А. О. Третьяк

ОСНОВНОЕ СОДЕРЖАНИЕ ТЕКСТА И ЕГО АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ

В настоящее время при обработке текста рядом автоматических систем возникает задача построения его информационного портрета, который в компактной форме характеризует основное содержание текста, то есть описанные в нем события и их участников. Наличие информационного портрета позволяет производить сравнение документов по содержанию (осуществлять автоматическую классификацию, категоризацию и рубрицирование документов); находить документы, близкие по содержанию к заданному; компактно представлять содержание документа в виде списка ключевых тем, затронутых в тексте, или же в форме реферата, то есть в виде набора репрезентативных предложений текста, содержащих упоминания о ключевых темах. Рассмотрим подробнее суть понятия «основное содержание текста» и подходы к его автоматическому выявлению.

Ученые, исследующие проблемы семантики текста, характеризуют его содержание как информацию, заключенную в тексте в целом. Однако, по мнению Л. Латышева, содержание текста находится вне самого текста, а именно в голове отправителя (того, кто создает и передает текст) и в голове адресата (того, кому текст предназначен, кто его воспринимает). Оно складывается из значений, непосредственно представленных в тексте языковыми знаками. Однако содержание не является простой суммой механического сложения значений. Их взаимодействие в тексте носит гораздо более сложный характер. В частности, взаимодействуя друг с другом, широкие, а иногда и расплывчатые значения языковых единиц конкретизируются. Такая конкретизация значения в тексте называется актуализацией значения [1, с. 101].

Известно, что любому тексту свойственно внутреннее и внешнее содержание. По определению А. И. Новикова, внутреннее содержание текста – это «мыслительное образование, которое формируется в интеллекте человека и соотносится с внешней формой не поэлементно, а в целом соответствует всей совокупности данных языковых средств» [2, с. 5]. Внутреннее содержание или семантика текста существует в виде той информации, которая возбуждается под воздействием совокупности языковых средств, поэтому, говоря о семантике текста, имеют в виду «содержание текста в отношении к средствам его выражения» [2, с. 33]. Такое понимание семантики текста является характерным для информационных подходов к ее изучению.

В зависимости от отношения языкового знака к его другим частям различают следующие типы содержания текста [1, с. 102–106].

1. Сигнификативное содержание, образуемое отношением знак – сигнификат. Оно включает в себя как объективные свойства денотата (денотат – предмет или явление, обозначаемое языком в конкретном речевом произведении), так и коннотации (коннотация – любой содержательный компонент, который, не относясь к денотативному содержанию, сопровождает и дополняет его). К сигнификативным коннотациям относятся не все дополнительные

компоненты содержания, а лишь те, которые являются фактом коллективного языкового сознания и как таковые зафиксированы в парадигматическом значении языкового знака (то есть до того, как он употреблен в речи).

2. Содержание на уровне интерпретатора является результатом отношения знак – интерпретатор. В процессе общения люди употребляют языковые знаки не только в соответствии с их семантическим значением, но и в их несобственных значениях, особым образом интерпретируя их. Автор может построить текст с расчетом на то, что реципиент сам дополнит его подразумеваемой информацией. В этих случаях содержание текста выходит за рамки суммы семантических значений образующих его знаков.

