УДК 519.246.8:519.766

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ТЕКСТОВ, НАПИСАННЫХ НА ЕСТЕСТВЕННЫХ ЯЗЫКАХ

Доценко С.В., Иванов А.А., Шишкевич В.Е.
Севастопольский национальный технический университет
Студгородок, г. Севастополь, Украина, 99053

Рассматривается возможность применения методов статистической обработки случайных стационарных числовых последовательностей для исследования текстов, написанных на естественных языках. Дан один из возможных методов преобразования текста в случайный процесс. Приводится пример обработки по данной методике текстов, написанных на английском и русском языках.

Достаточно длинные тексты, написанные на естественных языках, представляют собой случайные стационарные последовательности символов, которые являются буквами, пробелами, цифрами и знаками препинания, в основном не числами. В свою очередь, случайный процесс – это последовательность случайных чисел. Поэтому первая задача, которую следует решить при статистическом анализе текста – преобразовать символы в числа, и тем самым текст – в случайный процесс. Одно из важнейших требований к такому преобразованию – обеспечение его объективности. Это означает, что принципы числового преобразования не должны зависеть ни от вида языка, на котором написан текст, ни от того, кто конкретно производит это преобразование.

Перед статистическим анализом текст должен быть подвергнут предварительной обработке. Для удобства набора текста на компьютере применяются текстовые редакторы или процессоры, в которых

представление информации на экране аналогично представлению текстовой информации на бумаге. Обычно максимальная длина строки при внесении текста в редактор составляет не более 128 символов. Для удобства пользователя текст на экране представляется по ширине монитора, что составляет $N = 60…70$ символов, а это значительно меньше максимальной длины строки. Если текстовый редактор не отмечает переход символов на следующую строку (как правило, он вставляет символы «конец строки» или «возврат каретки»), то после окончания текстовой строки длиной N символов оставшиеся $128\text{-}N$ символов в строке дополняются пробелами, что недопустимо при статистическом анализе текстов.

При любой конечной длине строки часто возникает необходимость разбить слово и перенести его часть в начало следующей строки, так как оно не помещается целиком в данной. В этом случае перенесённое слово нарушает статистические связи в тексте, так как знак переноса нарушает взаимокорреляцию символов в данном слове и в тексте в целом. Кроме того, при компьютерной обработке знаки тире и переноса имеют один и тот же ASCII-код, поэтому количество тире в тексте увеличивается на количество переносов.

Далее, большинство текстовых редакторов, поддерживающих выравнивание текста по ширине, добавляют в строку лишние пробелы между словами, что приводит к изменению статистических свойств данного текста.

Следовательно, чтобы исключить влияние текстового редактора, в котором набирался текст, на статистические свойства исследуемого текста, необходимо удалить из данного текста следующие символы:

служебные (конец строки, перевод каретки и т.д.); переноса слов;

двойных пробелов.

Приписывание числовой меры конкретному символу естественного языка позволяет превратить любой текст в числовую последовательность, которая представляет собой случайный процесс. На первый взгляд кажется, что данная задача решена путём кодирования символов стандартной кодовой таблицей КОИ-8 или ASCII. Однако результаты анализа показывают, что такой подход не удовлетворяет указанному требованию[1]. Приведём простой пример: в английской раскладке клавиатуры ASCII-код для символа «О» равен 24, а для таблицы

русифицированных ASCII-кодов код 24 является кодом символа «Щ». Однако буква «О» в английском языке встречается значительно чаще, чем буква «Щ» в русском. Отсюда следует, что статистические свойства исследуемого текста, преобразованного в числовую последовательность, будут зависеть от кодовой таблицы символов.

В любом языке имеются некоторые устойчивые числовые статистические характеристики, которые практически не изменяются при переходе от одного осмысленного текста к другому. В частности, такой характеристикой является распределение вероятностей, т.е. пределов частот появления символов языка при бесконечном увеличении объёма текста. Экспериментально получено, что при объёме текста в 30 и более тысяч знаков частоту появления символа можно считать его вероятностью [2]. Статистические свойства языка, на котором написан данный текст, наиболее полно могут быть выявлены специальными видами кодирования его символов, одним из которых является метод, рассмотренный ниже. Он заключается в следующем: каждому символу сопоставляется значение вероятности его появления в тексте, написанном на каком-либо естественном языке. Например, при замене текста, написанного на английском языке, числовой последовательностью символ «О» заменяется значением вероятности его появления $p = 0{,}055$. Аналогично символ «О» в русскоязычном тексте заменяется на $p = 0{,}081$.

