Использование машинной лингвистики для анализа торгово-экономических новостей

УДК:004:339.5; ББК:73
DOI: 10.24412/2072-8042-2022-12-55-67

Андрей Николаевич СПАРТАК,
член-корреспондент РАН,
доктор экономических наук, профессор,
заслуженный деятель науки России,
Всероссийский научно-исследовательский конъюнктурный институт
(119285, Москва, ул. Пудовкина, 4) – директор;
Всероссийская академия внешней торговли
(119285, Москва, Воробьевское шоссе, 6А),
кафедра международной торговли и внешней торговли РФ – зав. кафедрой,
e-mail: Spartak@vniki.msk.ru

Иван Николаевич ОЛЕЙНИКОВ,
Всероссийская академия внешней торговли
(119285, Москва, Воробьевское шоссе, 6А),
Центр анализа данных – программист,
Email: oleynikov-in@ranepa.ru;

Александр Алексеевич ШАТИЛОВ,
Всероссийская академия внешней торговли
(119285, Москва, Воробьевское шоссе, 6А),
Центр анализа данных – аналитик,
Email: shatilov-aa@ranepa.ru;

Федор Михайлович ЯРОНСКИЙ,
Всероссийская академия внешней торговли
(119285, Москва, Воробьевское шоссе, 6А),
Центр анализа данных – аналитик,
Email: yaronskiy-fm@ranepa.ru

Аннотация

Мониторинг международных торгово-экономических новостей, выбор актуальных среди них и быстрый анализ текущей экономической ситуации в мире является необходимым процессом в работе экономических подразделений. Но ручной сбор необходимых статей среди огромного количества интернет-источников, визуальная проверка и классификация текстов по категориям представляется практически невозможной. Цель данной работы показать этапы сбора, подготовки новостных текстов к обработке, разметки и дальнейшей классификации торгово-экономических новостей с помощью моделей машинной лингвистики, описать, какие программные средства и подходы к моделированию были использованы в работе, продемонстрировать полученные результаты и практическое применение построенной системы.

Ключевые слова: обработка естественного языка, информационная система обработки текстов, торгово-экономические новости**,** модель, парсинг, разметка, сбор данных, обучение модели, база данных.

Using Machine Linguistics to Analyze Trade and Economic News

Andrey Nikolaevich SPARTAK,
Corresponding member of the Russian Academy of Sciences, Doctor of Economic Sciences, Professor, Honored Worker of Science of the RF, Russian Market Research Institute (VNIKI) (119285, Moscow, Pudovkina, 4) - Director, Russian Foreign Trade Academy (119285, Moscow, Vorob`evskoe shosse, 6A), Department of International Trade and Foreign Trade of the RF – the Head, E-mail: Spartak@vniki.msk.ru;

Ivan Nikolaevich OLEYNIKOV,
Russian Foreign Trade Academy (119285, Moscow, Vorobyovskoe Shosse, 6A), Data Analysis Center, Developer, Email: oleynikov-in@ranepa.ru;

Alexander Alekseevich SHATILOV,
Russian Foreign Trade Academy (119285, Moscow, Vorobyovskoe Shosse, 6A), Data Analysis Center, Analyst, Email: shatilov-aa@ranepa.ru;

Fedor Mikhailovich YARONSKIY,
Russian Foreign Trade Academy (119285, Moscow, Vorobyovskoe Shosse, 6A), Data Analysis Center, Analyst, Email: yaronskiy-fm@ranepa.ru

Abstract

Monitoring international trade and economic news, selecting the relevant ones and quickly analyzing the current state of the world economy is an important activity of economic departments. However, it is practically impossible to manually collect the necessary articles from a huge number of online sources, to visually check and then categorize them. The purpose of this paper is to show the stages of collection, preparation of news texts for processing, markup and further classification of trade and economic news using machine linguistics models, to describe which software tools and modeling approaches were used in the paper, and to demonstrate the results practical application of the built system.

