# Международный журнал информационных технологий и энергоэффективности

Том 6 Номер 4 (22)

2021

Международный журнал информационных технологий и энергоэффективности
Т.6 №4 (22) 2021 г.

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

## СОДЕРЖАНИЕ / CONTENT

## ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ


| | | |
|---|---|---|
| 1. | **Макаров И.О., Попрыгин А.Ю.** Выбор беспроводной технологии передачи данных и учет ее особенностей при обмене информацией на малых или средних расстояниях<br><br>**Makarov I.O., Poprygin A. Yu.** Selection of wireless data transmission technology and taking into account its features when exchanging information over short or medium distances | **3** |
| 2. | **Антонов А.А., Быков А.Н., Чернышев С.А.** Обзор существующих способ формирования онтологии предметной области при моделировании<br><br>**Antonov A.A., Bykov A.N., Chernyshev S.A.** Review of the existing methods of forming the ontology of the scope in modeling | **12** |
| 3. | **Балашов О.В., Букачев Д.С.** Подход к определению качественных характеристик объектов<br><br>**BalashovO.V., Bukachev D.S.** Approach to determining the qualitative characteristics of objects | **18** |
| 4. | **Кузьмин А.И.** Обзор и сравнение популярных инструментов для обработки естественного языка<br><br>**Kuzmin A.I.** Review and comparison of popular tools for natural language processing | **24** |

Макаров И.О., Попрыгин А.Ю. Выбор беспроводной технологии передачи данных и учет ее особенностей при обмене информацией на малых или средних расстояниях// Международный журнал информационных технологий и энергоэффективности. – 2021. – Т. 6 № 4(22) с. 3–11

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.73

# ВЫБОР БЕСПРОВОДНОЙ ТЕХНОЛОГИИ ПЕРЕДАЧИ ДАННЫХ И УЧЕТ ЕЕ ОСОБЕННОСТЕЙ ПРИ ОБМЕНЕ ИНФОРМАЦИЕЙ НА МАЛЫХ ИЛИ СРЕДНИХ РАССТОЯНИЯХ

**Макаров И.О., Попрыгин А.Ю.**
*Филиал ФГБОУ ВО «Национальный исследовательский университет «МЭИ» в г. Смоленске, Россия, (214013, г. Смоленск, Энергетический проезд, 1), e-mail: Knyaghichigor@mail.ru*

**В статье на примере рассматривается процесс выбора беспроводной технологии передачи данных на малых и средних расстояниях. В качестве примера взята автономная беспроводная система управления запирающими устройствами. В статье приведен алгоритм выбора беспроводной технологии передачи данных, рассмотрены основные характеристики наиболее распространенных в настоящий момент беспроводных технологий. Основное внимание уделяется вопросам безопасности и энергоэффективности, так как выбор беспроводной технологии осуществляется для автономной распределённой системы управления запирающими устройствами. Данная статья будет полезна специалистам, начинающим разработку беспроводной системы**

Ключевые слова: беспроводные технологии, BLE v 4.0, беспроводной контроллер для замков.

# SELECTION OF WIRELESS DATA TRANSMISSION TECHNOLOGY AND TAKING INTO ACCOUNT ITS FEATURES WHEN EXCHANGING INFORMATION OVER SHORT OR MEDIUM DISTANCES

**Makarov I.O., Poprygin A. Yu.**
*Smolensk Branch of the National Research University "Moscow Power Engineering Institute", Smolensk, Russia (214013, Smolensk, Energetichesky proezd, 1), e-mail: Knyaghichigor@mail.ru*

**The article examines the process of choosing a wireless personal area net. An autonomous wireless control system for locking devices is taken as an example. The article provides an algorithm for choosing a wireless technology for data transfer, considers the main characteristics of the most common wireless technologies at the moment. The focus is on the topic of security and energy efficiency, as the choice of wireless technologies is carried out for an autonomous distributed control system for locking devices. This article will be useful to those professionals who are starting to develop a wireless system.**

Keywords: wireless technologies, BLE v 4.0, wireless lock controller.

При обмене данными на малых или средних расстояниях можно использовать следующие беспроводные технологии: NFC (на малых расстояниях); Bluetooth, ZigBee, Thread и др. (WPAN технологии); Wi-Fi (WLAN технология). В большинстве случаев выбор беспроводной технологии передачи данных зависит от особенностей системы, в которой она будет использоваться. Поэтому рассмотрим все на конкретном примере.

В качестве примера возьмём автономную распределенную систему управления запирающими устройствами, которая позволяет управлять замками и защёлками со смартфона. Для выбора беспроводной технологии передачи данных нужно:

1. Проанализировать особенности системы, где нужно применить выбранную технологию.
2. Рассмотреть доступные технологии.
3. Руководствуясь выделенными особенностями системы на основе характеристик беспроводных технологий, выбрать наиболее подходящую из рассмотренных технологий.
4. Проанализировать выбранную технологию на наличие «слабых» мест в контексте рассматриваемой системы.
5. В случае обнаружения «слабых» мест, найти пути их решения или обхода.

К беспроводной технологии передачи данных в системе управления электромеханическими замками можно выдвинуть следующие требования:

- близкая (до 20 м.) дальность действия;
- высокая безопасность (так как происходит управление замками);
- высокая скорость соединения (так как система распределенная предполагается частое соединение/разъединение пользователей с системой);
- скорость обмена не ниже 100 кбит/с (передача данных между контроллером и пользователем не является потоковой, а осуществляется пакетами достаточно малого размера, поэтому требования к скорости относительно невысоки);
- низкое энергопотребление (так как система является автономной);
- наличие топологии соединения «точка-точка» или «звезда».

Наиболее важной из представленных требований в контексте рассматриваемой системы является безопасность обмена данными.

В современных смартфонах существует несколько беспроводных технологий передачи данных, основные из них и наиболее распространённые:Wi-Fi, Bluetooth, NFC. В таблице 1 представлены краткие описания и примеры конкретных решений для Wi-Fi, Bluetooth LE и NFC.

Таблица 1 – Краткое описание и примеры конкретных решений для Wi-Fi, Bluetooth LE и NFC[1]

| | Название модуля | Протокол | Дальность[2] | Потребляемый ток | Скорость соединения | Скорость обмена данными | Топология соединения | Цена[3] | Цена на российских ресурсах[4] |
|---|---|---|---|---|---|---|---|---|---|
| *Bluetooth* | | | | | | | | | |
| Общая характеристика | - | Bluetooth Classic | До 100 м. | До 100 мА. | 3-4 сек. [1] | До 2,1 Мбит/сек. | «звезда», «точка-точка» | - | - |
| | - | BLE v. 4.x | До 100 м. | До 5 мА. | 1-2 сек. [1] | До 270 Кбит/сек. [2] | «звезда», «точка-точка» | - | - |
| Примеры pcb модулей | E104-BT02 | BLE 4.2 | 70 м | Sleep mode – 6 мкА Work – 700 мкА | 1-2 сек. | 270 Кбит/сек. | «звезда», «точка-точка» | 2,08$ | 310 руб. |
| | E104-BT20 | Bluetooth v2.1+EDR | 50 | Work – 54,4 мА | 3-4 сек. | До 2,1 Мбит/сек. | «звезда», «точка-точка» | 2.24$ | 280 руб. |
| *Wi-Fi* | | | | | | | | | |
| Общая характеристика | - | - | До 100 м. | До 250 мА. | 5-6 сек. | До 600 Мбит/сек. [3] | «точка-точка», «звезда» | - | - |
| Примеры pcb модулей | E103-W01 | 802.11 b/g/n | 100 | Sleep mode – 900 мкА Work – от 15 мА до 170 мА. | 5-6 сек. | До 54 Мбит/сек. | «точка-точка», «звезда» | 1,50$ | 300 руб. |
| | ESP-03 | 802.11 b/g/n | 100 | Sleep – 860 мкА Work – до 215 мА | 5-6 сек. | До 54 Мбит/сек. | «точка-точка», «звезда» | 1,50 $ | 350 руб. |
| *NFC* | | | | | | | | | |
| Общая характеристика | - | - | До 10 см. | - | Доли секунды [1] | 106-424 кбит/с. [5] | «точка-точка» | - | - |
| Примеры модулей | TRF7970A | NFC-A NFC-B NFC-F NFC-V | До 10 см. | Sleep 1 мкА Work RX – до 10 мА Work RX and TX – до 130 мА. | Доли секунды | 106-424 Кбит/сек. | «точка-точка» | 1$ | 620 руб. |

[1] Часть показателей в общей характеристики и примеры WiFi, Bluetooth и NFC взяты из беспроводных pcb модулей (BLE и WiFi модули от E-BYTE [6], BLE и WiFi модули от Espressif Systems [7], NFC модули от Texas Instruments (TRF79**A)[8]).

[2] На открытой местности на высоте 2 м. (pcb антенна).

[3] Минимальная цена на Alibaba, AliExpress, eBay

[4] Минимальная цена за шт. на таких ресурсах как ChipDip и Platan.

С точки зрения безопасности, NFC является самой безопасной технологией из рассматриваемых, так как дальность его действия очень мала, из-за чего перехват данных обмена является практически невозможным. Wi-Fi является более защищенной технологией по сравнению с Bluetooth. Более высокая защищенность достигается за счет усложнения протоколов передачи данных и более сложным процессом соединения, что делает его более продолжительным.

Таким образом, исходя из требований, представленных в начале статьи, из рассматриваемых беспроводных технологий передачи данных больше всего подходит Bluetooth LE. Использование NFC неудобно из-за дальности действия (накладываются жесткие требования к расположению контроллера управления замком), несмотря на то, что Wi-Fi более защищен, потребление и скорость соединения также делают его использование неудобным.

Так как система автономная, то использование Bluetooth Classic менее предпочтительно, чем Bluetooth LE. Поэтому, выбирая из рассматриваемых технологий, предпочтительнее всего является Bluetooth LE.

Bluetooth LE удовлетворяет практически всем требованиям, предоставляемым в начале статьи. Так как наиболее важное требование рассматриваемой системы к беспроводной технологии является высокая защищенность обмена информацией, то необходимо рассмотреть вопрос безопасности технологии и возможности «атак» злоумышленников.

**Особенности Bluetooth LE**

Рассмотрим внутреннюю структуру Bluetooth LE.

