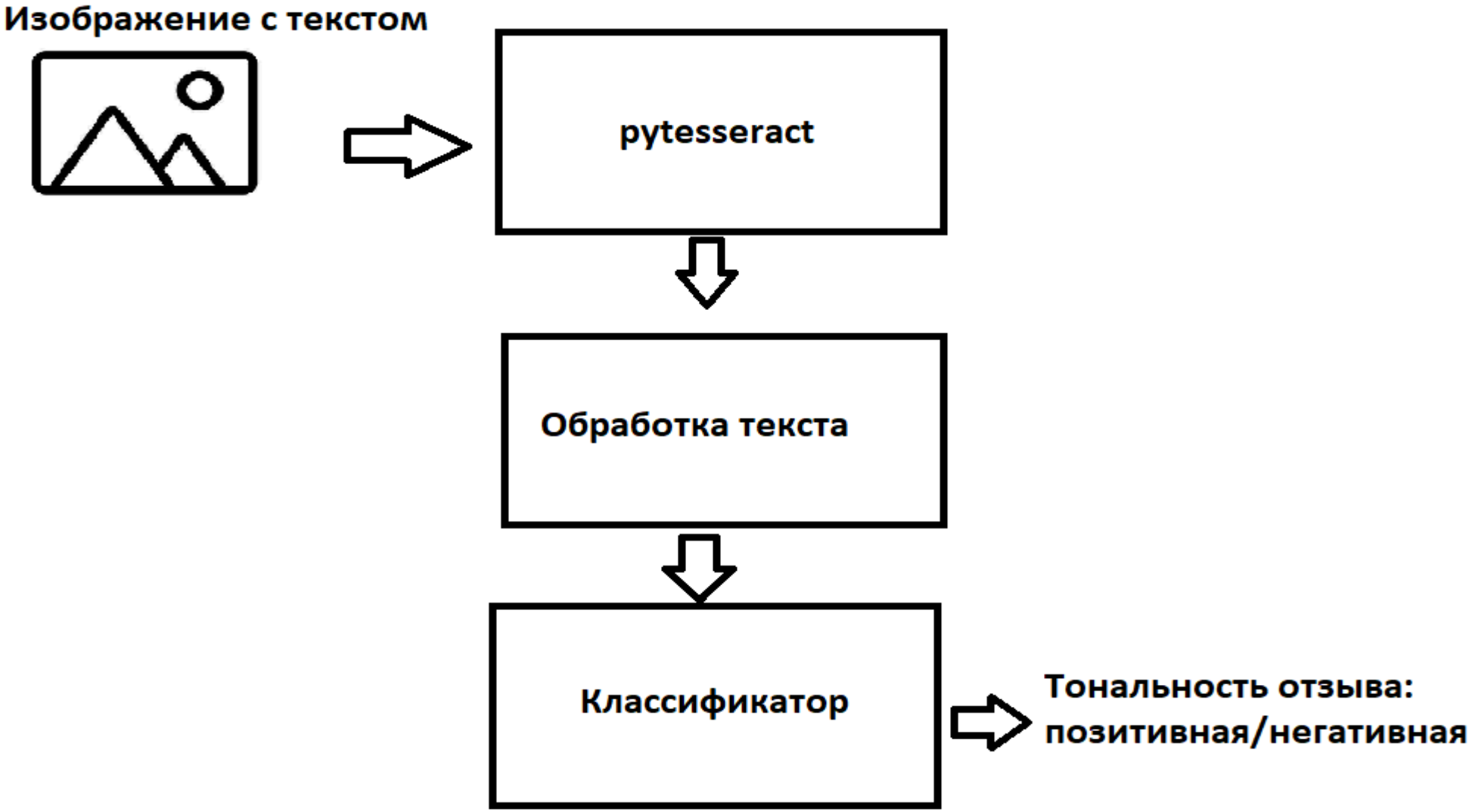
Распознавание и классификация текста

Рубанова Валерия, 5030102/00201

Pipeline проекта

- Входные данные: изображение с текстом
  - Распознавание текста: слова на изображениях распознаются и превращаются в текст
  - Классификация текста: полученный текст классифицируется по категориям.
- Выходные данные: категория текста

Pipeline проекта



Входные данные: отзывы с IMDB

The Trailer was better than the Movie ..
Avid_Movie_Viewer 30 January 2014

Warning: Spoilers

I had expectations from the trailer that this would be a tale of a protagonist that had some social shortcomings and how time travel would help him overcome them to fall in love.

Instead the movie appeared to be a quest by our protagonist to appear smooth and suave with women in manufactured relationship. Our protagonist Tim pursues two young women\girls in the film. One was puppy love that never comes to fruition and the other was completely manufactured by the time travel. I never saw a connection between Tim and Mary. We're supposed to believe that they fall in love with each other the very first day they meet. That was a bit much to believe. Worse is that he uses the time travel to manufacture the relationship. That's not love it's a guy manipulating someone.

Tim appeared like a needy guy who didn't want to do anything to screw up and end a relationship. He did screw up and end the relationships on multiple occasions but with the time travel he went back and fixed it. Funny but the relationship came off as not genuine.

The relationship between Tim and his dad appeared very genuine from beginning to end. That was the strength of this movie.

I actually thought our hero Tim would be forced to sacrifice his manufactured relationship with Mary with something important that would require him to time travel back before meeting Mary or having his children. There was a chance for this with his sister's accident or his father's death but Tim chooses to help himself instead which left me disappointed that everything was so easy.