Рассмотрим пример автокорреляционных функций (АКФ) и спектральных плотностей обработанных таким образом художественных текстов на русском и английском языках, представленных в виде случайных процессов (рисунок 1). Каждый из этих текстов содержал более 50 тысяч

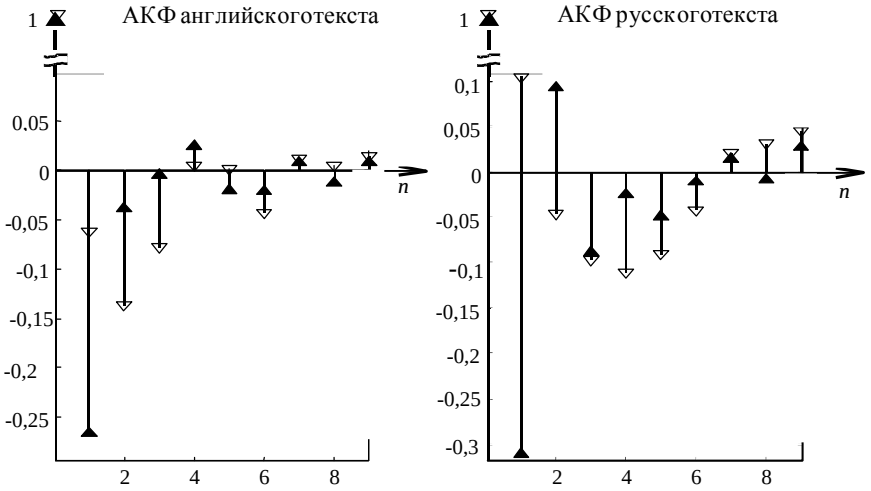

Рисунок 1 – Автокорреляционные функции произведений Набокова на английском и русском языках: ▲ - кодировка по вероятности символов; ▽ - кодировка ASCII

символов.

На рисунке 1 по оси абсцисс отложено число интервалов между буквами текста, а по оси ординат – коэффициент корреляции между числовыми кодами букв. Как видно из рисунка 1, корреляционная функция быстро убывает с ростом сдвига интервала между символами. Корреляция достаточно велика только для первого десятка символов. Далее значения корреляции быстро убывают и стремятся к нулю. Следовательно, любой текст как английского, так и русского языка при больших сдвигах слабо коррелирован.

Между видом корреляционной функции и временной структурой соответствующегоей случайного процесса существуетопределённая связь. В зависимости от того, составляющие каких частот и в каких пропорциях содержатся в случайном процессе, его корреляционная функция имеет тот или иной вид и, в частности, меняется её интервал корреляции. Следовательно, можно говорить о спектральном составе случайной функции. Спектральный анализ позволяет выявить наиболее характерные периоды в последовательностисимволов текста, определяемые внутренней структурой языка. На рисунке 2 изображены нормированные спектры

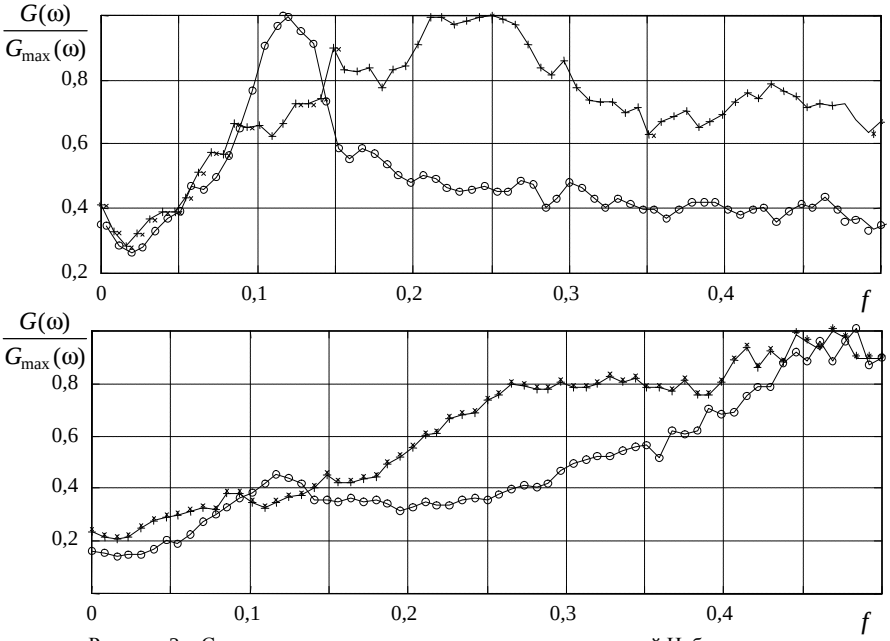

Рисунок 2 – Спектральные плотности мощности произведений Набокова:
* - английский язык; ° - русский язык

текстов произведений Набокова на русском и английском языках («Подлец» и «La Veneziana», каждое произведение содержит более 50 тысяч символов), использующих разную систему кодировки символов (вверху – ASCII-коды, внизу кодировка символов по их вероятности появления в тексте).

