Министерство образования Республики Беларусь
Учреждение образования БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет компьютерных систем и сетей

Кафедра информатики

РЕФЕРАТ

на тему

Обзор задач обработки естественного языка (Natural Language Processing)

Магистрант:

А.С. Долматович

МИНСК 2020

Содержание

Введение

Обработка естественного языка - одно из важнейших направлений в области искусственного интеллекта. Можно считать, что своими истоками оно идет с тех пор, когда появились первые идеи о том, что вычислительные машины могут быть использованы для решения задач, связанных с естественным языком, используемом в повседневной жизни каждого человека.

Одной из важных и первых задач стала задача машинного перевода - перевода текста автоматически с одного языка на другой. Для решения этой задачи было приложено немало усилий, и в 1950-х годах впервые были достигнуты значительные успехи по переводу текста с русского языка на английский.

Не менее важной задачей считалось создание системы, которая может вести осознанный диалог с человеком на понятном для него языке.

Также множество усилий было направлено на создание системы по принципу «вопрос-ответ». В целом, задача чем-то похожа на предыдущую, но в данном случае система должна была уметь отвечать на вопросы человека, которые представлялись в форме текста.

Актуальность направления обработки естественного языка связана с тем, что людям самостоятельно становится практически невозможно обрабатывать большие объемы информации, которые собираются и собирались последние десятилетия. Исходя из этого, можно выделить следующие типы задач, которые сейчас могут быть решены с помощью машинного обучения и обработки естественного языка:

- распознавание и синтез речи,
- информационный поиск,
- классификация и кластеризация текстов,
- резюмирование текстов,
- анализ информации из социальных медиа,

- создание систем «вопрос-ответ».

В последние годы все больше акцентируют внимание на анализе информации из социальных медиа. Но помимо этого задачи анализа естественного языка так же активно применяются в социологии, медицине и психологии [2].

Для решения этих задач в разное время применялись математические, логические, стохастические подходы. особой популярностью пользовались конечные автоматы, конечные преобразователи, логика предикатов и вероятностные подходы. Но в начале XXI века началась новая эра и задачи обработки естественного языка начали решать с помощью методов машинного обучения и нейронных сетей.

1. Что такое естественный язык?

Естественный язык — это важнейшее средство общения и выражения мыслей [1], это язык, который используют люди в общении друг с другом. В рамках данной работы естественный язык будет рассматриваться в виде текстового представления информации.

Если рассматривать текст с точки зрения информатики, то можно считать, что мы будем иметь дело с неструктурированной последовательностью символов. Для компьютера то, что человек может воспринимать как данность, является абсолютно непонятной и бесполезной информацией, которую он не в силах обработать.

Для того, чтобы решать задачи обработки естественных языков необходимо определиться с одним из возможных путей. Наиболее простой, но не всегда правильный путь — решать задачу без каких-либо знаний о языке. Это значит, что мы не принимаем во внимание ничего, кроме того, что текст — это цепочка символов, которые можно представить в виде байтов. Данный подход изначально не совсем верный, так как язык — это строгая система, у которой есть свои правила, свои уровни, своя фонетика, морфология, синтаксис и семантика. Это знание плавно вытекает во второй путь — решать задачу, принимая во внимание язык со всеми его правилами. Понимая, как он устроен, все его лингвистические особенности, выполняя полноценный лингвистический анализ текста.

Как правило, для решения поставленной задачи проводится морфологический анализ текстов, с которыми необходимо взаимодействовать. Для всех слов устанавливаются некоторые морфологические признаки или инварианты. Для различных наборов текстов эти признаки могут различаться. Затем обычно проводится синтаксический анализ текстов: слова объединяются в группы, между этими группами устанавливаются зависимости.

Большинство задач может быть решено абсолютно исключая знания о самом языке. Но данный подход не эффективен за счет того, что не выполняется лингвистический анализ текста и нейронная сеть не может установить какие-то правила, характерные для данного языка.

Для того, чтобы решить задачу обработки естественного языка, обычно необходимо выполнить следующие действия:

- отобрать тексты, необходимые для анализа,
- создать на их основе корпус текстов,
- составить словари идентификации.

Но существует ряд задач, где стоит разметить тексты лингвистом вместе с экспертом в предметной области и разделить все тексты между рядом классов. Все зависит от поставленной задачи.

