Классификация задач компьютерной лингвистики для создания инструмента по управлению потоками задач

П. А. Чересов, email: veterok.email@mail.ru[1]

[1]Московский Авиационный Институт

Аннотация. *В статье рассматриваются задачи компьютерной лингвистики и их разделение на базовые и прикладные. Распределение проводится с целью создания инструмента по управлению потоками задач, связанных с компьютерной лингвистикой. Он позволит снизить нагрузку на оборудование при наличии большого количества задач различной сложности.*

Ключевые слова: *компьютерная лингвистика, анализ текста, прикладные задачи компьютерной лингвистики, задачи этапов анализа текста*

Введение

В настоящее время накопилось очень большое количество текстовой информации на естественных языках. В связи с этим всё больше возрастает потребность автоматизации ее обработки. Решением этих задач занимается наука компьютерная лингвистика (КЛ) [1].

КЛ занимается разработкой теоретических основ и созданием алгоритмов и программного обеспечения (ПО) для решения языковых задач, обработкой текста на естественных языках и оптимизацией человеко-машинного взаимодействия [3]. Одним из самых важных направлений КЛ является автоматизированный анализ текста, который принято разделять на следующие этапы:

1. Графематический анализ – выделение из текста синтаксических и структурных единиц: абзацы, предложения, слова и знаки препинания.
2. Морфологический анализ – приведение слов к нормальной форме и назначение им соответствующих параметров таких как: часть речи, род, падеж, время и т. д.
3. Предсинтаксический анализ – разделение одной лексической единицы на несколько синтаксических и наоборот, а также проведение синтаксической сегментации.
4. Синтаксический анализ – определение роли слов и их связей друг с другом. Результат возвращается в виде набора деревьев [2].
5. Постсинтаксический анализ – нормализация полученных деревьев, т.е. сведение конструкций, выражающих одну и ту же мысль разными способами на разных языках, в одному и тому же виду.

6. Семантический анализ – выявление смысла слов и формирование между ними семантических связей.

Отдельным направлением КЛ является генерация текста, она используется при создании ответа пользователю, например, при переводе различных документов или в ходе диалога, т.е. в ходе выполнения прикладных задач КЛ.

С возрастанием поступающей информации на обработку и требуемого качества результатов увеличивается и нагрузка на оборудование. Это может привести к тому, что новые задачи невозможно будет начать или завершить, из-за того, что какая-то одна из них может занимать все ресурсы оборудования.

Решением этой проблемы является создание инструмента по управлению потоками задач, ориентированного на автоматическую обработку текстов на естественном языке. Необходимость создания отдельного инструмента заключается в сложности определения необходимых ресурсов для решения прикладных задач КЛ, т.к. это в большой степени зависит от входных данных – текстов с различным количеством предложений, слов, омонимов всех типов, разной синтаксической структуры и т.д., а не от размера файла с исходным текстом. Для этого необходимо проанализировать и классифицировать задачи КЛ с точки зрения вычислительной сложности.

1. Определение классов задач

Каждый из этапов анализа текста включает в себя набор специфических задач, например, выделение отдельных слов для графематического анализа, или разрешение морфологической омонимии для морфологического анализа. В рамках работы такие задачи называются базовыми. Практически любая прикладная задача КЛ реализуется с их применением.

Под прикладной задачей понимается набор отдельных базовых задач или некоторый специфический алгоритм, работающий с текстовыми данными, полученными в результате обработки на предыдущих этапах, с целью получения практически ценного результата. Они также могут включать в себя другие прикладные задачи.

Существуют следующие наиболее востребованные прикладные задачи КЛ:

- Машинный перевод.
- Автоматическое аннотирование текстов.
- Автоматическое извлечение ключевых слов.
- Информационный поиск.
- Реферирование текстов.
- Автоматизация создания и редактирование текстов.

– Генерация текстов на ЕЯ.
– Формирование ответов на вопросы.
– Организация диалога на ЕЯ.
– Извлечение информации из текстов.
– Классификация и кластеризация текстов.
– Анализ мнений и оценка тональности текстов.