This movie had potential but it took the road most traveled.

Worthless sentimental trash
simonmills47 9 October 2019

I have nothing more to say about this terrible film beyond the title of my review.

Bad Ripoff of the Truman Show
nickgrande 23 September 2022

Warning: Spoilers

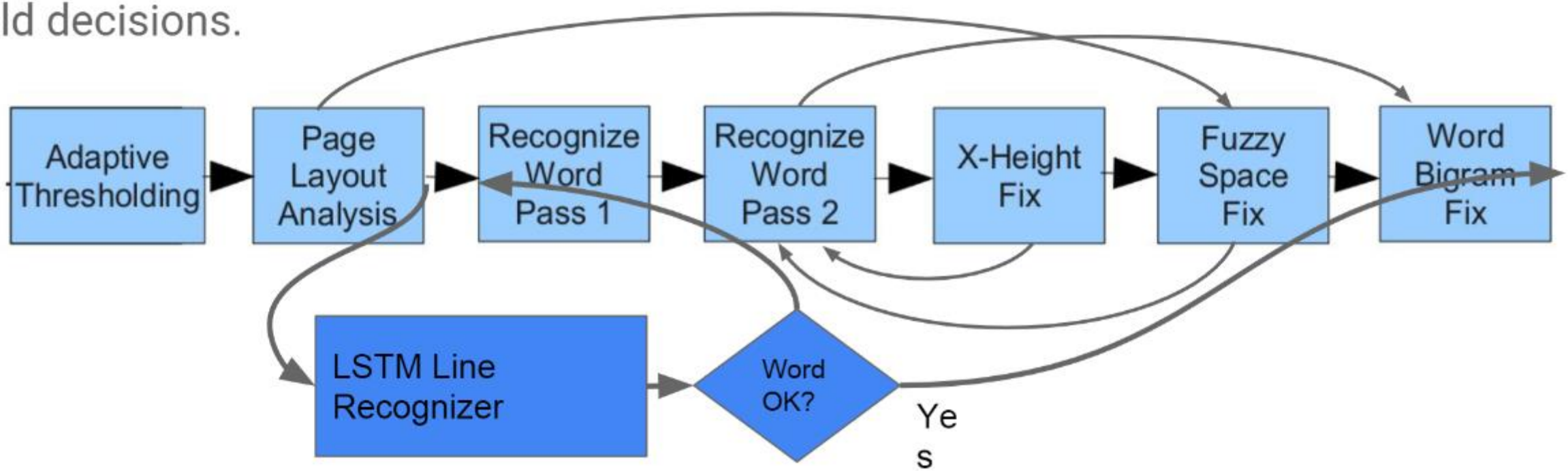
Harry styles can NOT act. Florence and Chris tried to save the movie but ultimately couldn't. The movie was very predictable, the "twist" was an exact copy of the truman show. And it was underwhelming. Harry Styles acting was laughable. There were so many plot holes and things that even at the end of the movie don't even make sense. Especially the last car scene seems like olivia just wanted to get action shots for the trailer. It's also kind of funny that she said the movie ratings reflect the skill of the director and this movie is getting bad reviews. The only way this movie gets good reviews will be from Harry Style fans that can't admit that he can't act. Movie sucks.

Как собирались данные

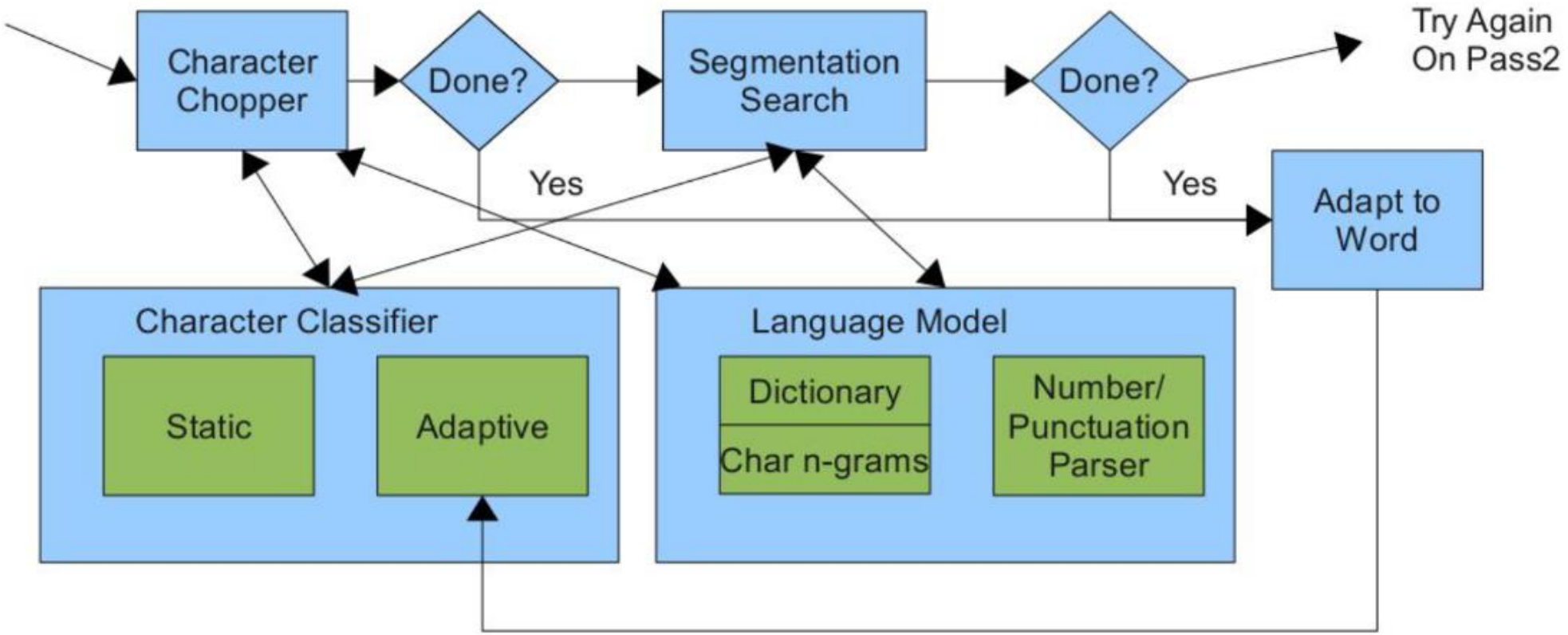
- Для обучения pytesseract – не собирались
- Для обучения классификатора использовался IMDB датасет с kaggle.
- Для тестирования использовались скриншоты с IMDB, которые я сама сделала и разметила.

