксимации темы.

В частности, задача аппроксимации темы состоит в том, чтобы создать представление темы для аппроксимации распределения темы текста. Распределение тем определяется независимыми от задачи моделями, которые обучаются на внешнем корпусе текстов [1]. Таким образом, наша структура обеспечивает контроль для разделения стилей тем и не требует человеческого вмешательства для аннотирования данных.

Цель исследования – изучить способы определения авторства и аппроксимации темы текста.

2 Механизмы конкурентного внимания и ограничения разделения-восстановления

Механизмы конкурентного внимания и ограничения разделения-восстановления

Часть разделения стиля темы пред- назначена для создания распределенных представлений темы и стиля соответствен- но. Стилевое представление используется для основной задачи: установления авторства, а тематическое – для вспомогательной задачи: аппроксимации темы.

Для достижения поставленных целей предлагаются две идеи: механизм конкурентного внимания и ограничение разделения-восстановления. Рисунки 1 и 2 иллюстрируют две идеи.

Конкурентное внимание – это расширение механизма внимания, который учится присваивать разные веса разным токенам. Здесь мы используем внимание, чтобы выделить общее текстовое представление, чтобы получить отдельные представления для темы и стиля. Если слову уделяется большое внимание для одной задачи, ему будет отведено низкое внимание для другой задачи. Другими словами, два представления конкурируют за внимание каждого слова.

Ограничение разделения-восстановления вводит затраты на восстановление в дополнение к оптимизации двух вышеупомянутых задач. Ожидается, что представления темы и стиля могут реконструировать представление текста, чтобы сохранить исходное значение.

3 Конкурентное внимание для разделения темы и стиля

Конкурентное внимание
для разделения темы и стиля

Рекуррентная нейронная сеть (RNN) подходит для обработки последовательных данных и для захвата долгосрочных зависимостей. RNN используется в качестве базовой архитектуры в этой работе [2].

1) LSTM на основе внимания

Слова текста преобразуются во вложения слов, которые представляют собой плотные векторы действительных значений. Входной текст может быть представлен в виде матриц

$$W = (w_1, \ldots, w_n) \in \mathrm{R}^{d \times r}$$

где d – размер встраивания, а n – количество токенов в тексте. В качестве основной ячейки памяти мы используем долговременную кратковременную память (LSTM). На временном шаге t LSTM берет скрытое состояние с предыдущего временного шага и встраивание слова с текущего шага в качестве входных данных и создает новое скрытое состояние [3].

$$h_t = LSTM(w_t, h_{t-1}) \tag{1}$$

Вся последовательность создает n скрытых состояний, представленных как $H = (h_1, \ldots, h_t)$. Сначала мы получаем u_i, скрытое представление h_i, через многослойный персептрон (MLP).

$$u_i = tanh(Mh_i + b) \tag{2}$$

Затем вводится контекстный вектор u_c для вычисления весов токенов.

$$\alpha_i = \frac{exp(u_i^T u_c)}{\sum_i (u_i^T u_c)} \tag{3}$$

где u_c является общим для всех текстов и случайным образом инициируется и обновляется во время обучения.

При векторе внимания окончательное представление текста является взвешенной суммой скрытых состояний [4],

$$h^* = \sum_i^n \alpha_i h_i \tag{4}$$

2) Конкурентное внимание

Стандартный механизм внимания подходит для получения единичного представления текста [5]. В нашем сценарии мы хотим получить два представления для темы и стиля соответственно. Наше решение состоит в том, чтобы использовать различное внимание к общему представлению, чтобы получить представление для конкретной задачи. Мы вводим механизм конкурентного внимания, чтобы усилить конкуренцию между двумя представлениями.

