А. В. Перунова, А. В. Глазкова

Тюменский государственный университет, г. Тюмень

УДК 004.912

ГЕНЕРАЦИЯ ТЕКСТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ С ПОМОЩЬЮ НЕЙРОННЫХ СЕТЕЙ

Аннотация. Генерация текста – одна из задач, которая может быть решена с использованием моделей глубокого обучения. В данной статье представлена модель нейронной сети для генерации текста на естественном языке с использованием рекуррентной нейронной сети (Recurrent neural network, RNN) и механизма внимания.

Ключевые слова: обработка естественного языка (Natural Language Processing, NLP), генерация естественного языка (Natural Language Generation, NLG), рекуррентные нейронные сети (Recurrent neural network, RNN), внимание (Attention).

Введение

Генерация нейронного текста – это генерация текста посредством нейронных сетей. Абстрактная модель генерации нейронного текста представлена на Рис. 1.

Рис. 1. Генерация нейронного текста.

«Состояние мира» – это представление окружающего, такое как изображение или письменное предложение; «кодер» – это модуль, который отображает состояние мира в смысловое пространство; «декодер» – переводит состояния из пространства смысла в определенный «язык».

Генерация нейронного текста имеет много применений, например:

- Машинный перевод, когда состояние мира состоит из предложения на исходном языке, которое кодируется в пространство значений и затем декодируется на один или несколько целевых языков.
- Генерация нейронного текста также может быть полезна для чат-ботов, где строка диалога кодируется в значащее пространство, а декодер отображает смысловое пространство в разумный ответ.
- Генерация субтитров для изображения, т.е. изображение кодируется в смысловое пространство, а затем декодируется в субтитры.
- Генерация уникальных текстов, что может использоваться с целью размножения текстов для наполнения сайта контентом (рерайт), написания книг и т.д.

Модель

Опираясь на абстрактную модель, была разработана нейронная сеть. В качестве кодера использована векторная модель языка, что позволяет двигаться в сторону большей «осмысленности» модели и «понимания» значения слов. В качестве декодера выступает рекуррентая нейронная сеть, так как она позволяет обрабатывать информацию циклично при движении от входа к выходу, причем выход зависит от предыдущих вычислений, обеспечивая эффект «памяти».

Архитектура сети имеет четыре слоя в каждом временном интервале: слой входного слова, слой проекции, рекуррентный слой, и слой $softmax$ [1].

Векторная модель языка

На слое проекции реализована операция поиска в таблице, которая преобразует слово в его векторное представление [2]. Векторная модель получена путем обучения модели FastText [3], где каждое слово представлено как «мешок» n-грамм – множество последовательных буквенных сочетаний длины n, и каждому слову добавлены специальные граничные символы "<" и ">", для разделения префиксов и суффиксов от других символов последовательности. Само слово также включается в набор его n-грамм,

чтобы иметь представление для него. Идея FastText основана на модели word2vec [4].

Рекуррентная нейронная сеть

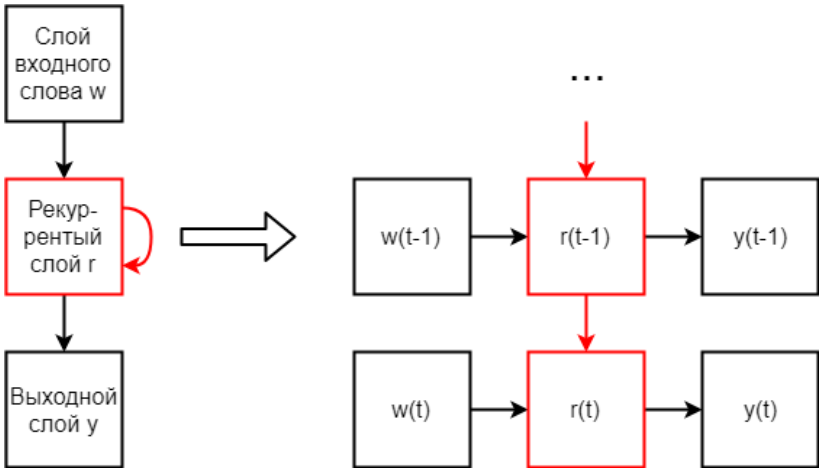

Рис. 2. Структура рекуррентной сети.

