УДК 002.53/55:001.814

Н.В. Воронков

РАЗРАБОТКА СИСТЕМ АВТОМАТИЧЕСКОГО ПРЕДМЕТНО-ОРИЕНТИРОВАННОГО РЕФЕРИРОВАНИЯ ТЕКСТОВ

Приводятся общая схема построения системы автоматического реферирования текстов, описание метода слов-подсказок и его применения для создания многополевого реферата текстового документа, рассматриваются проблемы, вызванные эвристическим характером метода, а также дается оценка качества работы экспериментальной системы многополевого реферирования, которая разработана с использованием метода слов-подсказок.

Введение

В настоящее время подавляющее большинство накопленных человечеством знаний доступно в виде различного рода документов (текстов) в печатной или электронной форме. Поэтому роль автоматической обработки текстов очень высока и продолжает расти. Поисковые серверы в интернете, несмотря на постоянное улучшение качества поисковых механизмов, возвращают при решении информационных задач огромное количество ссылок на документы, обработать которые с целью удовлетворения своей информационной потребности пользователь просто не в состоянии. С целью сокращения объема анализируемой пользователем информации в помощь ему создаются системы автоматического реферирования текстов (САР).

По принципу обработки текста САР можно разделить на следующие типы:

– выделяющие множество некоторых (наиболее информативных) предложений из исходного текста, возможно, с удалением вводных конструкций, неинформативных слов и словосочетаний и рассматривающие его в качестве реферата;

– выделяющие основные идеи исходного текста и в качестве реферата рассматривающие синтезированный на множестве этих идей текст.

Понятно, что выделение множества идей текста, выстраивание из них логических цепочек, их анализ и последующий синтез текста являются очень трудоемкими задачами, реализовать эффективное решение которых в виде промышленной быстродействующей системы пока все еще затруднительно. Поэтому будем рассматривать САР первого типа. При разработке таких систем ключевым является вопрос, каким образом выделить наиболее информативные предложения текста [1]. С этой целью обычно каждое предложение автоматически оценивается с точки зрения информативности по одному или нескольким алгоритмам (критериям), а в качестве итоговой оценки выступает некоторый интегральный показатель, зависящий от каждого из этих алгоритмов. Разработка таких критериев ввиду субъективизма человеческих (экспертных) оценок является сложной задачей и требует тщательной организации и разработки определенной методики их создания.

В данной статье будет рассмотрена система, в которой предложения оцениваются по нескольким критериям, в основе одного из которых лежит метод слов-подсказок.

1. Общая схема работы САР

Рассмотрим общую схему работы описываемой САР, которую можно представить как последовательность следующих основных этапов:

1) предварительное форматирование входного документа;

2) лингвистический (лексический, лексико-грамматический, синтаксический и семантический) анализ текста;

3) обработка лексико-статистической информации;

4) поиск слов-подсказок в предложении;

5) вычисление итогового веса всех предложений текстового документа в соответствии с начисленными им различными алгоритмами весами, определение по некоторому пороговому значению множества наиболее информативных предложений и выдача их пользователю в качестве реферата с соблюдением порядка следования предложений в документе, а также с учетом найденных слов-подсказок, отвечающих за взаимную зависимость близко расположенных предложений, что обеспечивает дополнительную связность полученного реферата.

Необходимость этапа предварительного форматирования объясняется существованием многих различных форматов документов, поэтому для упрощения процесса их последующей автоматической обработки возникает задача конвертации этих документов в некоторый единый формат, максимально сохраняющий стилистическую и структурную разметку документов. Кроме того, на данном этапе осуществляются разбиение текста на параграфы, выделение заголовков, подзаголовков и отдельных разделов текста. Также происходит фильтрация вспомогательного текста (кнопок, меню, скриптов и т. д.).

На этапе лингвистического анализа определяются границы слов и предложений, лексико-грамматические классы (теги) слов, строится синтаксическое дерево каждой фразы, выделяются синтаксические отношения, распознаются объекты и семантические отношения между ними типа «субъект – акция – объект» (САО), соответствующие грамматическим категориям русского языка «подлежащее – сказуемое – прямое дополнение» и отношения типа «причина – следствие» между самими САО-тройками [2, 3]. Понятно, что САО-тройка соответствует такому классическому элементу инженерии знаний, как факт, а отношение «причина – следствие» – правилу, отображающему закономерности предметной области.

