УДК 004.822

ПРОГРАММНЫЙ КОМПЛЕКС ДЛЯ ИССЛЕДОВАНИЯ СТРУКТУРЫ ЕСТЕСТВЕННОНАУЧНЫХ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

В. А. Крайванова
ФГБОУ ВПО «Алтайский государственный технический университет им. И.И. Ползунова»,
г. Барнаул

В статье описан программный комплекс, на котором моделируется процесс чтения текста с помощью скользящего окна для исследования зависимости между распределением частей речи по длине текста на естественном языке и содержанием фрагментов этого текста. Для каждого окна вычисляется ряд параметров. Текст представляется в виде кривой в пространстве параметров.

Ключевые слова: смысловая сегментация текста, Text Miming, кластерный анализ, распределение частей речи по тексту.

Большая часть полезной информации глобального, не временного значения представляет собой тексты средней и большой длины. Это описание различных явлений и законов природы и общества, нормативная литература и др. Существует ряд современных задач доступа к информации, для которых данное представление существенно снижает эффективность работы или делает достижение результатов невозможным в заданные сроки. Приведем три наиболее злободневных примера таких задач: использование больших текстов в качестве справочника, создание энциклопедических статей, обзоров, учебных материалов и викификация текстов, оценка качества текста.

Для решения этого широкого класса проблем Text Mining актуальным является задача анализа структуры и первичной сегментации текстов большого размера.

Значительная часть прикладных исследований научных текстов направлена на поиск документов и извлечение метаинформации: перекрестного цитирования, ключевых терминов и т.д. [1]. Форматы представления текстов, такие как TeX, позволяют достаточно эффективно извлекать некоторую семантическую информацию [TeXSem1, TeXSem2]. К сожалению, в большинстве случаев длинные тексты находятся в более простых форматах и при извлечении (например, из pdf) могут потерять даже разделение на разделы. Кроме того, описанные выше задачи принципиально сложнее, чем извлечение метаинформации, и требуют детального анализа структуры текста.

Целью данной работы является анализ зависимости между смысловым значением фрагментов в структуре текста и распределением частей речи, а также разработка алгоритма автоматического выделения таких фрагментов в тексте.

Процесс решения конкретной задачи смыслового анализа сходен с процессом использования текста конкретным человеком. При чтении человек воспринимает не весь тест целиком, а предложение за предложением в контексте уже прочитанного. Если подробный семантический анализ можно сравнить с изучающим чтением [2], то задача первичной семантической сегментации выполняется на уровне просмотрового чтения, цель которого – выбор наиболее интересных фрагментов для дальнейшего применения более сложных алгоритмов извлечения знаний.

Окно $W_{i,j}=\langle p_i,...,p_j\rangle$ - это непрерывная последовательность предложений текста W, где i - номер первого предложения, j - номер последнего предложения, $L=j-i$ - размер окна. В научном тексте предложения не всегда представляют собой фразы на естественном языке. В тексте могут встречаться выражения на искусственных языках, например, математические или химические формулы. Кроме того, предложения имеют различную длину и раз-

личное назначение в тексте. Чтобы сгладить возмущения в характеристиках синтаксического графа, вызванные этими факторами, размер окна L следует выбирать достаточно большим[3]. Объектом анализа является множество всех окон $W_{i,j}$ длины L.

В качестве параметров для анализа выберем следующие показатели распределения частей речи по тексту. Для каждого окна $W_{i,j}$ вычислим следующие шесть показателей.

1. Общее количество существительных P_{Noun}.
2. Количество различных существительных (разнообразие) $P_{DiffNoun}$.
3. Общее количество глаголов P_{Verb}.
4. Количество различных глаголов $P_{DiffVerb}$.
5. Общее количество прилагательных P_{Adj}.
6. Количество различных прилагательных $P_{DiffAdj}$.

Данные параметры для целого текста в числе других позволяют идентифицировать автора[4]. Так как общее количество существительных в текстах на русском языке значительно больше всех остальных параметров, для задачи фрагментации параметры нормируются в промежуток [0,1].

Для проведения исследования разработан программный комплекс, архитектура которого представлена на рисунке 1.

В программном комплексе реализованы следующие функции:

– синтаксический анализ текстов и сохранение результатов в базе данных;

– нанесение ручной разметки предложений-заголовков и предложений-определений;

– статистические подсчеты исследуемых параметров на основе результатов синтаксического анализа для фиксированного размера окна;

– вычисление среднего и дисперсии параметров для различных размеров окон;

– визуализация распределения текстовых окон на основе PCA[5];

– кластеризация текстовых окон на основе метода k-means, и просмотр результатов кластеризации.

