Сидоров Вадим Вячеславович

студент

ФГАОУ ВО «Новосибирский национальный

исследовательский государственный университет»

г. Новосибирск, Новосибирская область

СИСТЕМА СТРУКТУРНО-ЖАНРОВОЙ СЕГМЕНТАЦИИ ТЕКСТА НА ОСНОВЕ ЕГО ФОРМАЛЬНОЙ ЖАНРОВОЙ СТРУКТУРЫ

Аннотация: *в данной статье описывается система сегментации текста, использующая жанровую модель документов. Система позволяет определить принадлежность документа определенному жанру и разделить его на сегменты текста. Такой подход во время предварительной обработки документов позволяет уменьшить нагрузки на основной анализирующий модуль за счет уменьшения объема рассматриваемых текстов.*

Ключевые слова: *жанровая сегментация, жанровый сегмент, жанровая модель.*

DOI: 10.21661/r-112083

1. Введение

Извлечение информации из текста и ее последующая обработка являются актуальными задачами на сегодняшний день. Добавление в современные анализирующие программы блоков предварительного анализа, позволяющих проводить анализ структур документов, помогает обнаружить искомую информацию в тексте, не исследуя его полностью, что значительно сокращает время выполнения алгоритмов основного анализа за счёт уменьшения массива входных данных.

Каждый текстовый документ обладает определенной структурой [3, с. 1; 4, с. 3], что по Бахтину соответствует классическому определению жанра. Жанр – это типовая модель построения речевого целого [1, с. 307]. «Типичная воспроизводимая жанровая форма» документа, его жанровая модель, позволяет определить семейство документов, принадлежащих одному жанру [2, с. 7; 5, с. 172].

В данной статье будет рассмотрена система, позволяющая формировать жанровые структуры документов, а также проверять тексты на соответствие этим жанровым моделям.

2. Структурно-жанровая модель текста

Структура каждого текста определенного жанра может быть представлена тремя логическими уровнями: уровень жанровой модели, уровень жанровых сегментов и уровень маркеров.

Маркер M является конечной последовательностью символов текста. При обнаружении этой последовательности в документе можно описать экземпляр маркера двумя числами – начальным и конечным индексами найденной в тексте последовательности. Таких экземпляров может быть найдено несколько, однако должно выполняться условие: маркеры не должны пересекаться.

Имея наборы начальных и конечных маркеров можно определить жанровый сегмент S как пятерку вида $S = < M_B, M_E, I, P^I, T >$, где $M_B, M_E \in M$ – начальные и конечные маркеры, $I \subset S$ – множество вложенных сегментов, $P^I \subset I \times I$ – нерефлексивное, транзитивное, антисимметричное бинарное отношение частичного порядка над I, а T – тип сегмента. Экземпляр жанрового сегмента так же, как и экземпляр маркера, описывает вхождения данного сегмента в текст с помощью двух чисел – начального индекса и конечного индекса.

Жанровые сегменты могут быть нескольких типов. Для каждого типа сегмента задан набор аксиом и определено отображение, задающее правило построения F экземпляров сегментов. $F: FR_{Text} \times S \to S_{ex}$, где S_{ex} – экземпляр сегмента $s \in S$, построенного по фрагменту текста $fr_{Text} \in FR_{Text}$.

Простой сегмент имеет минимальную структуру, содержащую только множества начальных и конечных маркеров, по которым можно однозначно найти экземпляр данного сегмента в тексте. Сложный сегмент, в отличие от простого, дополнительно к наборам маркеров содержит в себе набор внутренних сегментов. Таким образом, можно определить иерархию сегментов, описав сложную жанровую структуру. Для формирования экземпляра сложного сегмента для

начала требуется определить экземпляры внутренних сегментов. Их совокупность определяет внутреннюю область сложного сегмента, которая должна принадлежать фрагменту текста, сопоставляемого экземпляру сегмента. Стоит отметить, что для каждого рассматриваемого в данный момент фрагмента текста формируется только один экземпляр сегмента.

Остальные типы сегментов являются вариацией сложного сегмента. Последовательные сегменты – это сложный сегмент, между внутренними сегментами которого не могут содержаться символы, в них не входящие. Повторяющийся сегмент – это сложный сегмент, содержащий внутри себя только один сегмент, который может встретиться в тексте один или более раз. Тип альтернативные сегменты может обозначить те вложенные сегменты, которые являются взаимозаменяемыми. Тип комбинация сегментов необходим в тех случаях, когда порядок взаиморасположения внутренних сегментов неизвестен. Факультативный сегмент – сегмент, который необязательно встречается в тексте. Сегмент типа отрицание сегмента может задаваться как внутренний сегмент другого сегмента и не должен быть найден в его границах.