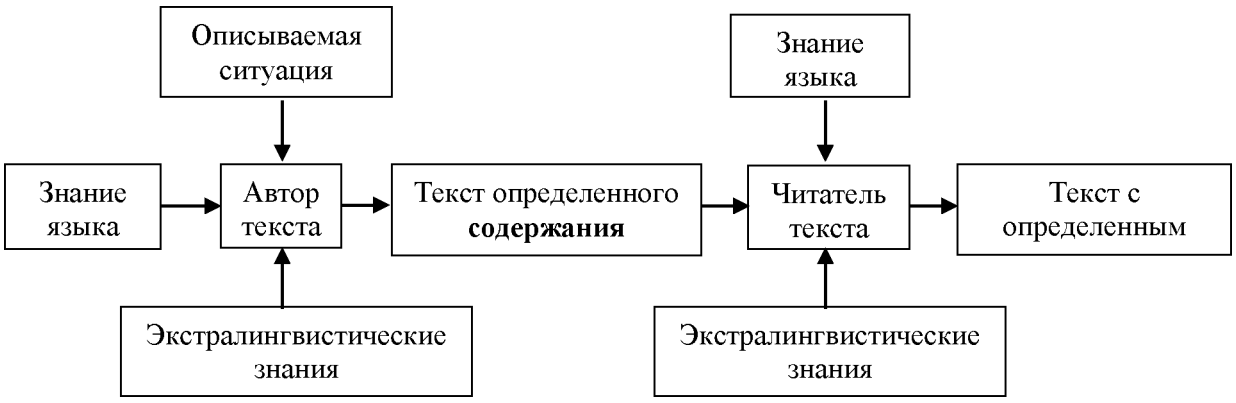

Рис. 1. Внутриязыковое содержание текста

3. Внутриязыковое содержание возникает в результате отношения знак – знак. Языковые знаки находятся в разнообразных отношениях сходства и различия, родства и взаимозависимости. Обычно эти отношения не играют самостоятельной роли в тексте и остаются незаметными для участников коммуникации. Но текст можно построить таким образом, что это отношение производности перестает быть незаметным и приобретает смысловую роль. Достигается это путем противопоставления языковых знаков на основе их сходства или различия. Все вышесказанное можно представить схематически следующим образом.

Говоря о смысле текста, обычно имеют в виду ту часть содержания высказывания, которая представляется наиболее важной, главной. Смысл имеет двойное значение: с одной стороны, это то, что автор хочет вложить в свое высказывание, с другой – то, что читатель / слушатель извлекает из него. Смысл получает выражение в тексте при помощи отбора конкретных слов и их распределения по тексту. Так как слово может быть полисемантично, то определение конкретного значения – содержания, непосредственно выраженного совокупностью знаковых единиц – происходит в тексте, когда слова взаимно обуславливают друг друга, в результате чего значение слова меняется в зависимости от его окружения. В тексте раскрывается система предметных отношений – денотатов, закодированных в словесных значениях [3, с. 104–105].

Смысл текста представляет собой воспринятую получателем структуру денотатов текста, на основе которой строится соответствующая модель, не в полной мере соответствующая модели автора текста. Реципиент восприни-

мает текст последовательно. Отдельные фрагменты текста формируют в его сознании совокупность частных моделей, образующих в конечном итоге общую, интегральную, окончательную модель, соответствующую смыслу текста для данного реципиента. Однако знания, получаемые реципиентом при восприятии текста, могут отличаться от знаний, заложенных в текст автором. Другими словами, реципиент может понимать не все и не так, как задумывал автор текста. Понимание текста обусловлено целым рядом экстралингвистических факторов (социокультурным и ситуативным контекстом, эмоциональным состоянием реципиента, его фоновыми знаниями). В смысловой цельности текста отражаются те связи и зависимости, которые имеются в самой действительности (общественные события, явления природы, человек, его внешний облик и внутренний мир, предметы неживой природы и т. д.). Естественно, что явления действительности могут оцениваться людьми по-разному. Если смысл для разных реципиентов различен, то это означает, что понимание ими данного текста различно, то есть в результате восприятия этого текста ими построены разные модели. Если реципиент не может построить модель текста, то смысл для него отсутствует, он не понимает текст. Таким образом, каждый текст имеет только одно содержание, определяемое автором, и может иметь для разных реципиентов различные смыслы.

Как отмечалось выше, понятие «содержание» связано с категорией информативности речи и присуще только тексту. Оно сообщает читателю индивидуально-авторское понимание отношений между явлениями, их значимости во всех сферах жизни. При этом формируются два вида информации: содержательно-фактуальная информация и содержательно-концептуальная информация. Две разновидности информации называют, соответственно, темой и основной мыслью текста. Не только тема, но и основная мысль объединяют предложения текста и придают ему смысловую цельность.