Keywords: natural language processing, text analysis information system, trade and economic news, model, parsing, markup, data collection, model training, database.

ВВЕДЕНИЕ

Информационный взрыв, который произошел в последние годы, сказался на всех сферах нашей жизни. Появилось огромное количество интернет-ресурсов, новостных и прочих сайтов. Информация дублируется и перефразируется на разных сайтах. Такой рост информационных источников приводит к тому, что ручной сбор и обработка новостных статей для последующего анализа становится абсо-

лютно невозможным или же требует большого количества человеческих ресурсов. Но это не единственная проблема в работе с информацией. После сбора с сайтов статьи необходимо обработать: классифицировать, разделить по категориям, проверить дублирование или же перефразирование данных, создать приемлемое хранилище для быстрого обращения к нужным категориям. Все эти этапы сложно выполнимы вручную, без машинной обработки данных, с помощью которой выполняется большая часть работы, но порой также требуется и ручная вспомогательная обработка.

Информационная система обработки новостных сообщений должна обладать следующими возможностями:

- Сбор новостей;
- Автоматизированная обработка новостей в реальном времени и запись найденных сущностей в базу данных;
- Регулярное обновление релевантных новостных событий в базе данных;
- Обновление и улучшение применяемых алгоритмов для анализа корпуса текстов;
- Сервис для удобного просмотра результатов работы.

Для построения информационной системы обработки торговых новостей, надо было решить последовательно несколько задач:

- автоматизированный сбор и фильтрация новостных сообщений;
- разметка данных для разных моделей анализа текстов;
- разработка моделей для лингвистического анализа новостей;
- разработка бэкэнд части сервиса новостей;
- разработка интерфейса сервиса новостей.

АВТОМАТИЗИРОВАННЫЙ СБОР И ФИЛЬТРАЦИЯ НОВОСТНЫХ СООБЩЕНИЙ

Для автоматизированного сбора и фильтрации новостных сообщений мы брали информацию с RSS-лент, т.к. большинство новостных источников предоставляют полный текст своих статей в данном виде. RSS – это специальный унифицированный формат разметки, основанный на языке XML, в котором в машиночитаемом виде хранятся отдельно заголовок, текст статьи и различная мета-информация. Парсинг[1] новостей из RSS-лент осуществляется с помощью библиотеки newspaper3k. В тех случаях, когда RSS-лента не предоставляется, приходится разрабатывать специализированные парсеры, написанные на языке программирования Python, и использующие библиотеку requests. Кроме того, веб-страница хранит информацию в виде DOM-дерева (Document Object Model). Для языка Python обычным способом работы с DOM-структурой является библиотека beautifulsoup4.

Хранение текстов новостей может осуществляться с помощью отдельных файлов, тогда индексация и поиск по новостным сообщениям могут быть произведены посредством штатных средств операционной системы. Однако это не всегда удобно, поэтому часто оказывается полезным использование специализированных средств для хранения информации, а именно баз данных. В тех случаях, когда помимо самих новостных сообщений присутствует дополнительная информация об именованных сущностях, тональности, издательстве или авторах, удобно хранить её в реляционных базах данных, где каждая таблица описывает свой класс сущностей (статья, автор, издания, страна), и данные таблицы связаны между собой посредством реляционных связей. Популярным решением является СУБД PostgreSQL [2].

РАЗМЕТКА ДАННЫХ

Для решения задач анализа новостей торгово-экономической тематики разработаны несколько моделей, и для каждой необходима своя качественная разметка. В проекте решались следующие задачи:

1. Классификация по торгово-экономическим темам. Определение темы новости: санкции, инвестиции, совместные проекты и программы, внешняя торговля и специальные отношения, не классифицированные по типу, при этом в одной новости может быть несколько тем.
2. Распознавание сущностей. Например, геополитические сущности, организации, страны, товары и пр.
3. Классификация отношений между сущностями.
4. Классификация товаров по СМТК.

