Поддержка перечислений для языка C++ в типе вопроса CorrectWriting

Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
Волгоградский государственный технический университет

Факультет Электроники и вычислительной техники
Кафедра Программное обеспечение автоматизированных систем

УТВЕРЖДАЮ
Зав. кафедрой ПОАС
_____ д.т.н., проф. А. М. Дворянкин
(подпись) (инициалы, фамилия)
«_____»_____2016

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

к выпускной работе бакалавра на тему
(наименование вида работы)
Название работы
Автор _____ Клевцов Вадим Александрович
(подпись и дата подписания) (фамилия, имя, отчество)
Обозначение МД-40-461-806-10.19-09.04.04-03-15-81
(код документа)
Группа ПрИн-466
(шифр группы)
Направление 09.03.04 Программная инженерия
(код по ОКСО, наименование направления, программы)

Руководитель работы _____ к.т.н О. А. Сычев
(подпись и дата подписания) (инициалы и фамилия)
Консультанты по разделам:

(краткое наименование раздела)	(подпись и дата подписания)	(инициалы и фамилия)
(краткое наименование раздела)	(подпись и дата подписания)	(инициалы и фамилия)

Нормоконтролер _____ О. Н. Ляпина
(подпись и дата подписания) (инициалы и фамилия)

Волгоград, 2016

Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
Волгоградский государственный технический университет

Кафедра Программное обеспечение автоматизированных систем

УТВЕРЖДАЮ
Зав. кафедрой ПОАС
_____ д.т.н., проф. А. М. Дворянкин
(подпись) (инициалы, фамилия)
«_____»_____2016

Задание на выпускную работу бакалавра
(наименование вида работы)

Студент Клевцов Вадим Александрович
(фамилия, имя, отчество)

Код кафедры 10.19 Группа ПрИн-466

Тема Название работы
Утверждена приказом по университету от «17» октября 2014 № 1529–ст
Срок представления готовой работы «01» января 2016 _____
(подпись студента)

Исходные данные для выполнения работы
задание, выданное научным руководителем с кафедры ПОАС, утвержденное приказом ректора

Содержание основной части пояснительной записки

Введение

1 Исследование подходов, методов и средств обработки естесвенных языков

Цель и задачи исследования

2 Исследование грамматики английского языка

Выводы

3 Разработка метода определения перечислений в английском языке на основе синтаксического анализа

Выводы

4 Реализация и интеграция метода в плагин Correct Writing, эксперемент, оценка достижения цели

Выводы

Заключение

Список использованных источников

Приложение А - Техническое задание

Перечень графического материала

1) 1: Название работы
2) 2-3: Актуальность

Руководитель работы _____ к.т.н О. А. Сычев
(подпись и дата подписания) (инициалы и фамилия)

Консультанты по разделам:

_____ _____ _____
(краткое наименование раздела) (подпись и дата подписания) (инициалы и фамилия)

Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
Волгоградский государственный технический университет
Кафедра «Программное обеспечение автоматизированных систем»

УТВЕРЖДАЮ
Зав. кафедрой ПОАС
_____ д.т.н., проф. А. М. Дворянкин
(подпись) (инициалы, фамилия)
«_____»_____2016

Название работы

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

МД-40-461-806-10.19-09.04.04-03-15-81-81

Листов 30

Научный руководитель
к.т.н, доц. каф. ПОАС
_____ О. А. Сычев
«_____»_____2016

Нормоконтролер
ст. преп. каф. ПОАС
_____ О. Н. Ляпина
«_____»_____2016

Исполнитель
студент группы ПрИн-466
_____ В. А. Клевцов
«_____»_____2016

Волгоград, 2016

Содержание

Введение

Дистанционные курсы, как и дистанционное обучение в целом набирают популярность в наше время. Существует множество систем для организации дистанционного обучения. Несмотря на внешние и внутренние различия данные системы функционируют по единому принципу: приобретение и закрепление знаний. Закрепление знаний чаще всего предполагает проведение тестирования или выполнение разного рода заданий.

Автоматизация тестирования позволяет снизить трудоемкость создания и проведения дистанционного обучения. Одним из примеров системы управления обучением является Moodle. Данное СУО позволяет автоматизировать процесс проведения и оценки результатов тестирования. Более того, за счет открытого исходного кода, система позволяет разрабатывать и адаптировать способ оценки теста и типы вопросов.

Часто в тестах используются вопросы с открытым ответом на естественном языке. Проблема всех естественных языков в степени их формализованости. Низкая степень формализованности увеличивает трудоемкость создания и оценки вопроса, так как требует оценки и ввода всех вариантов эталонного ответа. Более того в ответе так же могут содержаться перечисления, то есть кортежи элементов, порядок которых не изменяет семантику ответа.

Кроме определения уровня подготовки студента, тесты позволяют направить студента. Подсказки предлагаемые студенту зависят от возможностей системы, но обычно основаны на описании лексем содержащихся в ответе. Что увеличивает трудоемкость создания вопроса.

Данная работа посвящена созданию системы, позволяющей учитывать последовательности лексем порядок которых не важен.

1 Анализ современного состояния проблем в области автоматизированной обработки текстов на естественных языках

1.1 Анализ современного состояния проблем в области поддержки перечислений

Вопросы с открытым ответом являются популярным средством тестирования, так как исключают возможность выбора правильного ответа наугад. Однако не смотря на это существует ряд ограничений в использовании открытых ответов в процессе автоматизированного ответа:

а) необходимость ввода всех вариантов правильного ответа;

б) ввод вопроса в форме не позволяющей двусмысленного понимания.

Так же такого рода вопросы позволяет направлять студента при прохождении тренировочного тестирования, выдавая подсказки и исправляя ошибки.

