УДК 004.912; 004.822

М. Н. Мокроусов, кандидат технических наук, ИжГТУ имени М. Т. Калашникова

АВТОМАТИЧЕСКОЕ СЕМАНТИЧЕСКОЕ РЕФЕРИРОВАНИЕ ТЕХНИЧЕСКИХ ТЕКСТОВ НА ОСНОВЕ ПРАГМАТИЧЕСКОЙ СЕМАНТИЧЕСКОЙ МОДЕЛИ *

Введение

Рефератом называется связный текст, который кратко выражает центральную тему или предмет какого-либо документа, цель, применяемые методы и основные результаты. Проблемами автоматического создания сжатых текстов занимались такие исследователи, как Г. П. Лун [1], В. Е. Берзон [2], И. П. Севбо [3], Э. Ф. Скороходько [4], В. П. Леонов [5], Р. Г. Пиотровский [6] и многие другие. Все существующие методы автоматического реферирования можно разделить на две группы:

1) квазиреферирование (Sentence extraction);

2) генерация реферата с порождением нового текста (Abstraction);

Первое направление предполагает выделение фрагментов документа, наиболее насыщенных информационно. Особенностью такого метода является поверхностный анализ синтаксических отношений, в то время как особенности семантики естественного языка не учитываются.

Методы, направленные на порождение нового текста, используют более сложные семантические подходы, при этом текст реферата представляет собой уже новый документ – пересказ исходного текста. Такой метод позволяет получать рефераты, которые больше походят на рефераты, сделанные человеком.

В теории и практике существует большое количество методов автоматического реферирования: экстрактивный метод, модификации статистического метода Луна, ACSI-Matic, метод Освальда, метод статистических ассоциаций, логико-математический метод, дистрибутивный метод, метод содержательных аспектов, метод текстовых связей Берзона, графовые методы, метод LexRank, методы с использованием дистрибутивной семантики, методы с опорой на знания.

Большинство современных подходов, имеющих практическую реализацию, относятся к первому направлению. В традиционных методах чаще всего используются различные модификации Г. Луна, которые заключаются в отборе предложений с наибольшим весом для включения их в реферат. Вес предложения определяется как сумма частот, входящих в него значимых слов.

Результатом работы таких популярных систем, как Intelligent Text Miner for Text фирмы IBM, Inxight Summarizer, TextAnalyst является квазиреферат, то есть текст, составленный из предложений обрабатываемого документа.

Методы с опорой на знания используют грамматики и словари для морфологического и синтаксического разбора. Для оценки области, к которой принадлежит текст документа, используются онтологические справочники.

Прагматический человеко-машинный словарь

Для хранения онтологии предметной области предлагается использовать прагматический человеко-машинный словарь PraDict [7]. Элементами словарной статьи PraDict являются: *Имя понятия, Толкование понятия, Источник Толкования, Способы применения, Фразеологизмы, Атрибуты, Экземпляры, Состав понятия, Гиперссылки из Толкования и Способов применения,* а также *Мультимедиа* – графические, аудио, видео, текстовые файлы, дополнительно характеризующих понятие, *Перевод понятия* – перевод понятия на другой язык, поддерживаемый системой, *Прецедент* – особый пример использования понятия в текстах, который показывает возможное сочетание понятия с другими понятиями или классами понятий и необходим для моделирования рассуждений и построения исполняемых семантических моделей.

Каждый прецедент описывает возможный сценарий применения понятия в различных ситуациях. Для одного понятия могут быть созданы несколько сценариев. Каждый сценарий включает в себя предметы, действия и отношения с установленными для них свойствами, и графическую семантическую модель сценария (СеМС).

На рис. 1 показана панель «Прецеденты» в PraDict для понятия *to cross – пересекать* с построенным сценарием и графической моделью.

Для формализации семантической модели текста, описывающего ситуацию (СеМС), используется плекс-грамматика [8], где символы грамматических конструкций имеют не две (слева, справа), а *N* «точек примыкания». Значение свойства может быть текстовым или выражать эффект действия. Для действия to cross для типового свойства *effect* задано значение *Кто.место:=Куда*. В данном примере *Кто* – это одна из точек примыкания действия *to cross*, *место* – это одно из свойств предмета *smb*, *Куда* – другая точка примыкания действия *to cross*. Точки примыкания для каждого предмета являются типовыми, символы и их свойства берутся из текущего сценария.

Вместе с графической семантической моделью ситуации в PraDict хранится текстовый файл схемы, который необходим для последующего редактирова-

* Работа выполнена при частичной поддержке госзадания МОиН РФ (проект № 625) и гранта РФФИ № 15-07-08077 А.