Для корректной разметки задачи классификации по темам была разработана подробная инструкция, описывающая правила, позволяющие однозначно относить новость к заданным темам.

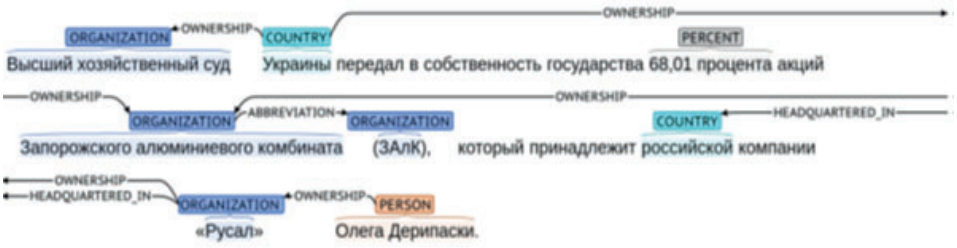

Рис. 1 - Пример интерфейса разметки в системе Brat
Fig. 1 - An example of a markup interface in the Brat system

Датасет[2] включал новости за 2022 год, предварительно отобранные моделью, чтобы исключить совсем нерелевантные новости.

Всего датасет содержит 9069 новостей на русском языке.

Также для проверки соответствия данных, размеченных в 2021 году и ранее, актуальной инструкции была подготовлена выборка из разных датасетов, содержащая всего 519 записей. Разметка была выполнена в системе Doccano[3].

Для распознавания сущностей и отношений в новостях были подобраны наборы данных с одинаковыми тегами разметки [1].

Датасеты содержат как разметку для извлечения именованных сущностей, так и для классификации и извлечения отношений между данными сущностями. Разметка сущностей и отношений между сущностями выполнена в системе разметке Brat[4] (см. рисунок 1).

Ввиду сложности разметки в такого рода задачах один текст размечался одним аннотатором.

Для решения задачи классификации сущностей-товаров по Системе Международной Товарной Классификации был подготовлен набор данных, включающий более 30 000 уникальных слов и словосочетаний на русском и английском языках. Хорошим источником уже размеченных данных является непосредственно классификатор СМТК, содержащий примеры по всем классам. Данные были получены парсингом pdf файла СМТК с помощью библиотеки pdfminer[5]

Англоязычные данные были получены переводом данных с русского языка моделью, обученной на парах текст-перевод. Все данные были проверены аннотаторами на соответствие сущности классу СМТК.

МАШИННО-ЛИНГВИСТИЧЕСКИЕ МОДЕЛИ

Модель классификации новостей по темам можно считать главной моделью, т.к. она определяет релевантность новостей, и к ее качеству предъявляются самые высокие требования. Для разработки модели классификации в качестве тренировочных данных были собраны несколько наборов данных:

- 6159 русских новостей, собранных с 2020 по 2021 год, размеченных экономистами;
- 7285 новостей на английском языке, собранных с 2017 по 2019 год, и размеченные в 2022 году;
- 8790 новостей на русском языке, собранных с конца 2021 по 2022 год, и размеченные в 2022 году.

Т.к. разметка проводилась в разное время, то в некоторых датасетах присутствовали неиспользуемые классы. Для приведения данных к нашей задаче некоторые классы были объединены в общий класс, а некоторые удалены полностью. Всего в датасетах было 13 классов.

Для проверки качества классификации модели, необходимо проверить ее работу на отложенной, валидационной, выборке. В качестве валидационной выборки для будущих моделей был размечен корпус английских и русских новостей за 2022 год. Корпус состоит из 3929 новостей.