Стек BLE состоит из двух основных частей: контроллера (Controller) и узла сети (Host). Контроллер включает в себя физический и канальный уровень и часто реализуется в виде системы-на-кристалле с интегрированным беспроводным трансивером. Часть стека, именуемая узлом сети реализуется программно на микроконтроллере приложений и включает в себя функциональность верхних уровней: уровень логической связи (Logical Link Control — LLC), протокол адаптации (Adaptation Protocol — L2CAP), протокол атрибутов (Attribute Protocol — ATT), протокол атрибутов профилей устройств (Generic Attribute Profile — GATT), протокол обеспечения безопасности (Security Manager Protocol — SMP), протокол обеспечения доступа к функциям профиля устройств (Generic Access Profile (GAP). Взаимодействие между верхней и нижней частями стека осуществляется интерфейсом Host Controller Interface (HCI). Схематично внутренняя структура BLE изображена на рисунке 1.

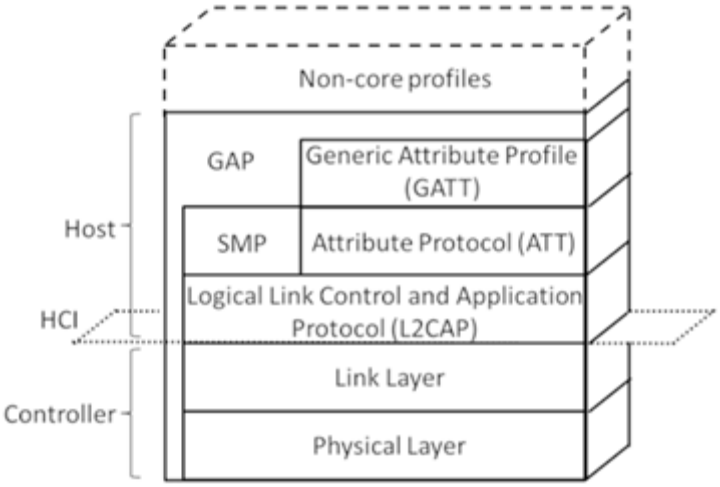


Рисунок 1 – Внутренняя структура BLE

На физическом уровне передача данных происходит в диапазоне частот ISM 2,4 ГГц с использованием частотной модуляции GFSK (Gaussian frequency-shift keying) на полосе частот 2 МГц, разделенной на 40 каналов. Три канала предназначены для широковещательной передачи данных (advertising channels), они выбраны таким образом, чтобы не пересекались с трёмя наиболее часто используемыми каналами в WiFi. Широковещательные каналы используются на GAP-уровне, а остальные 37 предназначены для передачи данных, используются на GATT-уровне [9].

Канальный уровень управляет скачкообразным переключением частоты, «рекламированием» (advertising), сканированием (scanning), инициированием подключения (initiating connections), подключением (connected). Канальный уровень BLE поддерживает шифрование и аутентификацию на основе алгоритма Cipher Block Chaining-Message Authentication Code (CCM) и блочного шифра AES-128. При использовании в соединении шифрования и аутентификации, к полезной нагрузке (PDU) добавляется четырехбайтное сообщение проверки целостности Message Integrity Check (MIC), после чего поля PDU и MIC шифруются. На данном уровне устройства подразделяются на ведомое (slave) и ведущее (master), причем ведомому устройству всегда доступна только топология «точка-точка».

Уровень L2CAP управляет потоком данных (разделяет данные, поступающие «сверху» на блоки нужного размера (22 байта)), реализует при необходимости повторную передачу данных.

GATT уровень представляет из себя иерархическую структуру хранения доступных клиенту данных [10]:

- Сервисы
- Характеристики
- Дескрипторы
- Возможные действия (read, write, notify, indicate)

ATT уровень реализует клиент-серверную архитектуру, предоставлением следующего протокола:

- команды от клиента серверу: Read (прочитать характеристику, сервис и т.п.), Write (изменить значение характеристики);
- команды от сервера клиенту: Notify (уведомить об изменении характеристики), Indicate (уведомление клиента об изменении характеристики и ожидание подтверждения о получении уведомления).

В роли сервера выступает текущее устройство независимо от того ведомое (slave) оно или ведущее (master).

SMP уровень отвечает за сопряжение устройств, обмен ключами при сопряжении и т.п. В общем случае сопряжение BLE устройств происходит в 3 этапа [10]:

1. Обмен данными о возможностях ввода/вывода устройств на канальном уровне.
2. Аутентификация соединения (формирование временного ключа для шифрования дальнейшего процесса сопряжения). Существуют несколько вариантов аутентификации:

a. Out Of Band – передача временного ключа по альтернативным каналам (например, NFC);

b. Presskey Entry – формирование ключа на основе вводимого пользователем пароля (последовательности из 6 цифр);

c. Just Work – не проводить процесс аутентификации (делает возможным атаку «человек посередине» MITM (Man In The Middle)).

3. Каждая из конечных точек соединения может передать другой конечной точке до трех 128-битных ключей, называемых Long-Term Key (LTK) – используется для шифрования на канальном уровне, Connection Signature Resolving Key (CSRK) – для подписи данных на уровне ATT, Identity Resolving Key (IRK) – для генерации частных адресов.

Уровень GAP определяет роль устройства, режим и процедуры обнаружения устройств и сервисов, управляет установлением соединения и безопасностью. В Bluetooth LE уровень GAP выделяет четыре роли для контроллера [10]:

1. Широковещательный (Advertiser). Может только передавать пакеты по рекламным каналам. Не поддерживает соединение с другими устройствами.
2. Наблюдатель (Scanner). Только прослушивает рекламные каналы (способен принимать пакеты, передаваемые Advertiser).
3. Периферийный (Peripheral). Способны поддерживать одно соединение с центральным устройством.
4. Центральный (Central). Способны поддерживать несколько соединений.

Роли центрального и периферийного узла предполагают, что устройство способно выполнять функции ведущего и ведомого, соответственно. Устройство может поддерживать несколько ролей, но одновременно активной может быть только одна из них. Обмен данными между устройствами происходит посредством изменения характеристик.

Исходя из вышесказанного о внутренней структуре Bluetooth LE технологии, можно сделать вывод, что безопасность обмена данными зависит от реализации уровней в тех или иных конкретных решениях.

**Возможные «атаки» на Bluetooth LE**

Рассмотрим некоторые возможные «атаки» на Bluetooth LE соединение [11]:

- Подслушивание (Sniffer): как следует из названия, подслушивание относится к стороннему устройству, прослушивающему данные, которыми обмениваются два сопряжённых устройства. Соединение между двумя сопряжёнными устройствами означает цепочку доверия. Цепь разрывается при удалении одного из устройств. Компания Nordic Semiconductor выпустила руководство для nRF Bluetooth Smart Sniffer, которое позволяет прослушивать даже зашифрованный канал связи между сопряженными устройствами [12]. Назначением данного устройства является отладка программ.
- Атаки «человек посередине» (MITM). Атаки «человек посередине» включают некоторое стороннее устройство, имитирующее легитимное, обманывая тем самым законные устройства. Имитатор заставляет поверить каждое из них в то, что они связаны друг с другом, когда на самом деле произошло подключение к имитатору (посреднику). Этот тип атаки позволяет злоумышленнику/имитатору получить доступ ко всем данным, которыми обмениваются устройства, а также манипулировать данными, удаляя или изменяя их, прежде чем они достигнут соответствующего устройства. Как было сказано раньше это возможно только при отсутствии фазы аутентификации при сопряжении.

Рассматриваемая система предоставляет высокие требования к безопасности, так как осуществляется управление замками. Из этого следует, что поверх основного стека протоколов необходимо добавить протокол обмена информацией между смартфоном и разрабатываемым устройством, который будет включать в себя свой «канальный» уровень с шифрованием передаваемой информации. Также нужно следить за тем, чтобы скорость обмена и соединения не сильно возрастала.

**Вывод**

Таким образом, на основе автономной распределенной системы управления запирающими устройствами был продемонстрирован процесс выбора беспроводной технологии передачи данных для управления устройством на малых или средних расстояниях.

В результате анализа доступных для смартфонов беспроводных технологий был сделан вывод, что для рассматриваемой системы оптимальным выбором является Bluetooth LE. Данная технология отличается низким энергопотреблением, близкой дальностью действия, а также приемлемыми скоростью соединения и обмена данными.

Ввиду особенностей рассматриваемой системы и выбранной технологии, необходимо предусмотреть следующие положения при использовании Bluetooth LE в качестве технологии передачи данных:

1. Соединение следует осуществлять без аутентификации, так как она слишком затягивает процесс соединения, который часто производится в распределённой системе.
2. Ввиду отсутствия механизма аутентификации и наличия повышенных требований к защищённости обмена данными, поверх основного стека протоколов следует предусмотреть дополнительный протокол обмена данными между смартфоном и контроллером замка, который будет включать в себя дополнительное шифрование данных перед отправкой по Bluetooth-каналу. Данная тема выходит за рамки текущей статьи.

**Список литературы**

1. Overview and Evaluation of Bluetooth Low Energy: An Emerging Low-Power Wireless Technology [Electronic resource]/ URL: https://www.mdpi.com/1424-8220/12/9/11734/htm (Date of treatment 23.09.2020).
2. Как выбрать лучший протокол Bluetooth для своего приложения [Электронный ресурс] / Режим доступа: https://spb.terraelectronica.ru/ news/6121 (Дата посещения 23.09.2020).
3. WiFi, Bluetooth или Zigbee – какой стандарт лучше? [Электронный ресурс] / Режим доступа: http://ua.automation.com/content/wifi-bluetooth-ili-zigbee-kakoj-standart-luchshe (Дата посещения 22.09.2020).
4. NFC: Разбор технологии Near Field Communication [Электронный ресурс] / Режим доступа: https://habr.com/ru/company/droider/blog/5041 96/ (Дата посещения 21.09.2020).
5. ГОСТ Р ИСО/МЭК 18092-2015 Информационные технологии (ИТ). Телекоммуникации и обмен информацией между системами. Коммуникация в ближнем поле. Интерфейс и протокол (NFCIP-1).
6. URL: http://www.ebyte.com/en/product-class.aspx (Date of treatment 23.09.2020).
7. URL: https://www.espressif.com/ (Date of treatment) (Date of treatment 23.09.2020).
8. URL: https://www.ti.com/ (Date of treatment) (Date of treatment 23.09.2020).
9. Bluetooth low energy technology [Electronic resource] / URL: https://www.compel.ru/wordpress/wp-content/uploads/2012/04/Bluetooth_ low_energy_technology.pdf (Date of treatment 23.09.2020).
10. Для мобильных стражей: беспроводной стандарт Bluetooth Low Energy в системах безопасности [Электронный ресурс] / Режим доступа: https://www.compel.ru/lib/53866 (Дата посещения 23.09.2020).
11. Что такое Bluetooth Low Energy (BLE) и как его взламывают [Электронный ресурс] / Режим доступа: https://hackware.ru/?p=9757 (Дата посещения 23.09.2020).
12. nRF Sniffer User Guide v 1.1 [Electronic resource] / URL: https://infocenter.nordicsemi.com/pdf/nRF_Sniffer_UG_v1.1.pdf (Date of treatment 23.09.2020).