ния схемы. С помощью прецедентов обеспечивается возможность описания прагматической модели понятия, которая включает в себя атрибуты, ограничения и реализацию процесса. Такая модель интерпретируется компьютером аналогично декларативным языкам программирования, а диаграфическая семантическая модель обеспечивает удобство ввода и верификации ситуации.

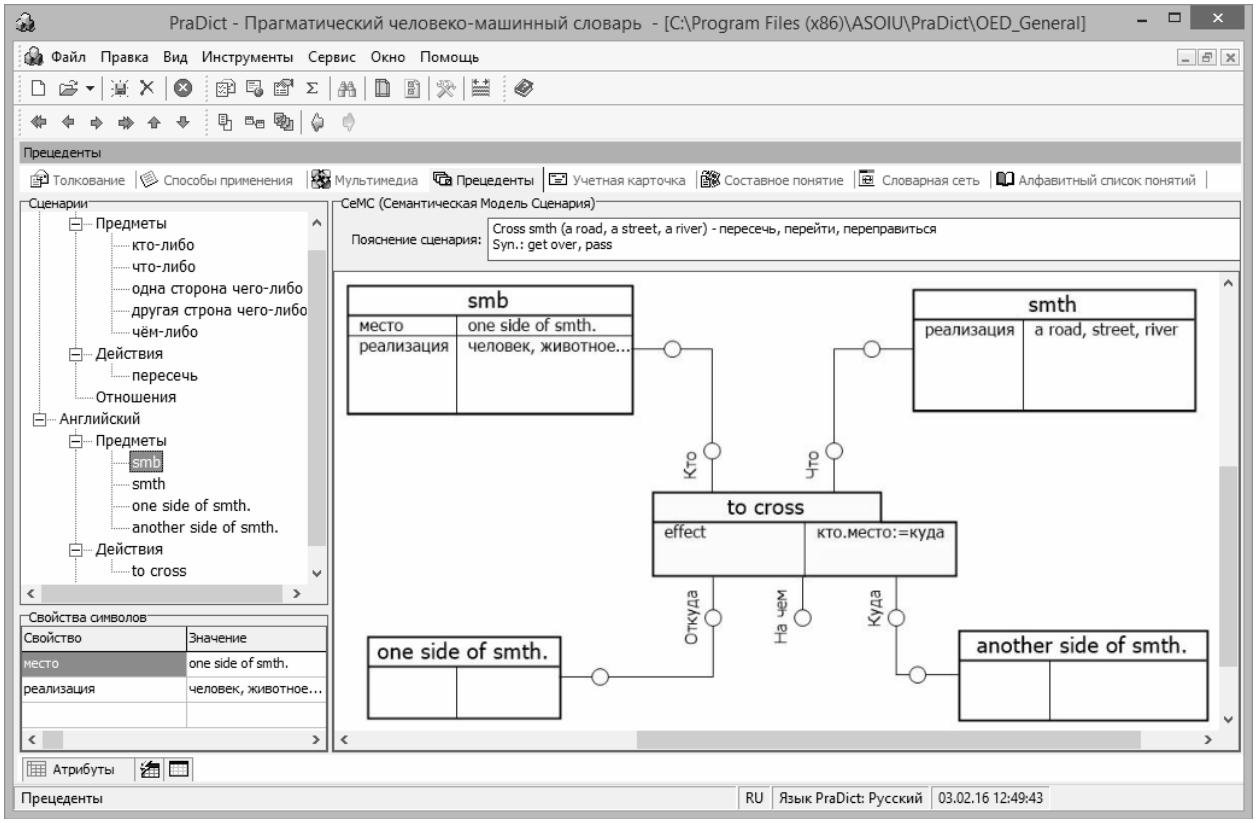

Рис. 1. Прецеденты в PraDict

Таким образом, реализация прецедента (в том числе правила) – это прагматическая интерпретация, т. е. изменение атрибутов в результате исполнения действий в условиях текущей ситуации.

На данный момент начальная онтология содержит 748 понятий, из которых 480 предметов, 106 действий, 154 свойства, 8 отношений.

Предлагаемый метод семантического реферирования основан на обходе прецедентов прагматической семантической модели текста, построенной по результатам синтаксического и семантического анализа исходного текста.

Прагматическая семантическая модель текста состоит из прецедентов процессов и предметов, участвующих в них. В атрибутах прецедентов записываются свойства, выраженные прилагательными, наречиями, числительными, устойчивыми оборотами, аббревиатурами, синонимами. Некоторые атрибуты предметов могут быть выражены обстоятельствами места, причины, цели, назначения. Отношения в такой модели, как правило, выражены предлогами и союзами.

При обходе модели по процессам, игнорируя атрибуты прецедентов, можно получить сжатый реферат, состоящий только из ключевых предметов и процессов. Процессы в модели упорядочены в хронологическом порядке. Свойства предметов и характеристики процессов скрываются во внутренней структуре прецедента и могут быть отображены в любой момент.

