Методика работы с коллекциями текстовой информации через анализ информационных портретов

© А.В. Антонов[1], Е.В. Ягунова[2]

[1]Корпорация «Галактика», г. Москва
[2]Санкт-Петербургский государственный университет
alexa@galaktika.ru, iagounova_elena@mail.ru

Аннотация

Рассматривается методика обработки текстовой информации с помощью Инфопортретов (ИП), автоматически определяемого набора наиболее значимых слов для выборки по запросу пользователя. Показывается, что ИП может выступать в качестве свертки, т. е. носителя наиболее важной информации о текстах выдачи. Методика включает отработку эксперимента с информантами для анализа особенностей структуры ИП и возможных путях извлечения из ИП этой информации.

1 Введение

Изменившиеся условия существования человека коренным образом перестроили процедуру анализа информации. Развитие технологий информационного и фактографического поиска открывает новое поле деятельности для специалистов в области компьютерной лингвистики текста. Раньше основным и единственным объектом лингвистического исследования был ***текст*** (его анализ, понимание). Но для того чтобы полноценно жить в информационном обществе, человек должен обрабатывать огромное количество информации. Лавина информации, содержащаяся в информационных потоках, не может быть воспринята и проанализирована человеком в силу его психофизиологических ограничений. Новый информационный объект – ***информационный поток*** – требует использования новых технологий, которые выступают в качестве посредника при извлечении адресатом коммуницируемого смысла. В данной работе ***информационный поток*** нами понимается как ***множество текстов***, ***выступающих как единый объект***: адресатов интересует смысл, заключенный сразу в сотнях и даже тысячах текстов.

При всем различии рассматриваемых информационных объектов – текст и информационный поток – нас интересует то, что они обладают информационной (смысловой) структурой и могут быть свернуты до набора слов и словосочетаний. Этот набор может выступать представителем (носителем) информационной структуры объекта (и текста, и информационного потока).

Свертки понимаются как результат компрессии текста в соответствии с некоторыми заданными правилами. Как известно, один из наиболее известных способов получения свертки текста – установление набора его ключевых слов (КС). Ключевыми словами (или аналогами ключевых слов) в разных контекстах называют, например,

- выписанные группой информантов слова, наиболее важные для решения поставленных в инструкции задач (обычно – понимания текста): уровень значимости слова определяется как относительная частота его встречаемости в протоколах информантов;
- автоматически выделяемые неслучайно встречающиеся в документах слова и словосочетания, важные для рассматриваемой выборки (выдачи) в рамках общего массива документов: уровень значимости слова рассчитывается на основании некоего алгоритма.

На конференциях Диалог-2008 и Диалог-2009 анализировались разные виды сверток (наборов слов и словосочетаний), полученных на материале единичных текстов с помощью разных видов экспериментов с информантами (см. подробнее, включая обзор литературы, в [3], [4]). Чтобы осуществить свертывание текста, этот текст, как правило, нужно понять. Поэтому естественно считать, что свертки представляют собой результат понимания текста или, иначе говоря, извлечения смысла из текста. С помощью дополнительного эксперимента изучалась возможность восстановления исходного смысла или информационной структуры текста.

Ресурс Галактика-Зум (http://galaktika-zoom.ru/, см. также http://webground.su) предоставляет возможности для проведения исследования на материале сверток (наборов) автоматически определяемых ключевых слов. Для каждой выдачи (в соответствии с запросом) этот ресурс вычисляет и предоставляет пользователю ***Информационный портрет*** (или Инфорпотрет), т. е. набор автоматически определяемых слов и словосочетаний, важных для рассматриваемой выборки (среза) в рамках общего массива документов. Инфопортрет как сверка множества текстов является основной возможностью для извлечения адресатом ***целостной информационной структуры***: большой объем не позволяет человеку оперировать непосредственно с каждым текстом.

Об основной идее Инфопортрета, критериях определения значимых слов, образующих Инфопортрет, подробно изложено в нескольких работах, в частности, [1].

Труды 12й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

2 Постановка целей и задач. Материал и методика эксперимента

2.1 Основные положения

Нами развиваются и экспериментально обосновываются следующие положения (ср. также [1]):

- Инфопортреты с достаточной точностью и полнотой включают значимые для данного запроса слова, т. е. представляют собой свертки, достаточным для понимания образом отражающие информационную структуру анализируемого набора текстов, что верифицируется с помощью эксперимента, в котором на основании этих сверток информанты извлекают требуемую (в инструкции к эксперименту) информацию;
- возможности ресурса Галактики-Зум в целом позволяют использовать его в исследовательских проектах при решении задач понимания и смысловой компрессии текстов [1]; к тому же этот ресурс оказался единственным из доступных авторам, который позволил провести исследование в рамках поставленной задачи [2].