**References**

1. Overview and Evaluation of Bluetooth Low Energy: An Emerging Low-Power Wireless Technology [Electronic resource]/ URL: https://www.mdpi.com/1424-8220/12/9/11734/htm (Date of treatment 23.09.2020).
2. Как выбрать лучший протокол Bluetooth для своего приложения [Электронный ресурс] / Режим доступа: https://spb.terraelectronica.ru/ news/6121 (Дата посещения 23.09.2020).
3. WiFi, Bluetooth или Zigbee – какой стандарт лучше? [Электронный ресурс] / Режим доступа: http://ua.automation.com/content/wifi-bluetooth-ili-zigbee-kakoj-standart-luchshe (Дата посещения 22.09.2020).
4. NFC: Разбор технологии Near Field Communication [Электронный ресурс] / Режим доступа: https://habr.com/ru/company/droider/blog/5041 96/ (Дата посещения 21.09.2020).
5. ГОСТ Р ИСО/МЭК 18092-2015 Информационные технологии (ИТ). Телекоммуникации и обмен информацией между системами. Коммуникация в ближнем поле. Интерфейс и протокол (NFCIP-1).
6. URL: http://www.ebyte.com/en/product-class.aspx (Date of treatment 23.09.2020).
7. URL: https://www.espressif.com/ (Date of treatment) (Date of treatment 23.09.2020).
8. URL: https://www.ti.com/ (Date of treatment) (Date of treatment 23.09.2020).
9. Bluetooth low energy technology [Electronic resource] / URL: https://www.compel.ru/wordpress/wp-content/uploads/2012/04/Bluetooth_ low_energy_technology.pdf (Date of treatment 23.09.2020).

10. Для мобильных стражей: беспроводной стандарт Bluetooth Low Energy в системах безопасности [Электронный ресурс] / Режим доступа: https://www.compel.ru/lib/53866 (Дата посещения 23.09.2020).
11. Что такое Bluetooth Low Energy (BLE) и как его взламывают [Электронный ресурс] / Режим доступа: https://hackware.ru/?p=9757 (Дата посещения 23.09.2020).
12. nRF Sniffer User Guide v 1.1 [Electronic resource] / URL: https://infocenter.nordicsemi.com/pdf/nRF_Sniffer_UG_v1.1.pdf (Date of treatment 23.09.2020).

Антонов А.А., Быков А.Н., Чернышев С.А. Обзор существующих способ формирования онтологии предметной области при моделировании // Международный журнал информационных технологий и энергоэффективности. – 2021. – Т. 6 № 4(22) с. 12–17

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.9

# ОБЗОР СУЩЕСТВУЮЩИХ СПОСОБОВ ФОРМИРОВАНИЯ ОНТОЛОГИИ ПРЕДМЕТНОЙ ОБЛАСТИ ПРИ МОДЕЛИРОВАНИИ

**[1]Антонов А.А., [2]Быков А.Н., [3,4]Чернышев С.А.**

*[1,2]Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, Россия (190000, ул. Большая Морская 67), e-mail: [1]aleksandr.antonov@dxc.com, [2]alexey_bykovoff@mail.ru*

*[3]Санкт-Петербургский государственный университет промышленных технологий и дизайна, Санкт-Петербург, Россия (191186, ул. Большая Морская 18)*

*[4]Санкт-Петербургский государственный экономический университет, Санкт-Петербург, Россия (191023, ул. Садовая 21), chernyshev.s.a@bk.ru*

**В статье рассматривается использование онтологий при создании систем поддержки принятия решений. Проводится анализ существующих подходов формирования онтологии предметной области. Особое внимание уделяется вопросу применения каждого подхода в той или иной сфере, а также его достоинствам и недостаткам**

Ключевые слова: онтология, предметная область, OWL, RDF, XML-схема, ER-модель, моделирование.

# REVIEW OF THE EXISTING METHODS OF FORMING THE ONTOLOGY OF THE SCOPE IN MODELING

**[1]Antonov A.A., [2]Bykov A.N., [3,4]Chernyshev S.A.**

*[1,2]Saint Petersburg State University of Aerospace Instrumentation Saint-Petersburg, Russian Federation (*190000, Bolshaya Morskaya str. 67), *e-mail: [1]aleksandr.antonov@dxc.com, [2]alexey_bykovoff@mail.ru*

*[3] Saint Petersburg State University of Industrial Technologies and Design, Saint-Petersburg, Russian Federation( 191186, Bolshaya Morskaya str. 18),*

[4]Saint Petersburg state university of economics, *Saint-Petersburg, Russian Federation (*191023, Sadovaya str. 21), *chernyshev.s.a@bk.ru*

**The article deals with the use of ontologies in creating decision support systems. The analysis of existing approaches to the formation of ontology of the scope is carried out. Special attention is paid to the issue of each approach in one or another field, and also to its advantages and disadvantages.**

Keywords: ontology, scope, OWL, RDF, XML-schema, ER-model, modeling.

**Введение**

В наше время становится популярным использование имитационных моделей [1] для создания сложных взаимосвязанных систем. Однако, это является трудозатратным процессом. Сложность обуславливается тем, что механизм принятия решения зависит от человеческого фактора, поэтому создать большую автоматизированную систему проблематично. При проектировании современных систем большую часть внимания уделяют формированию точной и качественной онтологии предметной области [2], так как она позволяет нивелировать фактор влияния человека на процесс принятия решений. Кроме того, и саму онтологию можно рассмотреть, как механизм принятия решений. Это обусловлено тем, что любая предметная область состоит из множества компонентов, или объектов, способных вступать во множество разнотипных отношений. Поэтому управление в этих системах можно рассматривать как упорядочивание, достижение слаженности отдельных объектов системы, приведение ее в соответствие с правилами предметной области и т. д.

Онтологии представляют собой описания знаний, сделанные достаточно формально, чтобы быть обработаны компьютерами. Основные ее компоненты: классы, отношения, функции, аксиомы и экземпляры.

Классы - это абстрактные группы, коллекции или наборы объектов. Они могут включать в себя экземпляры, другие классы, либо же сочетания и того, и другого [3].

Отношения - тип взаимодействия между понятиями предметной области.

Функции - конкретный случай отношения, в котором текущий элемент однозначной определяется предшествующим [4].

Аксиомы - это всегда истинные высказывания, которые используются в онтологиях для формирования ограничений на атрибуты и отношения, а также для проверки корректности информации.

Актуальность использования онтологии обуславливается рядом причин:

1. формирование коллективного понимания предметной области между людьми или программными агентами;
2. многократное использования знаний предметной области;
3. создание предметной области, не зависящей от ее компонентов;
4. исследование предметной области.

Важной составляющей в создании онтологии является способ ее формирования. Различные методы имеют ряд достоинств и ограничений, которые влияют на качество описания предметной области. Постоянный рост сложности систем, вынуждает более тщательно проводить формирование онтологии, ведь она позволяет более детально рассмотреть отдельные аспекты выбранной предметной области. Выбор метода описания предметной области является важным фактором в разработке системы, именно поэтому разработчик должен знать ограничения и особенности объекта исследования. И с учетом этих знаний правильно подобрать способ формирования онтологии.

**Обзор существующих способов формирования онтологии предметной области**

Разобравшись с областью применения онтологии, выделим основные способы формирования онтологии до некоторого времени развивавшихся отдельно:

Представление онтологии как формальной системы, в основе которых точные математические аксиомы.

Формирование онтологий как абстрактных понятий, которые могут быть выражены на естественном языке.

Первый тип является более практичным так как он оперирует точными математическими понятиями. Но в настоящее время математика отошла от практических основ и перешла к абстракции, поэтому почти каждая теория уже предполагает существование некой математической модели. Из-за чего дальнейшее развитие направления может привести к неверным выводам. Онтология позволяет создать математические модели из абстрактных понятий, в основе которых заложены аксиомы.

Хороший пример такого подхода — это теория игр [5]. Математической моделью является конфликтная ситуация в игре, основными объектами которой являются участники игры. Ключевая задача метода - это формирование стратегии принятия решений в условиях неопределенности, связанной с тем, что оппоненты преследуют противоположные цели, и результат любого действия зависит от хода противоположной стороны. Основное назначение системы — это принять оптимальное решение, которое реализует поставленную цель в наибольшей степени.

С другой стороны, абстрактные способы позволяют представить исследуемую область в виде упрощенной модели. Таким образом, происходит разделение области знаний на отдельные объекты и формирование отношений между ними, свойственные ей. К абстрактным способам относятся такие методы, как:

1. ER-модель;
2. XML схема;
3. Язык RDF;
4. Язык OWL.

ER-модель позволяет выделить ключевые объекты из предметной области и обозначить связи, которые могут существовать между ними, а также их ограничения [6]. Такой подход позволяет обеспечивать независимость модели данных от способов реализации. Основной характеристикой такой системы является слабая связь между объектами. Такой подход является узконаправленным и решает маленький диапазон задач, связанный с проектированием баз данных.

XML схема не разрабатывалась, как способ описания онтологий, но несмотря на это ее возможности позволяют представлять знания о предметной области в виде дерева знаний. XML схема описывает какие понятия должны быть включены в предметную область, как они связаны друг с другом, их свойства. Кроме того, она позволяет избежать внесение некорректных или излишних данных. Благодаря таким качествам, как простота, отсутствие ограничений на типы данных, структурный подход во взаимодействии с данными, она широко используется для описания онтологий веб ресурсов.