Учитывая скрытые состояния $H = (h_1, \ldots, h_n)$ конкурентные внимания представляют собой два вектора внимания $\alpha = (\alpha_1, \ldots, \alpha_n)$ и $\beta = (\beta_1, \ldots, \beta_n)$ для вычисления представлений для задачи *T1* и задачи *T2*. Сначала мы вычисляем α в соответствии со стандартным механизмом внимания [6]. S_α – отсортированные m $(1 \leq m \leq n)$ различных значений α $R\alpha = (r_1, \ldots, r_n)$ – ранг α_i среди m значений Для $1 \leq i \leq n$, пусть

$$\beta_i = \frac{S_\alpha[m + 1 - r_i]}{\sum_{j=1}^n S_\alpha[m + 1 - r_j]}$$

Окончательные представления для *T1* и *T2*:

$$h_1^* = \sum_i^n \alpha_i h_i,$$

$$h_2^* = \sum_i^n \beta_i h_i.$$

4 Ограничение разделения-восстановления

Ограничение разделения-восстановления

Теперь у нас есть отдельные представления по теме и стилю. Мы надеемся, что разделение не изменит смысла. Поэтому мы используем ограничение разделения-восстановления и ожидаем, что исходное представление может быть восстановлено с помощью представления темы и представления стиля [7].

1) Оригинальное представление

Объединяем скрытые представления $H = (h_1, \ldots, h_n)$, чтобы получить вектор h_1 размерности $d \times n$ в качестве исходного представления текста.

2) Скрытые представления

Мы используем представление темы h_{topic}^* и представление стиля h_{style}^*, созданное конкурентным вниманием, в качестве скрытых представлений.

3) Реконструированное представление

   h^*_{topic} и h^*_{style} объединяются, а затем сопоставляются с вектором $d \times n$, чтобы получить реконструированное представление h_r[2]

   т. е. $h_r = M'[h^*_{topic}, h^*_{style}] + b'$.

4) Потеря восстановления

   Сначала мы вычисляем манхэттенское расстояние D между h_r и h_0. Затем вычисляем потери

$$\beta_i = 1 - \frac{1}{L-1}, L \in [0, 1].$$

5 Многозадачное обучение для установления авторства

Многозадачное обучение
для установления авторства

Мы формулируем атрибуцию авторства с помощью многозадачного подхода к обучению на основе представления темы h^*_{topic} и представления стиля h^*_{style}.

1) Основная задача: установление авторства

   Мы используем h^*_{style} для указания авторства. h^*_{style} подключается к слою *softmax*, чтобы получить распределение по кандидатам в авторы. Перекрестная энтропия используется в качестве функции потерь для классификации.

2) Вспомогательное задание: приближение темы

   Учитывая текст, мы используем предварительно обученную модель, чтобы вывести его распределение тем θ по K темам. Тематическая модель основана на модели LDA, но с фоновой моделью для захвата общих слов, так что извлеченные темы обычно присваивают более высокие вероятности содержательным словам.

   Полносвязная сеть используется для сопоставления h^*_{topic} с вектором измерения K, а затем этот вектор нормализуется с помощью слоя *softmax*, чтобы получить приблизительное распределение тем θ'. Функция потерь для этой задачи представляет собой перекрестную энтропию между θ и θ'.

6 Практическая оценка предложенной модели

Практическая оценка
предложенной модели

Эксперимент проводится на наборе данных IMDb62, который содержит 62 000 рецензий на фильмы от 62 авторов, у каждого из которых по 1000 рецензий. Набор случайным образом разделён на обучающую выборку (80%) и тестовую выборку (20%).

Для проверки способности противостоять влиянию тем было выбрано подмноже- ство экземпляров из тестового набора, обозначенное как IMDb62-Hard. Специфика набора заключается в том, что у каждого автора есть не более одной рецензии на один фильм. Обзоры в тестовом наборе были выбраны таким образом, чтобы прокомментированные фильмы

появились и в обучающем наборе, но были прокомментированы другими пользователями, чтобы этот тестовый набор данных был более сложным по сравнению со всем тестом. Таким образом, у нас есть 6000 тестовых обзоров.

