УДК 004.912

Подход к описанию жанра текста на основе его формальной жанровой структуры

Сидоров В.В.

В статье представлено описание математической модели представления документов. Модель включает в себя описания различных типов сегментов, которые определяются в тексте с помощью маркеров. Для данной модели приведен алгоритм сегментирования текста документа.

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

1. Введение

Важной задачей на сегодняшний день являются обработка и извлечение информации из текста. Разработчики современных систем, предназначенных для анализа различных документов, стараются включить в них программные блоки, отвечающие за более глубокий анализ текста. Причина расширения анализирующих систем такими программными блоками заключается в стремлении находить и извлекать наиболее точную информацию из них.

Каждый текстовый документ обладает определенными чертами [4, 5], которые определяются спецификой контекста этого документа и формируются благодаря сходству тематического содержания, единству стиля и композиционного построения, что соответствуют классическому определению жанра речевого произведения, сформулированному М.М. Бахтиным [1]. Жанр – это типовая модель построения речевого целого. Для каждого документа должна быть определена конкретная жанровая модель, представляющая его «типичную воспроизводимую жанровую форму».

Каждая жанровая модель представляет собой общую структуру документов, принадлежащих этой модели [3, 6]. Таким образом, каждый документ может быть разбит на структурные части, называемые жанровыми сегментами, ограничивающие определенную логическую область в тексте с помощью маркеров [2]. Эти сегменты, в свою очередь, могут быть также представлены своей жанровой моделью, еще сильнее структурируя документ.

В данной статье будет рассмотрен способ описания формальных жанровых структур документа и предложен алгоритм структурно-жанровой сегментации текста относительно данных структур.

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. Маркеры представляют собой конечные последовательности символов, которые могут быть найдены тексте. Найденные маркеры (экземпляры) описываются начальной и конечной позицией в тексте.

Жанровый сегмент определяется наборами начальных и конечных маркеров. Экземпляр жанрового сегмента так же, как и экземпляр маркера, определяет границы данного сегмента в тексте. Жанровые сегменты могут быть нескольких типов. Для каждого типа сегмента задан набор аксиом и определено отображение, задающее правило построения экземпляров сегментов.

Жанровая модель является корнем иерархической жанровой структуры. Она содержит в себе набор главных сегментов, но не имеет границ, в отличие от сложного сегмента. Имея жанровую модель и документ, можно определить, принадлежит ли данный документ этой модели. Имея набор моделей, можно попытаться определить жанр документа, последовательно проверив его на соответствие этим жанровым моделям.

Разбиение документа на различные логические сегменты позволяет локализировать искомые блоки текста и рассматривать в конкретном контексте только необходимые области документа, игнорируя его оставшиеся части, являющиеся лишними в данном контексте. Применение этого метода при анализе больших объемов текстовых данных позволяет значительно сократить время выполнения специфичных анализирующих операций за счет уменьшения предназначенного для анализа массива данных.

2.1 Модель документа

Модель $Model$ является верхним уровнем жанровой структуры документа, она агрегирует в себе основные сегменты, которые, в свою очередь, имеют внутреннюю сложную структуру из других сегментов и маркеров, определяющих их границы (рис. 1).