Количество существительных в научных текстах существенно превосходит количество других частей речи. Большая часть глаголов в научном стиле функционирует в роли связочных. Это подтверждается проведенными экспериментами и согласуется с полученными ранее результатами для английского языка [6].

Размер окон L существенно влияет на результаты фрагментации. Пусть L_{text} – длина текста в предложениях. Рассмотрим возможные ситуации. $L<50$ - слишком велик шум, вызванный разнообразием предложений в тексте. $50 \le L<100$ - такой размер позволяет выявить небольшие фрагменты, если они содержатся в тексте (например, краткие врезки, отступления). $L < L_{text}/5$ - эти размеры окна позволяют выделить крупные описательные или повествовательные фрагменты в тесте.

$L_{text}/5 \le L$ - проявляются эффекты чрезмерной сглаженности, теряются детали фрагментации. При больших размерах окна текст в пространстве параметров (1)-(6) представляет собой кривую сложной формы (рисунок 2).

Размер окна позволяет регулировать размеры выделяемых фрагментов. Алгоритм фрагментации показывает лучшие результаты, когда размер окна соизмерим с размером выделяемых фрагментов. При таких размерах окна выявляется высокая дисперсия параметров (1) P_{Noun} и (3) P_{Verb}. Таким образом, размер окна следует выбирать в соответствии с поставленной задачей.

Чтобы обнаружить описанные далее закономерности использовалась кластеризация множества окон методом k-means. В качестве начального положения центроидов в пространстве параметров взято положение окон $W_{i,i+L}$ равномерно распределенных по длине текста:$\{W_{i,i+L} | \ i=[(0.5+a)C_w/k],\ a=0,1,\ldots,k-1\}$, где L – размер окон, C_w – количество окон, k – количество кластеров.

Исследование динамики параметров (1)-(6) на протяжении текста и взаимосвязи этой динамики со структурой текста с помощью кластеризации множества окон методом k-means показало возможность разделения статических и динамических фрагментов в научном тексте. Конкретные зависимости между параметрами (1)-(6) существенно зависят от конкретного текста. Кроме того, характеристики различных одного типа фрагментов могут существенно различаться, и, если количество кластеров позволяет, они оказываются в различных кластерах. Как правило, алгоритм кластеризации выделяет еще два кластера:

– разрушенный текст (фрагменты из очень коротких предложений; как правило, это подписи к рисункам, таблицы или многочисленные вставки на формальных языках);

– выбросы на границах фрагментов (при переходе из одного фрагмента в другой происходят существенные изменения в характеристиках (1)-(4)).

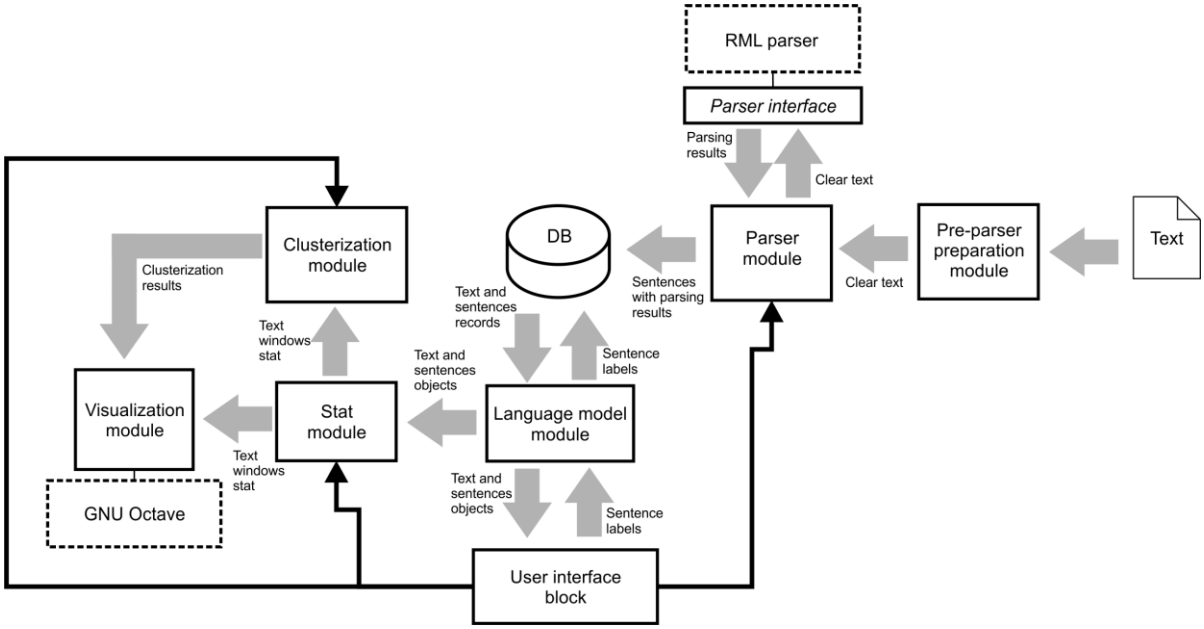

Рисунок 1 – Архитектура программного комплекса.