Этапы 1 и 2 будем считать подготовительными для САР. Таким образом, на вход модуля реферирования будут поступать предложения текстового документа с тегами их слов, выделенными синтаксическими и семантическими отношениями, а также раздел документа, в котором находится каждое предложение.

2. Обработка лексико-статистической информации

Задачами этапа обработки лексико-статистической информации являются выделение и накопление статистических весов информативных слов текста с целью выделения наиболее значимых слов [4], причем статистика должна собираться как на чисто лексическом уровне, так и на уровне семантических отношений между выраженными с помощью лексических единиц объектами, фактами и закономерностями предметной области. Для эффективной работы алгоритма выделен ряд тегов так называемых информативных слов и веса начисляются только словам, имеющим такие теги. Для повышения эффективности статистических оценок алгоритма используется ряд коэффициентов, повышающих веса слов, например, в зависимости от того, являются они частью именной группы либо частью семантического отношения САО. Также словам начисляются различные дополнительные веса в соответствии с тем, в каком поле семантического отношения они встречаются, например в субъекте, акции или объекте, а также являются ли они частью заголовка обрабатываемого документа и т. д. После обработки всего документа происходит нормализация весов слов таким образом, чтобы они находились в промежутке от 0 до 1. Затем производится выделение наиболее значимых слов текста посредством отсечения слов с низким весовым показателем по некоторой заранее заданной пороговой величине.

3. Использование слов-подсказок в САР

При использовании метода слов-подсказок решение о включении или невключении в реферат определенных предложений принимается на основании присутствия в этих предложениях некоторых характерных, специфических и тому подобных слов или фраз [5], например: «в данной статье рассматривается (описывается)», «компания является лидером в», «особенностью является» и т. д. Создание слов-подсказок представляет собой сложный процесс и является результатом совместной работы группы экспертов и инженеров по знаниям. Дело в том, что слова-подсказки в действительности являются довольно сложными правилами (так назы-

ваемыми патернами, представляющими собой формальную спецификацию свойства набора примеров, определенную в терминах некоторого формального языка [6]), которые, моделируя поведение эксперта, оперируют самой разной информацией, в том числе и собственно словами-подсказками.

Для реализации данного метода обычно разрабатывается специальный язык слов-подсказок, который обеспечивает возможность поиска в пределах предложения определенных слов, словосочетаний, обеспечивает проверку наличия в предложении заданных лексико-грамматических, синтаксических и семантических отношений. Этот метод по своей природе носит эвристический характер, так как одно и то же правило может интерпретироваться в разных случаях по-разному.

Слова-подсказки используются, например, для решения следующих задач:

– выделения наиболее информативных предложений текста;

– структуризации наиболее информативных предложений текста в соответствии с некоторым заранее заданным набором полей;

– удаления из реферата так называемых стоп-предложений (очень коротких предложений; предложений, не оканчивающихся знаками препинаний; служебных предложений типа «All rights reserved.» и т. д.);

– удаления вводных частей предложений.

В первых двух случаях каждому правилу обычно ставится в соответствие вес, отражающий информативность предложений, выделяемых данным правилом. Понятно, что использование слов-подсказок только для выделения наиболее информативных предложений документа приводит к построению так называемого классического (однополевого) реферата, а подключение их еще и на этапе структуризации реферата – к построению, например, так называемого многополевого реферата, в котором опять-таки представлены наиболее информативные предложения документа, но они распределены в реферате по заранее заданному набору его полей.

4. Использование метода слов-подсказок для создания многополевого реферата

Построение многополевого реферата является эффективным при наличии выраженной смысловой структурности и разнородности тематики реферируемых текстов. Примерами могут служить такие классы текстов, как патенты, пресс-релизы и т. д.

Рассмотрим систему многополевого реферирования применительно к текстам на английском языке. Экспертный анализ текстов патентного фонда США позволил выделить следующие основные с точки зрения использования их содержания при решении инновационных задач поля: Применение (описание области применения патента), Задачи (задачи, решаемые в патенте), Особенности (основные особенности изобретения), Метод (метод решения).