В любом случае, методы машинного обучения самостоятельно выводят какие-то правила и модели, которые в дальнейшем будут использованы для решения.

2. Возможные проблемы

Сегодня задачи анализа естественного языка довольно широко используются в коммерческих целях. По этой причине можно говорить о высоком уровне развития данного направления исследований. Но прежде, чем приступать к решению какой бы то ни было задачи, связанной с обработкой естественного языка, стоит понять, с какими проблемами можно столкнуться в процессе.

Наиболее важной из проблем можно считать многозначность. Она может быть устранена с помощью контекста и какой-то регулярности в использовании языковых конструкций. Но для этого необходимо принимать во внимания язык и все его лингвистические особенности.

Подход, выбранный для решения задачи анализа текста, в первую очередь зависит от предметной области самой задачи, языка, используемого для анализа, типа текстов. Например, проводить анализ постов или комментариев из социальных сетей совсем не то же самое, что анализировать научную литературу или художественный текст.

И несмотря на наличие множества публикаций, обучающих руководств, научных работ — нет универсальных правил и рекомендаций, которых необходимо передерживаться для того, чтобы эффективно решить какую-то задачу. Особенностью области машинного обучения можно считать то, что для достижения высоких результатов иногда необходимо тратить огромное количество времени на эксперименты с архитектурой сети или настройкой параметров. И самое главное, не всегда, затраченное время на эти эксперименты будет оправдано и принесет результат.

3. Способы решения

Решение большинства задач обработки естественного языка можно охарактеризовать следующими шагами:

1. Сбор данных,
2. Очистка данных,
3. Выбор представления данных,
4. Классификация,
5. Инспектирование,
6. Применение семантики.

3.1 Сбор данных

Решение любой задачи начинается с данных. В области машинного обучения исходные данные занимают важнейшее место. В зависимости от поставленной задачи, в качестве данных могут выступать посты в социальных сетях, комментарии, твиты, отзывы и др. В ряде задач данные для обучения можно найти в открытом доступе. Но если стоит несколько неординарная задача, то, вероятнее всего, данные необходимо будет собирать самостоятельно. Существует мнение, что иногда сбор данных является более трудоемкой задачей, чем само обучение нейронной сети на этих данных.

Одним из важнейших критериев считается размеренность данных. Это позволяет использовать обучение с учителем. Как подчеркивает Ричард Сочер, «обычно быстрее, проще и дешевле найти и разметить достаточно данных, на которых будет обучаться модель — вместо того, чтобы пытаться оптимизировать сложный метод обучения без учителя» [3].

3.2 Очистка данных

Чистые данные позволяют выделять наиболее значимые признаки и не переобучаться на нерелевантных данных, создающих шум. Это главная причина, ради чего необходимо позаботиться об очистке данных. Для того, чтобы очистить данные, необходимо выполнить следующее:

1. Удалить все лишние символы;
2. Произвести токенизацию текста;
3. Удалить нерелевантный слова;
4. Перевести все символы в один регистр (обычно в нижний)
5. Совместить слова, которые могли быть написаны с ошибками или слова, которые имеют альтернативное написание,
6. Произвести лемматизацию (приведение различных форм одного и того же слова к словарной форме).

3.3 Выбор представления данных

Модель машинного обучения принимает на вход числовые значения. Это значит, что прежде, чем обучать модель, необходимо как-то представить данные в том виде, чтобы машина смогла их понять и обработать.

3.3.1 «Мешок слов»

Можно проанализировать текст и составить словарь слов для данного текста. Следовательно, каждое предложение можно представить в виде массива длины словаря, где в нужных индексах будет отражено количество раз, сколько это слово встречалось в предложении.

При использовании такого метода при обучении может получится так, что модель переобучится на менее значимых словах (или шуме). Недостаток обусловлен тем, что данный метод не приоритезирует слова по значимости, а также не считает порядок слов в предложении.

3.3.2 TF-IDF

Чтобы дополнительно повысить или понизить приоритет слов в выборке можно воспользоваться скорингом TF-IDF, совместно используя его с мешком слов. Этот метод снижает значимость словам, которые являются шумом или встречаются слишком часто. А ключевым словам приоритет будет повышен. После такого анализа модель будет работать качественнее.