Tesseract System Architecture

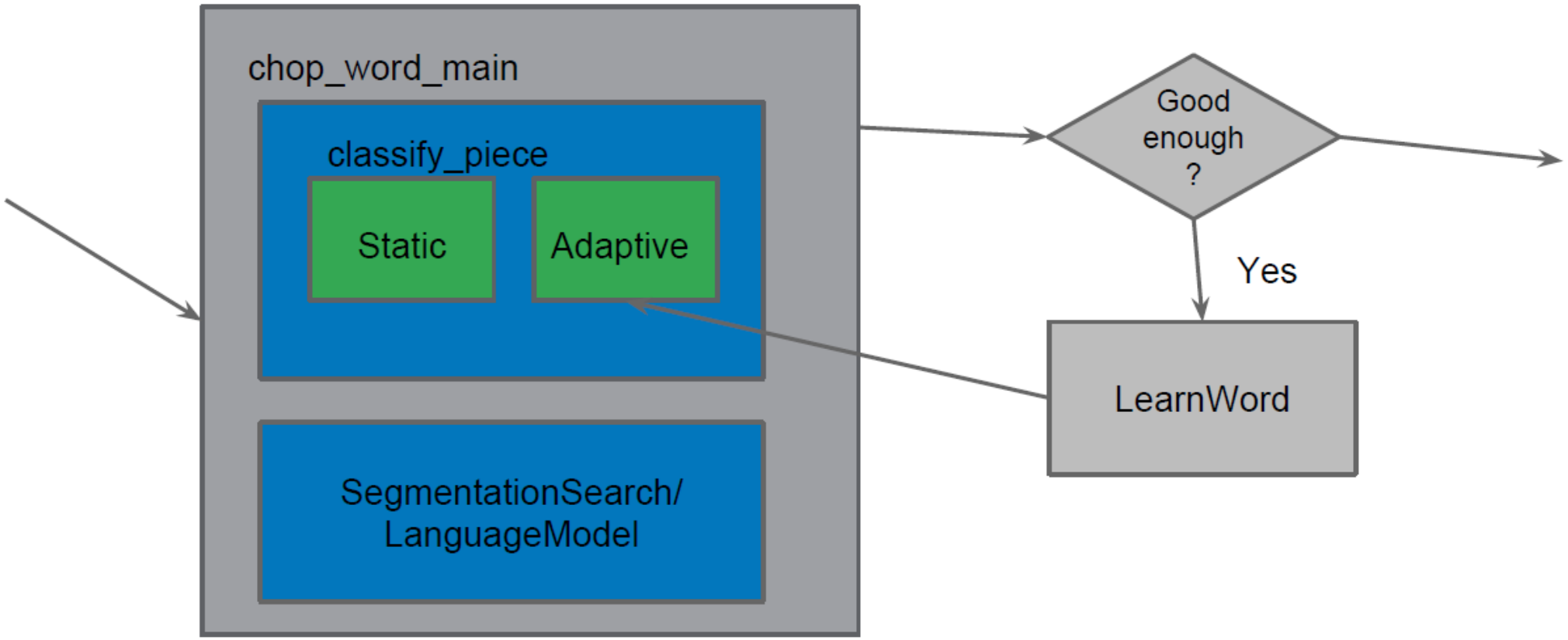
Nominally a pipeline, but not really, as there is a lot of re-visiting of old decisions.