Такого рода вопросы существуют во всех современных системах дистанционного обучения, однако ни одна система не позволяет использовать перечисления в ответе, кроме системы управления обучением Moodle. Данная система содержит плагин типа вопроса "CorrectWriting"предоставляющую базовую поддержку перечислений в ответе. Что позволяет расширить возможности обучения не увеличивая трудоемкость создания вопроса за счет использования модели ответа дополненной описанием перечислений в нем. Это описание позволяет подобрать порядок элементов для каждого перечисления, максимизирующий наибольшую общую подпоследовательность эталонного ответа и ответа студента.

На данный момент плагин поддерживает автоматизированное определение перечислений для языков C
C++ в следующих случаях:

а) последовательности объявления переменных одного типа;

б) последовательности математических операций(сложение,умножение,деление,разность,деление нацело);

в) последовательности логических операций(И, ИЛИ, эквивалентность, неэквивалентность);

г) последовательность битовых операций (И, ИЛИ, исключающее ИЛИ);

д) последовательность присваиваний;

е) последовательность объявление полей класса, структуры, объявления, перечисления;

ж) последовательность модификаторов видимости внутри класса;
Так как многие операции анализируемые плагином могут быть вложены друг в друга различными способами "CorrectWriting"использует стандартный для данного языка приоритет операций. Автоматизация определения перечислений основана на использования синтаксического анализа эталонного ответа, а именно на анализе результирующего синтаксического дерева.

Кроме того данный тип вопроса содержит форму редактирования перечислений в эталонном ответе, которая позволяет исправить неточности автоматизированного определения перечислений, и создать описание перечислений в сложных случаях.

В английском языке так же присутствуют перечисления. Для естественного языка единственным способом определения перечислений является синтаксический анализ. Далее мы рассмотрим существующие подходы к синтаксическому анализу естественных языков, а также существующие средства синтаксического анализа.

1.2 Существующие подходы к обработке текстов на естественных языках

В наше время обработка естественных языков используется в нескольких направлениях:

а) машинный перевод;

б) информационный поиск;

в) реферирование текста;

г) рубрицирование текстов;

д) обучение языку и др.

Несмотря на обилие направлений использования, подходов к обработке всего четыре, символьный, вероятностный, установления связей и гибридный.

1.3 Символьный подход

Подход основанный представлении языка как модели сложной, но прозрачной. Примерами такой модели могут послужить:

а) обучение на правилах;

б) индуктивное логическое программирование;

в) деревья разрешений;

г) концептуальная кластеризации;

д) алгоритмы типа k-средних.

Более подробно перечисленные методы будут описаны ниже. Общей чертой данных моделей является способ их получение, а именно обучение.

1.3.1 Обучение на правилах

Один из старейших методов обучение и построения моделей. Используется в случаях когда алгоритм не возможно закодировать иначе, то есть когда алгоритм содержит эвристические правила.

1.3.2 Индуктивное логическое программирование

Это раздел машинного обучения использующий в качестве примеров, фоновых знаний и гипотез логическое программирование. Логическое программирование — это парадигма программирования, которая основана на автоматическом доказательстве теорем. Логическое программирование основано на теории и аппарате математической логики с использованием математических принципов резолюций.

1.3.3 Деревья разрешений

Это средство принятия решений, используемое в прогнозировании и обработке данных. Структура дерева представляет собой «листья» и «ветки». На ребрах («ветках») дерева решения записаны атрибуты, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах — атрибуты, по которым различаются случаи. Для определения значения, необходимо спуститься по дереву до листа и вернуть его значение.

1.3.4 Концептуальная кластеризация

Кластеризация является еще одним способом обработки естественных языков. Кластеризация является задачей обучения без учителя. Суть метода заключается в обучении на большой выборке, позволяющий выделить объекты в однородные группы. Ниже приведена классификация методов кластеризации являющаяся общепринятой:

а) вероятностный подход:
   1) метод К-средних;
   2) метод K-medians;

3) EM-алгоритм;

4) алгоритм семейства FOREL;

5) дискриминантный анализ,

б) методы на основе систем искусственного интеллекта:

1) метод нечеткой кластеризации С-средних;

2) нейронная сеть Кохонена;

3) генетический алгоритм,

в) логический подход. Кластеризация на основе дерева решений,

г) теоретико-графический подход:

1) графические алгоритмы кластеризации;

д) иерархический подход. Используется в ситуации наличия подгрупп внутри групп:

1) агломеративные алгоритмы;

2) дивизивные алгоритмы,

е) остальные методы:

1) статистические методы кластеризации;

2) "ансамбль"кластеров;

3) алгоритмы семейства KRAB;

4) алгоритм, основанный на методе просеивания DBSCAN и др.

1.3.5 Алгоритмы типа k-средних

Наиболее популярный алгоритм кластеризации, стремящийся минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров.

Основная идея заключается в том, что на каждой итерации вычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике.

Алгоритм завершается, когда на какой-то итерации не происходит изменения центра масс кластеров. Это происходит за конечное число итераций, так как количество возможных разбиений конечного множества конечно, а на каждом шаге суммарное квадратичное отклонение не увеличивается, поэтому зацикливание невозможно.

1.4 Вероятностный подход

Подход использует различные математические техники, а также большие текстовые корпуса для разработки обобщенных моделей языковых явлений, базой для которой является реальные примеры найденные в текстовом корпусе не используя дополнительных знаний о языке или о внешнем мире. Основное отличие от символьного подхода использование реальных данных в качестве первичного источника информации.