С ростом популярности использования онтологий появилась необходимость стандартизовать способы их представления. Это стало началом развития языков, которые могли бы использоваться в различных системах. Одним из таких языков является Resource Description Framework (RDF). Основная цель RDF – предложить стандартную модель данных «объект – предикат – значение» для метаданных. К примеру, утверждение «Листья имеют зеленый цвет»

в терминологии RDF будет представлено в виде: объект – листья, предикат – имеют цвет и значение – зеленый. Главным преимуществом RDF над XML является то, что все объекты в RDF являются отдельными сущностями. Поэтому этап определения объектов и их отношений выполняется в меньшей степени, чем это требуется в XML. Несмотря на ряд преимуществ, RDF уже длительное время используется в основном учеными. Причина состоит в том, что синтаксис RDF вызывает множество споров со стороны рядовых пользователей. По их мнению, формы записи сложны, а описание ресурсов слишком неудобны для применения.

Другое решение проблемы совместимости описаний было вынесено на обсуждение World Wide Web Consortium. Ими был предложен Web Ontology Language (OWL) – язык описаний онтологий, который представляет в действительности модели данных (объект - свойств). В основе данного языка лежит RDF, который сам по себе основан на XML. Поэтому OWL реализует структуру онтологии, позволяющую описывать классы, свойства и отдельные экземпляры. Важной особенностью является поддержка систем, использующих предыдущие версии языка.

По сравнению с RDF у OWL имеется множество преимуществ. Например, локальное ограничение области распространения. Это дает возможность накладывать ограничения на свойства для конкретного класса, что делает онтологию более детализированной. Также в OWL присутствуют следующие операции над множествами:

1. Пересечение
2. Дополнение
3. Объединение
4. Непересекаемость

Другим важным понятием является мощность. Она позволяется наложить на свойства ограничение на число использований. Примером применения мощности может служить выражение – «У автомобиля должно быть не менее 4 колеса», где на свойство количество колес накладывается ограничение – не менее 4. Это позволяет сделать вывод, что язык OWL обладает всеми нужными качествами для описания онтологий, основное назначение которых – анализ текстовых данных.

Для лучшей наглядности и удобства сравнения нижа приведена таблица с краткими характеристиками по всем четырем методам.

Таблица 1 – Таблица сравнения методов

| **Модель** | **Особенность** | **Пример** |
|---|---|---|
| ER-модель | Независимость модели данных от способов реализации. Слабые связи между объектами. | Высокоуровневое проектирование баз данных. |
| XML схема | Формирует дерево знаний об объектах, избегая излишних данных | Описание онтологий веб ресурсов |
| Язык RDF | Каждый объект является сущностью. Минимальное количество связей. | Системы принятия решений. |
| Язык OWL | Оперирует классами. Возможность использования ограничений и др. функций. | Формирование онтологии крупных систем. |

**Заключение**

Подводя итоги, можно сказать, что онтология является наиболее подходящим средством для описания предметной области в информационных системах, так как позволяет детально описать их объекты, свойства объектов, их ограничения и взаимодействия. Несмотря на то, что каждый из рассмотренных способов имеет свои недостатки, все они нашли свое применение на практике. Однако язык OWL имеет наибольшую популярность среди пользователей. Это обусловлено тем, что он является универсальным языком и способен подстроиться практически под любую систему.

Таким образом, выбор способа формирования онтологии является одним из наиболее важных этапов, так как каждый способ специализируется на определенной области применения, и неудачный выбор может привести к некорректной работе модели.

**Список литературы**

1. Имитационное моделирование. История, принципы, примеры. URL: https://ek-ek.jimdofree.com/петухин/моделирование2/10-имитационное-моделирование-история-принципы-примеры/ (Дата обращения 10.09.2021)
2. Онтология предметной области «Удобство использования программного обеспечения». URL: https://www.ispras.ru/proceedings/docs/2018/30/2/isp_30_2018_2_195.pdf (дата обращения 11.09.2021)
3. Онтологии и тезаурусы: модели, инструменты, приложения. URL: https://intuit.ru/studies/courses/1078/270/lecture/6845 (дата обращения 13.09.2021)
4. Введение в логику. URL: https://intuit.ru/studies/courses/13859/1256/lecture/23987 (дата обращения: 13.10.2021)
5. Определение понятий: онтология, концепт, отношение, аксиома. URL: http://window.edu.ru/resource/583/64583/files/Dobrov_978-5-9963-0007-5%2FGlavy1-2_cC0007-5.pdf (дата обращения 16.09.2021)
6. Теория игр: Введение. URL: https://habr.com/ru/post/163681/ (дата обращения 16.10.2021)
7. Место онтологий в единой интегрированной системе ран. URL: https://www.benran.ru/SEM/Sb_03/15.htm (дата обращения 18.09.2021)

**References**

1. Simulation modeling. History, principles, examples. URL: https://ek-ek.jimdofree.com/петухин/моделирование2/10-имитационное-моделирование-история-принципы-примеры / (Available at: 09/10/2021)
2. Ontology of the subject area "Usability of software". URL: https://www.ispras.ru/proceedings/docs/2018/30/2/isp_30_2018_2_195.pdf (Available at: 11.09.2021)
3. Ontologies and thesauruses: models, tools, applications. URL: https://intuit.ru/studies/courses/1078/270/lecture/6845 (accessed 13.09.2021)
4. Introduction to logic. URL: https://intuit.ru/studies/courses/13859/1256/lecture/23987 (Available at: 10/13/2021)

5. Definition of concepts: ontology, concept, relation, axiom. URL: http://window.edu.ru/resource/583/64583/files/Dobrov_978-5-9963-0007-5%2FGlavy1-2_cC0007-5.pdf (Available at: 16.09.2021)
6. Game Theory: An Introduction. URL: https://habr.com/ru/post/163681 / (accessed 16.10.2021)
7. The place of ontologies in the unified integrated system of the Russian Academy of Sciences. URL: https://www.benran.ru/SEM/Sb_03/15.htm (Available at: 18.09.2021)

Балашов О.В., Букачев Д.С. Подход к определению качественных характеристик объектов // Международный журнал информационных технологий и энергоэффективности. – 2021. – Т.6 №4 (22). – с. 18-23

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.05

# ПОДХОД К ОПРЕДЕЛЕНИЮ КАЧЕСТВЕННЫХ ХАРАКТЕРИСТИК ОБЪЕКТОВ

**[1]Балашов О.В., [2]Букачев Д.С.**
*[1]Смоленский филиал АО «Радиозавод», Россия, (214027, г. Смоленск, улица Котовского, 2), e-mail: smradio@mail.ru*
*[2]ФГБОУ ВО Смоленский государственный университет, Смоленск, Россия (214000, г. Смоленск, ул. Пржевальского, 4), e-mail: dsbuka@yandex.ru*

**В статье рассмотрен способ определения нечёткости логико-лингвистической шкалы при определении качественных характеристик объектов управления и предложена методика выбора оптимального множества значений качественного признака.**

Ключевые слова: информация, неопределённость, выбор, функция принадлежности, отношение предпочтения.

# APPROACH TO DETERMINING THE QUALITATIVE CHARACTERISTICS OF OBJECTS

**[1]Balashov O.V., [2]Bukachev D.S.**
*[1]Smolensk branch of joint-stock company "Radio factory", Russia, (214027, Smolensk, street Kotovskogo, 2), e-mail: smradio@mail.ru*
*[2]Federal State Educational Institution of Higher Education Smolensk State University, Smolensk, Russia (214000, Smolensk, street Przewalski, 4), e-mail: dsbuka@yandex.ru*

**The article discusses a method for determining the fuzziness of the logical-linguistic scale when determining the qualitative characteristics of control objects and suggests a method for choosing the optimal set of values of a qualitative attribute.**

Keywords: information, indeterminacy, choice, membership function, preference relationship.

В настоящее время значительное число компаний заказывают и внедряют в информационные системы в качестве приложений системы поддержки принятия решений (СППР) руководителей различных уровней. Разработка программного обеспечения таких СППР требует специфических подходов к их проектированию, особенно к проектированию механизмов обработки информации. Опыт работ по созданию СППР показывает, что существует проблема формализации информации и приведения её к виду, удобному для машинной обработки. При создании баз данных систем поддержки принятия решений, а точнее в процессе разработки логико-лингвистических шкал, решается задача выбора того или иного лингвистического значения из заданного множества при описании характеристик реальных объектов путём обработки экспертной информации. Решая данную задачу, эксперты

испытывают определенные затруднения. Перед проектировщиками встаёт задача: исходя из анализа структуры множества шкальных значений нечёткой лингвистической шкалы (НЛШ), разработать правило, используя которое эксперт оценивал бы объекты с минимальными трудностями.

Процесс оценки свойств объектов можно рассматривать как некоторую процедуру измерения, поэтому рассмотрим ряд основных понятий теории измерений [1].

Пусть $Q$ – множество объектов с определённым на нем набором отношений $R_i$, ($i \in I$). Пара $\langle Q; R_i, i \in I \rangle$ определяется как система с отношениями. Если $Q$ интерпретируется как множество объектов реального мира, то система с отношениями (СО) называется эмпирической, если под $Q$ понимается числовая ось – числовой СО [1].

Под шкалой понимается гомоморфизм $l$ эмпирической СО $\langle Q; V_i, i \in I \rangle$ в числовую СО $\langle A; W_i, i \in I \rangle$. Элементы множества $A$ при этом называются множеством шкальных значений шкалы $l$. Более строго, под нечёткой лингвистической шкалой понимается гомоморфизм $l$ эмпирической СО $\langle Q; V_i, i \in I \rangle$ в СО $\langle \tilde{R}; W_i, i \in I \rangle$, где $\tilde{R}$ – совокупность поименованных нечётких чисел, заданных на $R_i$; СО $\langle \tilde{R}; W_i, i \in I \rangle$ называется лингвистической числовой СО [2]. Пример множества шкальных значений НЛШ «Рост» достаточно подробно приведен в [3].

Интегральная характеристика НЛШ – степень нечеткости, её содержательный смысл – это степень сомнений (колебаний) эксперта, которые он испытывает при описании объектов в данной шкале. Считается, что для каждой точки опорного пространства существует хотя бы одно шкальное значение, имеющее в этой точке ненулевое значение функции принадлежности, то есть

$$\forall u \in U \ \exists i \ (1 < i < t) : \mu_{ai} > 0. \qquad (1)$$

Это ограничение не является искусственным, так как, если оно не выполняется, то множество точек $U' = \{u \in U' : \forall i \ (i=1, \ldots, t) \ \mu_{ai} \ (u) = a\}$ можно без ущерба удалить из $U$, то есть в качестве универсума рассматривать множество $U \backslash U'$. Это говорит о том, что ни одному реальному объекту в точках множества $U$ не соответствует ни одно шкальное значение, то есть шкала плохо определена.