Из этого списка в данной статье были проанализированы три задачи: автоматическое аннотирование текстов, автоматическое извлечение ключевых слов и информационный поиск.

2. Выделение базовых задач этапов анализа текста

На каждом этапе анализа текста выполняется ряд базовых задач:

1. Графематический этап анализа текста преимущественно содержит задачи, направленные на выделение синтаксических и структурных единиц. В основном эти единицы составляют: абзац, предложения, слова и знаки препинания.

2. Морфологический этап анализа текста содержит такие базовые задачи, как определение нормальной формы слова, определение набора параметров для слова и разрешение морфологический омонимии. Набор параметров для слова включает следующее: часть речи, род, падеж, число, спряжение, склонение, время и т.д. Морфологические омонимия – это ситуация, когда одному слову можно соответствует несколько наборов морфологических характеристик.

3. Предсинтаксический этап анализа текста содержит такие базовые задачи, как: объединение отдельных лексических единиц, разделение одной лексической единицы, свёртка числительных и синтаксическая сегментация. Синтаксическая сегментация – это выделение фрагментов предложения, которые можно разобрать независимым образом [2].

4. Синтаксический этап анализа текста содержит следующие такие базовые задачи, как определение синтаксической роли слов и синтаксических связей между ними.

5. Постсинтаксический этап анализа текста содержит базовые задачи, направленные на уточнение смысла слов и нормализацию синтаксического дерева.

6. Семантический этап анализа текста содержит базовые задачи построения семантических сетей и онтологий.

Итоговая схема распределения базовых задач по этапам анализа текста представлена на рис. 1.

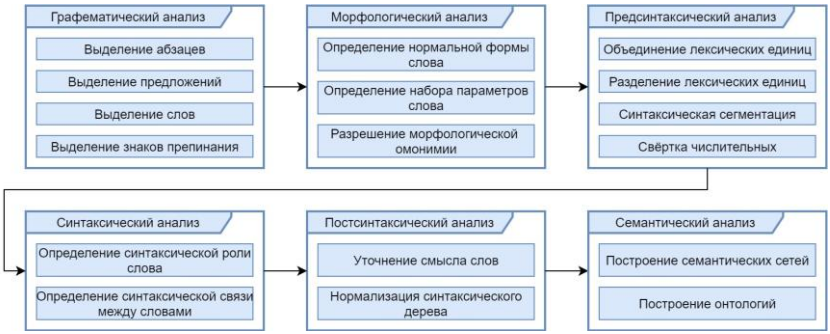

Рис. 1. Базовые задачи

На каждом этапе анализа текста должны в зависимости от решаемой прикладной задачи могут быть выполнены не все базовые задачи. На выполнение может подаваться только некоторый перечень, который необходим для какой-либо прикладной задачи.

3. Автоматическое аннотирование текстов

Аннотация – краткое изложение содержания какого-либо текста. У аннотации есть две функции [5, 6]:

– Сигнальная – кратко предоставить информацию об аннотируемом тексте чтобы читатель понял, нужна ли ему информация из полного текста.

– Поисковая – помогает найти требуемую информацию.

Автоматическое аннотирование текстов – извлечение из него информации о его содержании в кратком виде при помощи компьютерных технологий.

Алгоритмы автоматического аннотирования текстов принято делить на 2 группы [5]:

1. Вытягивающие алгоритмы. Находят в исходных текстах фрагменты наибольшей статистической и лексической важности, а затем объединяют их. Такие алгоритмы не требуют больших вычислительных мощностей, т.к. не проводят семантического анализа текста. Однако при этом уменьшается качество результата.

2. Генерирующие алгоритмы. Отличаются от вытягивающих присутствием семантического анализа, что увеличивает качество результата (отсутствие дублирования, полнота аннотирования, наличие семантических связей).

Итоговая схема распределения задачи аннотирования текстов по базовым задачам изображена на рис. 2.