Word recognizer



Adaptive classifier



LSTM

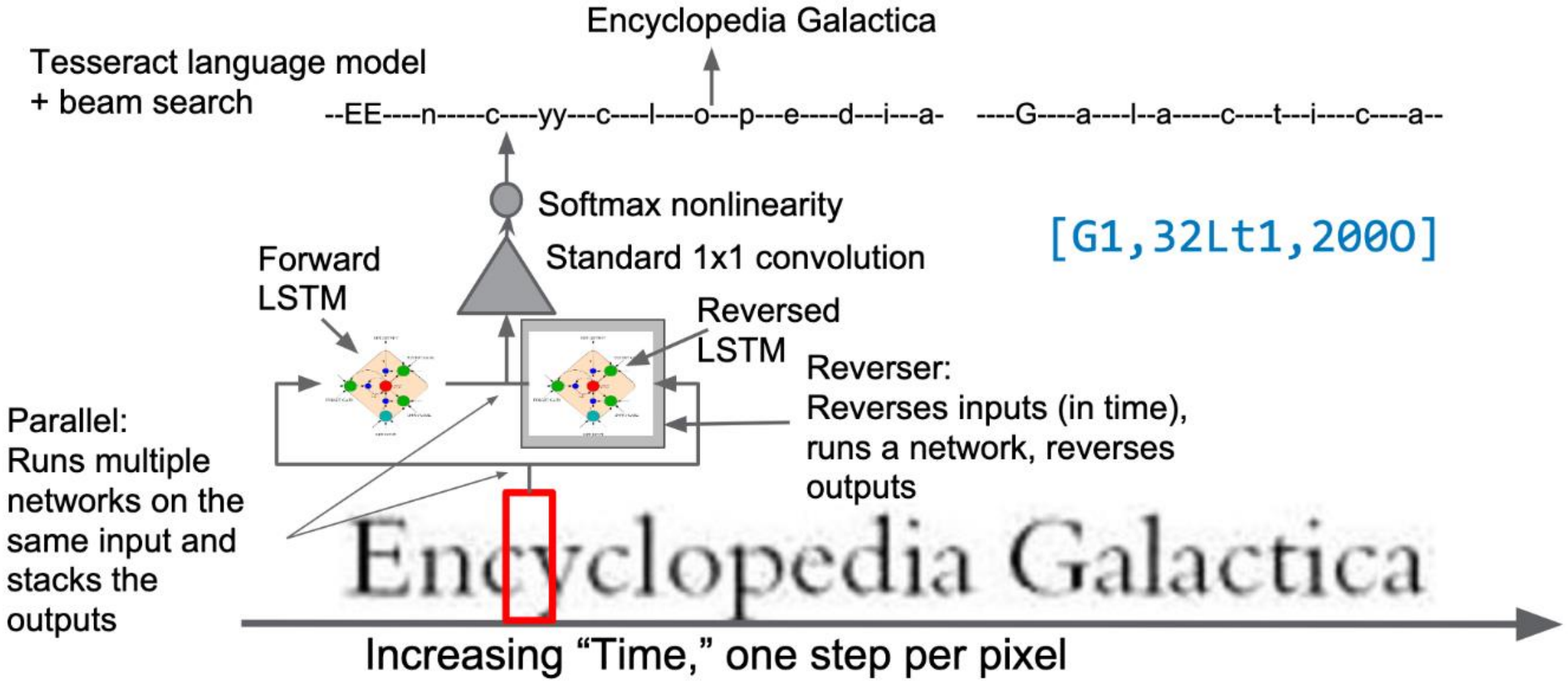
- Реализация основана на OCRopus

```
.Stacked: 0.000100 0.900000 11 13
.Stacked.Parallel: 0.000100 0.900000 11 40
.Stacked.Parallel.NPLSTM_SigmoidTanhTanh: 0.000100 0.900000 11 20
.Stacked.Parallel.Reversed: 0.000100 0.900000 11 20
.Stacked.Parallel.Reversed.NPLSTM_SigmoidTanhTanh: 0.000100 0.900000 11 20
.Stacked.SoftmaxLayer: 0.000100 0.900000 40 13
```

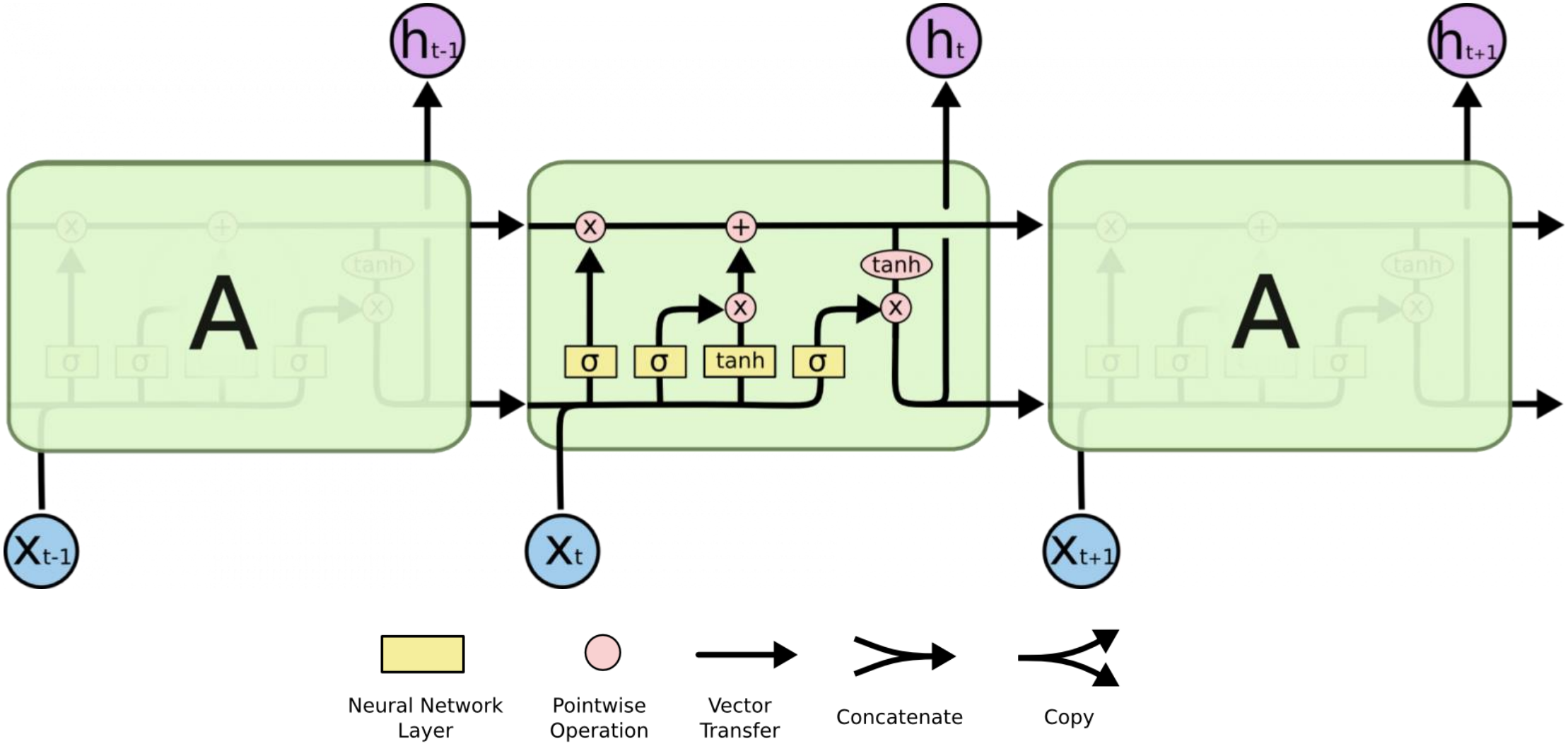
Stacked – два слоя друг на друге
Parallel – запускает несколько сетей параллельно на одном и том же входе

