Пореева Е.Э., Дымо А.Б.
(Николаев, Национальный университет кораблестроения)

АВТОМАТИЗИРОВАННАЯ СИСТЕМА РЕФЕРИРОВАНИЯ АНГЛОЯЗЫЧНОГО НАУЧНО-ТЕХНИЧЕСКОГО ТЕКСТА

Постановка задачи.

Данный доклад посвящен вопросам создания автоматизированной системы реферирования англоязычного научно-технического текста. В работе рассмотрена совокупность методов автоматического анализа и синтеза информации из нескольких источников. Показано, как эти методы способны повысить качество автоматического анализа текста. Обсуждаются пути создания самообучающихся систем автореферирования.

Актуальность проблемы.

Задаче компьютерного анализа текста на естественном языке посвящено множество работ, так как адекватное реферирование научно-технического текста является неотъемлемой частью научной жизни, отнимая много времени и усилий. Автоматические системы реферирования берут на себя частичное решение данной проблемы. Они применяют статистические методы для обработки текста и игнорируют лингвистические, что приводит к формированию бессвязных и неструктурированных текстов.

Вышеперечисленные проблемы не позволяют создать текст, четко и кратко излагающий основную идею, раскрывающий тему научной статьи (статей) и позволяющий пользователю получить четко сформированный ответ на интересующий его вопрос.

Проблемы бессвязности и неструктурированности текста до сих пор остаются нерешенными, так как для их разрешения необходимо проводить глубокий синтаксический, семантический, лексический и морфологический анализ, что является сложной и трудоемкой задачей.

Автоматизированная система реферирования: лингвистический аспект.

Решить проблему неструктурированности возможно с помощью автоматизированной системы реферирования, позволяющей

задавать пользователю структуру реферата;

- идентифицировать структуру реферируемых документов и выбирать их наиболее релевантные части для включения в реферат;

- задавать и настраивать параметры алгоритма структурной идентификации.

Методика работы такой системы должна быть аналогична человеческой, т.к. именно человек выполняет вышеперечисленные операции наиболее адекватно. Иными словами, она должна моделировать процесс реферирования текста человеком.

Данный процесс выглядит как:

- анализ темы и, как следствие, подбор базовых документов;
- анализ найденной информации:
  - поиск наиболее релевантных отрывков;
  - определение структуры базовых документов, их жанра и стиля;
- синтез итогового документа:
  - продумывание структуры итогового документа (построение плана);
  - компиляция релевантных частей текста, выбранных на этапе анализа базовых документов;
  - изменение структуры предложения для повышения адекватности текста структуре реферата и раскрытия темы;
- построение выводов.

Анализ темы подразумевает определение ее семантики, соотнесение ее с определенной областью в науке, поиск грамматического центра в предложении, а также выделение ключевых слов и словосочетаний, необходимых для дальнейшего поиска информации, подборки базовых документов и их анализа.

Семантика темы определяется на основе имеющихся знаний выполняющего реферирование. Подобный резерв можно обозначить термином «запас знаний» (‘stock of knowledge’). Ключевые слова и словосочетания определяются в процессе анализа грамматического центра предложения.

На этапе анализа найденной информации можно выделить два основных процесса, проистекающих параллельно. Это поиск наиболее релевантных отрывков и определение структуры, жанра и стиля базовых документов.

Поиск наиболее релевантных частей текста происходит на основе «первичного» запаса знаний (ЗЗ), т.е. на основе уже имеющихся знаний у выполняющего реферирование, и на основе «вторичного» ЗЗ, которая включает в себя информацию, полученную в процессе выполнения предыдущих этапов и данного. «Вторичный» ЗЗ содержит ключевые фразы и структуру анализируемых документов.

Структуру итогового документа определяет вид реферата. На основе выбранной структуры происходит компиляция текста из частей, выбранных в

процессе анализа и обработки найденной информации. Процесс перестройки предложений и текста происходит параллельно с процессом компиляции, и основывается на жанре и стиле итогового реферата, которые определяются выполняющим реферирование.

