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 | 3 |
| 1. Что такое естественный язык? | 5 |
| 2. Возможные проблемы | 7 |
| 3. Способы решения | 8 |
| 3.1 Сбор данных | 8 |
| 3.2 Очистка данных | 9 |
| 3.3 Выбор представления данных | 9 |
| 3.3.1 «Мешок слов» | 9 |
| 3.3.2 TF-IDF | 10 |
| 3.4 Инспектирование | 10 |
| 3.5 Применение семантики | 11 |
| 3.5.1 Word2Vec | 11 |
| 3.5.2 GloVe | 11 |
| 4. Рекуррентные нейронные сети | 12 |
| 4.1 LSTM | 13 |
| 4.2 GRU | 14 |
| Заключение | 15 |
| Литература | 16 |

**Введение**

Обработка естественного языка - важное направление исследований в области искусственного интеллекта. Истоки этого направления идут с тех пор, как появились первые вычислительные машины с идеей, что эти машины могут быть использованы для решения задач, связанных с естественным языком, который используют в своей жизни люди.

Одной из первых подобных задач являлась задача машинного перевода - автоматического перевода текста с одного языка на другой с помощью вычислительной машины. В середине 1950-х годов успешными оказались эксперименты по переводу с русского языка на английский.

Второй не менее важной задачей стояло создание диалоговых систем, которые ведут с человеком диалог на понятном для него естественном языке. В те же годы начали появляться первые подобные системы.

Третьей задачей служило создание систем по принципу «вопрос - ответ», которые должны были точно отвечать на вопрос человека, который, опять же, представлен в форме текста на естественном языке.

В современном мире актуальность направления прежде всего связана с необходимостью обрабатывать большие объемы информации, накопленной за последние годы или даже десятки лет. Тем самым круг задач по обработке естественного языка стал намного шире:

- задачи распознавания речи,

- задачи синтеза речи,

- задачи информационного поиска,

- задачи классификации текстов,

- задачи кластеризации текстов,

- задачи резюмированная текстов,

- задачи анализа информации из социальных сетей,

- задачи создания создания систем «вопрос-ответ».

В последнее время акцент все больше смещается на анализ сообщений из социальных медиа: тексты, которые пишутся обычными людьми. Так же активно анализ естественного языка применяется в медицине, социологии и психологии [2].

В разное время для выполнения данного спектра задач применялись логические, статические, математические, стохастические подходы: конечные автоматы, конечные преобразователи, мартовские цепи, логика предикатов, формальные грамматики и вероятностные подходы. Но начина с 2000-х годов данные задачи начали решать с помощью методов машинного обучения.

**1. Что такое естественный язык?**

Естественный язык — это язык в собственном смысле слова. Человеческий язык является важнейшим средством общения и выражения мыслей [1]. В контексте данной темы естественный язык можно рассматривать как текст.

С точки зрения информатики текст — это неструктурированная информация и последовательность символов. И это необходимо учитывать. По этой причине существует несколько возможных путей решения задач обработки естественных языков.

Первый — решать задачи условившись, что текст — это цепочка символов, байтов, слов и у нас нет никаких знаний о языке. Что в принципе изначально не совсем верно, так как язык — это строгая система, которая имеет свои уровни, фонетику, морфологию, синтаксис, семантику.

Второй — решать задачи, принимая во внимание знания, что такое язык, как он устроен, выполняя при этом лингвистический анализ текста.

Как правило, проводится морфологический анализ. То есть для слов устанавливаются некоторые инварианты или морфологические признаки. Очевидно, что для различных языков набор этих признаков может различаться. Далее проводится синтаксический анализ. Слова объединятся в синтаксические группы и между ними устанавливаются синтаксические зависимости.

Многие задачи обработки естественного языка могут быть решены без привлечения знаний о языке. Но все же эффективность решения подобных задач достигается именно с применением знаний о языке и выполнением лингвистического анализа текста.

В современном мире типичное лингвистическое исследование представляет собой следующую последовательность действий: отбираются тексты, которые планируется анализировать, создается корпус текстов, на основе предметной области пишутся некоторые правила или составляются словари идентификации. Эти правила и словари и используются для решения поставленной задачи.

Но существует и другой путь решения подобных задач: лингвист вместе с экспертом в предметной области размечают тексты, выделяют целевые конструкции или относят тексты к тем или иным классам, в зависимости от поставленной задачи. А далее методы машинного обучения самостоятельно выводят некоторые правила и модели, которые позволяют в дальнейшем их использовать для решения.