Для построения рефератов пресс-релизов можно выделить, как показывает анализ, следующий набор полей, отражающий техническую и деловую (бизнес-) информацию, представленную в текстах данного класса: Факты (общая информация), Применение (техническое применение), Особенности (особенности описываемого объекта, как правило, технические), Продукция и услуги (информация о продуктах, производимых компанией, а также об оказываемых ею услугах), Финансовая активность (финансовая активность компании), О компании (общая информация о компании).

При использовании метода слов-подсказок для создания многополевого реферата все подсказки группируются по полям (заранее заданным и зафиксированным) и выполняют функции определения принадлежности предложения к одному из полей, а также информативности предложения в пределах поля. Чтобы не допустить появления одного и того же предложения в реферате дважды, будем считать, что предложение может быть отнесено только к одному полю. Отсюда вытекает проблема, вызванная также эвристическим характером метода слов-подсказок, – неоднозначность отнесения к тем или иным полям одного и того же предложения. Рассмотрим, например, предложения пресс-релизов, выделяемые словами-подсказками *founded in* и *had revenue* (табл. 1).

Таблица 1

Пример предложения, принадлежащего нескольким полям

Поле реферата	Слово-подсказка	Предложение
О компании	founded + in	*Founded in* 1969 and based in Sunnyvale, California, AMD *had revenues* of $ 2.9 billion in 1999.
Финансовая активность	had + revenue	

В этом примере предложение может быть одновременно отнесено к двум полям – и к полю О компании, и к полю Финансовая активность. С целью снятия такого типа многозначности предлагается учитывать следующие показатели:

– вес каждой подсказки, отражающий качество и информативность выделяемой подсказками информации (под качеством понимается точность работы каждого правила, а под информативностью – важность информации, содержащейся в выделяемом предложении, применительно к полю, в котором находится шаблон слова-подсказки);

– приоритет полей, отражающий важность информации данного поля по отношению к другим полям при равных весах подсказок.

Качество и информативность подсказок в данном примере являются высокими, что ставит им в соответствие максимальный из возможных вес, равный 1, но более важной, по мнению экспертов, является финансовая информация, что должно быть отражено в приоритете полей и позволит отнести предложение данного примера к полю Финансовая активность.

Надо сказать, что такого рода многозначность порождается еще и эвристическим характером самой классификации полей, которая не гарантирует четких их границ (см., например, поля О компании и Продукция и услуги). Рассмотрим такое предложение: *«Abbott Laboratories* ***is a global, diversified*** *health care* ***company devoted to*** *the discovery,* ***development, manufacture and marketing of*** *pharmaceutical, diagnostic, nutritional and hospital products».* В этом предложении описывается продукция, которая производится компанией, но, с другой стороны, эта же информация указывает сегмент рынка, на котором представлена фирма, и не содержит конкретные марки продуктов, которые выпускаются фирмой. Поэтому данное предложение может быть отнесено как к полю О компании, так и к полю Продукция и услуги.

Таким образом, данный метод включает в себя долю субъективизма экспертов при отнесении слов-подсказок к тем или иным полям и оценивании их веса и пользователя, согласно предпочтениям которого может настраиваться приоритет полей.

Построение многополевого реферата кроме структуризации выделяемой информации имеет еще одно важное преимущество. Как показали исследования, более эффективным при построении классического реферата является предварительное построение многополевого реферата обрабатываемого текста. Далее производится выбор, например, по одному наиболее информативному предложению из каждого поля в соответствии с некоторым заранее заданным приоритетом, пока не наберется заданное пользователем число предложений. Эксперименты показали, что такой метод построения классического реферата позволяет увеличить оценку полноты для классического реферата в среднем с 44 до 56 % за счет более разнообразной информации, отражаемой в реферате.

5. Методика создания слов-подсказок для построения многополевого реферата

Методика создания слов-подсказок для рассматриваемого случая включает следующие шаги:

1. Создание корпуса текстов для заданного их класса. На этом этапе происходит подбор наиболее типичных текстов для выбранного класса.

2. Анализ структуры текстов с целью выбора и фиксации набора полей, характерных для данного класса текстов.

3. Разбиение текста на предложения.

4. Разметка экспертами предложений по их принадлежности к определенным полям. Для этого эксперт анализирует текст и для каждого предложения отмечает те поля, к которым можно отнести данное предложение.