Цель данной работы – апробировать методику экспериментального анализа особенностей Инфопортретов как сверток текстов. В результате такого анализа предполагается получение методики работы с информационными потоками – через систему, порождающую Инфопортреты как свертки, – аналогичной той методике, что используется для текстов.

2.2 Цели и задачи, материал и методика

2.2.1 Задача эксперимента

Основной задачей данного эксперимента было определить, является ли Инфопортрет реальной сверткой текста, т. е. сможет ли информант восстановить по нему информацию об объекте, описанном в данном тексте, в частности, информацию процедурно-временного характера. Для этого перед информантами ставится задача определения временного периода, к которому относится группа текстов. При этом из свертки должны быть удалены все непосредственные указания на временной период (месяц, квартал, конкретные даты).

2.2.2 Материал

Нами анализировались новостные тексты: их имеется достаточное количество по выбранной нами тематике, они, в основном, компактны и ограничены лексически.

В качестве запросов были выбраны запросы «ЕГЭ» и «единый государственный экзамен», т. е. выбирались тексты, содержащие данные слово или словосочетание. Основания для выбора именно таких запросов были следующие:

- «ЕГЭ» («единый государственный экзамен») может быть по праву названо одним из «ключевых слов» 2009 года; актуальность и востребованность этой темы позволили получить в выдаче большое количество текстов (см. табл. 1), причем выборочный анализ текстов выдачи показывает, что тематически они достаточно однородны;
- тема «ЕГЭ» (или «единый государственный экзамен») была выбрана из-за того, что в самой природе рассматриваемого объекта (и текстов, его описывающих) заключены периодизация и хорошо знакомый лингвистам принцип построения сюжета, причем эти периоды несут особую информационную нагруженность (подготовка – проведение – подведение итогов), что позволяет в процедуре проведения эксперимента с информантами через определения интервала эксплицировать основную информацию, содержащуюся в предъявляемых Инфопортретах.

2.2.3 Методика эксперимента

На вход системе были посланы запросы: (1) «ЕГЭ» и (2) «единый государственный экзамен». Результаты этих запросов система распределила по 9 выборкам, каждая из которых содержит документы, относящиеся к одному из прошедших месяцев 2009 года (от января по сентябрь включительно). В табл. 1 приведены объемы этих выборок в числе документов.

Таблица 1. Объемы выборок по запросу «ЕГЭ» и «единый государственный экзамен»

	«ЕГЭ»	*«единый государственный экзамен»*
месяц	Число документов	Число документов
Январь	566	196
Февраль	831	310
Март	800	227
Апрель	**1036**	**317**
Май	**964**	**342**
Июнь	**1225**	**333**
Июль	817	181
Август	739	148
Сентябрь	790	178
итого	7768	2232

При проведении эксперимента с информантами основной задачей было определить, насколько полно и точно свертка (ИП) – последовательность значимых слов и словосочетаний, ранжированных по убыванию значимости (см. пример в табл. 2) – отражает информацию, содержащуюся в множестве текстов, в частности, информацию процедурно-временного характера, т. е. перед информантами стояла задача на основании свертки определить, к какому периоду относятся тексты данной выборки. Каждому информанту выдавалась инструкция:

«Каждый из 9 листов соответствует выборке

[1] Субъективный подход к компрессии (задача, адресат и предметная область) реализуется в виде запроса пользователя, а объективных подход – в виде правил, согласно которым осуществляется компрессия (определение слов и словосочетаний, задающих свертку текста) (ср. [2])

[2] Например, на сайте Nigma.ru, который тоже дает свертку, нельзя отсечь документы по дате, что необходимо по условиям данного эксперимента

одного месяца 2009года. Ваша задача – оценить и отметить на каждом листе свой выбор:

1. предположительный период: подготовка к экзамену – проведение экзамена – подведение итогов;

2. месяц: от января до сентября 2009 года;

3. критерии, особенности, комментарии и т. д.»