Из-за того что часть новостей была размечена под другие задачи, а также сами новости датировались 2017-2020 годом, требовалась дополнительная проверка данных наборов. Данные могли содержать некоторый информационный и тематический сдвиг и не быть репрезентативными для поставленной задачи. Поэтому из русского и английского датасетов были выделены 251-356 (в зависимости от изначального размера) текстов для повторной разметки в 2022 году.

По итогам проверки наиболее репрезентативными оказались самые свежие (2021-2022 год) данные.

Разметка производилась несколькими аннотаторами, поэтому требовался механизм дополнительной оценки качества разметки. Сам датасет состоит из 6992 текстов, каждый из которых разметил от 1 до 7 аннотаторов. Для того чтобы дополнительно отсеять плохие данные, был применен механизм согласованности аннотаторов: остаются те результаты, в которых оценка одинаковая у нескольких аннотаторов. Дальнейшая разметка анотаторами показала, что с увеличением значения согласованности аннотаторов очень быстро падает и объем размеченных данных.

Поэтому было выбрано две различных стратегии решения задачи классификации новостей:

1. Использовать в качестве тренировочного набора повторно размеченные данные на момент 2022 года, а в качестве валидационной выборки использовать набор данных, размеченный заказчиком.
2. Использовать различные наборы старых/новых данных, а в качестве тестового набора использовать переразмеченные данные, и в качестве валидационного набора использовать данные, размеченные заказчиком.

Из-за того, что количество данных с согласованностью среди 3 и 4 аннотаторов достаточно небольшое относительно всего объема данных, было изначально предпринята идея использовать специализированные модели для работы с малыми данными (zero-shot). Но при тестировании этих моделей результаты показали, что специализированные модели для работы с малым количеством данных (zero-shot) уступают в качестве стандартным моделям, следовательно, объем данных достаточен для полноразмерных моделей (см. рисунки 2, 3).

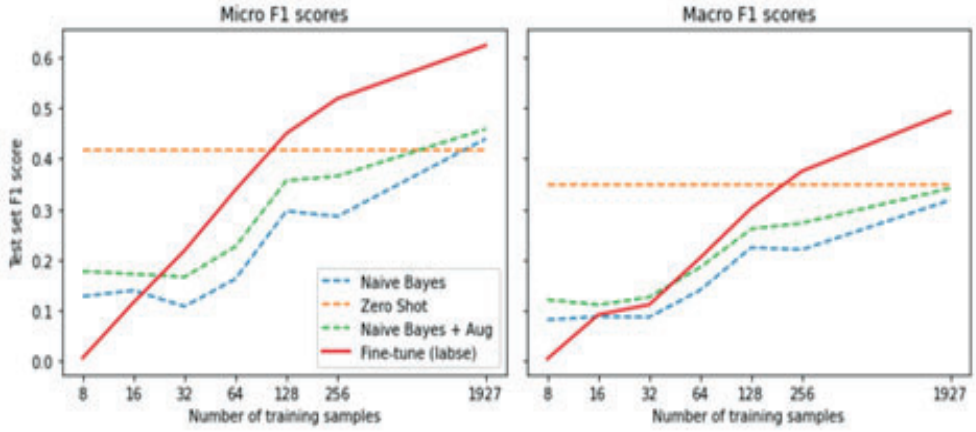

Рис. 2 - Оценка моделей с согласованностью 3
Fig. 2 - Evaluation of Models with Consistency 3