Многие исследователи считают, что следует различать понятия «тема текста» и «содержание текста». При формировании содержания текста автор выделяет элементы описываемой ситуации, приписывая им определенные для данной ситуации значения. В итоге содержание текста определяется его денотативной структурой, формируемой в сознании человека; оно реализуется в тексте путем отбора конкретных слов и их распределением по тексту. Причем подобный отбор осуществляется таким образом, чтобы потенциальный получатель текста мог сформулировать в своем сознании то же представление об описываемом фрагменте ситуации, которое вложил в текст автор. Содержание также определяется выделением существенных связей в тексте – смысловых опорных пунктов. Они являются носителями обобщенного смысла отдельных частей текста и образуют внутриречевую схему, в которую переводится содержание текста в процессе его понимания. Необходимо отметить, что содержание всего текста складывается из содержания входящих в него абзацев.

Тема является одним из важных структурных компонентов текста. Существуют разные трактовки этого понятия. Под темой понимают: 1) содержательный компонент коммуникации; 2) объединяющее начало текста; 3) основной тезис текста, подлежащий дальнейшему развертыванию;

4) смысловое ядро текста, конденсированное и обобщенное содержание текста; 5) основной компонент смыслового начала текста; 6) содержательный компонент ситуации [4, с. 60]. Понятие «тема текста» может рассматриваться в широком и узком смыслах. Широкое понимание темы включает понятие сообщаемой в тексте информации, а узкое – название объекта, о котором идет речь. Но в обоих случаях общей объективной чертой темы является предмет изложения [4, с. 62–63]. Следовательно, тему текста можно представить как отражение в сознании человека наиболее существенных составляющих ситуации, описанной в тексте. Тема текста – это тот инвариант, который остается одним и тем же, как для автора текста, так и для его читателя. Именно тема определяет предметно-тематическую область словаря, где должен осуществляться поиск указанных слов содержания. Тема устанавливает совокупность определенных слов и словосочетаний, отражающих наиболее существенные составляющие описываемой в тексте ситуации и наиболее важные отношения между ними. Т. А. ван Дейк и В. Кинч рассматривают тематическую структуру текста как иерархическую структуру, поскольку тема всего текста может быть описана с помощью его более мелких подтем, которые в свою очередь членятся на еще более мелкие микротемы. Каждое предложение связного текста связано с той или иной подтемой основной темы текста [5].

При автоматическом выделении основного содержания текста центральной является проблема критериев, используемых для выбора наиболее информативных слов и предложений из первичного документа. Чаще всего для этих целей используются статистические, позиционные и лингвосемантические методы [6]. Рассмотрим более подробно статистические методы, позволяющие достаточно быстро и эффективно определять основное содержание текста. Они основаны на использовании статистических параметров для оценки информативности различных элементов текста и учитывают, прежде всего, показатель частоты встречаемости слов в тексте. В результате ранжирования лексики в том или ином документе определяют слова с высоким рангом и их сочетаемость в различных фразах [7, с. 144]. Именно статистические методы используются в наши дни в промышленных компьютерных системах для решения таких задач, как аннотирование и реферирование текстов, их тематическая классификация и кластеризация, смысловой поиск и т.д., которые можно рассматривать в комплексе как задачу тематического анализа. Статистическая информация об отдельных лексических единицах легко извлекается компьютером из текста, и есть все основания полагать, что она адекватно отражает его основное содержание.

В настоящее время существует целая группа статистических методов. Так, согласно одному из них процедура выделения ключевых слов текста сводится к выполнению следующих этапов [8, с. 125].

1. Компьютер составляет по каждому абзацу текста алфавитно-частотный словарь словоформ.

2. Далее все алфавитно-частотные словари абзацев объединяются в единый распределительный алфавитно-частотный словарь словоформ текста. В нем указывается общая частота употребления словоформы в тексте, число абзацев, в которых она встретилась, а также частота употребления в конкретных абзацах.