3.4 Инспектирование

После того, как получена рабочая модель, необходимо произвести анализ ошибок, которые остались после обучения, чтобы ещё улучшить качество модели. Конечно, ни одна модель не сможет достичь уровня распознавания в 100%, однако улучшить качество можно за счёт решения неучтенных ошибок.

Для того, чтобы определить то, какие ошибки необходимо учесть при обучении модели необходимо составить список с предсказаниями и ошибками. После этого можно сгруппировать предсказания со схожими ошибками и получить наиболее распространённую. Сделать это можно используя матрицу ошибок. Основные типы: ложно-положительные и ложно-отрицательные. После определения наиболее значимой, либо той, с которой хотелось бы в дальнейшем встречать как можно реже, можно принимать решения о том, как с ней можно бороться.

3.5 Применение семантики

После обучения, модель может столкнуться с ситуацией, что при обучении не было использовано достаточного количества слов, чтобы успешно справляться с поставленной задачей. Возникнет сложность в классификации, даже если при обучении были использованы похожие слова. Чтобы решить данную проблему, необходимо захватить семантическое (смысловое) значение слова.

3.5.1 Word2Vec

Word2Vec — это способ представления слов из корпуса текста в n-мерном измерении или в пространстве смыслов, где каждое слово занимает своё место, согласно контексту применения этого слова в тексте. Тем самым после обучения получается трёхсотмерное измерение, значения в которых занимают слова. Причем наиболее похожие по смыслу слова находятся рядом друг с другом. Word2Vec обрабатывает большое количество текстов, чтобы получить такую структуру.

3.5.2 GloVe

Инструмент, аналогичный Word2Vec, за исключением того, что размерность генерируемого вектора может быть выбрана самостоятельно (50, 100, 200 и 300).

4. Рекуррентные нейронные сети

Рекуррентные нейронные сети — это такие нейронные сети, которые имеют свойство запоминать свои состояния, для того, чтобы вычислять значения слоёв на основе предыдущих входящих значений. Такая нейронная сеть имеет обратные связи. Данное отличительное свойство сетей позволяет провести аналогию с биологическим мозгом, так как он тоже способен сохранять свои состояния и реагировать на события, опираясь на предыдущий опыт.

Основной класс задач, решаемый такими сетями - это обработка корпусов текстов, т.к такие сети специально сделаны для того, чтобы обрабатывать последовательности данных неизвестной длины.

Каждая такая сеть имеет в себе особую структуру, похожую на звенья или модули, идущие один за другим. И каждый такой модуль представляет собой слой с какой либо выходной функцией активации.

Кроме этого, такая сеть может выполнять задачу предсказания слова в предложении, так как способна понимать контекст предложения. Рекуррентная нейронная может обрабатывать не только тексты, но и временные ряды, так как может обрабатывать их в двух направлениях.

Способ обучения такой сети очень схож с обучением обычной нейронной сети. Разница лишь в том, что рекуррентная сеть требует использования метода обратного распространения ошибки во времени вместо привычного обратного распространения ошибки. Как следствие, на рекуррентные сети можно применять те же самые методы для решения проблемы переобучения, как и с обычными сетями [5].

4.1 LSTM

LTSM — это развитие классической рекуррентной сети, однако эта сеть во много раз превосходит оригинал, причём делает это в большом количестве задач.

Внутреннее строение сети LTSM очень похоже на структуру обычной рекуррентной сети, однако взаимодействие компонентов отличается. LTSM имеет 4 слоя сети, которые имеют собственные, особые инструкции, в отличие от обычной рекуррентной сети (рисунок 1).

Рассмотрим взаимодействие компонентов LTSM послойно. Самый первый слой рассчитывает подготовительные данные для считывания. На втором слое происходит подсчёт информации, которая будет записана в память ячейки. Затем подготовительная информация и то, что было записано в память линейно перемножаются, чтобы получить новые данные для ячейки памяти, и после всех вычислений, на выход передаётся значение.

Главная часть ячейки LTSM — это память или её состояние. В верхней части ячейки располагается линия, которая постепенно принимает вычисления от фильтров, один за другим, как транспортная лента на производстве.