Таблица 2. Пример Инфопортрета (февраль)

государственный экзамен	все специальности
общеобразовательные предметы	Госэкзамен
ЕГЭ	успешная сдача
итоговая аттестация	репетиционный
вступительные испытания	Обществознанию
уважительные причины	обязательные экзамены
штатный режим	все выпускники
Аттестация	основные сроки
Аттестат	пересдать
досрочная сдача	те предметы
единый госэкзамен	экзаменационный
вторая волна	Двойка
дополнительные сроки	все экзамены
вступительные экзамены	первая волна
обязательные предметы	дополнительные занятия
экзаменационный лист	неудовлетворительный результат
единственная форма	

Помимо основной задачи эксперимента решались следующие методические задачи:

- выделить свертки, лучше всего отражающие рассматриваемую информацию, т. е. обеспечивающие «правильный» ответ информантов;
- ранжировать свертки по степени «правильности» ответов, для чего в анкете задано два параметра – период (главный параметр), месяц (уточняющий параметр) [3];
- определить причины «правильного» и «неправильного» выборов ответов.

Как уже говорилось, в нашем исследовании смысловой компрессии важно учитывать не только объективную, но и субъективную сторону, прежде всего, факторы адресата (в пользу которого осуществляется компрессия) и предметной области.

В качестве информантов (адресатов компрессии) выступили 16 (17) студентов и аспирантов СПбГУ гуманитарных специальностей в возрасте 20 – 24 лет. Они не являлись специалистами в предметной области (переход на систему ЕГЭ) ни в силу профессиональной деятельности, ни в силу жизненного опыта (т. к. сами сдавали традиционные экзамены). Процедуры принятия решения и используемые ими критерии не связаны со специальными знаниями и навыками (например, аналитической работой с информационными потоками). Смысловая структура текстов (выборок текстов) данной предметной области в большинстве случаев неоднородна и предполагает конкуренцию критериев, т. к. включает в себя в качестве подтем как минимум три: окончание школы – сдача ЕГЭ – поступление в вуз.

Второй эксперимент проводился через 2,5 месяца после первого с той же бригадой информантов (добавился один новый). Методика проведения эксперимента должна была минимизировать влияние индивидуальных ассоциативных связей. Собранная бригада участвовала в двух экспериментах. В промежутке результаты эксперимента с информантами не обсуждались. Сопоставительный анализ протоколов второго эксперимента исключает возможность влияния на его результаты первого эксперимента; таким образом, мы считаем, что экспериментальный дизайн удовлетворяет требованиям чистоты эксперимента.

3 Результаты

Первичной задачей анализа данных двух экспериментов было определить корректность Инфопортретов как достаточной свертки текста по правильности определения периода времени для заданных выборок. Кроме этого, было необходимо:

- выделить свертки, лучше всего отражающие рассматриваемую информацию, т. е. обеспечивающие «правильный» ответ информантов;
- ранжировать свертки по степени «правильности» ответов, для чего в анкете задано два параметра – период, месяц);
- определить критерии, помогающие и мешающие «правильному» принятию решения.

3.1 Определение периода

На основании результатов определения испытуемыми периода в эксперименте 1 можно выделить четыре класса (по убыванию числа правильных и согласованных ответов информантов, см. табл. 3)):

1 февраль, март, сентябрь (подготовка экзамена и подведение итогов);

2 январь (подготовка экзамена);

3 апрель, август (подготовка экзамена и подведение итогов);

4 май, июнь, июль (неопределенность проведение экзамена/подведение итогов).

Свертки, предъявленные испытуемым в ходе эксперимента 2, дали другое распределение правильных и согласованных ответов испытуемых. Если аналогичным образом сгруппировать свертки в классы по данным эксперимента 2, выделяется пять классов (см. табл. 3):

1 июль, февраль (подведение итогов и подготовка соответственно);

[3] Напомним, что параметры, заданные в анкетах испытуемых, нужны для определения правильности восстановления смысла (информационной структуры) выборки текстов

2 апрель (проведение экзаменов вместо подготовки);

3 январь, май, сентябрь (подготовка, проведение и подведение итогов соответственно);

4 июнь (подготовка экзамена вместо проведения, но сравнительно высокая согласованность);

5 август, март (подведение итогов и подготовка соответственно).

Месяц февраль сохранил лидерство среди сверток эксперимента 2 (класс 1), хотя, если сравнивать с данными эксперимента 1, количество правильных ответов несколько упало (с 81 % до 65 % случаев). Свертка март занимает полярные места в распределении сверток по правильности периода: лидирующее положение в эксперименте 1 и одно из самых низких в эксперименте 2.

Определение параметра «месяц» (частично рассмотрено далее) носит уточняющий характер.