Жанровая модель $Model$ является корнем иерархической жанровой структуры. Она содержит в себе только набор главных сегментов. $Model = < I, P^I >$, где $I \subset S$, $P^I \subset I \times I$ – нерефлексивное, транзитивное, антисимметричное бинарное отношение частичного порядка. Имея жанровую модель и документ, можно определить, принадлежит ли данный документ этой модели. Имея набор нескольких моделей можно попытаться определить жанр документа, последовательно проверив его на соответствие этим жанровым моделям.

3. Система структурно-жанровой сегментации

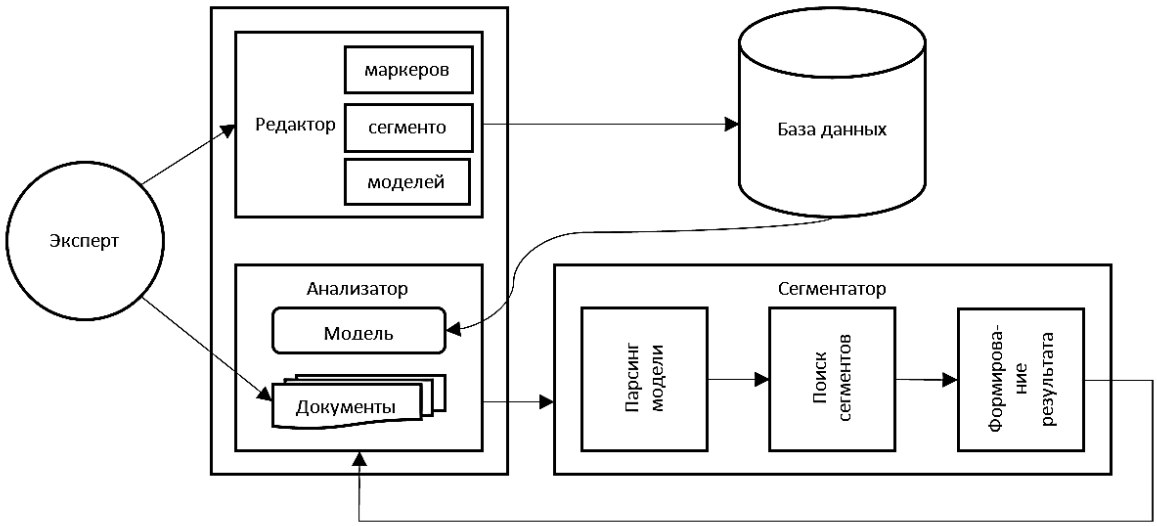

Рис. 1. Архитектура системы сегментации текста

Система структурно-жанровой сегментации текста состоит из трех частей (рис. 1): ядро (сегментатор), визуальный редактор и база данных жанровых структур. Сегментатор предназначен для разбиения документа на фрагменты с целью определения принадлежности текста к рассматриваемым жанрам. Это позволяет определять жанр исследуемого теста относительно уже известных жанров. Визуальный редактор позволяет конструировать сложные иерархические системы сегментов и добавлять их в базу данных. База данных, в свою очередь, содержит описание моделей, сегментов и маркеров и позволяет в любой момент получить доступ к уже созданным ранее моделям.

3.1. Визуальный конструктор

Визуальный редактор содержит четыре раздела: анализатор, раздел конструирования маркеров, раздел конструирования сегментов и раздел конструирования моделей.

Раздел конструирования маркеров позволяет создавать и сохранять маркеры. Раздел конструирования сегментов позволяет создавать и сохранять сложные иерархические структуры сегментов, используя созданные ранее сегменты и маркеры. Раздел конструирования моделей позволяет формировать конечные

жанровые структуры (модели) и использовать их в анализаторе для определения жанра документа.

Раздел анализатора позволяет загрузить документ и проверить его на соответствие определенному жанру, загруженному из базы данных, либо позволяет попытаться определить жанр автоматически (будет произведен анализ документа на соответствие всем жанровым моделям, находящимся в базе данных).

3.2. Сегментатор

Сегментатор состоит из двух модулей: основного модуля и парсера. Парсер используется для анализа входящего XML-файла и формирования структуры жанровой модели. Основной модуль сегментатора позволяет запустить анализ текста, определить для каждого сегмента его экземпляр и сформировать новую результирующую XML-структуру.