Tesseract 4.00 includes a new neural network subsystem configured as a textline recognizer. It has its origins in OCRopus' Python-based LSTM implementation, but has been totally redesigned for Tesseract in C++. The neural network system in Tesseract pre-dates TensorFlow, but is compatible

How Tesseract uses LSTMs...

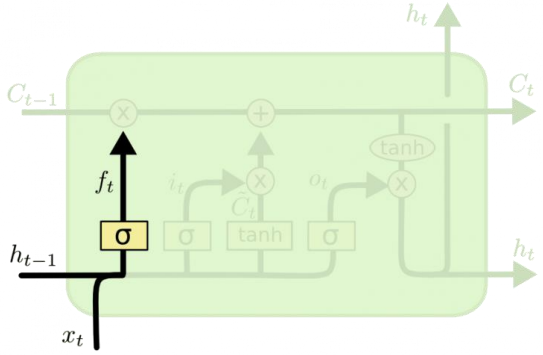


LSTM

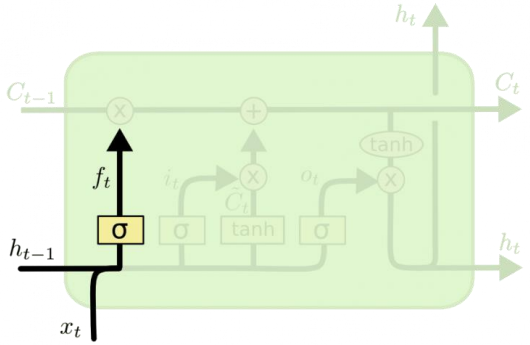

h_{t-1}
h_t
h_{t+1}
A
A
tanh
σ
σ
tanh
σ
X_{t-1}
X_t
X_{t+1}
Neural Network Layer
Pointwise Operation
Vector Transfer
Concatenate
Copy

LSTM

- Первый шаг в LSTM – определить, какую информацию можно выбросить из состояния ячейки. Это решение принимает сигмоидальный слой, называемый “слоем фильтра забывания” (forget gate layer)
- Следующий шаг – решить, какая новая информация будет храниться в состоянии ячейки. Этот этап состоит из двух частей. Сначала сигмоидальный слой под названием “слой входного фильтра” (input layer gate) определяет, какие значения следует обновить. Затем tanh-слой строит вектор новых значений-кандидатов, которые можно добавить в состояние ячейки.



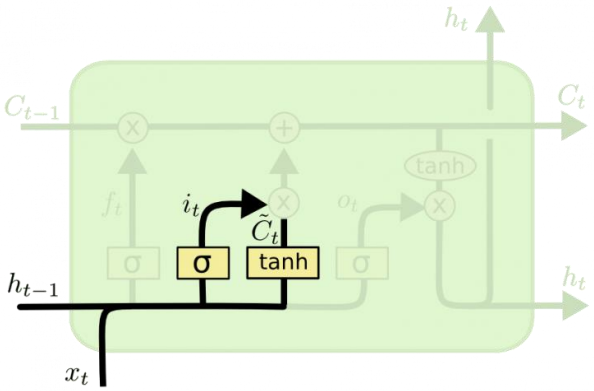
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

LSTM

- Настало время заменить старое состояние ячейки C_{t-1} на новое состояние C_t. Что нам нужно делать — мы уже решили на предыдущих шагах, остается только выполнить это.
- Мы умножаем старое состояние на f_t, забывая то, что мы решили забыть. Затем прибавляем i_t*\tilde{C}_t. Это новые значения-кандидаты, умноженные на t – на сколько мы хотим обновить каждое из значений состояния.