В вероятностном подходе существует несколько течений, среди которых особого внимания заслуживают модели, максимизирующие энтропию и скрытые марковские модели (СММ). СММ это конечный автомат, который имеющий множество состояний с определенными вероятностями переходов между ними. Каждое состояние производит один из наблюдаемых результатов с определенной вероятностью. Хотя результаты являются видимыми, но состояние модели скрыто от внешнего наблюдения. Главным преимуществом вероятностных моделей заключено в том, что они дают способ решения многих видов неоднозначных проблем, формулируемых так "с учетом N некоторых неоднозначных вводов выбрать один наиболее вероятный".

1.4.1 Методы максимизации энтропии

Данный метод классификации основан на понятии информационной энтропии. Информационная энтропия - это мера неопределенности информации. Данный термин был введен в оборот Шенноном, который предположил что прирост информации равен утраченной неопределенности и определил требования ее измерения:

а) мера должна быть непрерывной;

б) в случае когда все варианты равновероятны, увеличение количества вариантов ведет к увеличение значения функции;

в) должна быть возможность сделать выбор в два шага, в которых значение функции конечного результата должно являться суммой функций промежуточных результатов.

Поэтому функция энтропии H должна удовлетворять следующим условиям:

а) $H(p_1, \ldots, p_n)$ определена и непрерывна для всех $p_1, \ldots, p_n$, где $p_i \in [0,1]$ для всех $i = 1, \ldots, n$ и $p_1 + \ldots + p_n = 1$

б) Для целых положительных n, должно выполняться следующее неравенство: $H\underbrace{\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)}_{n} < H\underbrace{\left(\frac{1}{n+1}, \ldots, \frac{1}{n+1}\right)}_{n+1}$

в) Для целых положительных b_i, где $b_1 + \ldots + b_k = n$, должно выполняться равенство

$$H\underbrace{\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)}_{n} = H\left(\tfrac{b_1}{n}, \ldots, \tfrac{b_k}{n}\right) + \sum_{i=1}^{k} \tfrac{b_i}{n} H\underbrace{\left(\frac{1}{b_i}, \ldots, \frac{1}{b_i}\right)}_{b_i}$$

Шенон доказал, что есть только одна функция удовлетворяющая этим требованиям, она имеет вид: $-K \sum_{i=1}^{n} p(i) \log_2 p(i)$, где K - контстанта.

Для данного уравнения определить значение термина i, для естественного языка его можно выразить через отношение суммы планов содержания к сумме планов выражения. Это соотношение меняется взависимости от велечины информации.

1.4.2 Скрытые макрковские модели

Марковская модель является модель марковского процесса. В свою очередь марковский процесс - это случайный процесс, следующее состояние которого не зависит от предыдущих состояний при условии что известно текущее. Другими словами мпрковский процесс это модель авторегрессии первого порядка: $X_t = c + \alpha X_{t-1} + \varepsilon_t$.

Скрытая марковская модель — статистическая модель, имитирующая работу процесса, похожего на марковский процесс с неизвестными параметрами, и задачей является определение неизвестных параметров на основе наблюдаемых. Полученные параметры могут быть использованы в дальнейшем анализе.

1.5 Подход установления связей

Подход установление связей основан на моделях массивных связанных наборов простых и нелинейных компонентов. Эти компоненты работают параллельно. Приобретенное в результате обработки знание сохраняется в образце весов взаимосвязи компонентов.

1.6 Гибридный подход

Гибридные методы используют преимущества трех только что описанных подходов, минимизируя человеческие усилия, требуемые для типовой лингвистической конструкции и максимизируя гибкость, эффективность, и надежность применения NLP при человеко-компьютерном взаимодействии.

При всех подходах обработка языка, как правило, включает элементы машинного обучения: модель классификации и обучающую

последовательность. На основании описания атрибутов каждого объекта модель классификации относит каждый объект в какой-то класс, обучающая последовательность ставит в соответствие последовательности объектов последовательность классов.

1.7 Выбор программного средства синтаксического анализа

Синтаксический анализ является сложным и трудоемким процессом, в особенности если речь идет об анализе естественного языка. Именно это стало причиной принятия решения о использовании стороннего средства синтаксического анализа.

Было рассмотрено множество синтаксических анализаторов и выделены критерии сравнения. Наиболее развитыми на момент написания диссертации являются следующие анализаторы:

а) анализатор Стенфорского университета;
б) анализатор Института Брауна;
в) анализатор Института Беркли;
г) анализатор Института Токио.

1.7.1 Критерии сравнения

От выбора критериев сравнения зависит решения о выборе средства синтаксического анализа, что в конечном итоге повлияет на результат все проделанной работы.

Moodle является системой управления обучением с открытым исходным кодом, что накладывает ограничения на используемое с ним программное обеспечение. Поэтому лицензия средства синтаксического анализа является важным критерием отбора. Каждый из рассматриваемых средств синтаксического анализа имеет открытую лицензию, которая позволяет использовать это средство в связке с Moodle.

Следующим важным критерием отбора является наличие тестов для выбираемого программного средства. Так как наличие большого количества тестов и покрытие этими тестами кода, указывает на качество программного средства, а так же позволяет оценить его функциональные возможности.

Определение части речи, к которой принадлежит лексема, является минимальной необходимой функциональностью для использования для определения перечислений в тексте написанном на естественном языке. Наличие этого критерия является обязательный требованием к программному средству синтаксического анализа.

Еще одним критерием сравнения является язык на котором написано программное средство. Этот критерий повлияет на системные требования предъявляемые разрабатываемым программным средством.

Одним из важных критериев является способ взаимодействия разрабатываемого программного средства со средством синтаксического анализа.

В качестве дополнительного критерия выступает возможность определение средством синтаксического анализа связей между лексемами в предложении.