Вполне объективно месяцы апрель, май, июнь и июль характеризуются максимальным накалом страстей и конкуренцией рассматриваемых трех подтем (окончание школы – сдача ЕГЭ – поступление в вуз). Возможно, именно в реализации функции воздействия на адресата наряду с традиционной информационной функцией (т. е. в «публицистичности») заключается причина того, что в эксперименте 2 «период» определяется для этих месяцев гораздо согласованнее (по сравнению с экспериментом 1). Кульминация этого эффекта падает на месяц июль: наихудшее распознавание периода в эксперименте 1 и наилучшее в эксперименте 2 (44 % vs. 71 %).

Таблица 3. Результаты определения испытуемыми периода имя свертки (месяц, для которого осуществлялась выборка)

Эксперимент 1: запрос «ЕГЭ»

Период (согласно анкетам ии.)	январь	февраль	март	апрель	*май*	*июнь*	*июль*	август	**сентябрь**
Подготовка экзамена	**0,69**	**0,81**	**0,81**	***0,56***	*0,19*	*0,06*	*0,13*	0,31	0
Проведение экзамена	0,13	0,13	0,25	0,38	*0,50*	*0,50*	*0,44*	0,13	0,13
Подведение итогов	0,19	0,06	0	0,06	*0,31*	*0,44*	*0,44*	***0,56***	**0,88**

Эксперимент 2: запрос «единый государственный экзамен»

Период (согласно анкетам ии.)	январь	февраль	*март*	апрель	май	июнь	июль	*август*	сентябрь
Подготовка экзамена	**0,59**	**0,65**	*0,41*	0,29	0,35	***0,59***	0,00	*0,35*	0,35
Проведение экзамена	0,24	0,24	*0,12*	**0,65**	**0,59**	0,18	0,29	*0,12*	0,06
Подведение итогов	0,18	0,12	*0,47*	0,06	0,06	0,24	**0,71**	*0,47*	**0,59**

Эксперимент 1

Свертка «апрель»: несогласованность определения периода для свертки «апрель» в (56 % «подготовка к экзамену» и 38 % «проведение экзамена») связана с невозможностью правильного определения месяца (25 % «февраль» и 38 % «май»).

Свертка «май»: наблюдается правильный выбор и периода, и месяца, но низкая согласованность (50 % для периода и 25 % для месяца).

Свертка «июнь»: наилучший результат восстановления месяца – 50 % информантов; более того, это редкий случай, когда задачи определения периода и месяца оказались равносложными: 50 % – это и число испытуемых, правильно определивших период.

Свертка «июль»: колебания в определении периода для свертка (по 44 % между «проведением экзамена» и «подведением итогов»), по-видимому, связаны с особенностями выбора месяца (25 % «июль» и 31 % «август»).

Свертка «август»: колебания в выборе периода между «подготовкой к экзамену» и «подведением итогов» (31 % и 56 %) и невозможность определения месяца (максимальное для этой сверти число испытуемых – всего лишь 25 % – отнесла ее к «июлю»).

Эксперимент 2

Свертка «апрель»: сравнительно высокая согласованность определения периода – 65 % «проведение экзамена» (а не подготовка к нему, как было в эксперименте 1) – соотносится с гораздо лучшими результатами восстановления месяца (35 % «апрель» и 29 % «июнь»).

Свертка «май»: наблюдается правильный выбор периода, но низкая согласованность и наилучшее определение месяца (59 % для периода и 76 % для месяца).

Свертка «июнь»: невозможность определения месяца, определение периода как «подготовка к экзамену» в 59 % случаев (т. е. резкое отличие от «июня» в эксперименте 1).

Свертка «июль»: максимальная согласованность (в отличие от эксперимента 1) в определении периода «подведение итогов» и колебания в определении месяца (24 % «июль» и 47 % «август»).

Свертка «август»: колебания в выборе периода между «подготовкой к экзамену» и «подведением итогов» (35 % и 47 %) и невозможность определения месяца (максимальное для этой сверти число испытуемых – всего лишь 29 % – отнесло ее к «сентябрю»).

Анализ тем текстов разных выдач показывает,

что однозначное определение периода и месяца не обязательно должны соответствовать друг другу. Сроки проведения ЕГЭ колеблются от апреля до июля (согласно приказу «Об утверждении сроков и единого расписания проведения …» (http://www1.ege.edu.ru/content/view/475/36/): досрочное проведение – апрель, для основной массы выпускников 2009 года – июнь (а также 26 и 29 мая), для выпускников прошлых лет – июль.