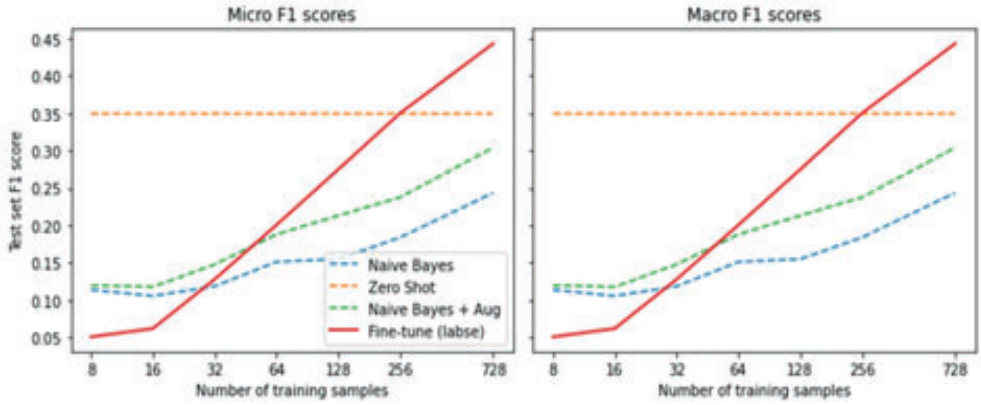

Рис. 3 - Оценка моделей с согласованностью 4
Fig.3 - Estimating Models with Consistency 4

В качестве полноразмерной модели была использована архитектура LABSE-en-ru[6], в качестве тестовой выборки данные, размеченные или собранные вручную, а в качестве валидационной – данные с согласованностью 4 среди аннотаторов.

Таблица 1

Метрики на тестовых данных

	precision	*recall*	*f1-score*	*support*
не по теме	0,92	0,73	0,82	369
другие отношения	0,15	0,71	0,25	14
торговля	0,75	0,63	0,69	68
проекты	0,69	0,47	0,56	57
санкции	0,77	0,80	0,79	55
инвестиция	0,67	0,42	0,52	38
micro avg	0,76	0,68	0,72	601
macro avg	0,66	0,63	0,60	601
weighted avg	0,83	0,68	0,74	601
samples avg	0,74	0,71	0,71	601

Следующий этап – распознавание сущностей и отношений между сущностями был выполнен с использованием модели распознавания сущностей. Учитывая, что новости могут быть как на английском, так и на русском языке, было принято решение использовать предобученную языковую модель BERT[7]. Для адаптации модели к нашим данным мы дополнительно дообучали модель на размеченных данных, описанных выше.

Модель показала удовлетворительный результат со значением метрики F1[8]=0.87, при этом качество распознавания сущностей, относящихся к классу «93 – Специальные операции и товары, не классифицированные по типу» не улучшилось, что, вероятно, связано с недостаточным количеством данных и особенностью данного класса, который, по сути, является достаточно широким.

Для решения задачи классификации отношений рассмотрены 2 подхода:

1. Модель исследователей из Института Южной Калифорнии [4] с разными методиками обучения. В наших экспериментах мы использовали три методики: entity-mask, typed-entity-marker, typed-entity-marker-punkt:

a. Entity-mask: эта техника вводит новые специальные маркеры [SUBJ-TYPE] или [OBJ-TYPE] для маскировки субъекта или объекта в оригинальном тексте, где TYPE заменяется соответствующим типом сущности. Такой подход предотвращает классификатор отношений от чрезмерного запоминания конкретных имен сущностей, что приводит к более обобщенным выводам.

b. typed-entity-marker: эта техника вводит размеченные именованные сущности в текст с помощью специальных токенов[9] <S:TYPE>, </S:TYPE>, <O:TYPE>, </O:TYPE>, где TYPE – это подходящая именованная сущность, размеченная с помощью классификатора.

c. typed-entity-marker-punkt – это вариант техники маркировки именованных сущностей и типов отношений, который маркирует область сущности и типы сущностей без введения новых специальных токенов. Идея заключается в том, чтобы заключить субъект и объект сущностей в рамки со знаками «@» и «#», соответственно. Для NER используются токены «*» для субъектов или «^» для объектов. Подобный подход позволяет уменьшить количество токенов в предложении, а также использовать редкие знаки в качестве маркеров NER и ER сущностей.

2. REBEL модель, представляющая задачу классификации отношений как задачу Seq2Seq, т.е. задачу машинного обучения, в которой входом и выходом модели является последовательность, в частности последовательность текста.