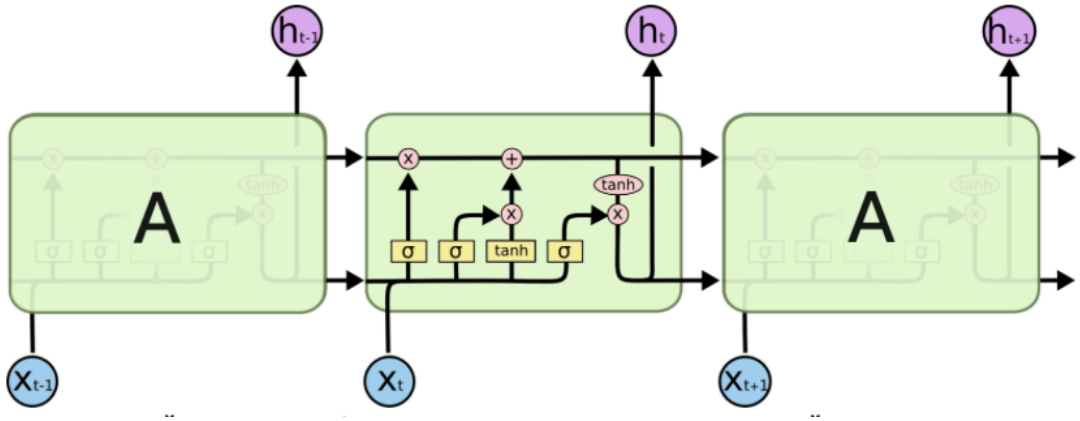
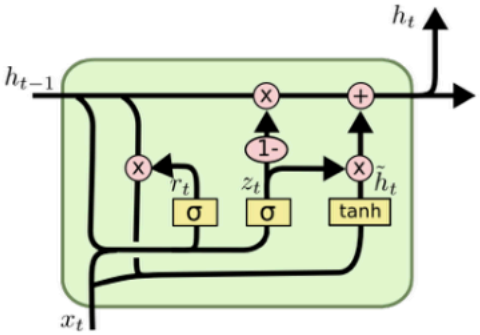


Рисунок 1 — Структура LSTM [4]

4.2 GRU

Существует аналог сети LTSM — это сеть GRU, что означает Gated Recurrent Unit. Этот вид сети использует меньше ресурсов и в нём находится меньше слоев, чем в LTSM, хотя фундамент у них общий [6]. Вид сети представлен на рисунке 2.



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Рисунок 2 — Структура GRU [4]

Заключение

Методы машинного обучения, хоть и находят все большее применение для различных задач обработки текстов, пока ещё остаются чрезвычайно сложными и трудоемкими для реального применения. Это объясняется не столько сложностью алгоритмов обучения, сколько, возможно, неудачными методологическими подходами к обучению.

Существует важная проблема проверки правильности работы обученной программы. Весьма важно, чтобы сама программа могла «понимать», что она не может справиться с задачей. Такое «понимание» может базироваться на том обстоятельстве, что для какого-либо шага нет однозначного решения или имеет место противоречие, конфликт некоторых правил. В этом случае программа должна запрашивать новые примеры или дополнительные знания экспертов-лингвистов.

Литература

1. Текстология [Электронный ресурс]. — Электронные данные. — Режим доступа: http://www.textologia.ru/slovari/lingvisticheskie-terminy/estestvenniy-yazik/?q=486&n=580. — Дата доступа: 10.02.2020

2. ПостНаука [Электронный ресурс]. — Электронные данные. — Режим доступа: https://postnauka.ru/video/92510. — Дата доступа: 15.02.2020

3. Хабр [Электронный ресурс]. — Электронные данные. — Режим доступа: https://habr.com/ru/company/oleg-bunin/blog/352614/. — Дата доступа: 15.03.2020

4. Хабр [Электронный ресурс]. — Электронные данные. — Режим доступа: https://habr.com/ru/company/wunderfund/blog/331310/. — Дата доступа: 19.03.2020

5. Научкор [Электронный ресурс]. — Электронные данные. — Режим доступа: https://nauchkor.ru/pubs/rekurrentnye-neyronnye-seti-v-zadache-analiza-tonalnosti-teksta-587d36595f1be77c40d58d52. — Дата доступа: 20.03.2020

6. Moluch [Электронный ресурс]. — Электронные данные. — Режим доступа: https://moluch.ru/archive/95/21426/. — Дата доступа: 25.03.2020