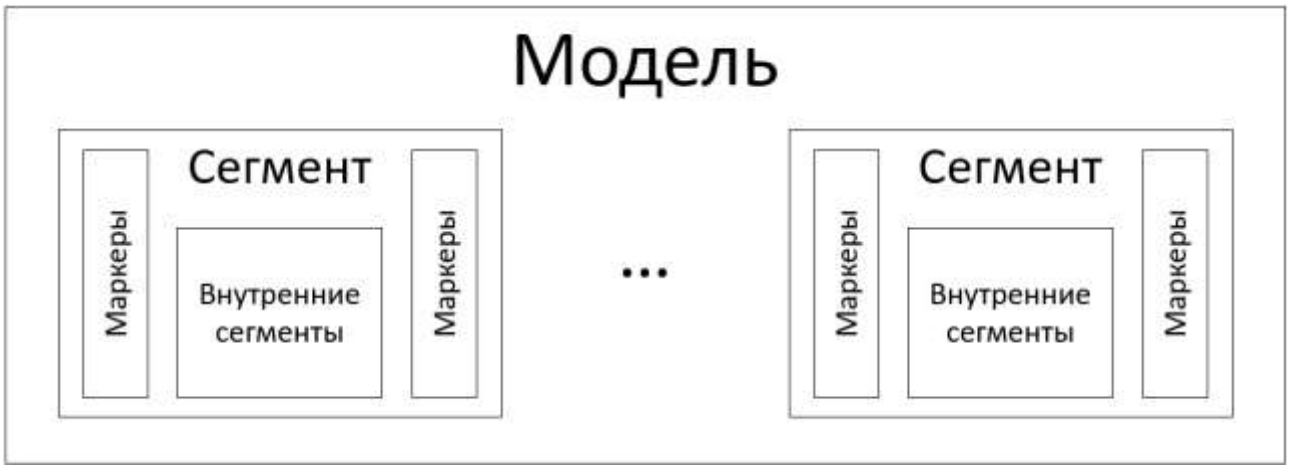

Рис. 1. Модель документа

$Model = < I, P^I >$, где I – множество сегментов, $P^I \subset I \times I$ – нерефлексивное, транзитивное, антисимметричное бинарное отношение частичного порядка.

2.2 Маркер

Алфавит A содержит в себе все заглавные и строчные буквы языка, а также специальные символы, такие как знаки препинания или символы конца и начала текста.

$A^* = \bigcup_{i=0}^{\infty} A^i$, где $A^i = \{\omega\upsilon \mid \omega \in A^{i-1}, \upsilon \in A\}, i \in \mathbb{N}$ – последовательности символов длины i, $A^0 = \{\varepsilon\}$, ε – пустой символ.

A^* - множество всех конечных последовательностей символов алфавита ($*$ - звезда Клини).

Текст $T \in A^*$ – произвольная конечная последовательность символов алфавита.

Фрагмент текста T – это $FR_T = \tau \in A^* \Leftrightarrow \exists\, \omega \in A^* \exists\, \upsilon \in A^*: \omega\tau\upsilon = T$.

Маркер M представляет из себя набор последовательностей символов. Можно считать, что он выделяет определенные части текста, совпадающие с этими последовательностями.

$M = < D >$, где $D \subset A^*$.

Примеры маркеров:

- «Начало текста» = <{¶$^{\text{H}}$}>; «Конец текста» = <{¶$^{\text{K}}$}> - состоят только из одного символа;

- «Конец предложения» = <{«. », «? », «! », «… »}> - можно составить полный список всех знаков препинания, на которые предложение может заканчиваться;
- «Адрес» = <{«Адрес», «Место проживания», «Город»}> - содержит основные слова, которыми можно обозначить место проживания;
- «Обращение» = <{«Уважаемый», «Уважаемая», «Здравствуйте,», «Добрый день»}> - этот список можно пополнить различными дополнительными определениями, такими как «Добрый вечер», «Доброго времени суток», однако полный перечень таких словосочетаний в данной статье приводить нецелесообразно;
- «Имя» = <{«Иван», «Андрей», «Мария», «Елена»,..}> - к сожалению, в данном случае нельзя составить полный список всех существующих имен или определить их какой-либо формулой, можно только перечислить наиболее распространенные;

2.3 Сегмент

В ходе исследования удалось выделить два основных типа сегментов: простой сегмент и сложный сегмент, который, в отличие от простого, содержит в себе набор внутренних сегментов. Более того, сложный сегмент может быть представлен следующими вариациями:

- последовательные сегменты;
- повторяющийся сегмент;
- альтернативные сегменты;
- комбинация сегментов;
- факультативный сегмент;
- отрицание сегмента.

В общем виде жанровый сегмент S описывается следующей системой:

$$S = < M_B, M_E, i_B, i_E, I, P^I, l_L, l_R, Type >,$$

где $M_B, M_E \in M$ – начальные и конечные маркеры,

$i_B, i_E \in \{0; 1\}$ – булевы значения, определяющие, требуется ли включить в экземпляр сегмента 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,

$l_L, l_R \in \{0; 1\}$ – булевы значения, определяющие, должны ли вложенные сегменты быть размещены строго в начале и конце искомого сегмента,

$Type \in \{Type_{Simple}, Type_{Complex}, Type_{Sequence}, Type_{Repeatable}, Type_{Alternative}, Type_{Combination}, Type_{Optional}, Type_{Negative}\}$ – тип сегмента.