Наилучший результат показала модель с вариантом техники маркирования typed-entity-marker с результатом в 0.956 пунктов по метрике[10] качества на тестовом датасете.

Модель REBEL изначально обучена на данных Википедии (около 500 различных связей) и показывает "state-of-the-art" результаты на датасете TACRED. Дообученная на наших данных модель показала результаты хуже, чем стандартная модель.

Для задачи классификации товаров по СМТК обучили модель на 36400 названиях продуктов на русском, английском языках 71 классов СМТК + 1 специальный класс.

Корпус был дополнительно обработан. Так, из названий товаров были удалены различные символы, которые не влияют на качество классификации. Также были удалены дубликаты.

Были рассмотрены следующие модели:

- TF-IDF[11] (на словах) + Logistic Regression
- TF-IDF (на символьных n-граммах[12]) + Logistic Regression
- TF-IDF (символьные n-граммы) + Decision Tree/Random Forest/Gradient Busting/SVM
- Word2vec embeddings + average (F1 macro[13] = 0.53)
- Fine-tune BERT[14] (F1 macro = 0.53, модель плохо предсказывает классы с небольшим количеством примеров).

Также было рассмотрено преобразование данных. Так, приведение слов к их лемме улучшило метрики, а BPE-токенизация улучшений не дала.

Добавление данных с классом "not-product" дало прирост в качестве модели.

Стоит отметить, что использование данных из классификатора СМТК улучшило метрику оценки качества моделей на 3%.

СЕРВИС ОБРАБОТКИ НОВОСТЕЙ

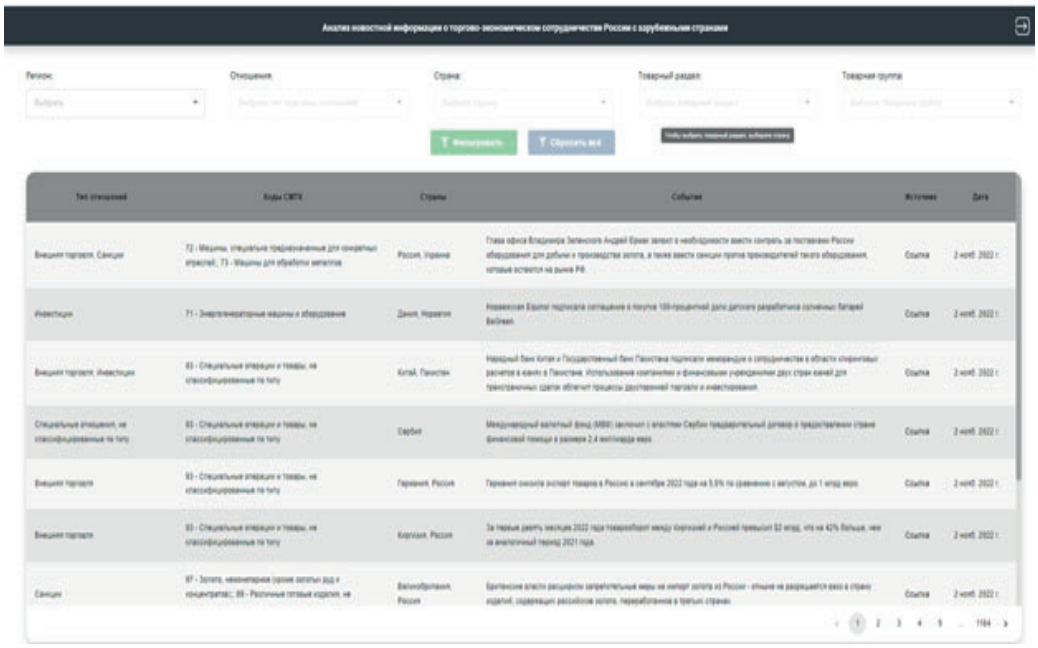

Рис. 4 - Интерфейс сервиса новостей.
Fig. 4 - News service interface

В рамках проекта анализа торгово-экономических новостей разработан сервис, позволяющий смотреть новости с возможностью фильтрации по регионам, странам, типу отношений и товарам по СМТК.