Под степенью нечёткости НЛШ понимается степень нечёткости множества $\tilde{R}$ её шкальных значений. Для определения степени нечёткости лингвистической шкалы $l_t$ ($t$ – число шкальных значений), вводится понятие степени её нечёткости) в точке $u \in U$ ($\eta(l_t, u)$).

На содержательном уровне под $\eta(l_t, u)$ понимается степень сомнения (колебаний) эксперта в выборе того или иного шкального значения данной шкалы в рассматриваемой точке опорного пространства при описании реальных объектов. Очевидно, величина этих колебаний обратно пропорциональна разности между максимальным и ближайшим к нему значениями функции принадлежности $\mu_{a1}(u), \ldots \mu_{at}(u)$, шкальных значений.

Для иллюстрации приведем следующий пример. Пусть на складе имеется запас ресурса с максимальным значением в 48 единиц. При оценке запаса $u_1 = 20$ единиц и $u_2 = 45$ единиц эксперты практически без колебаний выбирают одно из шкальных значений «средний» и «большой» соответственно. При описании запаса ресурса $u_3 = 30$ единиц эксперт начинает колебаться в выборе одного из названных шкальных значений, и эти колебания становятся максимальными в точке $u_4 = 35$ единиц. Таким образом,

$$\eta(l_t, u) = 1 - (\mu_{ai_1}(u) - \mu_{ai_2}(u)), \quad (2)$$

где

$$\mu_{ai_1}(u) = \max_{1 \le j \le t} \mu_{a_j}(u), \quad (3)$$

$$\mu_{ai_2}(u) = \max_{1 \le j \le t;\, j \ne i_1} \mu_{aj}(u). \quad (4)$$

Из формул (2) – (4) следует, что

$$\eta(l_t, u) = 1 \Leftrightarrow \forall j\ (1 \le j \le t, j \ne i)\ \exists i\ (1 \le i \le t) : \mu_{ai}(u) = 1,\ \mu_{aj}(u) = 0),$$

$$\eta(l_t, u) = 0 \Leftrightarrow \exists i_1, i_2\ (1 \le i_1, i_2 \le \mathrm{t}) :\ \mu_{ai_1}(u) = \mu_{ai_2}(u) = \max_{1 \le j \le t} \mu_{a_j}(u).$$

Возвращаясь к примеру ($l_t = l_3$= «Запас ресурса»), получается:

$$\eta(l_3, u_2) = 0;\ \eta(l_3, u_3) = 0{,}5;\ \eta(l_t, u_4) = 1.$$

Таким образом, $\eta(l_t, u)$ отражает степень колебаний эксперта при выборе того или иного шкального значения в точке $u \in U$.

Необходимо отметить, что такое определение степени нечёткости лингвистической шкалы в точке универсума (формулы (2) – (4)), не учитывают влияния оставшихся ($t$–2) шкальных значений в данной точке и наиболее адекватно отражают случай пересечения функций принадлежности двух из них.

В силу ряда субъективных причин (ограниченного числа экспертов, различий в уровне их квалификации, неудачном выборе множества шкальных значений и др.) может наблюдаться ситуация плохого определения лингвистической шкалы $l_t$, когда в некоторой точке $u \in U$ могут пересекаться функции принадлежности трех, четырех и больше (вплоть до $t$) шкальных значений. Можно пойти двумя путями: усложнять формулы (2) – (4) или рассматривать некоторое подмножество всех шкал – $G$-шкалы, определяемые следующим образом:

$$l_t \in F \Leftrightarrow \forall u \in U:$$

либо $\exists\, \mu_{ai_1}(u), \mu_{ai_2}(u), \mu_{ai_3}(u)\ (1 \le i_1,\ i_2,\ i_3 \le t) :\ \mu_{ai_1}(u) \ne 0,\ \mu_{ai_2}(u) \ne 0,\ \mu_{ai_3}(u) \ne 0;$

либо $\forall \mu_{ai}(u)\ (\mu_{ai_1}(u) > \mu_{ai_2}(u) > \ldots \mu_{ai_k}(u) > 0,\ k \ge 3,\ 1 \le i_1, \ldots, i_k \le t)$:

$$\mu_{ai_1}(u) - \mu_{ai_3}(u) > c(\mu_{ai_1}(u) - \mu_{ai_2}(u)),\ c = const,\ c > 2.$$

$G$-шкалы не только удобны для теоретического анализа, но и наиболее часто встречаются на практике, так как описанные выше требования на содержательном уровне означают необходимость того, чтобы используемые понятия (шкальные значения) достаточно различались между собой семантически, не описывали одни и те же объекты, не являлись омонимами.

Используя понятие степени нечёткости шкалы в точке $u \in U$, определяется степень нечёткости $\nu(l_t)$ лингвистической шкалы как средняя степень её нечёткости во всем множестве $U$:

$$\nu(l_t) = \frac{1}{|U|}\int_U \eta(l_t, u)du\,, \tag{5}$$

где $|U|$ – мощность универсума $U$, $\eta(l_t, u)$ определяемая формулами (2) – (4).

Очевидно, что:

$\forall l_t$: $\nu(l_t) = 0 \Leftrightarrow \forall u \in U\ \forall j\ (1 \le j \le t)\ \exists i_1^*\ (1 \le i_1^* \le t,\ j \ne i_1^*)$ : $\mu_{ai_1}(u) = 1,\ \mu_{aj}(u) = 0$ (то есть в абсолютно чётком случае);

$\forall l_t$: $\nu(l_t) = 1 \Leftrightarrow \forall u \in U\ \exists i_1^*, i_2^*, (1 \le i_1^*, i_2^*, \le t)$ : $\mu_{ai_1^*}(u) = \mu_{ai_2^*}(u) = \max\limits_{1 \le j \le t} \mu_{aj}(u)$.

С помощью введенного понятия степени нечёткости НЛШ можно определить степень нечёткости множества как частный случай $\nu(l_t)$.

Совокупность результатов оценок некоторого объекта в НЛШ $l_{t_1}^1, \ldots, l_{t_k}^k$ называются его лингвистическим описанием по признакам с номерами $1, \ldots, k$ [5]. Множество лингвистических описаний объектов интерпретируется как база данных информационной системы СППР. Рассмотрим показатели качества работы информационной системы – это средние индивидуальные потери информации и шумы, возникающие при поиске информации в ней, под которыми понимается следующее.

При общении с СППР лицо, принимающее решения (ЛПР) формирует запрос (например, «Выдать описания всех объектов, имеющих значение характеристики «Запас ресурса», равное «НИЗКИЙ») и получает из базы данных СППР некоторое количество описаний объектов, удовлетворяющих поисковому предписанию. При этом, если бы ЛПР знал реальные значения характеристик, выданных на запрос объектов и объектов, описания которых хранятся в базе данных, он, возможно, забраковал бы часть выданных объектов (информационный шум), а часть объектов, описания которых не были выданы на запрос, наоборот, принял бы (потери информации). Механизм возникновения таких потерь и шумов связан с размытостью элементов шкалы, с тем, что источник информации и ЛПР могли бы выбрать для описания одного и того же реального объекта разные шкальные значения НЛШ. Такие потери информации и шумы можно назвать индивидуальными. Обозначим через $B(l_t)$ и $S(l_t)$ объемы средних индивидуальных потерь информации и шумов, возникающих при поиске информации по признаку с множеством значений, состоящим из имен шкальных значений НЛШ $l_t$.

Для подсчета объема средних потерь информации и шумов вводится упрощение: считается, что описание $I(Q)$ реального объекта $Q$, хранящегося в СППР, содержит только одну поисковую характеристику, имеющую $t$ значений $a_1, a_2, \ldots a_t$, представляющих собой множество имен шкальных значений лингвистической шкалы $l_t$. Для функций принадлежности, входящих в набор множества шкальных значений, кроме условия (1) выполняется требование ортогональности [5]:

$$\forall u \in U \sum_{j=1}^{t} \mu_{aj}(u) = 1 \tag{6}$$

Это предположение связано с выбором той или иной интерпретации понятия «функция принадлежности». По приведенной в [5] классификации такое требование определяет так называемую вероятностную интерпретацию, что, однако, не сводит данное понятие ни к функции распределения, ни к плотности вероятности [5, 6]. Выдвинутое требование выполняется или нет в зависимости от метода построения функции принадлежности, в частности, при следующем определении:

$$\mu_{aj}(u) = \frac{n_1}{n_1 + n_2},$$

где $n_1$ – число экспертов, относящих u к множеству $a_j$;

$n_1 + n_2$ – общее число экспертов.

При опросе возможны только ответы вида «да, $u \in a_j$» или «нет, $u \notin a_j$». Приведенные рассуждения позволяют сделать вывод о том, что ограничение (6) не является очень искусственным.

При проектировании СППР, для оценки объёма средних потерь информации и шумов в ходе работы с экспертами была сформулирована и доказана следующая теорема.

**Теорема**. Пусть $l_t$ – некоторая лингвистическая шкала, имеющая $t$ шкальных значений, функции принадлежности которых удовлетворяют условиям (1), (6). $\nu(l_t)$ – степень нечеткости, $B_x(l_t)$ и $S_x(l_t)$ – средние индивидуальные потери информации и шумы, возникающие при поиске информации по признаку с множеством значений $X$, совпадающим с множеством имен шкальных значений $l_t$; $U$ – универсум.

$N(u)$ – число объектов, описания которых хранятся в базе данных СППР, имеющих фактическое значение характеристики, равное $u$ – есть константа. Для ЛПР значения признака представляют одинаковый интерес, то есть вероятности запросов по каждому значению признака равны. Тогда

$$B_x(l_t) = S_x(l_t) = \frac{2N}{3t}\,\nu(l_t),\ N = const.$$

Кратко необходимо пояснить, что при её доказательстве использовалось (L–R)-представление нечётких множеств и формулы аналитической геометрии. При выполнении условий теоремы

$$\nu(l_t) = \frac{1}{|U|}\frac{1}{2}\sum_{j=1}^{t} d_j\ ;\ B_x(U) = S_x(U) = \frac{N}{3t}\sum_{j=1}^{t} d_j\ .$$

Отсюда следует, что уменьшение степени нечёткости на $d\%$ при фиксированном числе значений характеристики ведёт к такому же сокращению средних индивидуальных потерь информации и шумов, возникающих при работе с данной характеристикой. Уменьшение же степени нечёткости при одновременном увеличении числа значений характеристики делает эту зависимость ещё более сильной.