5. Создание новых слов-подсказок. Размеченный корпус обрабатывается существующими словами-подсказками (если они уже есть) и выбираются предложения, которые не выделены САР в отмеченные экспертами поля. Далее эксперт для каждого предложения выделяет ключевые слова, а также определенные синтаксические и семантические отношения, которые могут использоваться в качестве элементов правил для выделения этого предложения. Затем на их основе инженер по знаниям пишет соответствующие правила. Так, например, в процессе анализа предложений патентов, отнесенных экспертами к полю Применение, могут быть выделены характерные слова (табл. 2).

Таблица 2

Примеры предложений для поля Применение с характерными словами

Предложение	Характерные слова
In another aspect, the invention pertains to a refrigeration process and apparatus.	...invention...pertains to...
The apparatus of the present invention is suited to a variety of procedures for varying the composition and physical properties of a solution or suspension of macromolecules, while maintaining a constant volume.	...invention...is...suited to...
The present invention is unusually well suited for the production of titanium from titanium tetrachloride and zirconium from zirconium tetrachloride.	...invention…is...suited for...

Для выделения приведенных в табл. 2 предложений могут быть составлены слова-подсказки, поиск которых осуществляется по следующим алгоритмам:

А. Искать слово, имеющее каноническую форму *invention* (т. е. это само слово *invention* и все его словоформы в пределах части речи), за которым на расстоянии не более четырех слов идет слово, имеющее каноническую форму *pertain* и за которым на расстоянии не более двух слов следует предлог *to*.

Б. Искать семантическое отношение типа САО, для полей которого выполняются условия:

– поле Объект отношения содержит слово, имеющее каноническую форму *invention*;

– поле Акция отношения содержит слово, имеющее каноническую форму *be*, за которым следует слово *suit*, имеющее лексико-грамматический тег причастия прошедшего времени;

– поле Предлог отношения содержит предлоги *to* или *for*.

6. Тестирование и корректировка новых слов подсказок. Полученными подсказками обрабатывается некоторый корпус документов, и эксперт оценивает качество работы каждого правила с точки зрения качества (точности) и информативности выделяемых правилом предложений. Затем производится коррекция правила, возможно, его обобщение, и тестирование повторяется вновь итеративно, пока не будут получены результаты, на основании которых можно принять решение о добавлении новых правил в базу правил САР или об отказе от каких-то из них.

6. Вычисление итоговых весов предложений

После обработки предложений всей совокупностью алгоритмов реферирования производится вычисление итоговых весов предложений, отражающих их информативность в пределах обрабатываемого документа [7]. Для этого каждый из алгоритмов получает некоторую «долю» в итоговом весе в соответствии с точностью и качеством алгоритма. Само вычисление итогового веса производится в два этапа:

– нормализация весов предложений, полученных по каждому из алгоритмов в пределах промежутка от 0 до 1, причем вес наиболее важного предложения, выделенного каждым алгоритмом, не обязательно должен быть равен 1;

– итоговый вес предложения вычисляется по формуле

$$SW_i = \sum_j a_j \cdot w_{ij},$$

где SW_i – вес *i*-го предложения;

j – количество алгоритмов, по которым производится оценка информативности предложений;

a_j – доля *j*-го алгоритма в итоговом весе предложения (предполагается, что $\sum_j a_j = 1$);

w_{ij} – информативность *i*-го предложения в соответствии с *j*-м алгоритмом.

При выполнении описанных выше условий очевидно, что итоговый вес предложения будет находиться в пределах от 0 до 1.

В процессе отображения реферата показывается некоторое заданное заранее число самых информативных предложений каждого поля, которое может меняться пользователем в процессе просмотра реферата.

7. Оценка качества работы экспериментальной системы многополевого реферирования

Результат тестирования качества работы разработанной экспериментальной системы, использующей описанный выше подход для реферирования, например, пресс-релизов, отображен в табл. 3.

Таблица 3

Качество работы экспериментальной САР для документов типа «пресс-релиз»

Поле / Качество	О компании		Продукция и услуги		Финансовая активность		Применение		Особенности	
Показатель	R	P	R	P	R	P	R	P	R	P
Величина	0,48	0,8	0,27	0,84	0,82	1	0,65	1	0,29	0,73
F-Measure	0,600		0,409		0,900		0,788		0,415	

Примечания:

R – показатель полноты (recall), традиционно определяемый как отношение количества корректно выделенных САР предложений из определенного поля к количеству предложений из этого документа, относящихся к данному полю [8];

P – показатель точности (precision), традиционно определяемый как отношение количества корректно выделенных САР предложений из определенного поля к количеству всех предложений, выделенных САР для этого поля [8];

F-Measure – интегральный показатель оценки качества работы САР (такого типа единый показатель часто используется для оценки качества работы информационных систем и отдельных алгоритмов обработки информации вместо совокупности двух указанных показателей [9]), определяемый по формуле $\text{F-Measure} = \frac{2 \cdot \text{R} \cdot \text{P}}{(\text{R} + \text{P})}$.