1.7.2 Синтаксический анализатор Института Брауна

Данный синтаксический анализатор разрабатывается начиная с 2000 года. И сильно изменился с первой версии. Это синтаксический анализатор основанный на методе максимизации энтропии и модели самообучения. Данное программное средство разработано на языке Python и предполагает использование по средствам вызова из командной строки. К исходному коду парсера прилагается набор тестов, позволяющих оценить качество написанного программного средства.

Возможности данного пасера ограничиваются определением членов предложения и принадлежащих им лексем.

1.7.3 Синтаксический анализатор Института Беркли

Данное программное средство разрабатывается Институтом Беркли с 2001 года. В основу парсера лег символьный подход к синтаксическому анализу, алгоритм K-best. Данное программное средство написано на языках Java и Scala. И так же как рассмотренное ранее средство синтаксического анализа предполагает запуск с аргументами из командной строки.

Возможности данного парсера аналогичны возможностям ранее рассмотренного программного средства.

1.7.4 Синтаксический анализатор Института Токио

Институт Токио занимается разработкой средства синтаксического анализа английского языка с 2005 года. Программное средство использует подход установления связей в процессе анализа текста, алгоритм построения структуры предложения управляемой главным членом предложения. Язык написания данного программного средства C++. В отличие от двух предыдущих средств синтаксического анализа данное помимо запуска из командной строки предлагает возможность запуска в качестве сервера, с которым возможно общение по протоколу HTTP.

Возможности данного средства синтаксического анализа расширяют возможности средств описанных выше, за счет определения связей между членами предложения.

1.7.5 Синтаксический анализатор Стэнфордского университета

Данное средство синтаксического анализа естественных языков ведет свою историю с 1990 года. Этот синтаксический анализатор

является самым развитым на данный момент. Данное программное средство использует гибридный подход к синтаксическому анализу. В качестве языка написания данного программного средства выступает язык Java. Тестовая база данного синтаксического анализа включается в себя более 10000 тестов. Так же как и предыдущий анализатор данный предлагает два режима работы, в качестве средства командной строки, а так же в качестве отдельного сервера, с возможность выполнения запроса к нему по протоколу HTTP.

Возможности данного программного средства синтаксического анализа аналогичны возможностям предыдущего средства. С одной лишь разницей, данное средство имеет более высокие показатели качества за счет использования более продвинутого подхода к синтаксическому анализу.

1.7.6 Выводы

В результате изучения представленных выше программных средств синтаксического анализа, было принято решение выбрать для использование программное обеспечение разработанное на базе университета Стенфорда. Причинами для такого решения послужили следующие аргументы:

а) синтаксический анализатор Стенфорского университета использует обобщенное описание синтаксиса языка, что позволяет использовать его с другими естественными языками;

б) это наиболее активно развивающийся синтаксический анализатор;

в) данный анализатор имеет наибольшую тестовую базу, что позволяет удостовериться в качестве его работы на наибольшем количестве тестовых ситуаций.

2 Выделение структур английского языка удовлетворяющих перечислениям

2.1 Определение структур языка являющихся перечислениями

Для построения модели перечисления в контексте естественного языка, необходимо определить структуры языка подходящие на роль перечислений. В естественных языках перечислениями являются однородные члены предложения, а также сложно сочиненные члены предложения.

В английском языке существуют следующие однородные члены предложения:

а) сказуемые;
б) подлежащие;
в) определения;
г) дополнения;
д) обстоятельства.

Рассмотрение каждой структуры в отдельности позволит построить модель перечисления на естественном языке, и выработать алгоритм определения перечислений для естественного языка.

2.1.1 Сказуемое

Сказуемое является главным членом двусоставного предложения, обозначающим действие или признак того, что выражено подлежащим.

Сказуемое имеет лексическое значение (именует то, что сообщается о реалии, названной в подлежащем) и грамматическое значение (характеризует высказывание с точки зрения реальности или ирреальности и соотнесенности высказывания с моментом речи, что выражается формами наклонения глагола, а в изъявительном наклонении — и времени).

Рассмотрим как пример следующее предложение: «We went to the cafe and buy a cup of coffee.». Для большей наглядности используем изображение 2.1.1.

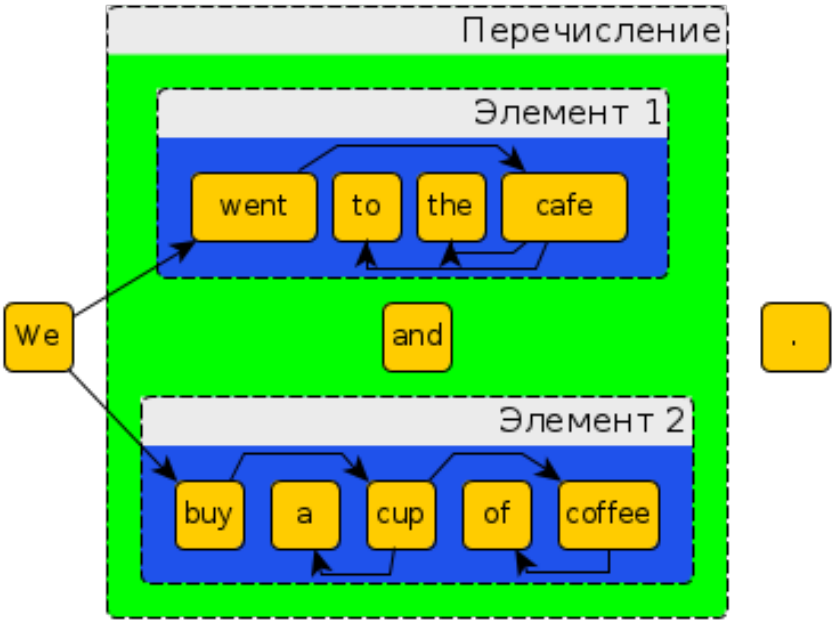

Рисунок 1 — Предложение с однородными сказуемыми, осложненными зависимыми членами предложения