Таким образом, исходя из вышеизложенного, можно предложить следующую методику выбора множества значений качественного признака:

- сформировать все возможные множества значений признака;

- каждое множество значений признака представить в виде множества шкальных значений лингвистической шкалы;
- для каждого множества значений вычислить степень нечёткости признака по формулам (2) – (5);
- в качестве оптимального множества значений, минимизирующего неопределённость при описании объектов, выбрать то множество, для которого степень нечёткости минимальна;
- в качестве оптимального множества значений, повышающего качество поиска информации, выбрать то множество, для которого отношение степени нечёткости к числу значений признака минимально.

**Список литературы**

1. Ларичев О.И. Теория и методы принятия решений. – М.: Логос, 2002.
2. Борисов А. Н., Алексеев А.В., Крумберг О. А. и др. Модели принятия решений на основе лингвистической переменной. Рига, Зинатне, 1982.
3. Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений. М. Мир, 1976.
4. Балашов О.В., Букачев Д.С. Кондратова Н.В. Способы формализации задачи принятия решений для проектирования систем поддержки принятия решений // Международный журнал информационных технологий и энергоэффективности. – 2018. – Т.3, №1(7). – С. 25-34.
5. Аверкин А.Н., Батыршин И.З., Блишун А.Ф. и др. Нечёткие множества в моделях управления и искусственного интеллекта/ Под ред. Поспелова Д.А. М. Наука. Гл.ред.физ.-мат. лит., 1986.
6. Борисов А.Н., Алексеев А.В., Меркурьева Г.В. и др. Обработка нечёткой информации в системах принятия решений. М. Радио и связь, 1989.

**References**

1. Larichev O.I. Teoriya i metody prinyatiya reshenij. – M.: Logos, 2002.
2. Borisov A. N., Alekseev A.V., Krumberg O. A. i dr. Modeli prinyatiya reshenij na osnove lingvisticheskoj peremennoj. Riga, Zinatne, 1982.
3. Zade L. Ponyatie lingvisticheskoj peremennoj i ego primenenie k prinyatiyu priblizhennyh reshenij. M. Mir, 1976.
4. Balashov O.V., Bukachev D.S. Kondratova N.V. Sposoby formalizacii zadachi prinyatiya reshenij dlya proektirovaniya sistem podderzhki prinyatiya reshenij // Mezhdunarodnyj zhurnal informacionnyh tekhnologij i energoeffektivnosti. – 2018. – T.3, №1(7). – pp. 25-34.
5. Averkin A.N., Batyrshin I.Z., Blishun A.F. i dr. Nechyotkie mnozhestva v modelyah upravleniya i iskusstvennogo intellekta/ Pod red. Pospelova D.A. M. Nauka. Gl.red.fiz.-mat. lit., 1986.
6. Borisov A.N., Alekseev A.V., Merkur'eva G.V. i dr. Obrabotka nechyotkoj informacii v sistemah prinyatiya reshenij. M. Radio i svyaz', 1989.

Кузьмин А.И. Обзор и сравнение популярных инструментов для обработки естественного языка // Международный журнал информационных технологий и энергоэффективности. 2021– Т. 6 № 4(22) с. 24–33

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 681.3.019

# ОБЗОР И СРАВНЕНИЕ ПОПУЛЯРНЫХ ИНСТРУМЕНТОВ ДЛЯ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

**Кузьмин А.И.**
*Филиал ФГБОУ ВО «Национальный исследовательский университет «МЭИ»в г. Смоленске, Россия,(214013, г. Смоленск, Энергетический проезд), e-mail: timonkyz2@mail.ru*

**В работе представлено общее сравнение и анализ существующих инструментов для обработки естественного языка. Составлены диаграммы сравнения для более популярных библиотек для работы с естественным языком. Приведены преимущества и недостатки самых популярных инструментов.**

Ключевые слова: естественный язык, обработки текста, анализ текста

# REVIEW AND COMPARISON OF POPULAR TOOLS FOR NATURAL LANGUAGE PROCESSING

**Kuzmin A.I.**
*Smolensk Branch of the National Research University "Moscow Power Engineering Institute", Smolensk, Russia (214013, Smolensk, Energetichesky proezd, e-mail: timonkyz2@mail.ru*

**The article presents a general comparison and analysis of existing tools for natural language processing (NLP). Comparison diagrams for the more popular libraries for natural language processing are compiled. The advantages and disadvantages of the most popular tools are given.**

Keywords: natural language, text processing, text analysis

## Введение

Понятие Data Mining, появившееся в 1978 году, приобрело высокую популярность в современной трактовке примерно с первой половины 1990-х годов. До этого времени обработка и анализ данных осуществлялись в рамках прикладной статистики, при этом в основном решались задачи обработки небольших баз данных.

Data mining (DM) это процесс вычислительного извлечения новой информации из Big Data [1], а различные отрасли генерируют огромные объемы данных, открывая эру "больших данных". Это создает широкие возможности для разработки и внедрения новых алгоритмов интеллектуального анализа. Широкий спектр методов извлечения ценных сведений из различных типов и моделей данных подпадает под понятие "data mining".

Согласно определению в статье "Data clustering: a review", "кластеризация — это классификация шаблонов (наблюдений, элементов данных или векторов признаков) в группы (кластеры) без наблюдения" [2].

Классификация схожа с кластеризацией, поскольку она разделяет данные на группы, называемые классами, но в отличие от кластеризации, анализ классификации требует знания и спецификации того, как определяются эти классы.

Теория статистического обучения стремится "обеспечить основу для изучения проблемы вывода, то есть получения знаний, составления прогнозов, принятия решений или построения моделей на основе набора данных", - утверждает Буске и др. [3].

Заметным переходом, демонстрирующим мощь новых алгоритмов и данных, стало использование подходов интеллектуального анализа данных для изучения не только первичных характеристик, но и характеристик, специфичных для контекста. Например, первоначальные подходы к поиску данных, которые строили одну модель [4]. В отличие от этого, последние подходы изучают множество контекстно-специфических моделей, позволяя строить сети, специфичные для разнообразных процессов [5].

Text mining (TM) — это область интеллектуального анализа данных, целью которой является извлечение новой ценной информации из неструктурированных (или полуструктурированных) источников [6]. Text mining извлекает информацию из документов и агрегирует извлеченные фрагменты по всей коллекции исходных документов получения новой информации. Это предпочтительный взгляд на данную область, который позволяет отличить текстовый майнинг от обработки естественного языка (NLP) [7]. Таким образом, получив на вход набор документов, методы интеллектуального анализа текста стремятся обнаружить новые закономерности, взаимосвязи и тенденции, содержащиеся в документах. В достижении общей цели обнаружения новой информации помогают инструменты NLP, которые варьируются от относительно простых задач обработки текста на лексическом или грамматическом уровнях (таких как токенизация или тегирование части речи) до относительно сложных алгоритмов извлечения информации (таких как распознавание именованных сущностей (NER) для поиска концепций, нормализация для сопоставления их с их уникальными идентификаторами или извлечение отношений и системы анализа настроений). Чем выше сложность задачи, тем больше вероятность интеграции методов интеллектуального анализа данных (таких как классификация или статистическое обучение).

К подобластям text mining, кратко обобщенным, относятся:

- Информационный поиск(IR) занимается проблемой поиска релевантных документов в ответ на конкретную информационную потребность (запрос).
- NER лежит в основе автоматического извлечения информации из текста и занимается проблемой поиска ссылок на объекты (упоминаний) присутствующие в тексте на естественном языке, и их маркировки с указанием местоположения и типа.
- Идентификация именованных сущностей позволяет связать интересующие объекты с информацией, которая не указана в тексте.
- Извлечение ассоциаций - одна из высокоуровневых задач. Она использует результаты предыдущих подзадач для получения списка ассоциаций между различными сущностями, представляющими интерес

Всеобъемлющая проблема анализа текстов заключается в том, чтобы включить многочисленные доступные ресурсы знаний в конвейер NLP. Например, в медицинской области, в отличие от общей области поиска текстов, имеется доступ к большому количеству обширных, хорошо проверенных онтологий и баз знаний. Например, использование онтологий позволило использовать неструктурированные клинические записи для получения

практических данных о безопасности высокоэффективного непатентованного препарата для лечения заболеваний периферических сосудов [8].

**Основные этапы обработки естественного языка**

Можно выделить общие этапы, которые объединяют обработку текста разными инструментами. Для понимая процесса необходимо подробнее остановиться на первоначальных этапах:

- Стэмминг.
- Лемматизация.
- Тегирование частей речи.

Стэмминг — это процесс сокращения слова до его основы, т.е. корневой формы. Корневая форма не обязательно является словом сама по себе, но она может быть использована для образования слов путем присоединения нужного суффикса.

Например, слова рыба и рыбалка образуются от корня "рыба", что является правильным словом. С другой стороны, слова study, studies и studying превращаются в studi, что не является английским словом.

Чаще всего алгоритмы стемминга (они же стеммеры) основаны на правилах отсечения суффиксов. Наиболее известным примером является стеммер Портера, представленный в 1980-х годах и в настоящее время реализованный в различных языках программирования.

Традиционно поисковые системы и другие приложения применяют стемминг для повышения вероятности совпадения различных форм слова, рассматривая их почти как синонимы, поскольку концептуально они "принадлежат" друг другу.

Цель лемматизации - сгруппировать различные формы слова, называемые леммами. Этот процесс в чем-то схож со стеммингом, поскольку он объединяет несколько слов в один общий корень. Результатом лемматизации является правильное слово, и отсечение суффиксов не даст такого же результата. Например, лемматизатор должен преобразовать gone, going и went в go. Для достижения своей цели лемматизация требует знания контекста слова, поскольку процесс зависит от того, является ли слово существительным, глаголом и т.д. [9]

Метки частей речи (POS) — это процесс отнесения слова к его грамматической категории, чтобы понять его роль в предложении. Традиционными частями речи являются существительные, глаголы, наречия, союзы и т.д. [10].

Теггеры частей речи обычно принимают на вход последовательность слов (т.е. предложение) и выдают на выходе список кортежей, где каждое слово связано с соответствующим тегом.

Тегирование частей речи предоставляет контекстуальную информацию, необходимую лемматизатору для выбора подходящей леммы [11,12].