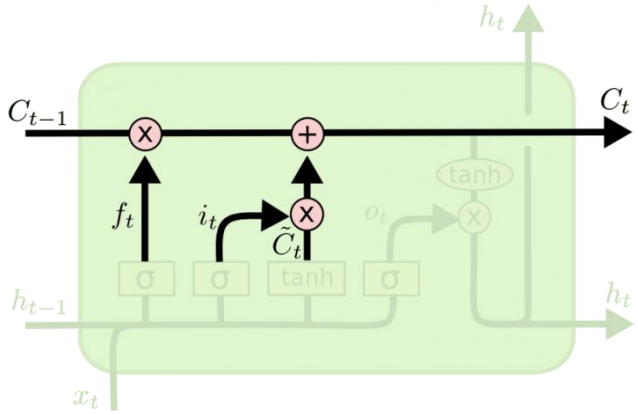


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

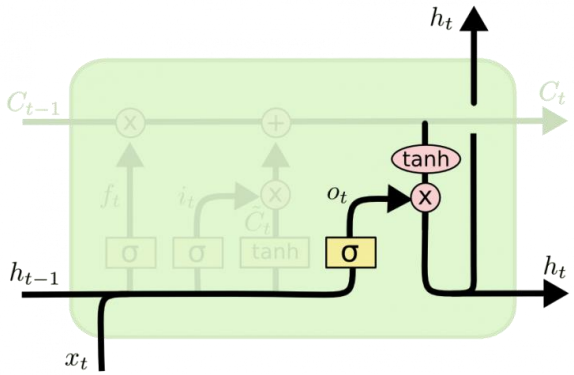
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

LSTM

- Наконец, нужно решить, какую информацию мы хотим получать на выходе. Выходные данные будут основаны на нашем состоянии ячейки, к ним будут применены некоторые фильтры. Сначала мы применяем сигмоидальный слой, который решает, какую информацию из состояния ячейки мы будем выводить. Затем значения состояния ячейки проходят через tanh-слой, чтобы получить на выходе значения из диапазона от -1 до 1, и перемножаются с выходными значениями сигмоидального слоя, что позволяет выводить только требуемую информацию.
- Мы, возможно, захотим, чтобы наша языковая модель, обнаружив существительное, выводила информацию, важную для идущего после него глагола. Например, она может выводить, находится существительное в единственном или множественном числе, чтобы правильно определить форму последующего глагола.



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

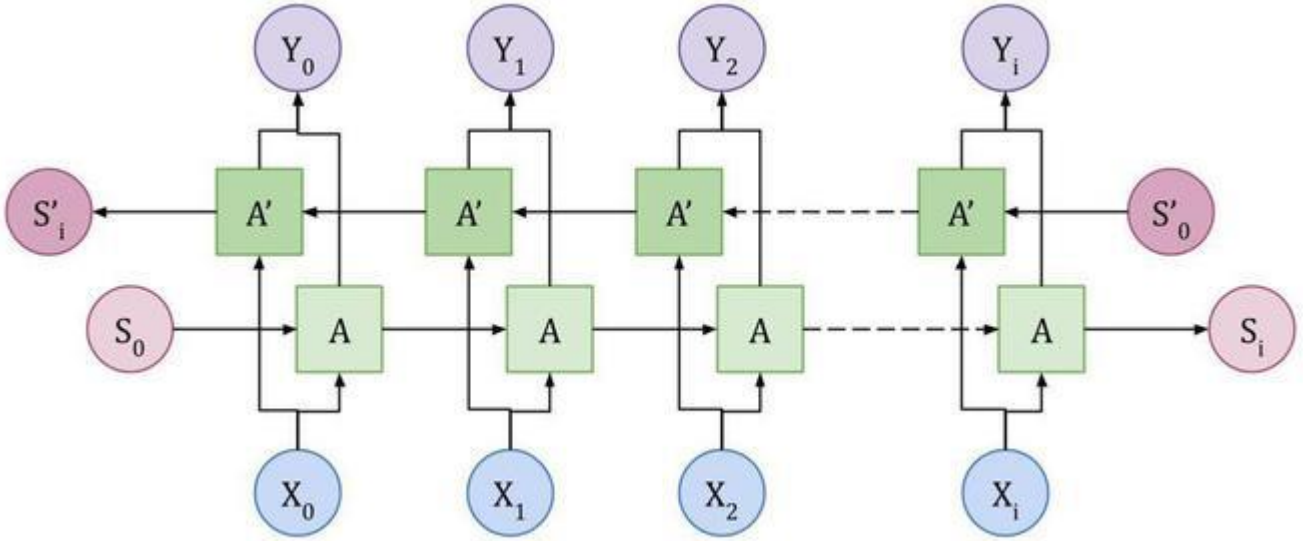


$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

Forward and reversed LSTM

- Архитектура двунаправленного LSTM состоит из двух однонаправленных LSTM, которые обрабатывают последовательность как в прямом, так и в обратном направлениях. Эту архитектуру можно интерпретировать как наличие двух отдельных сетей LSTM: одна получает последовательность токенов как есть, а другая — в обратном порядке.
  LSTM модель в tesseract является seq-2-seq моделью.



1x1 convolution & softmax

- Свертка 1 x 1 — это свертка с некоторыми особыми свойствами, заключающимися в том, что ее можно использовать для уменьшения размерности, эффективных низкоразмерных эмбеддингов и применения нелинейности после сверток.
- Функция softmax принимает входной вектор и нормализует его в распределение вероятностей, где сумма всех элементов выходного вектора равна 1. Это делает ее подходящей для задач классификации нескольких классов, поскольку она присваивает вероятность каждому классу.

$$s\left(x_i\right) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

Tesseract language model & beam search

- Language model – своя для каждого языка, по дефолту загружается eng.traineddata
- Лучевой поиск — это эвристический алгоритм поиска, который исследует граф, расширяя перспективные узлы в ограниченном наборе.