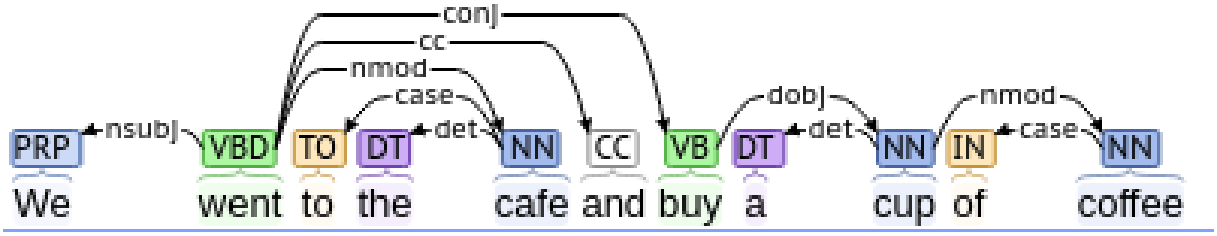

Рисунок 2 — Предложение с однородными сказуемыми, осложненными зависимыми членами предложения

На рисунке 2.1.1 выделены элементы перечисления, и стрелками показаны важные для нас отношения членов предложения. От подлежащего «We» две стрелки идут к однородным членам «went» и «buy» , осложненым обстоятельством места и дополнением связи с ними также показаны на рисунке. На рисунке 2.1.1 представлено дерево построенное Стенфорским парсером для данного предложения, оно выглядит иначе, основынм отличием является то что он подлежащего

построена связь только к первому однородному сказуемому, в свою очередь существует связь соединяющая однородные члены между собой. Специфика построения синтаксических деревьев будет учтена при создании модели перечисления.

2.1.2 Подлежащее

Подлежащее называет то, о ком или о чём говорится в предложении. Подлежащие не разрывно связано со сказуемым. Рассмотрим в качестве примера следующее предложение: «Hot coffee and green tea the are best monday morning drinks.». Связи элементов представлены на рисунке 2.1.2, в данном случае два однородных подлежащих «coffee» и «tea» , осложненные определениями, связаны со сказуемым «are» . На рисунке 2.1.2 изображено посттроенное дерево. Тут как и в первом случае присутсвуют отличия, демонстрирующие отличие логического подхода построения связей между элементами от лингвистического подхода синтаксического подхода. Данное различие так же необходимо учесть при создании модели.

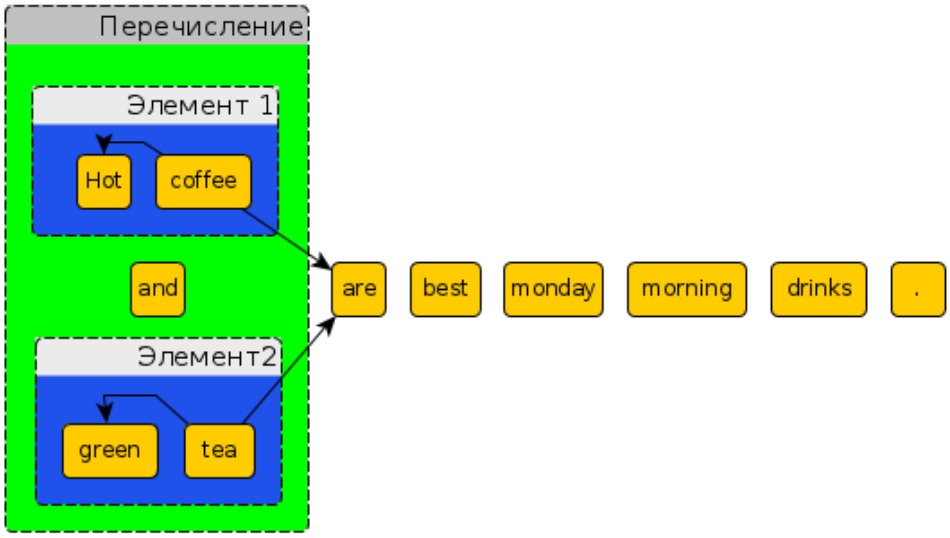

Рисунок 3 — Предложение с однородными подлежащими, осложненными зависимыми членами предложения

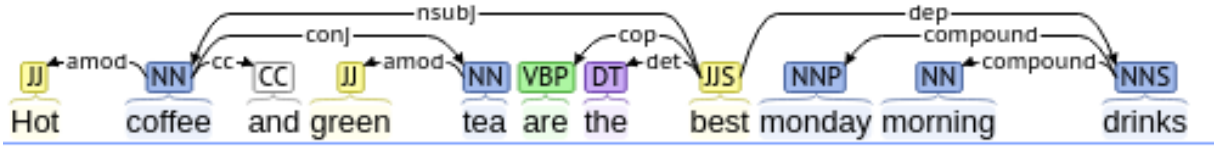

Рисунок 4 — Предложение с однородными подлежащими, осложненными зависимыми членами предложения

2.1.3 Определение

Определение — второстепенный член предложения, обозначающий признак, качество, свойство предмета. Рассмотрим на примере следующего предложения «The weather today is a little cloudy, extremely raining, freezing-cold and densy foggy.». На рисунке 2.1.3 представлены члены предложения и важные для исследования связи между ними. Связи отвечающие за однородные члены идут от лексеммы «weather» к лексеммам являющихся определениями «cloudy», «raining»,

«freezing-cold» и «foggy», так же как и в примерах выше присутствуют зависимые лексемы. Результат определния связей между лексемами представлен на рисунке 2.1.3, прослеживается общая для каждого из примеров логика, что связь между подлежащим и однородными определениями парсер представляет в виде связей двух видов:

а) связь подлежащего и первого из однородных определений;

б) связи между первым и остальными определниями.

