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 ВРБ-40-461-806-10.19-09.03.04-02-15
(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

ВРБ-40-461-806-10.19-09.03.04-02-15-81

Листов 17

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

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

1.1 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 тектса;
д) обучение языку и др.

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

1.2 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.

1.2.1 Обучение на правилах

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

1.2.2 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.

1.2.3 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.

1.2.4 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.

1.2.5 Алгоритмы типа k-средних

Наиболее популярный алгоритм кластеризации, стремящийся минизировать 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.

1.3 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".

1.3.1 Методы максимизации энтропии

Данный метод классификации основан на понятии информационной энтропии. Информационная энтропия - это мера неопределенности вероятностного распределения.

Еще не нашел хорошего описания этого алгоритма.

1.3.2 Скрытые макрковские модели

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.

1.4 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.

1.5 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.

1.6 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.

1.6.1 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.

1.6.2 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.

1.6.3 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.

1.6.4 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.

1.6.5 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.

1.6.6 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

Определение структур языка являющийся перечислениями сводится определению структур языка, перестановка элементов которых, не изменяет семантического смысла предложения и не нарушают синтакического строения предложения. Примерами таких структур служат однородные члены предложения:

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

Другим примером служит сложносочиненные преложения.

2.2 Построение модели