АВТОМАТИЧЕСКОЕ ПОСТРОЕНИЕ ТАБЛИЧНОГО РЕФЕРАТА ГРУППЫ ТЕКСТОВ ОДНОЙ ТЕМАТИКИ

AUTOMATIC CONSTRUCTION OF TABLE ABSTRACT OF COMMON SUBJECT TEXTS

А.В. Зубов
mailto: proscien@mslu.by
Минский государственный лингвистический университет

В статье предлагается лингвистический алгоритм автоматического построения табличного реферата группы текстов одной тематики (поездки, встречи и т.п.). Система автоматически выделяет главные опорные слова каждого текста и делит их на субъекты, объекты, предикаты, места действия и время действия. Совокупность таких разрядов слов для группы текстов и составляет их реферат.

Успехи в создании программ автоматического анализа слов и предложений естественных языков, наличие программ автоматического таггирования текстов позволяют создать информационные системы оперативной обработки информации, появляющейся в таких источниках как газеты, Интернет и т.п. Такие системы, накапливая информацию на одну тему ("передвижение некоторого государственного деятеля", "выборы в какой-то стране", "военные учения", "участие страны / команды / игрока в спортивных соревнованиях по какому-то виду спорта'" и т.д.), позволяют далее извлекать из них большой объем информации – как оперативной, так и отсроченной (с изменениями за неделю, месяц, год, несколько лет). Основой такой системы может стать предлагаемый табличный реферат группы текстов одной тематики.

Процесс его создания включает следующие пять задач:

1. Ввод очередного текста в компьютерную память.
2. Выделение из данного текста наиболее значимых для содержания слов.
3. Сегментацию каждого предложения на формальные группы: группу подлежащего, группу сказуемого, группу допол-

нения, обстоятельственные группы.

4. Построение табличного реферата данного текста, включающего следующие графы: "субъект", "предикат", "объект", "обстоятельство места", "обстоятельство времени".

Последовательное заполнение такой таблицы для каждого нового текста позволяет решить 5-ую задачу – построение табличного реферата группы текстов одной тематики. Из такого реферата можно в любое время получить, например, ответы на следующие вопросы:

1. Когда руководитель X посетил страну У?
2. В каких странах побывал руководитель за время с W по Z?
3. С кем встречался руководитель X в стране У?
4. Какие вопросы обсуждал руководитель X в стране У? и т.д.

Нами был проведен эксперимент по созданию такой системы. Материалом для ее построения послужил 21 текст на английском языке, взятый из сети Интернет и содержащий информацию о визитах Президента России В. Путина в разные страны в 2000–2002 годах.

Отмеченные выше 5 задач, которые необходимо решить в процессе построения системы, были реализованы следующим образом. В ходе выполнения первых двух задач исходные тексты подвергались следующей обработке (при помощи ранее созданных программ):

1. Тексты приводились к формату "только текст" ("plain text").
2. Тексты разбивались на слова и предложения, выделялись устойчивые словосочетания (например, "look for", "in front of").
3. Производилось таггирование текстов (сопровождение слов текстов лексическими тэгами (признаками)).
4. Затем производилось разрешение анафоричности слов текстов. А именно, анафорические местоимения (местоимения, которые ссылаются на эксплицитно выраженные именные группы) на основе словарей (абстрактных понятий, географических названий, фамилий, мужских и женских имен и т.д.) были заменены их антецедентами (или референтами). Например, словоформа Putin заменяла словоформу he, которая является ее антецедентом.

Далее обработанные тексты подавались на вход модуля, отвечающего за построение словарей **главных опорных слов**. Процедура создания таких словарей начиналась с построения распределительного алфавитно-частотного словаря. Подобный словарь содержит информацию о частоте употребления каждой словоформы в тексте (F), общем количестве абзацев, в которых встретилась данная словоформа (m), и конкретных номерах абзацев. Все словоформы словаря упорядочены по алфавиту и убыванию частоты их употребления в каждом тексте.

Затем распределительный алфавитно-частотный словарь был подвергнут следующей обработке:

1. На первом этапе из словаря была исключена служебная и некоторая общеупотребительная лексика. Например, из словаря были удалены артикли (a, the), предлоги (to, in, for), союзы (and, but), местоимения (his, you), общеупотребительные (go, come, have, be), вспомогательные (do, have, be) и модальные (must, should) глаголы, прилагательные (private, inevitable), числительные (one, two, first, second).
2. На втором этапе были объединены разные грамматические формы одного и того же слова. Например, словоформа says имеет частоту F=1 и встретилась в одном абзаце текста. Словоформа said имеет частоту F=4 и встретилась также в одном абзаце этого же текста. После объединения грамматических форм в словаре остается одна словоформа say с суммарной частотой F=1+4=5 и общим количеством абзацев m=2. Или, например, словоформа Putin имеет частоту F=9 и встретилась в четырех абзацах текста (0, 1, 2, 3), а словоформа Putin's имеет частоту F=2 и встретилась в двух абзацах этого же текста (4, 5). В результате объединения получается одна словоформа Putin с суммарной частотой F=9+2=11 и общим количеством абзацев m=6.
3. Наконец, на последнем этапе из распределительного алфавитно-частотного словаря были удалены словоформы, встретившиеся только в одном абзаце (m=1) и передающие, следовательно, основное содержание только данного абзаца.