Процесс перестройки текста обычно включает в себя построение высказываний, выводов, вступлений, и происходит на базе изменения структуры предложения на синтаксическом уровне с использованием процессов слияния, расширения и обобщения. Избавление от тавтологии на всех языковых уровнях происходит с использованием процессов синонимии и вариантности.

Практически все вышеперечисленные процессы составление реферата человек выполняет подсознательно, т.е. автоматически, на основе накопившихся и сформировавшихся в процессе приобретения опыта знаний и правил.

Моделирующая процесс система ввиду своей автоматизированности оставляет часть действий для человека, а часть выполняет автоматически. Так, за пользователем системы остается анализ темы и подбор базовых документов, а также, самое основное, задание структуры итогового документа. На долю автоматической части системы остается:

- поиск наиболее релевантных отрывков;
- определение структуры базовых документов, их жанра и стиля;
- компиляция релевантных частей текста, выбранных на этапе анализа базовых документов.

Определить структуру базовых документов возможно с помощью таблицы соответствий предложений и таблицы согласованности абзацев. Информация из данных таблиц содержится в первичном ЗЗ и представляет собой множество пар вида регулярное выражение-элемент структуры. Пример приведен в таблице 1.

Таблица 1. Соответствие элементов структуры и регулярных выражений

элемент структуры	***регулярное выражение***
problem statement (actuality, innovation)	in the following .*(problem | <subject>) (is first)? formulated
solution	describe? the advance of <subject> (may be)? represented (through | by)
conclusions	.*has been performed

Для научных текстов легко построить достаточно полную таблицу соответствия ввиду однообразности их стиля. Используя таблицу, распознаются элементы структуры, присутствующие в базовых документах и одновременно отмечаются во вторичном ЗЗ предложения, представляющие эти элементы. Отмеченные предложения автоматически рассматриваются как релевантные для задачи компиляции. Вместе с этим, применяются адаптивные средства

фильтрации, которые позволяют пользователю в каждом конкретном случае указать критерии выбора информации, на основе которых составляется итоговый документ. Например, возможно отбросить либо наоборот включить элементы структуры, распознанные как примеры, графики и т.п.

В дополнение к таблице соответствия как методу обеспечения структурированности, используются дополнительные правила обеспечения связанности текста результата:

- правила сохранения предложений (таблица 2) - множество регулярных выражений для нахождения важных с точки зрения обеспечения связанности предложений;
- правила согласованности предложений (таблица 3) и абзацев (таблица 4) - множество продукций для нахождения, соответственно, предложений и абзацев, согласованных с рассматриваемым.

Таблица 2. Правила сохранения предложений

^(all)though
there <to be> (strong)? (evidence | fact | prove)
this .* is <action> .* in another (paper | work)

Таблица 3. Правила согласованности предложений

because of (this | that) :: –<sentence>
in addition to: -<sentence> + <broadening>

Таблица 4. Правила согласованности абзацев

In addition to: Example-
This is (was) (merely)? is (one)? example: Example- or Example+

Для описания регулярных выражений и продукций, представленных в таблицах 1-4 используется синтаксис из [2, 101-233] и [3].

Автоматизированная система реферирования: взгляд с точки зрения программной реализации.

Задача реферирования разбивается на задачи синтаксического анализа предложений, структурного анализа предложений и абзацев и компоновки. Синтаксический анализ может быть выполнен с помощью контекстно-свободного анализатора Link [3], идея которого восходит к модели непосредственно составляющих Бархударова [1, 20].

Структурный анализ есть:

1) сопоставление частей поступающих предложений регулярным выражениям, соответствующим определенным структурам;
2) сопоставление целых предложений и абзацев правилам согласованности.

Компоновка предполагает сравнение грамматических центров предложений, соответствующих одному элементу структуры.