**Сравнение популярных инструментов NLP**

Существует множество инструментов и библиотек, предназначенных для решения задач NLP. Краткое сравнение основных инструментов будет представлено далее, но нужно понимать, что все библиотеки, которые рассматриваем, имеют лишь частично пересекающиеся задачи. Поэтому иногда их трудно сравнивать напрямую. Некторые особенности опустим и сравним между собой только те библиотеки, в которых имеется аналогичный функционал.

- *NLTK (Natural Language Toolkit)* используется для решения таких задач, как токенизация, лемматизация, стемминг, синтаксический разбор, POS-тегирование и т.д. В этой библиотеке есть инструменты практически для всех задач NLP.
- *Spacy* является основным конкурентом NLTK. Эти две библиотеки могут использоваться для решения одних и тех же задач.
- *Scikit-learn* предоставляет большую библиотеку для машинного обучения. Здесь также представлены инструменты для предварительной обработки текста.
- *Gensim* - пакет для моделирования тем и векторного пространства, сходства документов.
- Общая задача библиотеки *Pattern* - служить в качестве модуля веб-майнинга. Таким образом, она поддерживает NLP только в качестве побочной задачи.
- *Polyglot* - еще один пакет python для NLP. Он не очень популярен, но также может быть использован для широкого круга задач NLP.
- *UDPipe* - На основании информационных банеов проекта Universal Dependencies создана библиотека UDPipe, позволяющее производить токенизацию, лемматизацию, морфологический анализ, а также строить деревья зависимостей между словами. UDPipe реализовано в виде бесплатных библиотек и пакетов на разных языках программирования, содержащие предварительно обученные языковые модели.

Чтобы сделать сравнение более наглядным, было подготовлена таблица, в которой указаны плюсы и минусы различных инструментов. [13]

Таблица 1 – Преимущества и недостатки библиотек для решения задач NLP

| | Преимущества | Недостатки |
|---|---|---|
| Natural Language Toolkit | 1)Наиболее известная и полная библиотека для работы NLP;<br>2)Имеется множество сторонних дополнений;<br>3)Возможность использовать множество подходов к каждой задаче NLP;<br>4)Быстрая токенизация предложения;<br>5)Поддержка наибольшего количества языков по сравнению с другими библиотеками; | 1)Сложность в изучении и использовании;<br>2)Довольно медленная обработка по сравнению с другими библиотеками;<br>3)При токенизации предложений NTLK разбивает текст только на предложения, не анализируя семантику и структуру в целом;<br>4)Выполняется обработка строк, что не очень характерно для объектно-ориентированного языка Python;<br>5) Не использует нейронные сети;<br>6)Отсутствуют интегрированные векторы слов; |
| spaCy | 1)Самый быстрый NLP-фреймворк;<br>2)Понятный в использовании, потому что для каждой задачи имеется определённый высоко оптимизированный инструмент;<br>3)Объект процессов более объектно-ориентированный, | 1) Менее гибкая по сравнению с NLTK;<br>2)Токенизация предложений медленнее, чем в NLTK;<br>3)Поддерживает маленькое количество языков; |

| | | |
|---|---|---|
| | по сравнению с другими библиотеками;<br>4)Использует нейронные сети для обучения некоторых моделей; | |
| Scikit-learn NLP toolkit | 1) Большое количество алгоритмов для построения моделей;<br>2) Содержит функции для работы с Bag-of-Words моделью;<br>3) Имеет подробную документация и интуитивно понятные методы в классах | 1)Плохой препроцессинг, что вынуждает использовать ее в связке с другой библиотекой (например, NLTK);<br>2) Не использует нейронные сети для препроцессинга текста. |
| Genism | 1)Работает с большими датасетами;<br>2)Поддерживает глубокое обучение;<br>3)Предоставляет возможность работы с word2vec, tf-idf vectorization, document2vec. | 1)Библиотека заточена под модели без учителя;<br>2)Не содержит достаточного функционала, необходимого для NLP, что вынуждает использовать ее вместе с другими библиотеками; |
| Pattern | 1)Позволяет тегировать части речи, искать n-граммы, анализировать настроения, WordNet, работать с моделью векторного пространства и кластеризацию и SVM;<br>2)Есть веб-краулер. DOM парсер, некоторые API. | 1)Наличие веб-майнера; фреймворк также может быть недостаточно оптимизирован для некоторых специфических задач NLP |
| Polyglot | 1)Поддерживает большое количество языков(16-196 языков для различных задач) | 1)Не так популярен, как, например, NLTK или Spacy;<br>2)Может быть медленным в работе<br>3)Слабая поддержка сообщества |
| UDpipe | 1)Поддержка более 50 языков<br>2)Можно обучать собственные модели непосредственно из R<br>3)Доступны готовые модели для загрузки<br>4)При тестировании для голландского, французского, испанского, итальянского, португальского языков UDPipe в целом показал себя лучше, чем SpaCy<br>5)Возможность токенизации тегирования частей речи, тегирования морфологических признаков, лемматизации, парсинга зависимостей | 1)Медленее spaCy примерно в 5 раз |

**Подробное сравнение библиотек UDPipe и spaCy.**

Для решение конкретных задач, чаще всего используются две основные библиотеки, которые необходимо рассмотреть подробнее. Для сравнения *UDPipe* и *spaCy* необходимо выделить блоки, по которым можно определить критерии сравнения инструментов [14,15].

Традиционный поток обработки естественного языка состоит из нескольких строительных блоков, которые могут быть использованы для создания на его основе приложения для обработки естественного языка. А именно:

1. Токенизация.
2. Маркировка частей речи.
3. Лемматизация.
4. Маркировка морфологических признаков.
5. Синтаксический разбор зависимостей.
6. Распознавание сущностей.
7. Извлечение смысла слов и предложений.

Можно сравнить данные инструменты по ряду критериев:

- языки, на которых работают инструменты;
- простота использования;
- возможности аннотации;
- точность аннотации моделей;
- скорость аннотации.

Поскольку модели spaCy и UDPipe для испанского, португальского, французского, итальянского и голландского языков были построены на данных из одного и того же древа универсальных зависимостей, можно сравнить точность различных этапов обработки НЛП (токенизация, POS-тегирование, тегирование морфологических признаков, лемматизация, разбор зависимостей)[16].

Оценка традиционно производится путем исключения некоторых предложений из обучающей части и просмотра того, насколько хорошо модель справилась с этими исключенными предложениями, которые были помечены людьми, поэтому их называют "золотыми".

Ниже приведена статистика точности для различных задач NLP с использованием скрипта оценки общих задач Conllu 2021 на удержанных тестовых наборах [17,18]. Эти графики в основном показывают, что:

- UDPipe обеспечивает лучшие результаты для французского, итальянского и португальского языков, равные результаты для испанского языка, менее хорошие результаты для разбора зависимостей и тегов, специфичных для древовидного дерева, для голландского языка, но лучшие результаты для универсальных тегов частей речи.
- Для английского языка можно сравнить только теги XPOS из Penn Treebank, и spaCy показывает менее хорошие результаты, чем мы ожидали, при сравнении с моделью UDPipe.

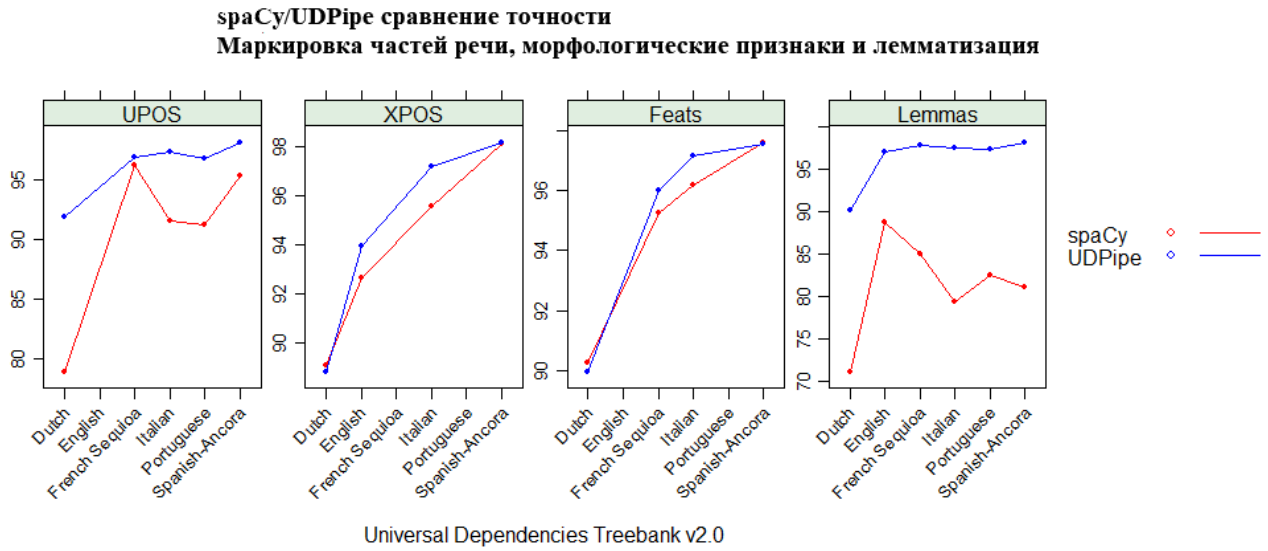


Рисунок 1 – Сравнение точности работы библиотек на разных этапах обработки

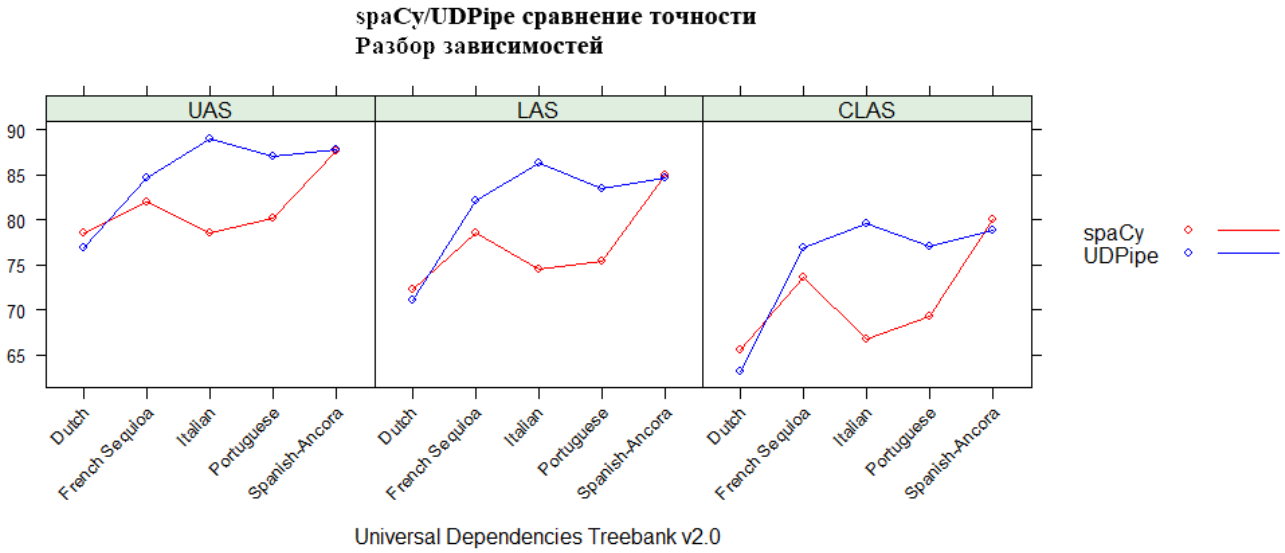


Рисунок 2 – Сравнение точности работы библиотек при разборе зависимостей

Таблица 2 – Сравнение библиотек UDPipe и spaCy [19,20].