Алгоритм реферирования состоит из следующих этапов:

1) строится прагматическая семантическая модель заданного текста;

2) для предметов, процессов и их атрибутов вычисляются коэффициенты значимости по отношению к тексту;

3) обход прагматической семантической модели по прецедентам-действиям генерирует простейший синтаксически корректный вариант реферата;

4) реферат может быть дополнен атрибутами прецедентов, значимость которых выше некоторого порога.

Рассмотрим следующий фрагмент текста (https://ru.wikipedia.org/ wiki/Жёсткий_диск), описывающий устройство жесткого диска.

Накопитель на жестких магнитных дисках или НЖМД – устройство хранения информации, основанное на принципе магнитной записи. Является основным накопителем данных в большинстве компьютеров. Жесткий диск состоит из гермозоны и блока электроники.

Гермозона включает в себя корпус из прочного сплава, собственно диски (пластины) с магнитным

покрытием, блок головок с устройством позиционирования, электропривод шпинделя. Блок головок – пакет рычагов из пружинистой стали (по паре на каждый диск). Одним концом они закреплены на оси рядом с краем диска. На других концах (над дисками) закреплены головки. Диски (пластины), как правило, изготовлены из металлического сплава. Диски жёстко закреплены на шпинделе. Во время работы шпиндель вращается со скоростью несколько тысяч оборотов в минуту. При такой скорости вблизи поверхности пластины создается мощный воздушный поток, который приподнимает головки и заставляет их парить над поверхностью пластины. Форма головок рассчитывается так, чтобы при работе обеспечить оптимальное расстояние от пластины. Шпиндельный двигатель жесткого диска трехфазный, что обеспечивает стабильность вращения магнитных дисков, смонтированных на оси (шпинделе) двигателя. Статор двигателя содержит три обмотки, включенные звездой с отводом посередине, а ротор – постоянный секционный магнит. Для обеспечения малого биения на высоких оборотах в двигателе используются гидродинамические подшипники.

На рис. 2 показан фрагмент прагматической модели, включающий первые 11 предложений.