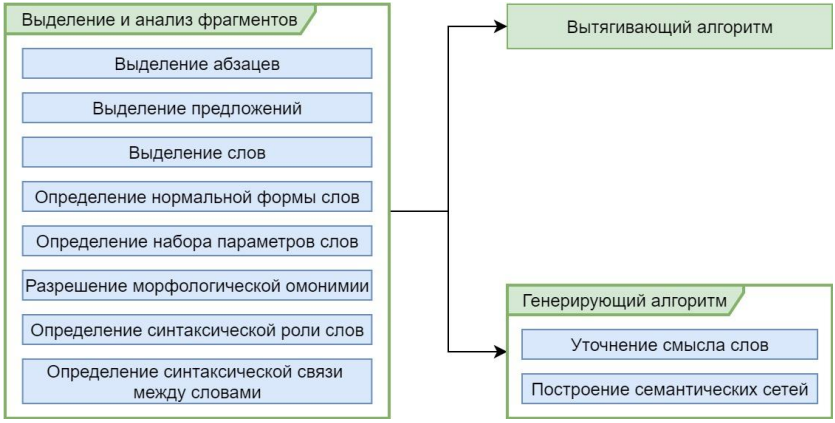

Рис. 2. Автоматическое аннотирование текстов

На этапе выделения и анализа фрагментов выполняется больше всего базовых задач. Под фрагментами могут пониматься предложения, а иногда и целые абзацы, в соответствии с этим выполняются базовые задачи по их выделению. Выделение слов нужно для последующего морфологического и синтаксического анализа.

Несмотря на то, что прикладная задача выделения фрагментов содержит больше всего базовых, более ресурсоёмкой является задача генерации. Связано это с тем, что для её выполнения приходится решать задачи семантического этапа анализа – самого сложный из всех. Использование вытягивающего алгоритма не требует решения дополнительных базовых задач.

4. Автоматическое извлечение ключевых слов

Ключевые слова – особо важные слова в тексте, набор которых даёт читателю описание содержания текста, в котором они находятся [6]. Соответственно, автоматическое извлечение ключевых слов – процесс поиска группы ключевых слов в требуемом тексте.

Данный процесс разделён на 3 этапа:

– Приведение текста в удобный для распознавания ключевых слов формат. Для этого проводится графематический, морфологический и, в случае гибридной модели, синтаксический анализ. На этом же этапе удаляются все стоп-слова – слова, которые не несут смысловой нагрузки [7].

– Распознавание потенциальных кандидатов на ключевые слова.

– Фильтрация. Каждому из слов присваивается некоторый набор признаков, который затем сравнивается с эталонными значениями, на

основании которых и определяется принадлежность этих слов к множеству ключевых. В зависимости от метода извлечения ключевых слов на этом этапе, возможно, нужно будет использовать словари [6].

В автоматическом извлечении ключевых слов существуют 2 модели:

1. Статистическая модель. Преимуществом выступает универсальность и независимость от лингвистических баз знаний.

2. Гибридная модель. Методы извлечения ключевых слов дополняются применением лингвистических процедур.

Итоговая схема распределения задачи автоматического извлечения ключевых слов по базовым задачам изображена на рис. 3.

Рис. 3. Автоматическое извлечение ключевых слов

Самой простой прикладной подзадачей, является «Удаление стоп-слов». Она является важной частью при извлечении ключевых слов, т.к. стоп-слова зачастую могут встречается в тексте довольно часто и их извлечение в качестве ключевых априори будет является неверным результатом.

Гибридная модель дополняет статистическую модель дополнительными процедурами. Это указано на схеме путём разветвления от прикладной задачи «Статистическая модель», в рамках которой выполняются базовые задачи морфологического анализа.

Гибридная модель требует дополнительно выполнить базовые задачи предсинтаксического и синтаксического анализа. Для выполнения этой задачи требуется больше ресурсов, по сравнению с остальными.

5. Информационный поиск

Информационный поиск является одной из самых важных прикладных задач КЛ. Благодаря ему пользователь может найти нужную ему информацию среди огромного количества различных документов, ресурсов, текстов, которыми заполнен интернет [8, 9]. Основным

инструментом информационного поиска в интернете являются вербальные поисковые системы (например, Google, Yandex).

Большое значение имеют возможности языков запросов, которыми оперирует пользователь. Её структура в основном включает следующее компоненты [8]:

1. Поисковые элементы.
2. Средства морфологической нормализации.
3. Булевские операторы.
4. Дополнительные условия поиска (ограничение области поиска, поиск в определённых частях документа и т. д.)