Проблема реферирования является для автоматизированной системы P-проблемой, т.е. разрешимой за полиномиальное время (в нашем случае за $O(Pc*Sc^2*N^5*Rl*SPc)$, где Pc - количество абзацев в тексте, Sc – количество предложений в тексте, N – количество слов в предложении, Rl – суммарная длина регулярного выражения, полученного дизъюнкцией выражений из таблицы соответствия, SPc – количество пар предложений, подлежащих компоновке). Такая характеристика системы, безусловно, является важной. Однако, большое количество определяющих параметров и немалые показатели степени снижают потенциальные области применения. Улучшение же производительности позволит применять предлагаемый метод реферирования в онлайновых научных электронных библиотеках и архивах, до сих пор использующих менее адекватные, но более производительные статистические методы со временем работы $O(Sc*N*SPc)$.

Представляется возможным подойти к проблеме увеличения производительности системы реферирования с трех сторон, соответствующих синтаксическому анализу, структурному анализу и компоновке. Однако, наиболее очевидный способ лежит в сведении времени сопоставления предложения элементу структуры от $O(N^2*Rl)$ к $O(N)$. Характеристика $O(N)$ в данном случае является целью, т.к. она является скоростью работы человека на той же задаче. Если задаться целью определить алгоритм сопоставления, в какой-то мере соответствующий процессу, происходящему в человеческом мозге, то можно увидеть путь, ведущий к такой цели. Наиболее удачными средствами моделирования работы мозга есть нейронные сети, описанные, например, в [4].

Нейронная сеть есть весьма удачная аналогия человеческому мозгу применительно к задаче сопоставления предложений элементам структуры по двум причинам. Первая причина – это то, что как и мозг, сеть в состоянии выдать состояния активации (сигналы) в выходном слое нейронов за время $O(N)$ в случае последовательной подачи слов из предложений во входной слой и даже за время $O(1)$ при подаче всего предложения целиком. Второй причиной удачности аналогии есть свойство сетей обучаться и аппроксимировать. Сеть с прямой подачей будучи обучена с помощью разработанной системы автоматизированного реферирования на некотором обучающем множестве предложений, может затем выдавать заключения о предложениях не входящих в него основываясь на их

“похожести”, т.е. выполнять ту же эвристическую процедуру, что и человек в процессе реферирования.

Нейронная сеть Кохонена [5],[6] является наиболее подходящим способом реализации самообучающейся распознающей сети. В начале своего существования ее топология будет представляться множеством нейронов, соединенных между собой случайным образом и случайным-же образом активирующихся при подаче предложений на вход. После начального обучения сети нейроны будут организованы таким образом, что при подаче предложений, принадлежащих одной структуре, будут активироваться только нейроны из некоторой, вполне определенной пространственной области. Таким образом, по принадлежности активированных нейронов областям будет определяться принадлежность предложения элементу структуры. Использование алгоритма самообучения позволит распознавать не только предложения, точно соответствующие структуре, но и похожие.

Заключение.

Создаваемая система реферирования англоязычного научно-технического текста, опираясь на методы реферирования, успешно используемые человеком обладает следующими характеристиками:

- создает релевантные, связные и структурированые рефераты;
- позволяет адаптировать процесс реферирования для получения желаемого результата;
- обладает потенциальными возможностями самообучения.

Литература

1. Бархударов Л.С. Структура простого предложения современного английского языка. М.: Издательство “Высшая школа”. 1966. - 200с.
2. Хопкрофт Дж., Мотвани Р., Ульман Д. Введение в теорию автоматов, языков и вычислений. - М.: Издательский дом “Вильямс”. 2002. - 528 с.
3. Sleator D.D., Temperley D. Parsing English with a Link Grammar. - Pittsburg: School of Computer Science, Carnegie Melon University. - 1993. - 14p.
4. Krose B., P. van der Smagt An introduction to Neural Networks. - Amsterdam: The University of Amsterdam. - 1996. - 135p.
5. Kohonen T. Self-organized formation of topologically correct feature maps //Biological Cybernetics, №43. - 1982. - pp.59-69.
6. Kohonen T. Self-Organization and Associative Memory. Berlin: Springer-Verlag.