Кроме того, каждый тип сегмента имеет определенный набор аксиом, определяющих этот тип.

Жанровый экземпляр сегмента – это $S_{ex} = (b, e), b \in \mathbb{N}, e \in \mathbb{N}$.

Для каждого сегмента и для каждого текста определен конкретный экземпляр сегмента, определяющий положение этого сегмента в тексте. b, e – позиции начала и конца сегмента, нумерующиеся с 0. Таким образом, если $T' \in A^i$, то экземпляр сегмента S'_{ex} для T' – это пара чисел (b', e') таких, что $0 \leq b' \leq e' < i$ и $\tau \in FR_{T'}: \omega\tau\upsilon = T', \omega \in A^{b'}, \upsilon \in A^{i-e'-1}$ – искомый фрагмент текста.

Для каждого типа сегмента задано определенное отображение, определяющее правила построения F экземпляров сегментов:

$F: FR_T \times S \to S_{ex}$, где S_{ex} – экземпляр сегмента $s \in S$, построенного по фрагменту текста $fr_T \in FR_T$.

2.3.1 Простой сегмент

Простой сегмент является основным атомарным типом сегмента.

$S_{Simple} \subset S, F_{Simple}: FR_T \times S_{Simple} \to S_{ex}$.

Для данного типа сегмента определены следующие аксиомы:

1. $I = \emptyset$;
2. $P^I = \emptyset$;
3. $l_L = 0$;
4. $l_R = 0$;
5. $Type = Type_{Simple}$.

Пример:

«Предложение» = <{«Начало текста», «Конец предложения»}, {«Конец текста», «Конец предложения»}, 0, 1, $\emptyset, \emptyset, 0, 0, Type_{Simple}$>.

Тогда из текста «¶[H]*На комоде лежала какая-то книга. Он каждый раз, проходя взад и вперед, замечал ее; теперь же взял и посмотрел. Это был Новый завет в русском переводе. Книга старая, подержанная, в кожаном переплете.*¶[K]» (Ф.М. Достоевский, "Преступление и наказание") могут быть выделены следующие сегменты «Предложение»:

- *На комоде лежала какая-то книга.*
- *Он каждый раз, проходя взад и вперед, замечал ее; теперь же взял и посмотрел.*
- *Это был Новый завет в русском переводе.*
- *Книга старая, подержанная, в кожаном переплете.*

2.3.2 Сложный сегмент

Сложный сегмент, в отличие от простого, имеет структуру, которая содержит в себе вложенные сегменты. Для всех вариаций сложного сегмента подразумевается наличие следующей аксиомы: $I \neq \emptyset$.

$S_{Complex} \subset S, F_{Complex}: FR_T \times S_{Complex} \rightarrow S_{ex}$.

Для сложного сегмента определены следующие аксиомы:

1. $I \neq \emptyset$;
2. $Type = Type_{Complex}$.

Пример:

Расширим пример сегмента «Предложение». Пусть этот сегмент теперь должен содержать в себе слово «книга».

Определим новый простой сегмент, а также новые маркеры:

Маркер «Книга (м)» = <{«Книга», «книга»}>.

Сегмент «Книга» = <{«Книга (м)»}, {«Книга (м)»}, 1, 1, $\emptyset, \emptyset, 0, 0, Type_{Simple}$> - блок текста, состоящий только из слова «Книга» или «книга».

Теперь преобразуем простой сегмент «Предложение» в сложный, добавив в него вложенный сегмент «Книга»:

«Предложение» = <{«Начало текста», «Конец предложения»}, {«Конец текста», «Конец предложения»}, 0, 1, {«Книга»}, ∅, 0, 0, $Type_{Complex}$>.

Тогда в тексте «¶[H]*На комоде лежала какая-то книга. Он каждый раз, проходя взад и вперед, замечал ее; теперь же взял и посмотрел. Это был Новый завет в русском переводе. Книга старая, подержанная, в кожаном переплете.*¶[K]» можно найти следующие сегменты «Предложение»:

- *На комоде лежала какая-то книга*.
- *Книга старая, подержанная, в кожаном переплете*.