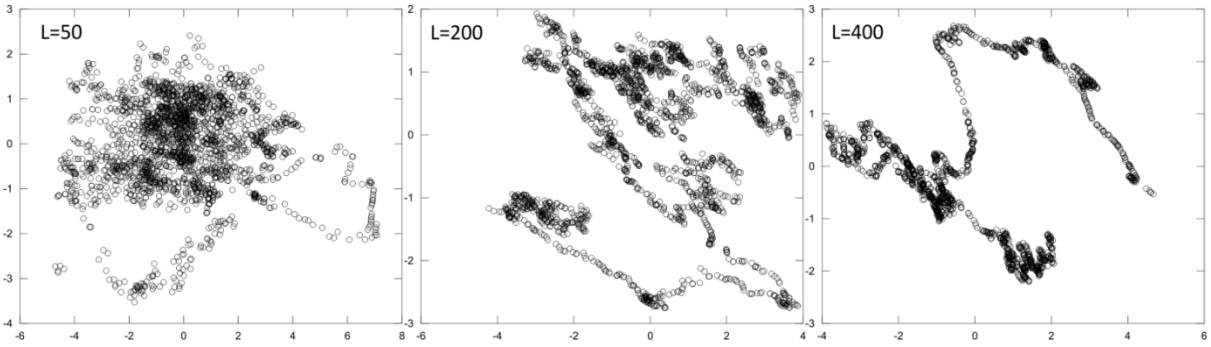

Рисунок 2 – PCA-визуализация текста [7] в пространстве параметров (1)-(6) для различных размеров окна.

Кластеризация не позволяет получить четких границ фрагментов. Это следствие того, что в скользящем окне должно накопиться свойство кластера. Для уточнения границ необходима комбинация результатов кластеризации для большого и маленького размеров окна L.

Подводя итог вышесказанному, можно сделать следующие выводы. Текст на естественном языке представляет собой уникальную фигуру в пространстве параметров (1)-(6). Распределение параметров (1)-(6) по длине текста позволяет разбивать текст на фрагменты путем кластеризации окон.

В дальнейшем планируется продолжить работу в следующих направлениях. Исследование влияния каждого отдельного параметра на результат фрагментации и подтверждение результатов на объемной выборке. Разработка алгоритма уточнения границы фрагментов. Распространение предложенных механизмов анализа на другие статистические параметры текстов, таких как распределение видов синтаксических связей. Применение описания текстов на основе распределения структурных паттернов для сравнительного анализа текста. Разработка алгоритмов иллюстрирования текстов графовыми моделями на основе комбинированных шаблонов.

СПИСОК ЛИТЕРАТУРЫ

1. The intelligent search engine “Exactus”, available at: http://expert.exactus.ru/

2. Кортава Т. В.. Материалы по курсу "Русский язык и культура речи". Лекция 3. Стили речи [Электронный ресурс]. Режим доступа: http://www.geogr.msu.ru/student/uch_materials/kortava_lek3_stili_rechy.doc

3. Krayvanova V., Kryuchkova E.(2013), Application of automatic fragmentation for the semantic comparision of texts, In 15th International conference SPECOM 2013 Proceedings, September 1-5, Pilsen, Czech Republic.

4. Львов А. Лингвистический анализ текста и распознавание автора [Электронный ресурс], 2008. Режим доступа: http://fantlab.ru/article374.

5. Shlens J. A tutorial on Principal Components Analysis, 2009 available at: http://www.snl.salk.edu/~shlens/pca.pdf.

6. Хомутова Т. Н. Научный текст: интегральный анализ лексики, Язык и культура. 2010. №4. [Электронный ресурс], Режим доступа:: http://cyberleninka.ru/article/n/nauchnyy-tekst-integralnyy-analiz-leksiki

7. Котова Д.Л., Девятова Т.А., Крысанова Т.А., Бабенко Н.К., Крысанов В.А., Методы контроля качества почвы: Учебно-методическое пособие [Электронный ресурс] 2007, режим доступа: http://window.edu.ru/library/pdf2txt/575/59575/29643

Крайванова Варвара Андреевна – к.ф.-м.н., тел.: +7-913-230-3419, e-mail: krayvanova@yandex.ru.