**2. Возможные проблемы**

Сегодня анализ естественного языка довольно широко используется в коммерческий системах, что говорит о высоком уровне развитости этого направления исследований.

Одна из основных проблем анализа естественного языка — это многозначность, существующая на всех уровнях. Многозначность может устранятся за счет контекста и некоторой регулярности в использовании конструкций языка.

Методы анализа текстов сильно зависят от предметной области, языка или жанра текста: анализ новостных сообщений не то же самое, что и анализ художественного текста или текста из социальных сетей.

Не смотря на наличие огромного количества различного рода публикация и обучающих руководств на данную тему, на сегодняшний день практически не существует универсальных и полноценных рекомендаций о том, как эффективно справляйся с задачами обработки естественного языка.

**3. Способы решения**

Сегодня существует много открытых библиотек и платформ для обработки естественного языка.

Наибольшую популярность имеет использование методов машинного обучения для снижения трудозатрат на размёту текстов; использование методов машинного обучения с учителем, без учителя или с частичным привлечением учителя.

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**

Любая задача машинного обучения начинается с данных. Для задач обработки естественного языка это могут быть посты в социальных сетях, твиты, комментарии, адреса электронной, отзывы почты и др.

Данные могут быть предоставлены открытыми ресурсами или же могут собираться самостоятельно. Иногда сбор данных является более затратной задачей, чем создание модели для какой-то конкретней задачи.

Если мы используем обучение с учителем, данные должны иметь метки. И, 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**

Основная причина, ради которой необходимо производить очистку данных — знание о том, что чистый датасет позволит вычленить значимые признаки и не переобучиться на нерелевантный данных. Обычны очистка данных заключается в следующем:

1. Удаление всех нерелевантный символов;

2. Токенизация текста;

3. Удаление нерелевантный слов;

4. Перевод всех символов в один регистр;

5. Совмещение слов (например, слов, написанных с ошибками или имеющих альтернативное написание);

6. Лемматизация (приведение различных форм одного и того же слова к словарной форме).

**3.3 Выбор представления данных**

В качестве ввода модели машинного обучения принимают числовые значения. Таким образом, необходимо как-то представить данные в нужном виде, чтобы алгоритм смог их понять.

***3.3.1 «Мешок слов»***

Мы можем построить словарь всех уникальных слов и ассоциировать уникальный индекс каждому слову в словаре. Таким образом, каждое предложение может быть отображено списком, длина которого равно числу уникальных слов в нашем словаре, а в каждом индексе в этом списке будет храниться число, равное тому, какое количество раз текущее слово встретилось в предложении. Но данная модель полностью игнорирует порядок слов в предложении.

На очень больших объемах данных может сложиться ситуация, когда модель может переобучаться на словах, которые являются шумом. Этот момент основан на том, что в мешке слов все данные являются равнозначными.

***3.3.2 TF-IDF***

Чтобы помочь модели сфокусироваться на значимых словах, можно использовать скоринг TF-IDF поверх «мешка слов». Идея TF-IDF заключается в том, что слова анализируются и им устанавливается некоторый приоритет. Например, слова, встречающие слишком часто и добавляющие шум понижаются в приоритете.

**3.4 Инспектирование**

Даже если удалось получить достаточно высокую точность модели, необходимо понять, какие типы ошибок совершает модель и с какими видами ошибок хотелось бы встречаться реже всего. Не существует такой модели, которая давала бы 100% результат, особенно если это касается задач обработки естественного языка.

Для того, чтобы понять, какие ошибки встречаются, можно составить матрицу ошибок, которая сравнивает предсказания, сделанные моделью с реальными данными. Таким образом можно будет понять, каких ошибок встречается больше. Основные типы: которые можно выделить с помощью матрицы ошибок — ложно-положительные и ложно-отрицательные. В зависимости от поставленной задачи можно попытаться снижать уровень той или иной ошибки.

**3.5 Применение семантики**

После обучения, модель может столкнутся 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 инструмент для поиска отображений для слов. Word2Vec обучается на огромном количестве текстов, запоминая, какие слова встречаются в схожих контекстах. После обучения Word2Vec генерирует вектор из 300 измерений для каждого слова в словаре, в котором слова со схожими значениями располагаются ближе друг к другу.