Как видно из таблицы 1, улучшения на IMDB62 и IMDB62-Hard небольшие. Основная причина состоит в том, что все тексты имеют сходную тематику. Большинство обзоров концентрируются на таких аспектах, как актеры/актрисы/режиссеры, сюжеты, музыка и личные чувства. Это не очень хорошая межтематическая настройка. Помимо языковых стилей, личные интересы авторов, например особые предпочтения в отношении некоторых режиссеров или жанров фильмов, служат сигналами для их различения.

7 Определение темы

Определение темы

В табл. 2 показаны наиболее вероятные темы, основные слова темы и внимание к теме на уровне токенов для образца текста в IMDb62. Текст имеет высокие вероятности по теме 96 (P = 0,46), теме 18 (P = 0,29) и теме 10 (P = 0,1) среди изученных тематических моделей. Под словами темы показаны весовые коэффициенты внимания к теме на токенах, основанные на механизме конкурентного внимания.

Из таблицы видно, что слова с высоким весом внимания темы также имеют более высокие вероятности в показанных языковых моделях темы, таких как шоу, музыка, телевидение. Это указывает на то, что аппроксимация темы успешно направляет представление темы к распределению темы текста.

С другой стороны, мы видим токены с малым весом внимания к теме. Многие из них являются общеупотребительными служебными словами. Некоторые распространенные глаголы, такие как like, be и местоимения, также имеют низкий вес внимания к теме. Эти слова не зависят от темы, но могут в некоторой степени отражать личный стиль.

8 Заключение

В этой статье представлен многозадачный подход к обучению для разделения стилей темы для определения авторства, предложены механизмы конкурентного внимания и ограничение разделения-восстановления для разделения темы и стиля. Ожидается, что метод многозадачного обучения, основанный на аппроксимации темы, будет особенно эффективен в межтематических условиях, однако он позволяет лучше определять авторство даже на специфическом наборе данных, с незначительной вариативностью тем. Предложенная модель так же способна эффективно различать формальный и неформальный стили речи, что способствует более точному определению темы текста.

Список литературы

1. Soler J., Wanner L. On the relevance of syntactic and discourse features for author profiling and identification. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Vol. 2. Short Papers. 2017. P. 681–687.

2. Gómez-Adorno H., Posadas-Durán J.P., Sidorov G. Document embeddings learned on various types of n-grams for cross-topic authorship attribution. Computing. 2018. P. 1–16.

3. Батурин М.М., Белов Ю.С. Использование сверточных, рекуррентно-сверточных нейронных сетей и метода опорных векторов для определения авторства текста // Научные исследования в современном мире. Теория и практика: сборник избранных статей Всероссийской (национальной) научно-практической конференции. СПб., 2022. С. 47–49.

4. Zhang R., Hu Z., Guo H. Syntax encoding with application in authorship attribution. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018. P. 2742–2753.

5. Sundararajan K., Woodard D. What represents 'style' authorship attribution? Proceedings of the 27th International Conference on Computational Linguistics. 2018. P. 2814–2822.

6. Батурин М.М., Белов Ю.С. Использование сверточных нейронных сетей, долгой краткосрочной памяти и оценок внимания для различения авторства текста // E-Scio [Электронный ресурс]. URL: http://e-scio.ru/wp-content/uploads/2022/01/Батурин-М.-М.-Белов-Ю.-С.pdf (дата обращения: 23.05.2022).

7. Shen T., Lei T., Barzilay R. Style transfer from nonparallel text by cross-alignment. Proceedings of Advances in Neural Information Processing Systems 30. 2017. P. 6833–6844.

8. Shrestha P., Sierra S., González F. Convolutional Neural Networks for Authorship Attribution of Short Texts. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Vol. 2. Short Papers. 2017. P. 669–674.

9. Stamatatos E. Authorship Attribution Using Text Distortion. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Vol. 1. Long Papers. 2017. P. 1138–1149.