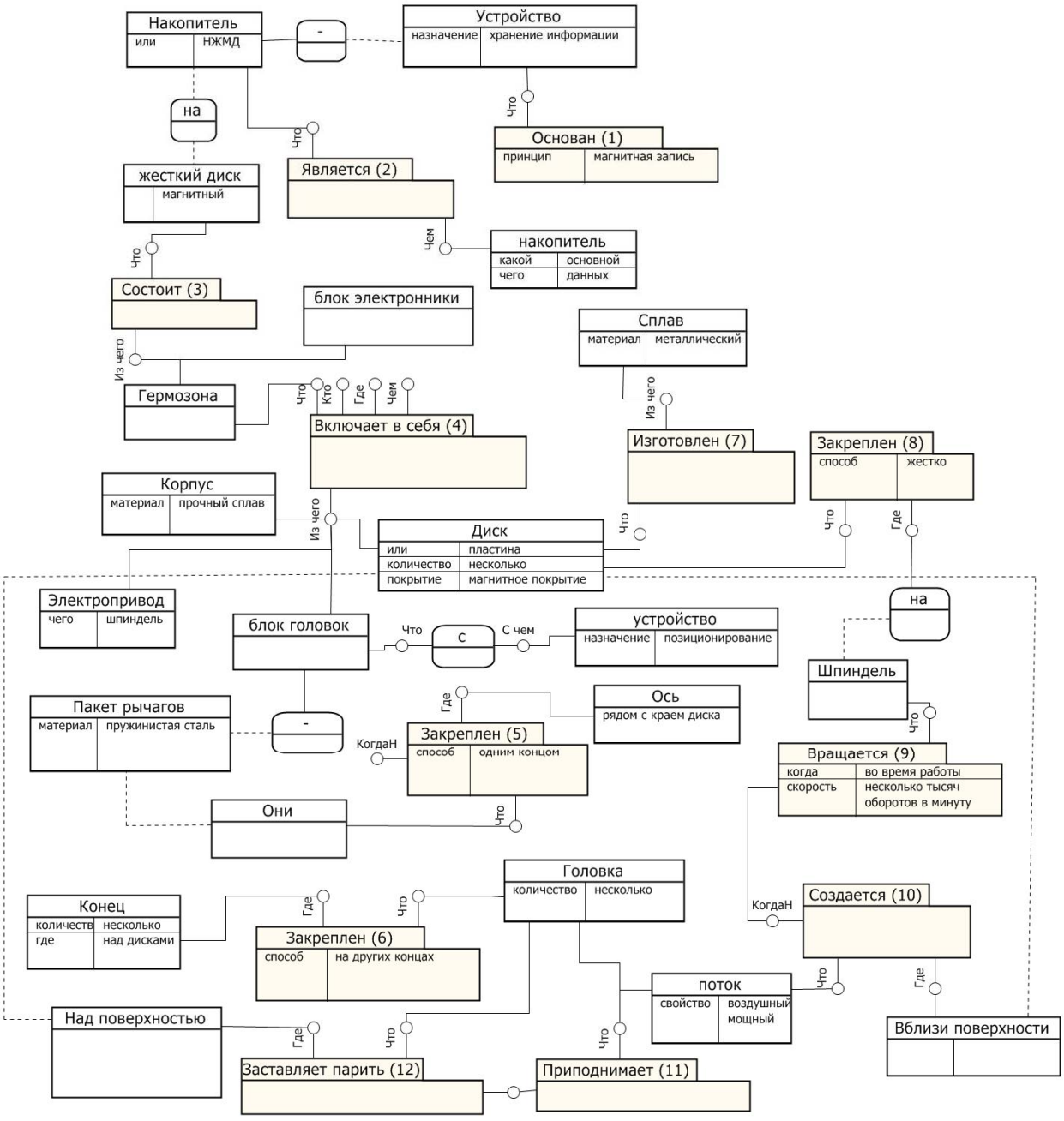

Рис. 2. Прагматическая семантическая модель текста

При обходе по процессам данного фрагмента модели может быть построен следующий реферат.

Накопитель на жестких дисках является основным накопителем данных. Жесткий диск состоит из гермозоны и блока электроники. Гермозона включает в себя корпус, диски, электропривод, блок головок. Блок головок – пакет рычагов. Они закреплены на оси рядом с краем диска. Головки закреплены над дисками. Диски изготовлены из сплава. Диски закреплены на шпинделе. Шпиндель вращается, и создается поток вблизи поверхности дисков. Поток приподнимает головки, заставляет парить головки над поверхностью дисков...

Как видно, такой реферат содержит минимум информации о происходящих процессах и дает общее представление о сюжете. Проход по модели при генерации реферата осуществляется по «действиям», пронумерованным в порядке их встречаемости в исходном тексте. Наличие связей между предметами и процессами, которые именованы вопросами, позволяет выстраивать синтаксически корректную структуру предложения. Наличие атрибута *количество* означает множественное число предмета.

При необходимости можно настроить генерацию таким образом, что в текст будут включены все или некоторые атрибуты прецедентов. Если атрибутов несколько, то в текст реферата могут быть выведены наиболее значимые слова.

Заключение

Предлагаемая методика автоматического реферирования текстов включает четыре этапа:

1) автоматизированный анализ входного текста, включающий предобработку текста, морфологический и синтаксический анализ;

2) синтез прагматической семантической модели текста с использованием процессно-ориентированной онтологии предметной области и прагматического толкового словаря, содержащего прецеденты понятий;

3) генерация реферата путем обхода прагматической семантической модели по прецедентам процессов, расположенных в модели в хронологическом порядке, с сокрытием свойств процессов и предметов;

4) взвешивание элементов текста, в качестве которых выступают слова, словосочетания, предложения, абзацы, и корректировка полученного результата.

Таким образом, предложенный подход к реферированию совмещает в себе как статистический, так и семантический анализ текстов.

Получено 08.02.2016

Библиографические ссылки

1. *Luhn H.* The automatic creation of literature abstracts. In IBM Journal of Research and Development. – 1958. – Vol. 2(2). – Pp. 159–165.

2. *Берзон В. Е.* Синтаксические сверхфразовые связи и их инженерно-лингвистическое моделирование / отв. ред. Р. Г. Пиотровский. – Кишинев : Штиинца, 1984. – 167 с.

3. *Севбо И. П.* Структура связного текста и автоматизация реферирования. – М. : Наука, 1969. – 135 с.

4. *Скороходько Э. Ф.* Семантические сети и автоматическая обработка текста. – Киев : Наук. думка, 1983. – 220 с.

5. *Леонов В. П.* О методах автоматического реферирования. – НТИ. Сер. 2. – 1975. – № 6. – С. 16–20.

6. *Пиотровский Р. Г.* Инженерная лингвистика и теория языка : монография.– Л. : Наука, 1979. – 112 С.

7. *Мокроусов М. Н.*, *Кучуганов В. Н.* Прагматическая компонента текста и человеко-машинный словарь : тр. конгресса по интеллектуальным системам и информационным технологиям «IS&IT-15» : в 3 т. – Таганрог : Изд-во ЮФУ, 2015. – Т. 1. – С. 222–227.

8. *Кучуганов В. Н.* Элементы теории ассоциативной семантики // Управление большими системами. – Вып. 40. – М. : ИПУ РАН, 2012. – С. 30–48.