Морфологическая нормализация является рекомендуемым требованием при информационном поиске, но не является обязательным. Вместо этого можно сделать следующее:

– Каждую словоформу можно рассматривать как отдельную лексическую единицу. Однако это ведёт к значительному увеличению объёма данных, а также вызывает неудобства для пользователя.

– В поисковом запросе можно вводить только часть слова. В этом случае механизм поиска будет искать слова, которые имеют эту часть.

В итоге, получается схема распределения, изображённая на рис. 4.

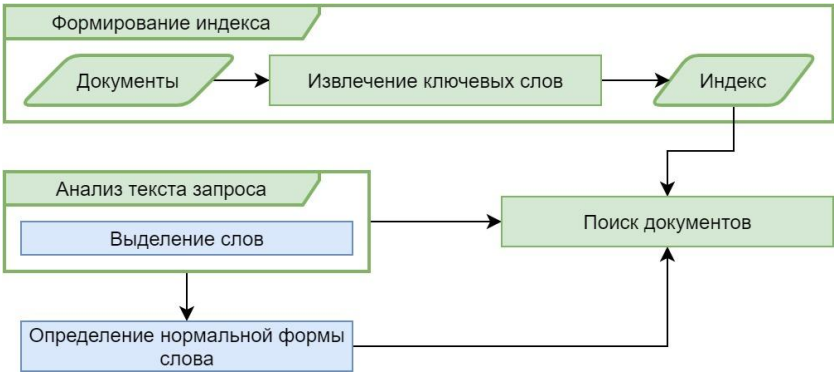

Рис. 4. Информационный поиск

На данной схеме параллелограммами обозначаются ресурсы, которые можно обработать. Формирование индекса включает в себя задачу КЛ извлечения ключевых слов, которая уже была рассмотрена в предыдущем разделе.

«Анализ текста запроса» включает в себя базовую задачу выделения слов, без которой не обойтись. Определение нормальной формы слов текстового запроса не является обязательным.

Заключение

Анализ задач КЛ необходим для создания инструмента по управлению потоками задач, связанных с автоматической обработкой текста. Для этого для некоторых задач КЛ было сделано распределение их по базовым и прикладным.

Разрабатываемый инструмент требуется для более равномерного распределения нагрузки на систему, которая может привести к невозможности выполнения других задач. На основании предложенной классификации будут определены весовые коэффициенты и приоритеты задач при решении задачи диспетчеризации.

Разделение базовых задач в соответствии с этапами анализа текста позволит учесть особенности потребления ресурсов с учетом лингвистических особенностей текстов, что даст возможность более оптимальным образом обеспечить необходимую обработку текстовых данных.

Список литературы

1. Дюжева А. Н. Информационные технологии в лингвистике. – 2021.
2. Большакова Е. И. и др. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. – 2011.
3. Аслонов Ш. Ш. и др. Компьютерная лингвистика и филология: проблемы и решения //Гуманитарный трактат. – 2020. – №. 84. – С. 17-19.
4. Макшанова Т. В. ВИДЫ АННОТАЦИЙ //ББК 72 Н108. – 2018.
5. Бисикало О. В., Назаров И. А. Обзор методов автоматического аннотирования текстов //Научные труды Винницкого национального технического университета. – 2013. – №. 2.
6. Ванюшкин А. С., Гращенко Л. А. Методы и алгоритмы извлечения ключевых слов //Новые информационные технологии в автоматизированных системах. – 2016. – №. 19. – С. 85-93.
7. Шереметьева С. О., Осминин П. Г. Методы и модели автоматического извлечения ключевых слов //Вестник Южно-Уральского государственного университета. Серия: Лингвистика. – 2015. – Т. 12. – №. 1. – С. 76-81.
8. Захаров В. П. Лингвистические средства информационного поиска в Интернете //Библиосфера. – 2005. – №. 1. – С. 63-71.
9. Урвачева В. А. Обзор методов информационного поиска //Вестник Таганрогского института имени АП Чехова. – 2016. – №. 1. – С. 457-463.