Заметим, что не важно, в начале или в конце располагается искомое слово, однако, если изменить параметр l_L на true, то сегмент «Предложение» обязан будет начинаться с этого слова, и подойдет только такой вариант:

- *Книга старая, подержанная, в кожаном переплете*.

2.3.3 Последовательные сегменты

Последовательные сегменты являются вариацией сложного сегмента. Их отличие от сложного сегмента заключается в том, что между ними не может быть разрывов, а также первый вложенный сегмент должен располагаться строго в начале искомого сегмента, а последний строго в конце. Таким образом, тип последовательные сегменты является своего рода контейнером, содержащим в себе некоторые сегменты и предотвращающем появление дополнительного текста вокруг них.

$S_{Sequence} \subset S, F_{Sequence}: FR_T \times S_{Sequence} \to S_{ex}$.

Для данного типа сегмента определены следующие аксиомы:

1. $I \neq \emptyset$;
2. $l_L = 1$;
3. $l_R = 1$;
4. $Type = Type_{Sequence}$.

В предыдущем примере сложный сегмент «Предложение» содержал в себе (кроме сегмента «Книга») дополнительный текст – «*старая, подержанная, в кожаном переплете*.». Сегмент типа последовательные сегменты не может содержать в себе лишний текст, поэтому в примере, предложенном выше, нельзя найти ни один сегмент последовательного типа:

«Предложение» = <{«Начало текста», «Конец предложения»}, {«Конец текста», «Конец предложения»}, 0, 1, {«Книга»}, ∅, 1, 1, $Type_{Sequence}$>. На самом деле, существует только два «Предложения», подходящие под это определение: предложение, состоящее из слова «Книга» (с заглавной буквы) и из слова «книга» (со строчной).

2.3.4 Повторяющийся сегмент

Повторяющийся сегмент также является вариацией сложного сегмента. Он описывает сегмент, который может быть встречен в тексте подряд один или более раз. Таким образом, этот тип сегмента является «надстройкой» над другим сегментом, означающий возможное повторение оного.

$S_{Repeatable} \subset S,\ F_{Repeatable}: FR_T \times S_{Repeatable} \to S_{ex}.$

Для данного типа сегмента определены следующие аксиомы:

1. $I = \{s\}, s \in S;$
2. $P^I = \emptyset;$
3. $Type = Type_{Repeatable}.$

В вышеописанном примере делового письма упоминался блок «Адресат». Также утверждалось, что адресатов может быть несколько. Такой случай можно описать сегментом повторяющегося типа:

«Адресаты» = <Ø, Ø, 0, 0, {«Адресат»}, Ø, 0, 0, $Type_{Repeatable}$>.

Пустые множества начальных и конечных маркеров означают, что данный сегмент не имеет определенных границ и может иметь произвольное положение в тексте. Его границы в таком случае будут определены границами внутренних вложенных сегментов.

2.3.5 Альтернативные сегменты

Тип сегмента альтернативные сегменты может обозначить те сегменты, которые являются взаимозаменяемыми, или те, для которых не важно, какой из них должен быть найден в данном контексте.

$S_{Alternative} \subset S,\ F_{Alternative}: FR_T \times S_{Alternative} \to S_{ex}$.

Для данного типа сегмента определены следующие аксиомы:

1. $I \neq \emptyset$;
2. $P^I = \emptyset$;
3. $Type = T_{Alternative}$.

Пример:

Имеются следующие сегменты: «E-mail адрес», «Телефон».

Можно описать сегмент «Контактная информация», содержащий либо адрес электронной почты, либо телефон, следующим образом:

«Контактная информация» = <Ø, Ø, 0, 0, {«E-mail адрес», «Телефон»}, Ø, 0, 0, $Type_{Alternative}$>.

2.3.6 Комбинация сегментов

Для тех случаев, когда порядок взаиморасположения нескольких сегментов неизвестен, но известно, что они должны присутствовать в рассматриваемом жанре документа, предусмотрен тип комбинация сегментов.

$S_{Combination} \subset S, F_{Combination}: FR_T \times S_{Combination} \to S_{ex}$.

Для данного типа сегмента определены следующие аксиомы:

1. $I \neq \emptyset$;
2. $P^I = \emptyset$;
3. $Type = Type_{Combination}$.