Целевая функция: Connectionist Temporal Classification Loss

- CTC (Connectionist Temporal Classification) Loss – это функция потерь, используемая для обучения нейронных сетей в задачах распознавания речи или оптического распознавания символов (OCR). Она позволяет моделям с пропусками выравнивать входные последовательности с переменной длиной с целевыми последовательностями фиксированной длины.
- CTC Loss основан на идее введения специального символа "blank" (пустота) для отражения пропусков или повторов в выходной последовательности. Она позволяет модели генерировать различные выходные последовательности, представляющие одну и ту же входную последовательность, и затем вычисляет вероятность соответствия целевой последовательности каждой из сгенерированных последовательностей.
- Функция потерь CTC основывается на вычислении вероятности каждой возможной выходной последовательности, принимая во внимание все возможные способы порождения данной последовательности из входной. Затем она минимизирует суммарную отрицательную логарифмическую вероятность правильной последовательности.

Интуиция за CTC

- Добавляется «пропуск» как возможный выходной лейбл
- Создается нейронная сеть, которая выводит лейбл для каждого инпута
- Сколлапсировать получившиеся последовательности
  - AAAAA→A
  - AAAABB→AB
  - AAAA_ _ _BB→AB
  - AAA_ _AAA→AA
- Сравнить получившиеся последовательности с метками

CTC

- На каждом временном шаге сеть фактически выводит распределение вероятностей по всем возможным меткам и пустому символу. Если мы случайным образом выберем метку независимо от каждого из этих распределений, какова вероятность того, что мы получим выходную последовательность, которая схлопнется в основную истинную последовательность? Потери CTC представляют собой отрицательный логарифм этой вероятности.

Пример плохой работы pytesseract

Powerful :

okpilak’ 16 Yanuary. 2623 «

This:rijowie.i& about how jéuritalism works: The: work of journalists, not jist a calling

Agad-ontelevisiény, Buk thé effort involved-in uncovering’ a Story ‘and getting the

collaboration’of: what thay-are reporting on. This isn't-ther story of the S-o'clock evening:

news. that.is heard. and quickly forgotten. The-strength of. the movie-is how they-showed

. the: difficulty ia geiting the womén to come forward,. and. the ‘substantial: power:thdse.in

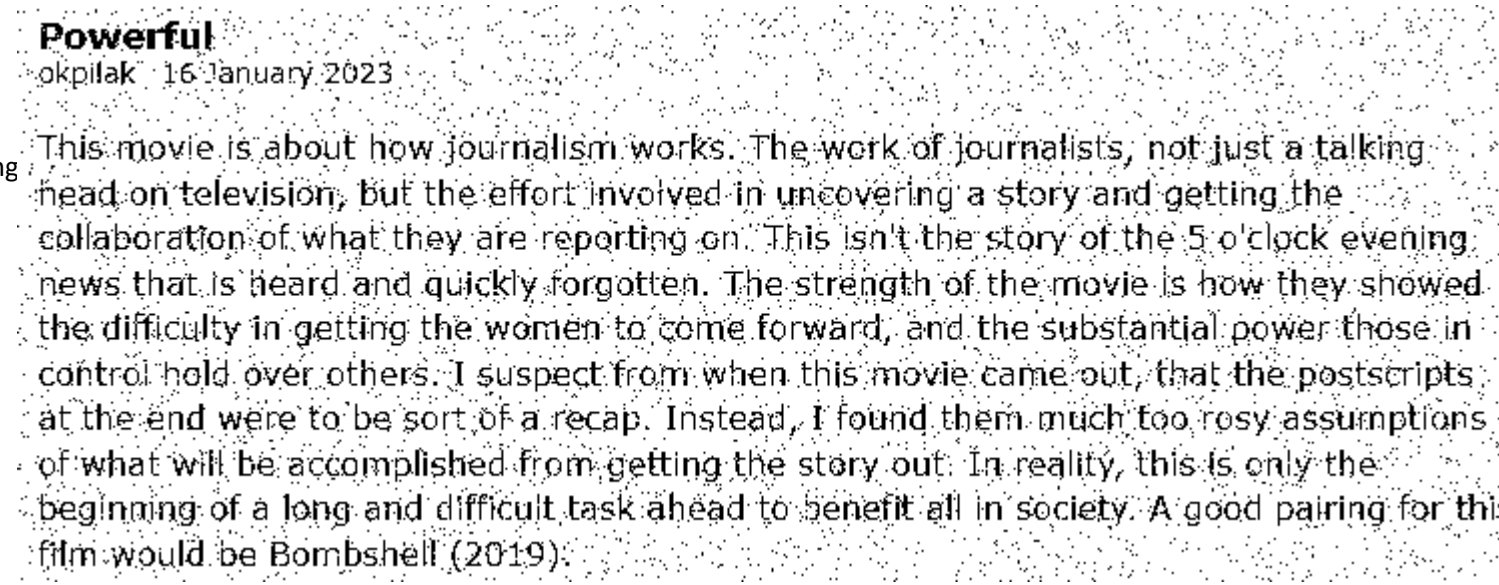
contrét‘hold.over:others:.1 suspect: from. when this iovie. came. out; that the. poststripts ;

at the 2nd wére to’be'sort bf a-recap. Instead, f foutid ther. much: tao, Tosy: assurptions