3. Затем производится сокращение распределительного алфавитно-частотного словаря словоформ текста до словаря потенциальных ключевых (опорных) слов. Эта процедура включает в себя следующие операции:

– из распределительного словаря удаляется вся служебная и общеупотребительная лексика;

– в оставшейся части словаря суммируются частоты всех грамматических форм одного и того же слова;

– в этом же словаре суммируются частоты синонимов (в том числе и контекстуальных);

– из оставшейся части распределительного словаря удаляются слова, которые встретились только в одном абзаце.

4. На последнем этапе словарь потенциальных опорных слов текста делится на две части. В первую часть входят главные опорные слова (ГОС), а во вторую – второстепенные опорные слова (ВОС). Данная процедура осуществляется с учетом коэффициента важности слова, который вычисляется по формуле $K_в$=F*m/N*n, где F – абсолютная частота употребления слова с учетом всех грамматических форм и синонимов; m – количество абзацев, в которых встретилось слово; N – общее число словоупотреблений в тексте; n – общее число абзацев в тексте. Предварительно, в зависимости от длины текста (в словах и абзацах) по специальной формуле определяется средняя (пороговая) величина коэффициента важности [9, с. 45]. Коэффициент важности каждого слова словаря сравнивается с этой величиной. Если коэффициент важности слова выше пороговой величины, то оно относится к числу главных опорных слов, если меньше пороговой величины – к числу второстепенных опорных слов текста. Слова, входящие в каждую из выделенных групп, не однородны по своему содержанию. Они обозначают основные компоненты описанной в тексте ситуации: адресата и адресанта (субъекты), совершаемые ими действия в определенное время и в определенном месте, а также объекты действительности, о которых идет речь в тексте.

Необходимо заметить, что для адекватного формализованного представления основного содержания текста необходимо оперировать не отдельными словами, а определенным набором ключевых слов, который целесообразно представить в виде таблицы основного статического содержания текста (ТОСС). Каждая ТОСС и представляет собой информационный портрет текста.

ЛИТЕРАТУРА

1. *Латышев, Л. К.* Технология перевода : учеб. пособ. для студентов вузов / Л. К. Латышев. – М. : Академия, 2005. – 320 с.

2. *Новиков, А. И.* Семантика текста и ее формализация / А. И. Новиков. – М. : Наука, 1983. – 97 с.

3. *Соколов, А. Н.* Семантика, логика и интуиции в мыслительной деятельности человека : психологические исследования / А. Н. Соколов, Л. Л. Гурова, Н. И. Жинкин. – М. : Педагогика, 1979. – 184 с.

4. *Вишнякова, С. А.* Теоретические основы обучения моделированию научного текста (Русский как иностранный, основной этап обучения) / С. А. Вишнякова. – СПб. : Европейский дом, 2001. – 258 с.

5. *Дейк, Т. А. ван.* Стратегии понимания связного текста / Т. А. ван Дейк, В. Кинч // Новое в зарубежной лингвистике. – М. : Прогресс, 1988. – Вып. 23. – С. 153–211.
6. *Горяник, Л. В.* Тематический фильтр текстов / Л. В. Горяник, Г. В. Дорохина // Искусственный интеллект. – Донецк, 2004. – № 4. – С. 580–586.
7. *Блюменау, Д. И.* Информационный анализ/синтез для формирования вторичных документов : учеб.-практич. пособ. / Д. И. Блюменау. – СПб. : ПРОФЕССИЯ, 2002. – 240 с.
8. *Зубов, А. В.* Компьютерная лингвистика / А. В. Зубов, И. И. Зубова // Основы лингвистической информатики : в 2 ч. – Минск : МГПИИЯ, 1992. – Ч. 2. – 203 с.
9. *Зубов, А. В.* Вероятностно-алгоритмическая модель порождения текста / А. В. Зубов // Проблемы порождения текста : в 4 ч. – Минск : МГПИИЯ, 1989. – Ч. 2. – 68 с.