Пример:

Многие документы содержат раздел общей информации, содержащий в себе такие данные как ФИО, Возраст, Пол, Город. Предположим, структуры этих сегментов уже определены. Однако, неизвестно в каком порядке эти сегменты будут располагаться в документе. Поэтому, их можно объединить как комбинацию сегментов.

«Общая информация» = <Ø, Ø, 0, 0, {«ФИО», «Возраст», «Пол», «Город»}, Ø, 0, 0, $Type_{Combination}$>.

2.3.7 Факультативный сегмент

Факультативный сегмент, также, как и повторяющийся сегмент, является надстройкой над сегментом. Он позволяет пометить тот сегмент, присутствие которого в тексте необязательно.

$S_{Optional} \subset S, F_{Optional}: FR_T \times S_{Optional} \to S_{ex}$.

Для данного типа сегмента определены следующие аксиомы:

1. $I = \{s\}, s \in S$;
2. $P^I = \emptyset$;
3. $Type = Type_{Optional}$.

Предположим, что в предыдущем примере сегмент «Город» может быть факультативным.

Определим факультативный сегмент «Город*» = <Ø, Ø, 0, 0, {«Город»}, Ø, 0, 0, $Type_{Optional}$>.

Тогда новый сегмент «Общая информация» будет выглядеть так:

«Общая информация» = <Ø, Ø, 0, 0, {«ФИО», «Возраст», «Пол», «Город*»}, Ø, 0, 0, $Type_{Combination}$>.

2.3.8 Отрицание сегмента

Отрицание сегмента также является надстройкой над конкретным сегментом. Добавление этой надстройки к сегменту означает, что этого сегмента в данном блоке текста быть не должно.

$S_{Negative} \subset S, F_{Negative}: FR_T \times S_{Negative} \to S_{ex}$.

Для данного типа сегмента определены следующие аксиомы:

1. $M_B = \emptyset$;
2. $M_E = \emptyset$;
3. $i_B = 1$;
4. $i_E = 1$;
5. $I = \{s\}, s \in S$;
6. $P^I = \emptyset$;
7. $l_L = 0$;
8. $l_R = 0$;
9. $Type = Type_{Negative}$.

Предположим, имеется база текстов, полученных из разных источников, включая Интернет. Если требуется проанализировать все тексты, которые не являются web-страницами, можно добавить отрицание сегмента «HTML»:

«HTML» = <{«<html>»}, Ø, 1, 1, $\emptyset, \emptyset, 0,0, Type_{Simple}$> - простой сегмент.

«-HTML» = <Ø, Ø, 1, 1, «HTML», $\emptyset, 0,0, Type_{Negative}$> - его отрицание.

Добавив отрицание к любой структуре сегментов, например, сформировав сложный сегмент, содержащий «-HTML» на первом месте в множестве вложенных сегментов, можно распознавать тексты, не являющиеся web-страницами.

2.4 Пример «Резюме»

Общая структура резюме (рис. 2) состоит из трех разделов: раздела общей информации, который может быть представлен сегментом сложного типа, раздела основной информации, представляемого комбинацией сегментов, и раздела дополнительной информации, представляемого простым сегментом с факультативной надстройкой (штриховая пунктирная линия).

Раздел общей информации является сложным сегментом и содержит подразделы главной и общей информации. Стоит отметить, что резюме не может начинаться с обращения, поэтому перед подразделом главной информации находится отрицание простого сегмента «Обращение» (точечная пунктирная линия).

В начале подраздела главной информации находится ФИО составителя резюме, и сразу за ним должна быть указана целевая должность. Таким образом, удобнее всего определить сегмент подраздела главной информации как сегмент типа последовательные сегменты. Подразделы ФИО и Должность являются простыми сегментами.

Подраздел общей информации представим комбинацией факультативных сегментов над сегментами Возраст, Адрес и E-mail. Так как в резюме могут быть указаны несколько записей Адреса и E-mail, лучше всего представить их как повторяющиеся сегменты. Сегмент Возраст имеет следующую структуру: он является сегментом альтернативного типа, показывая, что на его месте может быть обнаружен как сегмент Возраст составителя, так и сегмент Дата рождения.