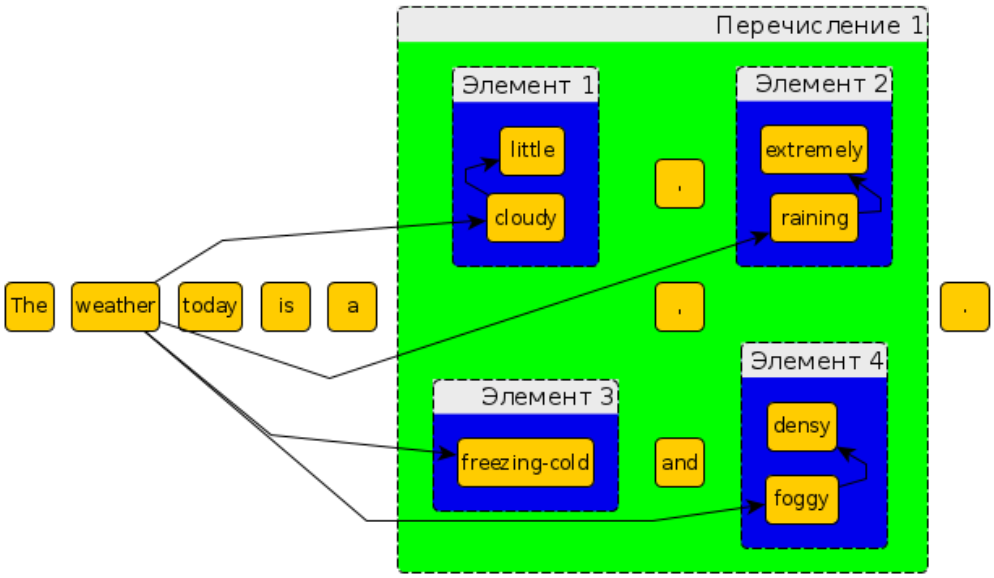

Рисунок 5 — Предложение с однородными определениями

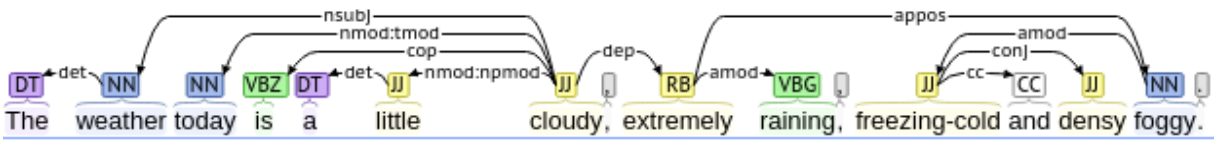

Рисунок 6 — Предложение с однородными определениями

2.1.4 Дополнение

Дополнение — второстепенный член предложения, выраженный существительным или местоименным существительным. Дополнение обозначает предмет или лицо, являющееся объектом действия, выраженного сказуемым. We need your first and last names, back-white photo and home and mobile number.