Адрес сервиса в сети интернет:

https://trade-news.vavt.ru

Сервис показа новостей состоит из двух частей:

- фронтэнд часть, написанная на языке программирования Typescript с использованием библиотеки React, и отвечающая за интерфейс и взаимодействие с пользователем (см. рисунок 4);
- бэкэнд часть, написанная на языке программирования Python с использованием фреймворка Django в качестве API сервиса, и отвечающая за обработку на сервере запросов, которые приходят от пользователей.

Архитектура бэкэнд части автоматизированной обработки новостей включает следующие элементы:

1. База данных Postgresql.
2. Сервис сбора новостей.

3. Сервис обработки новостей моделями.
4. Сервис с загруженными моделями.
5. Сервис показа новостей.

Все сервисы используют веб-сервер nginx для работы с запросами.

Данные с бэкэнд части передаются через API[15] в формате JSON[16].

Работа сервиса осуществляется в виде контейнера. Контейнеризация программных продуктов обеспечивает независимую работу модулей и исключает какие-либо конфликты между различными библиотеками и программами. В данной работе контейнеризация выполняется с помощью программы Docker.

При сборке образа контейнера выполняется установка библиотеки, написанной специалистами ЦАД, которая позволяет скачивать данные из объектных S3-хранилищ. С помощью этой библиотеки осуществляется скачивание всех моделей, необходимых для работы сервиса.

При завершении сборки модели загружаются в оперативную память.

Работа с моделями выполняется через API, который разработан с помощью библиотеки FastAPI.

Каждой модели в сервисе соответствует свой модуль:

- Модуль классификации новостей также позволяет получать вектор новости с помощью атрибута feature_extractor. Векторы новостей используются в сервисе для семантического поиска дубликатов новостей.
- Модуль распознавания сущностей.
- Модуль распознавания отношений между сущностями.
- Модуль классификации сущностей-товаров по СМТК.

Для получения предсказаний текст подается в метод infer соответствующего класса. Модель выдает предсказания в виде словаря, где ключами являются названия классов, а значениями вероятность этого класса.

Обращение к вышеописанным модулям осуществляется в главном модуле, в котором публикуется метод infer, через который можно обращаться к моделям. Адрес моделей имеет следующий шаблон:

https://[host]:[port]/infer/[app_name],

где host – адрес хоста, на котором разворачивается сервис;

port – порт хоста;

app_name – тип модели или задачи (например, clf_news, ner, relation_extraction).

Таким образом, конвейер обработки новостных данных состоит из следующих этапов:

1. Предварительно все новости собираются из RSS-лент и с помощью специализированных скрейперов, программ получения веб-данных путем извлечения их со страниц веб-ресурсов. Новости собираются раз в час и записываются в Postgres-базу данных вместе с метаинформацией.

2. Отдельный процесс опрашивает базу данных и ищет ранее необработанные новости для выделения новостей торгово-инвестиционной тематики с помощью модели машинного обучения.

3. Для новостей торгово-инвестиционной тематики производится дальнейший анализ именованных сущностей и отношений между ними для выделения товаров, инвестиционных договоров и соглашений, объектов инфраструктуры, а также отношений между данными видами сущностей и субъектами в виде стран и организаций. Результаты анализа записываются в базу данных в соответствующие таблицы.

4. Производится лемматизация[17] извлеченных сущностей и их сопоставление с нормализованными сущностями данного типа из базы данных (т. е. сопоставляем выделенную именованную сущность «бананов» со значением «банан» типа PRODUCT и соответствующим кодом СМТК из базы данных).