Понятно, что значение F-Measure находится в промежутке от 0 до 1 и чем выше показатели полноты и точности, тем выше показатель F-Measure.

Заключение

Метод слов-подсказок при построении САР в сочетании с лексико-статистическими алгоритмами дает качество автоматического построения реферата, приемлемое для промышленных систем САР. Однако эвристический характер этого метода приводит к необходимости тщательной разработки методики создания и тестирования слов-подсказок.

Согласно методу, представленному в статье, была разработана экспериментальная версия САР для многополевого реферирования пресс-релизов компаний по трем бизнес-полям и двум техническим полям. Качество реферирования (F-Measure) оказалось близко к 0,6, что вполне приемлемо для экспериментальной версии.

Предварительное создание многополевого реферата дает преимущество при построении из него однополевого путем отбора наиболее информативных предложений каждого поля и по-

следующего их объединения. Это дает возможность получить более разностороннюю информацию о реферируемом тексте, что на практике позволяет повысить оценку полноты уже существующей однополевой САР с 44 до 56 %, т. е. в 1,27 раза.

В дальнейшем более детально будет изучено влияние структуры документов на информативность предложений. Также будет производиться развитие возможностей языка слов-подсказок с целью расширить границы применения слов-подсказок за пределы одного предложения и использовать информацию, выделенную одними словами-подсказками, в других словах-подсказках, что позволит повысить качество формирования полей в многополевом реферате.

Список литературы

1. Mani, I. Advances in Automatic Text Summarization / I. Mani, M.T. Maybury. – Cambridge: The MIT Press, 1999. – 434 p.

2. Document semantic analysis/selection with knowledge creativity capability utilizing subject-action-object (SAO) structures / L.S. Batchilo, I.V. Sovpel, V.M. Tsourikov. US Patent 6167370, December, 2000. – 120 p.

3. Совпель, И.В. Инженерно-лингвистические принципы, методы и алгоритмы автоматической переработки текста / И.В. Совпель. – Минск: Вышэйшая школа, 1991. – 120 с.

4. Воронков, Н.В. Автоматическое топик-ориентированное реферирование / Н.В. Воронков, И.В. Совпель // Материалы междунар. конф. «Когнитивное моделирование в лингвистике». – Казань, 2002. – С. 94–102 .

5. Воронков, Н.В. Использование эвристических оценок в задаче автоматического реферирования текстов / Н.В. Воронков // Материалы 1-й Междунар. конф. «Информационные системы и технологии». – Минск, 2002. – С. 122–127.

6. Городецкий, В.И. Современное состояние технологии извлечения знаний из баз и хранилищ данных. Ч. 1 / В.И. Городецкий, В.В. Самойлов, А.О. Малов // AI News. – 2002. – № 3. – С. 3–12.

7. Computer based summarization of natural language documents / L.S. Batchilo, I.V. Sovpel, V.M. Tsourikov. US Patent Appl. № 20030130837, 2003.

8. Salton, G. Introduction to Modern Information Retrieval / G. Salton, M.J. McGill. – New York: McGraw-Hill, 1983.

9. Rijsbergen, C. J. van. Information Retrieval / C. J. van Rijsbergen. – London: Butterworths, 1979.

Поступила 08.09.05

ИП «Инвеншион Машин»,
Минск, ул. Революционная, 11
email: nvoronkov@gmail.com,
nvoronkov@imb.invention-machine.com

N.V. Voronkov

SYSTEM DEVELOPMENT FOR AUTOMATIC DOMAIN-ORIENTED TEXT SUMMARIZATION

The article presents a general scheme of automatic text summarization system. A description of cue-words-based method is given as well as its application to generation of multi-field summary of text documents. Problems related to heuristics nature of the method are described. Quality estimation of an experimental multi-field summary generation system based on the above-mentioned method is given.