Каждый сегмент содержит в себе свой экземпляр. Изначально он пуст, но после анализа может быть определен конкретными значениями. Такие сегменты как отрицание сегмента могут не иметь каких-либо позиций в тексте и в этом случае содержат пустой экземпляр. При формировании результата для каждого сегмента модели его экземпляр проверяется на пустоту. При положительном результате такой сегмент не включается в XML-структуру. Также, если сегмент содержит внутренние сегменты, для них рекурсивно запускается такая же процедура.

Ниже предложен алгоритм сегментирования текста на основе заданной жанровой модели документа.

Ввод: структура модели в XML формате (содержит иерархию сегментов, каждый сегмент содержит свои наборы маркеров), текст. $\mathrm{i} = 0$.

1. Взять фрагмент текста FR^{i} ($\mathrm{FR}^{0} = \mathrm{Text}$) сегмента i.

1. Построить вектора экземпляров начальных и конечных экземпляров маркеров $\mathrm{V}_{\mathrm{b}}^{\mathrm{i}}, \mathrm{V}_{\mathrm{e}}^{\mathrm{i}}$ для FR^{i}.

2. Определить область $\mathrm{S}_{\mathrm{i}} = (\mathrm{a}, \mathrm{b}) \subseteq \mathrm{FR}^{\mathrm{i}}, \mathrm{a} = \min\left(\mathrm{V}_{\mathrm{b}}^{\mathrm{i}}\right), \mathrm{b} = \max(\mathrm{V}_{\mathrm{e}}^{\mathrm{i}})$ сегмента i.

3. Для n внутренних сегментов:

– $i = i'$;

– рекурсивно с первого пункта запустить алгоритм с параметрами: $FR^{i'} = S_i$;

– переопределить начало внутренней области: $S_i^a = S_{i'}^b + 1$.

4. $V_b^i \to \forall v \in V_b^i \; v < S_{i1}^a$. $V_e^i \to \forall v \in V_e^i \; v > S_{in}^b$ – отфильтровать экземпляры маркеров. Экземпляры маркеров сегмента не должны лежать в области его внутренних сегментов.

5. $S_i = (a, b), b = \min(V_e^i), a = \max(V_b^i)$ – определить итоговую область сегмента. Если нельзя, то EXIT.

Вывод: XML структура, содержащая облегченную иерархию сегментов. Если анализ текста окончился неудачей, и основной алгоритм преждевременно завершил работу, в файле вывода это будет отражено, но структура все равно будет сформирована, определяя те сегменты, которые были успешно найдены до завершения работы.

4. Заключение

В статье рассмотрена система, позволяющая формировать модели документов, а также производить анализ текста на его соответствие этим моделям. Система может быть использована для предварительной обработки документов. Она позволяет разбить документ на определенные области, что упрощает поиск необходимой информации в нем. Также, системы обработки, анализа и синтеза текста могут использовать результаты работы данной системы, чтобы уменьшить объемы входных данных. Данная система была апробирована на жанровой модели «Резюме». Полученные результаты показали работоспособность предложенной системы. В дальнейшем она будет апробирована на расширенном наборе жанровых моделей. Планируется интеграция системы с различными системами обработки текста. Также планируется расширение системы до мультипользовательской и дополнение структуры базы данных таким образом, чтобы хранящиеся в ней модели могли быть разделены по темам, предметным областям. Такой

подход позволит ускорить определение жанра документов, так как анализ будет происходить только в определенном контексте предметной области.

Список литературы

1. Бахтин М.М. Эстетика словесного творчества / Сост. С.Г. Бочаров, примеч. С.С. Аверинцев и С.Г. Бочаров. – М.: Искусство, 1979. – 423 с.

2. Кибрик А.А. Модус, жанр и другие параметры классификации дискурсов // Вопросы языкознания. – 2009. – №2. – С. 3–21.

3. Кононенко И.С. Жанровые аспекты классификации веб-сайтов / И.С. Кононенко, Е.А. Сидорова. – С. 32–40.

4. Кононенко И.С. Обработка делового письма в системе документооборота / И.С. Кононенко, Е.А. Сидорова // Труды международного семинара «Диалог’2002» по компьютерной лингвистике и ее приложениям. – Т. 2. – М.: Наука, 2002. – С. 299–310.

5. Щипицина Л.Ю. Жанры компьютерно-опосредованной коммуникации. – Архангельск: Поморский университет, 2009. – 238 с.