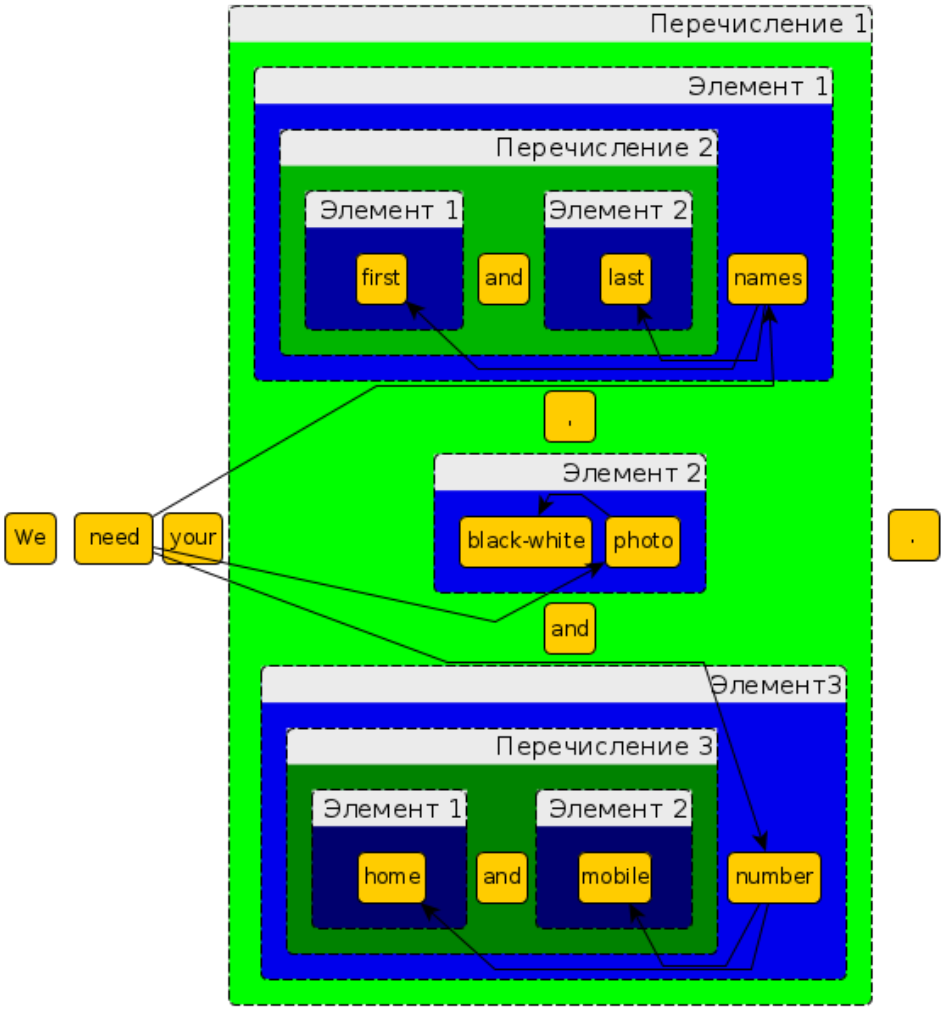

Рисунок 7 — Предложение с однородными сказуемыми, осложненными зависимыми лексемами

2.1.5 Обстоятельства

I go to the cinema, cafe and to the park.

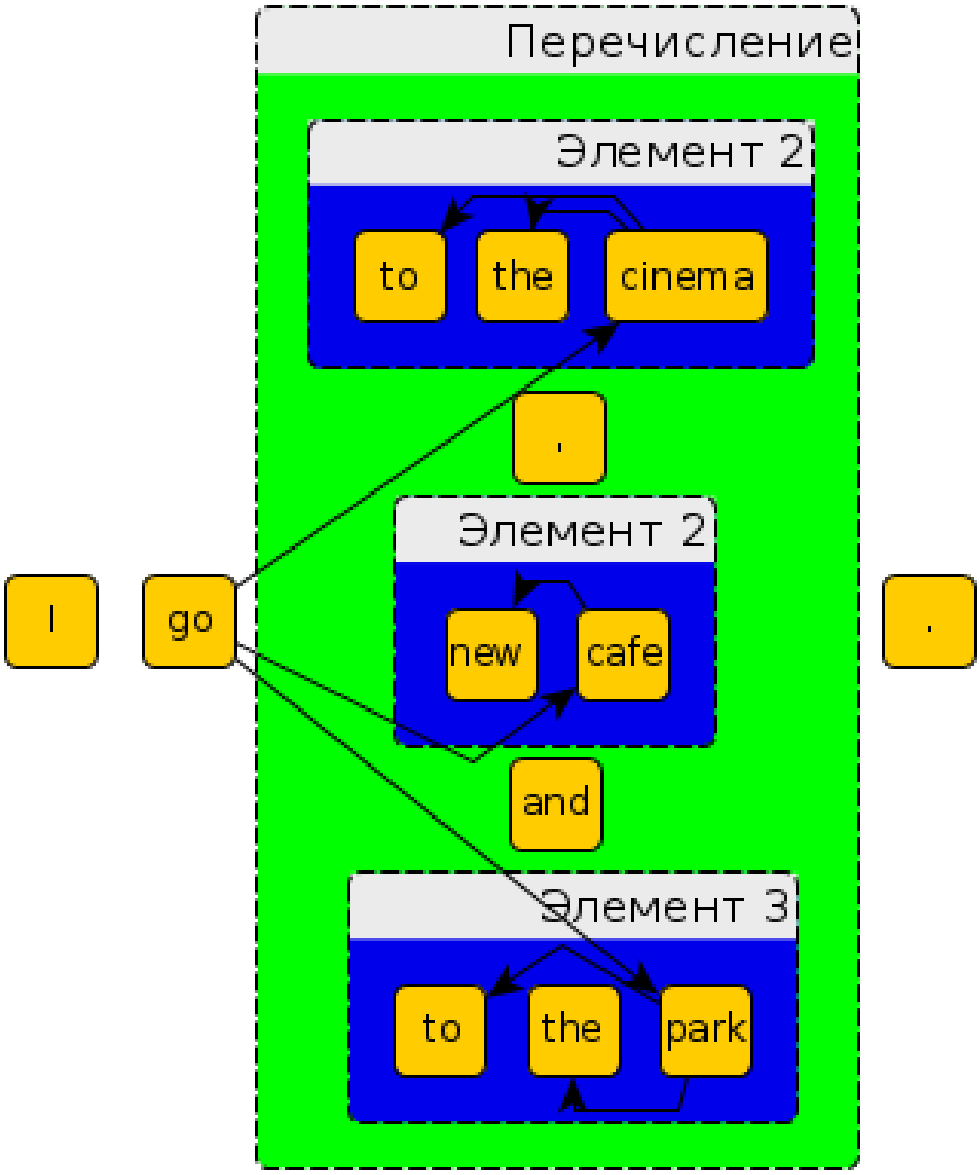

Рисунок 8 — Предложение с однородными сказуемыми, осложненными зависимыми лексемами 1

2.1.6 Сложносочиненые предложения

I see a big group of people which contains three subgroups: children, women and men, who were dressed in blue overalls or strange suits with red and green lines.