| | UDPipe | spaCy |
|---|---|---|
| Языки | 1)Поддержка более 50 языков;<br>2)Возможность обучить собственные модели в R; | 1)Поддерживает 8 языков<br>2)Сообщество постепенно добавляет поддержку новых языков<br>3)Обучать модели только через Python |
| Простота использования | 1)Установка в одну команду для R;<br>2)Имеются встроенные гибкие модели; | 1)Установка более сложная по сравнению с UDPipe, нужны дополнительные зависимости для Python |
| Точность | 1)При тестировании для голландского, французского, испанского, итальянского, португальского языков UDPipe в целом показал себя лучше, чем SpaCy | 1)На определенных языках и моделях показывает результат лучше чем UDPipe, например, на голландском языке |
| Возможности аннотации | 1)Возможность токенизации тегирования частей речи, тегирования морфологических признаков, лемматизации, парсинга зависимостей | Аналогичные возможности, за исключением того, что spaCy может лемматизировать только англоязычные данные |
| Скорость работы | 1)Медленная обработка по сравнению с spaCy | 1)Быстрее примерно в 5 раз чем UDPipe<br>2)Обширное сообщество, которое оптимизирует работу библиотеки |

Определенно, самыми популярными пакетами для NLP сегодня являются UDPipe и Spacy. Они являются основными конкурентами в области NLP. Разница между ними заключается в общей философии подхода к решению задач.

UDPipe более гибкий, также поддерживает большее количество языков. С его помощью можно попробовать различные методы и алгоритмы, комбинировать их и т.д. Spacy же предоставляет одно готовое решение для каждой проблемы, не нужно думать о том, какой метод лучше: авторы Spacy уже позаботились об этом. Кроме того, Spacy работает очень быстро (в несколько раз быстрее, чем UDPipe). Одним из недостатков является ограниченное количество языков, поддерживаемых Spacy. Однако количество поддерживаемых языков постоянно увеличивается.

**Вывод.**

Таким образом в статье рассмотрены популярные библиотеки для обработки естественного языка, приведены преимущества и недостатки каждой библиотеки, а также проведён их сравнительный анализ. Также было проведено подробное сравнение двух основных библиотек по пяти критериям. В результате сравнительного анализа была выделена библиотека UDPipe, которая несмотря на относительно небольшую скорость работы, обладает большей точностью при работе с различными языковыми группами, и в целом является библиотекой с более гибким и обширным функционалом.

**Список литературы**


1. Witten I. H., Frank E. Data mining: practical machine learning tools and techniques with Java implementations //Acm Sigmod Record. – 2002. – Т. 31. – №. 1. – С. 76-77.
2. Jain A. K., Murty M. N., Flynn P. J. Data clustering: a review //ACM computing surveys (CSUR). – 1999. – Т. 31. – №. 3. – С. 264-323.
3. Bousquet O., Boucheron S., Lugosi G. Introduction to statistical learning theory //Summer school on machine learning. – Springer, Berlin, Heidelberg, 2003. – С. 169-207.
4. Lee I. et al. A probabilistic functional network of yeast genes //science. – 2004. – Т. 306. – №. 5701. – С. 1555-1558.
5. Myers C. L., Troyanskaya O. G. Context-sensitive data integration and prediction of biological networks //Bioinformatics. – 2007. – Т. 23. – №. 17. – С. 2322-2330.
6. Feldman R. et al. The text mining handbook: advanced approaches in analyzing unstructured data. – Cambridge university press, 2007.
7. Hearst M. A. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. – 1999.
8. Leeper N. J. et al. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes //PloS one. – 2013. – Т. 8. – №. 5. – С. e63499.
9. Müller T., Schütze H. Robust morphological tagging with word representations //Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2015. – С. 526-536.
10. D. Tunkelang, Stemming and Lemmatization [Site] URL: https://queryunderstanding.com/stemming-and-lemmatization-6c086742fe45 11.01.22
11. Juršic M. et al. Lemmagen: Multilingual lemmatisation with induced ripple-down rules //Journal of Universal Computer Science. – 2010. – Т. 16. – №. 9. – С. 1190-1214.
12. McGillivray B., Passarotti M., Ruffolo P. The Index Thomisticus Treebank Project: Annotation, Parsing and Valency Lexicon //Trait. Autom. des Langues. – 2009. – Т. 50. – №. 2. – С. 103-127.
13. Straka M., Straková J. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe //Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. – 2017. – С. 88-99.
14. Kwartler T. Text mining in practice with R. – John Wiley & Sons, 2017.
15. A comparison between spaCy and UDPipe for Natural Language Processing for R users URL: https://www.bnosac.be/index.php/blog/75-a-comparison-between-spacy-and-udpipe-for-natural-language-processing-for-r-users (дата обращения 12.01.22)
16. Colic N., Rinaldi F. Improving spaCy dependency annotation and PoS tagging web service using independent NER services //Genomics & informatics. – 2019. – Т. 17. – №. 2. Manning C., Schutze H. Foundations of statistical natural language processing. – MIT press, 1999.
17. Kharis M. et al. How to Lemmatize German Words with NLP-Spacy Lemmatizer? //International Seminar on Language, Education, and Culture (ISoLEC 2021). – Atlantis Press, 2021. – С. 189-193..
18. Антропова О. И., Огородникова Е. А. Экстернальная оценка предварительно обученных моделей UDPipe в применении к извлечению гипер-гипонимических словесных пар из словарных определений //AIP Conference Proceedings. – AIP Publishing LLC, 2020. – Т. 2313. – №. 1. – С. 070020.
19. Schmitt X. et al. A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate //2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). – IEEE, 2019. – С. 338-343.
20. Straka M., Straková J., Hajič J. UDPipe at SIGMORPHON 2019: Contextualized embeddings, regularization with morphological categories, corpora merging //arXiv preprint arXiv:1908.06931. – 2019.

**References**


1. Witten I. H., Frank E. Data mining: practical machine learning tools and techniques with Java implementations //Acm Sigmod Record. – 2002. – Т. 31. – №. 1. – pp. 76-77.
2. Jain A. K., Murty M. N., Flynn P. J. Data clustering: a review //ACM computing surveys (CSUR). – 1999. – Т. 31. – №. 3. – pp. 264-323.
3. Bousquet O., Boucheron S., Lugosi G. Introduction to statistical learning theory //Summer school on machine learning. – Springer, Berlin, Heidelberg, 2003. – pp. 169-207.
4. Lee I. et al. A probabilistic functional network of yeast genes //science. – 2004. – Т. 306. – №. 5701. – pp. 1555-1558.
5. Myers C. L., Troyanskaya O. G. Context-sensitive data integration and prediction of biological networks //Bioinformatics. – 2007. – Т. 23. – №. 17. – pp. 2322-2330.
6. Feldman R. et al. The text mining handbook: advanced approaches in analyzing unstructured data. – Cambridge university press, 2007.
7. Hearst M. A. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. – 1999.
8. Leeper N. J. et al. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes //PloS one. – 2013. – Т. 8. – №. 5. – pp. e63499.
9. Müller T., Schütze H. Robust morphological tagging with word representations //Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2015. – pp. 526-536.
10. D. Tunkelang, Stemming and Lemmatization [Site] URL: https://queryunderstanding.com/stemming-and-lemmatization-6c086742fe45 11.01.22
11. Juršic M. et al. Lemmagen: Multilingual lemmatisation with induced ripple-down rules //Journal of Universal Computer Science. – 2010. – Т. 16. – №. 9. – pp. 1190-1214.
12. McGillivray B., Passarotti M., Ruffolo P. The Index Thomisticus Treebank Project: Annotation, Parsing and Valency Lexicon //Trait. Autom. des Langues. – 2009. – Т. 50. – №. 2. – pp. 103-127.
13. Straka M., Straková J. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe //Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. – 2017. – pp. 88-99.
14. Kwartler T. Text mining in practice with R. – John Wiley & Sons, 2017.
15. A comparison between spaCy and UDPipe for Natural Language Processing for R users URL: https://www.bnosac.be/index.php/blog/75-a-comparison-between-spacy-and-udpipe-for-natural-language-processing-for-r-users (дата обращения 12.01.22)
16. Colic N., Rinaldi F. Improving spaCy dependency annotation and PoS tagging web service using independent NER services //Genomics & informatics. – 2019. – Т. 17. – №. 2. Manning C., Schutze H. Foundations of statistical natural language processing. – MIT press, 1999.
17. Kharis M. et al. How to Lemmatize German Words with NLP-Spacy Lemmatizer? //International Seminar on Language, Education, and Culture (ISoLEC 2021). – Atlantis Press, 2021. – pp. 189-193..
18. Antropova O. I., Ogorodnikova E. A. Extrinsic evaluation of UDPipe pre-trained models in application to hyper-hyponymic verbal pairs extraction from dictionary definitions //AIP Conference Proceedings. – AIP Publishing LLC, 2020. – Т. 2313. – №. 1. – pp. 070020.
19. Schmitt X. et al. A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate //2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). – IEEE, 2019. – pp. 338-343.
20. Straka M., Straková J., Hajič J. UDPipe at SIGMORPHON 2019: Contextualized embeddings, regularization with morphological categories, corpora merging //arXiv preprint arXiv:1908.06931. – 2019.