Рекуррентная нейронная сеть имеет три типа слоев в каждом временном интервале (Рис. 2) [5]: слой входных слов w, рекуррентный слой r и выходной слой y. Активация входного, рекуррентного и выходного слоев в момент времени t обозначается как $w(t)$, $r(t)$ и $y(t)$ соответственно. $w(t)$ является векторным представлением текущего слова. $y(t)$ можно рассчитать следующим образом:

$$x(t) = [w(t)r(t-1)]$$
$$r(t) = sigmoid(U \cdot x(t))$$
$$y(t) = softmax(V \cdot r(t))$$

где $x(t)$ – вектор, который объединяет $w(t)$ и $r(t-1)$, а U и V – матрицы весов, которые будут обучаться.

В качестве слоя входных слов выступает слой проекции, описанный выше.

Внимание

Чтобы модель могла опираться на контекст, в ячейку RNN добавлен механизм внимания. Механизм внимания выступает «фильтром» поступающей информации: на какие «важные» части предложения

необходимо сконцентрироваться, то есть как выбрать из большого объема поступающих данных то, что нужно обработать прямо сейчас (Рис. 3) [6].

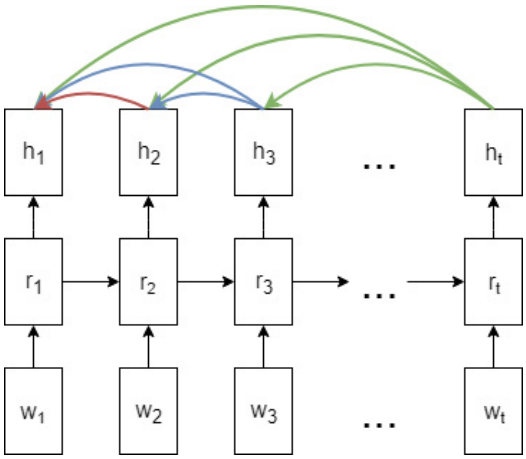

Рис. 3. Механизм внимания.

Пусть $Y \in \mathbb{R}^{k \times L}$ – матрица, состоящая из выходных векторов контекста $[h_1, \ldots, h_L]$, созданных RNN при чтении L слов контекста, где k – гиперпараметр, обозначающий размер входных векторов, $e_L \in \mathbb{R}^L$ – вектор из единиц, а h_N – последний выходной вектор из RNN [7].

Механизм внимания будет производить α – вектор весов внимания и взвешенного представления контекста r:

$$M = \tanh(W^y Y + W^h h_N \otimes e_L) \qquad M \in \mathbb{R}^{k \times L}$$

$$\alpha = softmax(w^T M) \qquad \alpha \in \mathbb{R}^L$$

$$r = Y\alpha^T \qquad r \in \mathbb{R}^k$$

где W^y, $W^h \in \mathbb{R}^{k \times k}$ – матрицы весов, которые будут обучаться, $w \in \mathbb{R}^k$ – обученный вектор параметров, а w^t – его транспонирование. Внешнее произведение $W^h h_N \otimes e_L$ повторяет линейно преобразованное h_N столько раз, сколько слов в контексте (т.е. L раз).

Описание эксперимента

Для получения результатов вектора FastText длиной 100 были собраны на корпусе детской литературы различных жанров. Нейронная сеть была обучена на серии книг Гарри Поттер английской писательницы Дж. К. Роулинг.

Текст был переведен в нижний регистр, и разбит на токены – слова и символы пунктуации. На основе датасета собран словарь, состоящий более чем из 70 тыс. слов. Для обучения сети данные разбиты на пакеты по 50 слов (такая длина обусловлена необходимостью подать в сеть достаточно информации, а так как средняя длина предложений в датасете – 15 слов, сеть может учитывать контекст нескольких последовательных повествовательных предложений, а также коротких диалогов).