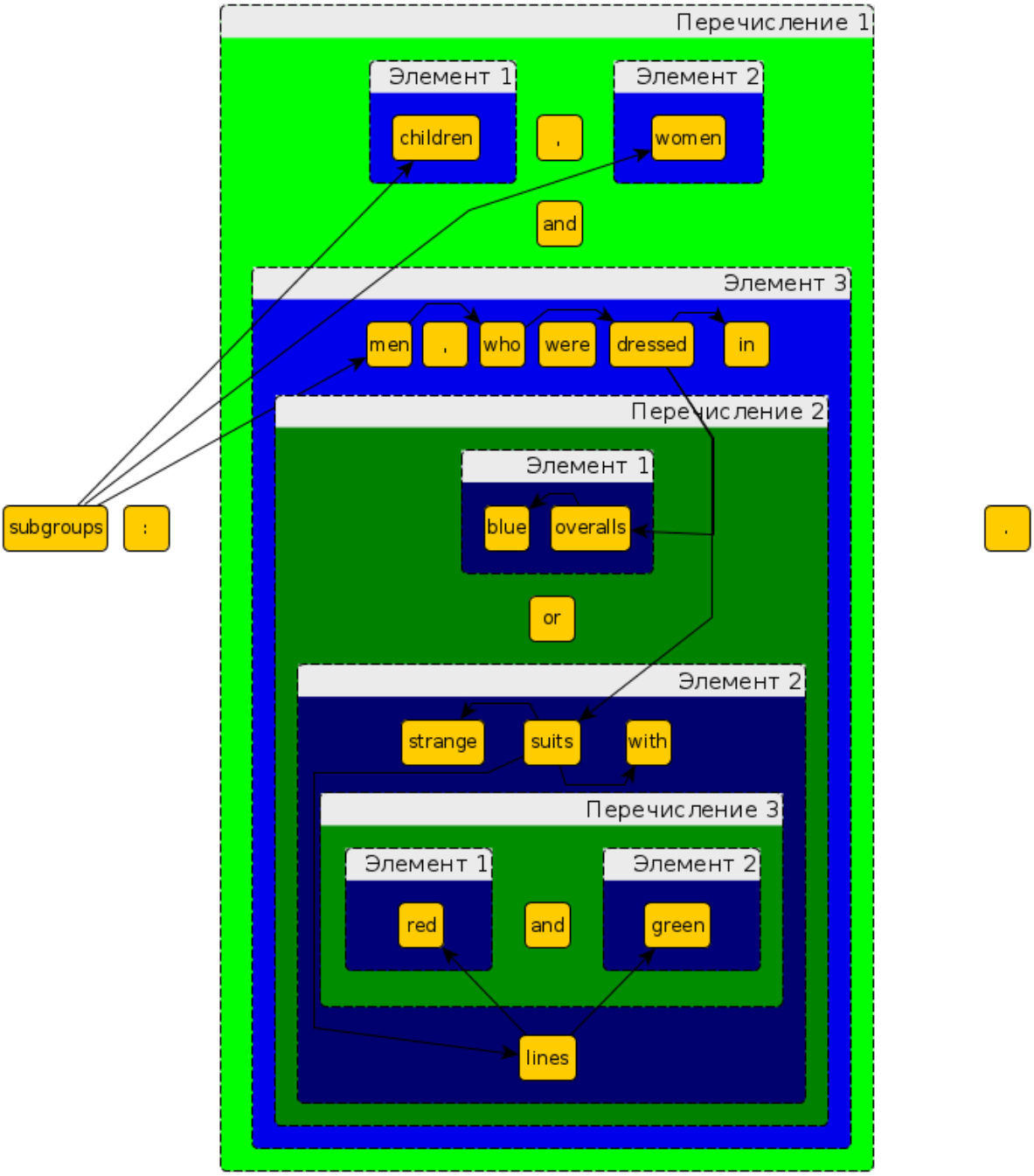

Рисунок 9 — Предложение с однородными сказуемыми, осложненными зависимыми лексемами

2.2 Построение модели

Список использованных источников

1 Найденова К. А. , Невзорова О. А. , Машинное обучение в задачах обработки естественного языка: Обзор современного состояния исследований / К. А. Найденова, О. А. Невзорова; Ученые записки Казанского государственного университета. Физико-математические науки. - Казань, 2008. - Том 150, книга 4, 20 с.

2 Аматов, А.М. Информационная энтропия как фактор конвергенции синтаксических структур в языках разных типов(на примере русского и английского языков) / А. М. Аматов // Вест. Волгогр. гос. ун-та, Сер. 2, Языкозн. / ВолГУ. - Волгоград - №2 - С. 142-146

3 Gernot A. Fink, Markov Models for Pattern Recognition: From Theory to Applications / Gernot A. Fink; Springer Berlin Heidelberg. - Berlin, 2007. - 248 p.

4 Кластеризация [Electronic resource]. – Mode of access : http://www.machinelearning.ru/wiki/index.php?title=(date of access 12.03.2015).

5 Бериков В. С., Лбов Г. С. Современные тенденции в кластерном анализе // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы», 2008. — 26 с.

6 Quinlan, J. R., (1986). Induction of Decision Trees. Machine Learning 1: 81-106, Kluwer Academic Publishers

7 Adam Coates and Andrew Y. Ng. Learning Feature Representations with K-means, Stanford University, 2012

8 Девятков В. В. Системы искусственного интеллекта / Гл. ред. И. Б. Фёдоров. — М.: Изд-во МГТУ им. Н. Э. Баумана, 2001. — 352 с.

9 Pauls A. , K-Best A∗ Parsing. / A. Pauls, D. Klein; University of California, Berkeley. - California, 2009. - 9 p.

10 Durrett G. , Neural CRF Parsing. / G. Durrett, D. Klein; University of California, Berkeley. - California, 2015. - 11 p.

11 Liang P. , Agreement-Based Learning. / P. Liang, D. Klein, M. I. Jordan; University of California, Berkeley. - California, 2008. - 8 p.

12 Petrov S. , Improved Inference for Unlexicalized Parsing. / S. Petrov, D. Klein; University of California, Berkeley. - California, 2007. - 8 p.

13 Durrett G. , Learning-Based Single-Document Summarization with Compression and Anaphoricity Constraints. / G. Durrett, T. Berg-Kirkpatrick, D. Klein; University of California, Berkeley. - California, 2016. - 11 p.