Оставшиеся после перечисленных выше трех процедур обработки распределительного алфавитно-частотного словаря словоформы составили словарь **потенциальных опорных слов** текста. Данный словарь содержит список словоформ с информацией о частоте употребления каждой словоформы в тексте (F) и общем количестве абзацев, в которых встретилась данная словоформа (m).

Затем для каждого слова из словаря потенциальных опорных слов был вычислен **коэффициент важности (коэффициент семантической значи-**

мости). Вычисления производились по формуле $K_B = F * m / N * n$, где F – абсолютная частота слова в тексте; m – общее число абзацев, в которых встретилось слово; N – общее число слов в тексте; n – общее число абзацев в тексте.

Используя разработанные автором формулы, для каждого текста с помощью компьютера были определены критические (пороговые) значения коэффициентов важности слова K_B^1 и K_B^2. Они и позволили разделить словарь потенциальных опорных слов текста на две части и получить список **главных опорных слов** текста и **второстепенных опорных слов** текста. Для создания реферата были использованы лишь главные опорные слова текстов.

Следующий этап построения табличного реферата группы текстов связан с автоматической сегментацией предложений исходных текстов с целью выделения формальных групп подлежащего, сказуемого, дополнения, обстоятельства места и обстоятельства времени (см. 3-ью задачу). Этот этап состоял из двух подэтапов:

1. Лексико-семантический анализ текстов при помощи ранее разработанных программ с целью выделения групп подлежащего, сказуемого и дополнения. На вход подавались заранее таггированные тексты. Производилась обработка каждого предложения текстов, в результате которой выделялись следующие поля: подлежащее, сказуемое, дополнение, предлог косвенного дополнения, косвенное дополнение и обстоятельство.
2. На следующем подэтапе разработанный модуль производил анализ полученных ранее полей с целью выделения отмеченных выше формальных групп.

Основной сложностью на втором подэтапе было выделение формальных групп дополнения, обстоятельства места и времени. В результате лингвистического анализа левой и правой контактной дистрибуции имен существительных были получены списки маркеров левой и правой границ именной группы. К ним относятся:

1. Единичные словоформы, общие для левой и правой границы (смешанные границы).
2. Бинарные сочетания, общие для левой и правой границы (смешанные границы).
3. Единичные словоформы, характерные только для левой (или для правой) границы.
4. Бинарные сочетания, характерные только для левой (или для правой) границы.

Все эти списки были упорядочены по признакам инклюзивности (включения в именную группу) и эксклюзивности (не включения в именную группу). При более глубоком анализе маркеров именной группы были получены маркеры формальных групп дополнения, обстоятельства места и времени.

Процесс решения 4-ой задачи опирался на словари главных опорных слов каждого текста и выделенные на предыдущем этапе формальные группы подлежащего, сказуемого, дополнения, обстоятельства места и времени. Формальные группы подлежащего, сказуемого и дополнения проверялись на наличие в них главных опорных слов. В случае нахождения хотя бы одного главного опорного слова в одном из вышеуказанных трех полей, группа подлежащего заносилось в поле "субъект", группа сказуемого – в поле "предикат", группа дополнения – в поле "объект", группа обстоятельства места – в поле "обстоятельство места" и группа обстоятельства времени – в поле "обстоятельства времени".

После построения реферата каждого вновь введенного текста происходила обработка словарей главных опорных слов всех текстов, уже находящихся в компьютере, с целью создания единого словаря, содержащего слова, которые являются главными опорными словами для всей группы текстов. Для этого использовался модуль объединения словарей главных опорных слов, который производил трансформации исходных словарей, позволяющие определить количество текстов, в которых встретилась словоформа, и ее общий суммарный вес. Например, если словоформа Putin встретилась в 16-ти текстах, то ее вес равен сумме всех весов, которые слово имело в текстах.

Далее модуль обрабатывал все полученные главные опорные слова и оставлял только те из них, которые встретились более чем в одном тексте. Критерием для подобного отбора служил принцип, использованный при построении словарей главных опорных слов каждого текста, а именно, из словаря были удалены словоформы, встретившиеся только в одном тексте и передающие, следовательно, основное содержание только одного текста.

Затем для оставшихся словоформ по полученной экспериментальным путем формуле подсчитывался их вес, при этом учитывался суммарный вес ключевых слов, количество текстов, в которых встретилось слово, и количество всех текстов. В конечный словарь заносились только те словоформы, которые имели наибольший коэффициент важности и сумма коэффициентов важности которых составляла не больше 50% от суммы коэффициентов важности всех оставшихся словоформ.

Табличный реферат группы текстов получается путем проверки присутствия главных опорных слов в поле "субъект" или "предикат" всех полученных ранее рефератов. Ниже приведен фрагмент табличного реферата группы исследуемых текстов.

Субъект	Предикат	Объект	Обстоятельство места	Обстоятельство времени
Putin	visits	Neutral Austria		
Russian President Vladimir Putin	made	an official visit	to Austria	between 8 and 9 February
Russia	will respect	Austria's decision		
the Austrian President Thomas Klestil	highlighted	the development of economic links between the two countries		
Russia	will be forced to respond to	the legal action		
Mr Putin	said	his visit		
Putin	might use	his visit		
Russia's President Vladimir Putin	will visit	Stockholm		

Рис. 1. Фрагмент табличного реферата группы текстов