Результаты

Для получения результатов были взяты случайные части датасета, на котором обучалась сеть. Данные подавались в сеть, как первоначальный контекст, при непосредственной генерации текста, в качестве контекста выступали ранее уже сгенерированные слова. Результаты генерации сети представлены в таблице 1.

Таблица 1. Результаты генерации.

Вход в сеть	Выход сети
Ночь опускалась на замок	Ночь опускалась на замок. Когда они добрались до ворот, Гарри и Рон крались следом за ней. - Не знаю, – сказала Гермиона, когда они отошли от Хагрида – что это с ним? Рон и Гермиона переглянулись.
Он снова поднял палочку и	Он снова поднял палочку и взмахнул. - Экспекто патронум! Экспекто патронум! – из палочки вылетела струйка серебристого пара, и дементор замедлил движение, словно окаменев. - Не смейте! – крикнул Гарри. Он схватил Малфоя за шиворот и протолкнул его в угол. - Он обвинил меня в том, что я сделал, – сказал он.

В толпе	В толпе засверкал свет, и Гарри увидел, как побелели костяшки пальцев, сжимавших в руках волшебную палочку. Он бросился на нее, повалил на землю и приложил к горлу. Гарри ощутил страшное дыхание. Его колотило, и он старался, чтобы его не было.
Он взмахнул волшебной палочкой, но	Он взмахнул волшебной палочкой, но змея, схватив палочку, повернулась к Рону. Она была так рассержена, что Гарри казалось, будто она не могла взять эту палочку.
- Но почему? Почему не	- Но почему? Почему не подождал? Гарри, почему? - Не знаю, – ответил Гарри, – я просто хотел бы, чтобы ты отгородился от нее, – признался Гарри. - Я не хотел, чтобы он умирал, – сказал он. Гарри обернулся к нему, но тут же отвел взгляд. Гарри не знал, что он думает, что он не вернется, что он сожалеет.

Однако если дать сети генерировать достаточно большой текст, можно наблюдать ее недостаток: персонажи, их местоположение и действия резко меняются от предложения к предложению (Выход сети: «<n токенов> … Он взял Гарри за локоть и вышел, расталкивая толпу второкурсников. Гарри и Рон отдали Стэну, а они, похоже, не забыли Ириску. Гарри неслышно перебежал через кухню и преградил ему дорогу до улицы.»).

Заключение

Разработанная модель может сохранять логику повествования и строить диалоги на относительно коротких текстах (длиной 3-4 предложения), но ей не хватает глобального контекста и сохранения структуры повествования, как в настоящих художественных произведениях. Поэтому, дальнейшая работа будет направлена на улучшение модели в пользу глобального понимания «смысла» того, что она производит.

СПИСОК ЛИТЕРАТУРЫ

1. Junhua Mao, Wei Xu, Yi Yang etc. Explain Images with Multimodal Recurrent Neural Networks Attention // arXiv, 2014. [Электронный ресурс. https://arxiv.org/abs/1410.1090] (дата обращения: 05.05.2019).
2. Тарасов, Д.С. Генерация естественного языка, парафраз и автоматическое обобщение отзывов пользователей с помощью рекуррентных нейронных сетей.
3. Piotr Bojanowski, Edouard Grave etc. Enriching Word Vectors with Subword Information // arXiv, 2017. [Электронный ресурс. https://arxiv.org/abs/1607.04606] (дата обращения: 05.05.2019).
4. Tomas Mikolov, Ilya Sutskever, Kai Chen etc. Distributed Representations of Words and Phrases and their Compositionality // arXiv, 2013. [Электронный ресурс. https://arxiv.org/abs/1310.4546] (дата обращения: 05.05.2019).
5. J. L. Elman. Finding structure in time. Cognitive science, 14(2):179–211, 1990.
6. Николенко С., Кадурин А., Архангельская Е. Глубокое обучение – СПб.: Питер, 2018 – 480 с.
7. Tim Rocktaschel, Edward Grefenstette, Karl Moritz etc. Hermann Reasoning about Entailment with Neural Attention // arXiv, 2015. [Электронный ресурс. https://arxiv.org/abs/1509.06664] (дата обращения: 05.05.2019).