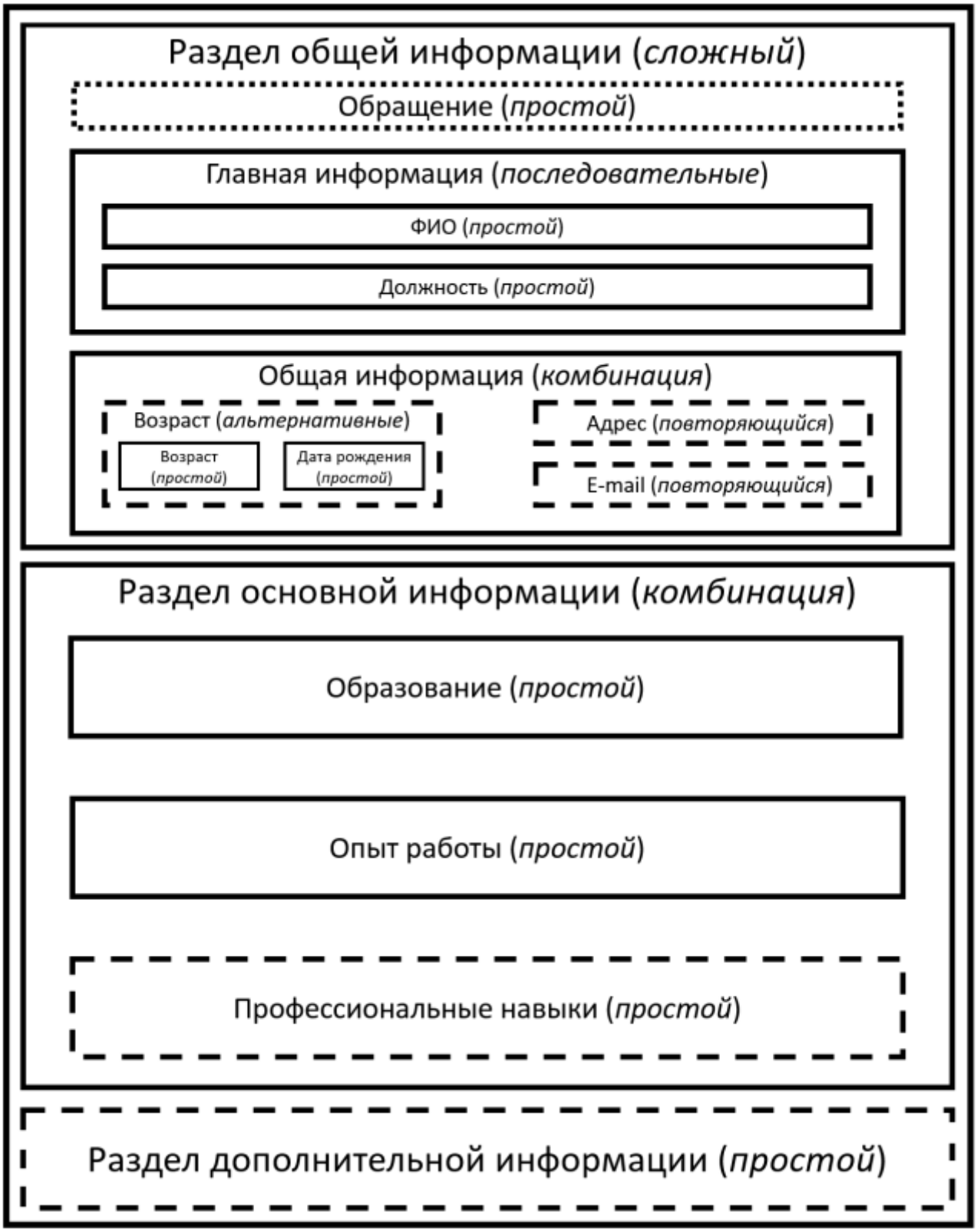

Рис. 2. Структура резюме

Раздел основной информации является комбинацией сегментов и включает в себя такие обязательные подразделы как Образование и Опыт работы (простые сегменты) и раздел Профессиональные навыки (факультативный сегмент).

Раздел дополнительной информации является факультативной надстройкой над простым сегментом.

3. Алгоритм построения жанровой модели текста

Построение жанровых моделей документов позволяет проводить углубленный анализ, синтез текстов. Ниже приведен пример алгоритма сегментирования текста, также позволяющий определять принадлежность документа определенному жанру.

