Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики
Факультет информационных технологий и программирования
Кафедра компьютерных технологий

# Разработка модели криптовалюты на основе алгоритма консенсуса реализующего метод доказательства доли владения

Агапов Г.Д.
Научный руководитель: Чивилихин Д.С.

Санкт-Петербург
2018

## ОГЛАВЛЕНИЕ

Стр.

# ВВЕДЕНИЕ

В последние годы такая область знаний, как обработка естественного языка (Natural Language Processing или NLP) находит все больше и больше применений в различных сферах человеческой деятельности. К этой области в частности относятся задачи информационного поиска, извлечения информации, распознавания и синтеза речи, построения систем машинного перевода, вопросно-ответных систем, а также множество других задач, приложений.

Построения лингвистических систем “с нуля” — процесс крайне трудозатратный, требующий как правило усилий значительного числа специалистов достаточно широкого профиля. Основная сложность заключается в разработке и наполнении моделей, описывающих язык, позволяющих с ним работать. Именно поэтому при разработке лингвистическим систем обыкновенно прибегают к использованию готовых инструментов и ресурсов общего назначения, многие из которых есть в открытом доступе. Примерами таких инструментов и ресурсов могут служить: инструмент для извлечения морфологической информации MyStem [1], инструмент анализа лексики на основе моделей дистрибутивной семантики Word2Vec [2], открытый корпус текстов на русском языке OpenCorpora [3] и другие.

Наконец, широко распространенным видом ресурсов для работы с лексикой языка являются тезаурусы. С тех пор, как в 1990 году вышла первая версия тезауруса для английского языка Princeton WordNet (PWN), электронные тезаурусы нашли применение во многих приложениях NLP. Были разработаны аналогичные ресурсы для целого ряда языков, в том числе был предпринят ряд попыток построения тезауруса русского языка (РуТез, RussNet), последней из которых является YARN (Yet Another RussNet) — тезаурус современного русского языка, разработка которого началась в 2013 году.

Настоящая работа посвящана одной из задач, возникшей при построении тезауруса YARN, задачи выравнивания — сопоставления его понятий понятиям тезауруса PWN, т.е. нахождения для понятий лексики русского языка (хранящихся в YARN) соответствий в лексике английского языка. Подобная задача впервые решалась исследователями, работавшими над проектом EuroWordNet, и с тех пор является классической при построении тезаурусов

для новых языков. Ценность её решения заключается в возможности объединения тезаурусов различных языков в единую сеть, в которой, имея понятие одного языка можно будет легко получить доступ к соответствующим понятиям других языков. Как следствие, информацию о таких связях можно использовать в построении систем машинного перевода (для увеличения точности перевода), систем извлечения смысловой информации из текста и других.

В первой главе даются основные понятия, используемые в работе, проводится краткий обзор предметной области. Во второй главе формулируется задача выравнивания, рассматриваются предшествующие попытки её решения, ставятся цели, задачи, преследуемые авторами настоящей работы.

Во второй и третьей главах предлагается метод решения задачи выравнивания, условно разделенный на два этапа: автоматическое выравнивание (т.е. алгоритмический предпроцессинг) и выравнивание с применением техник краудсорсинга. В этих главах также приводятся результаты тестирования полученного метода.

# ГЛАВА 1. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

В настоящей главе проводится краткий обзор предметной области. Даются определения основных понятий, связанных с тезаурусами. Рассматриваются ресурсы YARN и PWN, работе с которыми посвящена настоящая работа.

## 1.1. Блокчейн

### Резюме

В изложенной главе были даны определения основных понятий, используемых в настоящей работе, в частности понятий тезауруса, синсета, гиперонимии, меронимии, полисемии.

Дан короткий обзор проекта Princeton WordNet, существующих тезаурусов русского языка, а также интересующего нас главным образом тезауруса YARN.

# ЗАКЛЮЧЕНИЕ

В настоящей работе была рассмотрена задача выравнивания синсетов тезаурусов Princeton WordNet и YARN. Был предложен метод для решения этой задачи, состоящий из двух этапов: автоматический предпроцессинг и выравнивание с применением техник краудсорсинга. Каждый из этапов был подробно рассмотрен в главах **??**, **??**.

Полученный метод был протестирован на случайном подмножестве существительных из BCs, были получены хорошие результаты как для автоматического выравнивания (см. раздел **??**), так и для выравнивания с применением краудсорсинга. Итоговая точность метода на тестовой выборке составила (88 %). Были рассмотрены направления для дальнейших исследований (нацеленных на повышение точности).

После достижения достаточной точности порядка (95%) (как с помощью применения подходов, рассмотренных в **??**, так и других), планируется провести процедуру выравнивнивания для всех существительных BCs, интегрировать полученные связи в структуру тезауруса YARN.

В дальнейшем планируется адаптировать полученный в настоящей работе метод для проведения выравнивания по произвольному множеству синсетов PWN.

## СПИСОК ИСТОЧНИКОВ


1. Mystem. [Электронный ресурс]. URL: https://tech.yandex.ru/mystem/.
2. Word2Vec. [Электронный ресурс]. URL: https://code.google.com/archive/p/word2vec/.
3. OpenCorpora. [Электронный ресурс]. URL: http://opencorpora.org/.