5. Для сущностей типа PRODUCT определяется их принадлежность в Международной торговой классификации товаров. Результаты анализа записываются в базу данных в соответствующие таблицы.

6. На регулярной основе, 1 раз в сутки, производится обновление материализованного представления новостных событий в базе данных.

7. Веб-сервис на основе фреймворка Django предоставляет доступ к результатам автоматического анализа новостей через API и состоит из двух частей:

- Открытой. Доступ для просмотра для всех посетителей сайта. Интерфейс показывает релевантные, проверенные вручную новости.

- Закрытой. Доступ возможен только по логину и паролю. Позволяет редактировать и проверять автоматически собранные новости и сохранять их в таблицу с релевантными новостями.

8. Кроме того, доступ можно получить посредством написания запросов и просмотра таблиц в самой базе данных.

Такая структура позволяет получать разные выборки по торгово-экономическим новостям, фильтровать новостные статьи по разным параметрам: локация, тема, сущности, отношения, СМТК.

Как мы продемонстрировали выше, несмотря на то что обработка языковых данных является непростой и трудоемкой задачей, машинная лингвистика развивается и позволяет очень сильно упростить масштабные задачи по сбору и обработке новостных данных. Даже несмотря на необходимость сочетать машинное обучение с ручным трудом, происходит значительная экономия времени, а также результаты и объём выборок несопоставимы с тем, что можно собрать и обработать только усилиями человека.

ПРИМЕЧАНИЯ:

1 Парсинг – сбор и сортировка данных по заданным критериям.

2 Датасет – обработанная и структурированная информация в табличном виде.

3 Doccano – инструмент для разметки данных с открытым исходным кодом.

4 BRAT – система для разметки сущностей и отношений в тексте https://brat.nlplab.org/

5 https://github.com/euske/pdfminer

6 LABSE – архитектура нейросетевой модели глубокого обучения (Language-agnostic BERT Sentence Embedding, https://arxiv.org/abs/2007.01852)

7 BERT – языковая модель, основанная на архитектуре трансформер, предназначенная для предобучения языковых представлений с целью их последующего применения в широком спектре задач обработки естественного языка.

8 F1 – мера точности теста, изменяется в диапазоне [0, 1], где 0 – не распознано ни одного значения, 1 – все значения распознаны верно.

9 Токен –- единица текста.

10 Метрика – количественный показатель, определяющий эффективность модели.

11 TF-IDF – (от англ. TF — term frequency, IDF — inverse document frequency). статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса.

12 n-грамма – последовательность из n элементов (в нашем случае токенов).

13 Macro – расчет показателей для каждого класса и поиск их невзвешенного среднего с учетом пропорции классов.

14 BERT – языковая модель, основанная на архитектуре трансформер, предназначенная для предобучения языковых представлений с целью их последующего применения в широком спектре задач обработки естественного языка.

15 API – механизмы, которые позволяют двум программным компонентам взаимодействовать друг с другом, используя набор определений и протоколов

16 JSON – текстовый формат обмена данными, основанный на JavaScript.

17 Лемматизация – приведение слова к его нормальной форме, лемме. Например, “модели” - “модель”, “торговали” - “торговать“

ИСТОЧНИКИ:

1. Gordeev, D., Davletov, A., Rey, A., Akzhigitova, G., Geymbukh, G.: Relation extraction dataset for the Russian language. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog” [Komp’iuternaia Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog”]. Russia Moscow, June 17–20, 2020 - Url: https://docs.yandex.ru/docs/view...

2. Барахнин В.Б. и др. Проектирование структуры программной системы обработки корпусов текстовых документов /Бизнес-информатика. – 2019. – Т. 13. – № 4, с. 60-72 @@ Baraxnin V. B. i dr. Proektirovanie struktury` programmnoj sistemy` obrabotki korpusov tekstovy`x dokumentov /Biznes-informatika. – 2019. – T. 13. – № 4, s. 60-72