Изначально имеется иерархическая структура сегментов (заданная в XML формате) и текст, для которого будет выполняться поиск. Структура сегментов представляет собой жанровую модель, содержащую в себе список внутренних сегментов. Соответственно, каждый сегмент также содержит в себе вложенные сегменты (если таковые имеются) и множества начальных и конечных маркеров.

Изначально полагается, что модель является корневым сегментом. Тогда для корневого сегмента:

1. Берется фрагмент текста (в начале фрагмент полагается равным тексту).
2. Создается два вектора экземпляров начальных и конечных маркеров (экземпляр маркера содержит индексы начала и конца данного маркера в тексте, вектор экземпляров содержит все возможные позиции в тексте для всех заданных маркеров). Этот шаг пропускается для корневого сегмента, так как модель содержит только список внутренних сегментов и не имеет границ.
3. Если сегмент содержит внутренние сегменты:

    3.1. Относительно векторов экземпляров маркеров находятся максимальные начало и конец сегмента, определяя внутреннюю область, в которой будет производиться поиск внутренних сегментов.

    3.2. Для каждого сегмента:

    3.2.1. Алгоритм рекурсивно запускается с первой операции. Фрагмент текста выбирается через ранее определенную внутреннюю область. Если для данного внутреннего сегмента уже был определен предыдущий соседний внутренний

сегмент, то начало анализируемого текста смещается до конца найденного сегмента.

3.2.2. У сегмента берется его экземпляр. Он определяет область, в которой был найден этот сегмент. Эта область соединяется с областью, полученной ранее для предыдущих сегментов.

3.3. Вектор экземпляров маркеров фильтруется таким образом, чтобы его экземпляры не попадали в область найденных сегментов.

4. Относительно векторов экземпляров маркеров определяются наилучшие минимальные границы сегмента.

5. Создается экземпляр сегмента – структура, содержащая позиции начала и конца данного сегмента. Если создать экземпляр не удалось, то алгоритм завершает работу.

После завершения работы основного алгоритма для всех сегментов формируется XML структура, содержащая облегченную иерархию сегментов (без векторов маркеров и без сегментов, которые не имеют конкретной позиции в тексте, например, отрицание сегмента), в которой определены позиции начал и концов этих 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 работе приведено математическое описание модели жанра документа, ориентированное на автоматическую обработку текста. Данная модель может применяться как для задач анализа жанровой структуры текста и жанровой классификации, так и для синтеза, где жанровая модель позволяет выстроить повествовательную структуру текста и оформить ее в соответствии с заданным форматом. Жанр в предложенном подходе сопоставляется формальной структуре текста, снабженной жанровыми маркерами. Приведенный алгоритм демонстрирует применимость рассмотренной модели для задач жанровой сегментации текста, что является важным компонентом при решении широкого класса задач, связанных с анализом текста. Апробация данного алгоритма на корпусе текстов жанра резюме показало его работоспособность, и, следовательно, корректность предложенного формализма описания модели. В дальнейшем планируется провести исследование применимости модели для более широкого набора жанров, чтобы выявить ее различительную способность для решения классификационных задач.

Список литературы

1. Бахтин М.М. Проблема речевых жанров // Эстетика словесного творчества. – М.: Искусство, 1986. – С. 250–296.
2. Блюменау Д.И., Гендина Н.И., Добронравов И.С., Лахути Д.Г., Леонов В.П., Федоров Е.Б. Формализованное реферирование с использованием словесных клише (маркеров) // Научно-техническая информация. Сер. 2. – 1981. – № 2. – С. 16–20.
3. Кибрик, А. А. Модус, жанр и другие параметры классификации дискурсов // Вопросы языкознания. – 2009. – №2. – С. 3–21.
4. Кононенко И.С., Сидорова Е.А. Жанровые аспекты классификации веб-сайтов // Программная инженерия № 8. 2015. С. 32–40.
5. Кононенко И.С., Сидорова Е.А. Обработка делового письма в системе документооборота // Труды международного семинара Диалог’2002 по компьютерной лингвистике и ее приложениям. Т.2. – Москва: Наука, 2002. –С. 299–310.
6. Щипицина Л.Ю. Жанры компьютерно-опосредованной коммуникации. – Архангельск: Поморский университет, 2009. – 238 с.