***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 — подкласс нейронный сетей с обратными связями, которые используют предыдущие состояния сети для вычисления текущего. Если проводить аналогию с человеческим мозгом, можно считать, что рекуррентные нейронные сети добавляют «память» к искусственным нейронным сетям. Данный класс нейронный сетей чаще всего применяют для задач классификации текста по причине того, что рекуррентные нейронные сети хороши для задач обработки последовательностей нефиксированный длины.

Любая рекуррентная нейронная сеть имеет форму цепочки повторяющихся модулей нейронной сети. Структура одного такого модуля очень проста. Например, это может быть один слой с функцией активации tanh.

Для последовательностей, отличный от временных рядов, модель RNN часто работает лучше, если она обрабатывает последовательность не только от начала до конца, но и в обратном направлении. Например, для задач предсказания следующего слова в предложении, полезно знать не контекст вокруг слова.

Для обучения рекуррентных нейронный сетей используется алгоритм обратного распространения ошибки во времени (backpropagation through time), который является вариантом алгоритма обратного распространения ошибки (backpropagation). Поскольку обратное распространение ошибки по времени использует внутри себя обычный метод обратного распространения ошибки, то в нем применимы те же методы для избежания проблемы переобучения [5].

**4.1 LSTM**

LSTM — модификация рекуррентной нейронной сети, которая во многих задачах значительно превосходит стандартную версию [4].

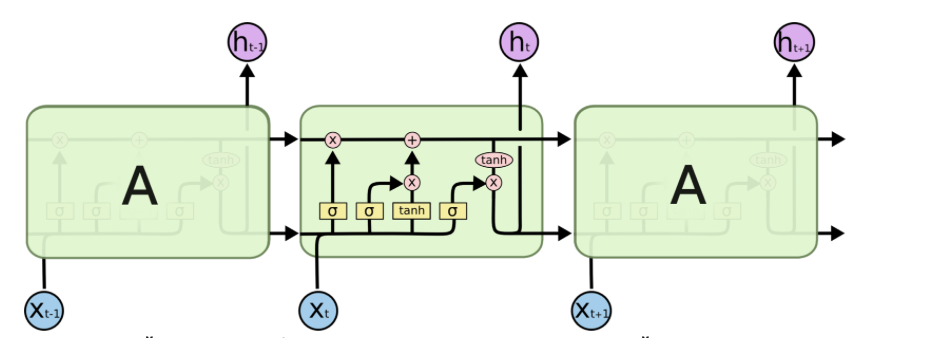
Структура LSTM также напоминает цепочку модулей, но сами модули выглядят немного иначе, чем у обычных рекуррентных сетей. Вместо одного слоя нейронной сети они содержат четыре, которые взаимодействуют между собой по особым правилам (рисунок 1).

Рисунок 1 — Структура LSTM [4]

Первый слой нейронной сети вычисляет множители к компонентам вектора памяти. На втором шаге вычисляется новая информация, которая называется наблюдением. Эта информация записывается в память. На третьем шаге вычисляется линейная комбинация того, что находится в памяти, и наблюдения. Таким образом получается новое состояние памяти. На последнем шаге вычисляется значение выходного нейрона.

Ключевой компонент LSTM — это состояние ячейки (cell state). Состояние ячейки напоминает конвейерную ленту, которая проходит напрямую через всю цепочку, участвуя лишь в нескольких преобразованиях.

**4.2 GRU**

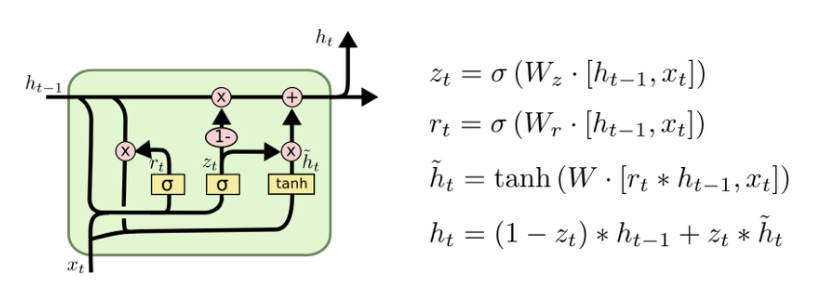
В 2014 году была представлена модель GRU (Gated Recurrent Unit), основанная на тех же принципах, что и LSTM, но использующая меньше фильтров и операций для вычисления [6]. Структура GRU приведена на рисунке 2.

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