На рисунках 3 и 4 показаны спектральные плотности мощности различных произведений Набокова, написанных на русском и английском языках соответственно, а также средние значения спектров по данным произведениям. Как видно из графиков, спектральные плотности текстов как на русском, так и на английском языках, существенно зависят от вида кодировки символов. При статистической обработке тех же текстов, использующих другую (не ASCII и не по вероятности) кодировку символов алфавита, получатся спектральные зависимости, отличные от приведенных выше. Несмотря на то, что художественные произведения написаны одним и тем же автором, но из-за того, что использовались языки различной лингвистической группы, имеются различия в виде спектральных кривых.

Таким образом, можно сделать следующие выводы.

1. Возможно объективное преобразование текстов, написанных на естественных языках в числовые последовательности

2. Существуют устойчивые информационно-статистические характеристики данных таких числовых последовательностей, которые зависят от того, на каком языке был написан текст.

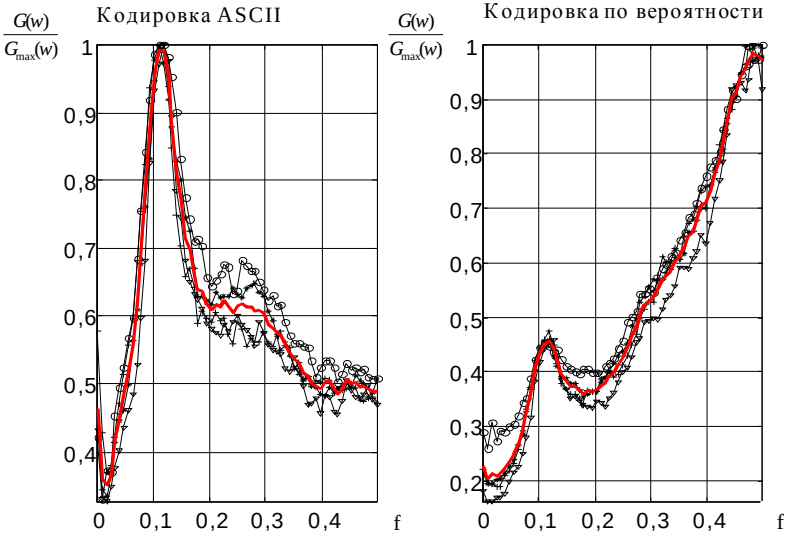

Рисунок 3 – Спектральные плотности мощности произведений Набокова:

— – среднее значение спектральной плотности мощности; -v - Владимир Набоков. «Машенька»: 164040 символов; -* - Владимир Набоков. «Король, дама валет»: 363243 символов; -o - Владимир Набоков. «Защита Лужина»: 340344 символов; -+ - Владимир Набоков. «Камера обскура»: 284714 символов.

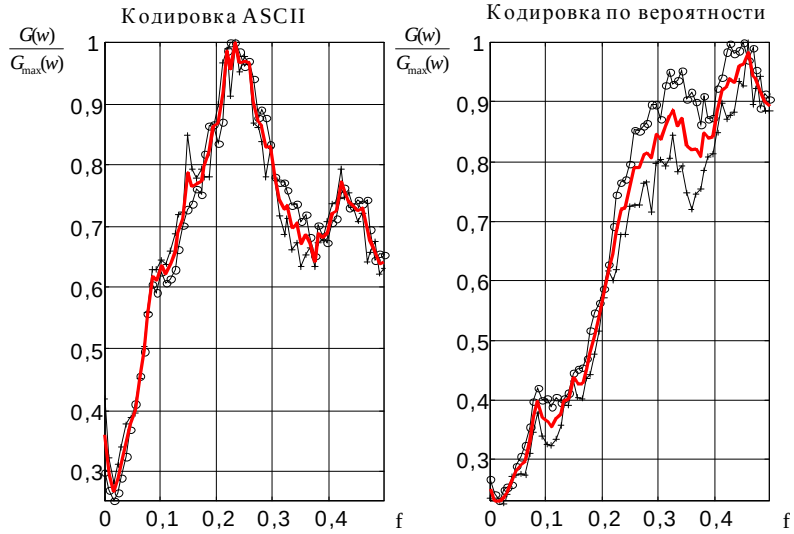

Рисунок 4 – Спектральные плотности мощности произведений Набокова:
— - среднее значение спектральной плотности мощности ; -o - Владимир Набоков. «Transparent things»: 158738 символов; -+ - Владимир Набоков. «La Veneziana»: 62866 символов.

Библиографический список

1. Либерти Д. Освой самостоятельно C++ за 21 день / Д. Либерти. — М.: Изд-во “Вильямс”, 2000. — 815 с.

2. Яглом А.М. Вероятность и информация / А.М. Яглом, И.М. Яглом. — М.: Наука, 1973. — 512 с.

Поступила в редакцию 23.04.2001 г.