of what ‘will be accomplished From getting-the story out: in-reality, thiss onlythée

beginning: of a long. and difficult task Shéad-to Sjetiefit all in’ socieky: A'go6d pairing for, this

film would ‘be Bombshell (2049): i oe



Powerful

okpilak 16 January 2023

This movie is about how journalism works. The work of journalists, not just a talking head on television, but the effort involved in uncovering a story and getting the collaboration of what they are reporting on. This isn't the story of the 5 o'clock evening news that is heard and quickly forgotten. The strength of the movie is how they showed the difficulty in getting the women to come forward, and the substantial power those in control hold over others. I suspect from when this movie came out, that the postscripts at the end were to be sort of a recap. Instead, I found them much too rosy assumptions of what will be accomplished from getting the story out. In reality, this is only the beginning of a long and difficult task ahead to benefit all in society. A good pairing for thi film would be Bombshell (2019).

Классификатор

- MultinomialNB (Multinomial Naive Bayes) - это алгоритм классификации, основанный на методе наивного байесовского классификатора. Он используется в машинном обучении для решения задач классификации с дискретными (частотными) признаками.
- MultinomialNB особенно полезен, когда признаки представлены в виде целых чисел, таких как количество слов в документе или частота появления разных слов. Это часто применяется в приложениях анализа текста, таких как классификация документов, спам-фильтры или анализ тональности текстов.
- Однако MultinomialNB имеет предположения о распределении данных, такие как мультиномиальное распределение, простое предположение о независимости признаков и отсутствие взаимодействия между ними. Если эти предположения не выполняются, то MultinomialNB может давать плохие результаты. Поэтому перед использованием MultinomialNB всегда необходимо проверить, соответствуют ли его предположения данным задачи.

Целевая функция

- Наивный байесовский классификатор оптимизирует условную вероятность класса C, заданную признаками x1, x2, ..., xn. Другими словами, целевая функция наивного байесовского классификатора - это вероятность P(C|x1, x2, ..., xn) того, что объект с признаками x1, x2, ..., xn принадлежит к классу C.
- Для определения этой вероятности, наивный байесовский классификатор использует теорему Байеса и предполагает независимость между признаками. Таким образом, он вычисляет вероятность P(C) априори для каждого класса C и вероятности P(xi|C) для каждого признака xi в каждом классе C. Затем он использует эти вероятности для вычисления условной вероятности P(C|x1, x2, ..., xn) для каждого класса C и выбирает класс с наибольшей вероятностью в качестве ответа.

Качество работы

- Точность модели на тестовой выборке из 40 изображений до предобработки текста: 87.5%
- После – 90%
- Ошибки 1 рода (false positive) возникают, когда алгоритм классификации неправильно предсказывает, что отзыв является позитивным, хотя на самом деле он негативный.
  - Ошибка 1 рода в моей реализации возникает на 4 тестовых изображениях.
- Ошибки 2 рода (false negative) возникают, когда алгоритм классификации неправильно предсказывает, что отзыв является негативным, хотя на самом деле он позитивный.
  - Ошибок 2 рода в моей реализации нет.

Примеры верного срабатывания

"Terrible, terrible joy."
benjaminskylerhill 10 December 2022

With less than 1/4 of the budget of the soulless Disney live action remake earlier this year, Del Toro & Co. Have managed to craft a version of Pinocchio with more personality, heart, and soul than Disney could have dreamed of crafting.

It injects new life into the character by telling a story that is vastly different both narratively and thematically than any version we've seen on screen before.

This tale deals with the malleability of identity, unconditional love, the impressionable nature of children, and the close link between joy and sorrow. And it does so with dark wit, refreshingly complex three-dimensional characters, and stunningly haunting stop-motion animation.

However, I do think this could have benefited a bit from cutting down on the plethora of plot points and having Pinocchio and Geppetto spend more time together. This version lacks the tight focus and brisk pacing of the 1940 version.

But aside from this, I was thoroughly entranced by this dark fairy tale. It has a spine and a soul, and unlike it's titular protagonist, it's far from wooden.

The absolute worst.
Viation 7 September 2019

As a man in my early twenties, RomComs are my guilty pleasure. I like them, I really do. And I've watched really bad ones and still liked them. Whether it was for their relatable characters or their main cast, there's almost always something positive I can "cling" to, even in the bad RomComs. But this one was too much. While the gist of the story would theoretically make for a good movie, gigantic plot holes, illogical use of the main character's time traveling ability and the incredibly dumb (that painfully, cringe inducing kinda dumb, not the cute kinda dumb) behavior of the main character make this unwatchable for me. I realize I'm not necessarily the target audience but I would recommend staying away from this even if you are.

Примеры ложного срабатывания

Bad (clearly a man wrote the screenplay)
oliviakd 6 December 2020

Warning: Spoilers

The protagonist faces no consequences for manipulating women or any of his actions in general. I'm not sure what the point of this movie is, it ends with the protagonist learning he needs to reflect on the good parts of his life as if there were any bad moments. The only thing that really stood out for me was how this film managed to anger and bore me simultaneously.

Incredibly boring
cdmarker2001 7 November 2021

The whole movie was honestly lacking in all areas... it wasn't exciting, dramatic, suspenseful, funny... nothing. The plot was dull, slow paced, the acting was bland. The time traveling aspect is the only intriguing part and even that wasn't enough to keep things interesting. 2 hours I would like to get back.

I don't understand how this has as high of a rating as it does.