Выборочный анализ текстов выдач по рассматриваем запросам (месяцы апрель – июль) показывает, что выдача на запрос «единый государственный экзамен» в большей степени ориентированы на «проблемные» случаи, а на запрос «ЕГЭ» – на типичные. Для «апреля» (эксперимент 2) проблемным является досрочное проведение ЕГЭ (ср. высокую согласованность и сравнительно неплохим восстановлением месяца). Для июня и июля – сдача ЕГЭ выпускниками прошлых лет (неравное положение выпускников 2009 года и прошлых лет, т. е. более сложные условия для последних). Поэтому «июнь» дает большее внимание к подготовке, а «июль» – к подведению итогов.

4 Заключение

Результаты эксперимента подтвердили обоснованность работы с Инфопортретом (ИП) выборки – сверткой множества текстов – как единым объектом. На данном этапе исследования инфопортретов была отработана методика проведения экспериментов и сформирован круг вопросов, которые возможно изучать на рассматриваемом материале.

1. ИП можно рассматривать как свертку текста, т. е. даже наивный адресат (а не подготовленный работник аналитического или информационного отдела) может восстановить по свертке (сверткам) информацию об исходном объекте. Под исходным объектом понимается множество текстов выдачи, т. е. срез информационного потока, полученный в соответствии с заданным запросом в рамках работы система Галактика-Зум. Таким образом, мы получили методику работы с информационными потоками – через систему, порождающую Инфопортреты как свертки.

2. Новостные тексты различаются по коммуникативной цели, типу и структуре текста. С помощью разработанной методики можно анализировать специфику преимущественно информационных и преимущественно публицистических текстов (названия условные и уточняются по ходу работы).

3. Разные лексические варианты анализируемого термина («ЕГЭ, единый государственный экзамен, единый госэкзамен, единый экзамен») в запросе порождают выдачи, различающиеся по составу информационных и публицистических текстов. Запрос с самым кратким вариантом – «ЕГЭ» – обеспечивает выдачу преимущественно информационных текстов, а запросу «единый государственный экзамен» (максимально развернутое атрибутивное сочетание) соответствует выдача с большим количеством публицистических текстов.

Применяя нарративную метафору, можно рассмотреть девять сверток (для каждого из периодов, которому соответствовала одна выборка) как компоненты единой смысловой структуры высокого уровня, характеризующейся динамичной сменой ситуаций (при том, что каждая из этих ситуаций сама имеет сложную смысловую структуру). Тогда свертку «январь» можно описать как ***преамбулу*** (фазу ориентации), «февраль» – как основу ***завязывания сюжета***, «сентябрь» – как ***коду*** (мораль всей истории). Именно эти компоненты нарратива ведут себя сходным образом и для запроса «ЕГЭ», и для запроса «единый государственный экзамен». Наиболее сюжетными и неоднозначными оказались свертки «апрель – июль», на которых происходит развитие сюжета. Анализ результатов экспериментов демонстрирует разные сюжетные линии. Степень «публицистичности» (воздействия на адресата, например, убеждения) задает разные направления: типичное положение дел (для «информационных текстов») или проблемные случаи (для «публицистичных текстов»).

Степень «публицистичности» соотносится с преимущественной стратегией работы информантов: на основании анализа свертки как целостного объекта или выделения наиболее важных слов в ИП. Информационно нагруженные тексты характеризуются:

- более определенными наборами (более высокими значениями коэффициента значимости);
- в таких наборах информантам легче выделить слова, важные для последующего принятия решения, критерии выбора близки у разных информантов.

Публицистически окрашенные тексты, напротив, отличаются:

- менее определенными наборами (более низкими значениями коэффициента значимости);
- в этих случаях информанты указывают на важность разнообразных слов, критерии определения этих слов во многом определяются личными предпочтениями.

Литература

[1] Антонов А.В., Ягунова Е.В. Лингвистический анализ информационного портрета как свертки множества текстов. Постановка эксперимента // Новые информационные технологии в автоматизированных системах: материалы тринадцатого науч.-практ. семинара. – М., 2010. – С. 50-59.

[2] Леонтьева Н.Н. О методах смысловой компрессии текста // Интернет и современное общество: труды X Всерос. объединенной конф., СПбГУ. СПб., 2007.

[3] Ягунова Е.В. Вариативность стратегий восприятия звучащего текста (экспериментальное исследование на материале русскоязычных текстов разных функциональных стилей). – Пермь, 2008.

[4] Ягунова Е.В. Набор опорных слов как вид свёртки текста (в сопоставлении с набором

ключевых слов) // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Межд. конф. «Диалог» (Бекасово, 4 – 8 июня 2008 г.). – М.: РГГУ, 2008. – Вып. 7 (14).