ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 22 МАТЕМАТИКИ Выпуск 4
2015

Н. В. Вильбоа, **В. О. Мироншин** (Москва, МИЭМ НИУ ВШЭ, ТВП). **Исследование информационных характеристик естественных языков.**

Исследование вероятностных и статистических свойств языка имеет важное значение для различных теоретических и практических применений, в частности, теории кодирования и теории информации, лингвистике и экономике. Знание свойств информационных характеристик текста, таких как энтропия, распределение m-грамм, слогов и словосочетаний, позволяет эффективнее реализовывать различные алгоритмы перебора, а также оценивать количество осмысленных текстов. Понятия количества информации и энтропии позволяют определять характеристики процесса передачи информации, в том числе по каналу связи при наличии ошибок. Важное применение статистические характеристики языка находят в лингвистике. Так, имея хронологическую статистику значений информационных характеристик текстов для исследуемого языка в различные периоды времени, можно описать динамику развития языка [1], а также строить прогнозы.

Следует также отметить, что информационные характеристики используются не только в фундаментных научных исследованиях, но и в прикладных областях. Например, в работах [10], [11] изучаются проблемы информационной энтропии, приводятся результаты исследований информативности публикуемых статей в СМИ с учетом знаний законов лингвистики и теории информации.

Такое широкое применение энтропии, собственной информации и избыточности требует построение эффективных методов вычисления этих характеристик, а также повышения точности соответствующих вычислений.

В работе проведён обзор имеющихся на сегодняшний день решений, представлено сравнение точности расчетов энтропии на основе указанных подходов, а также сформулированы дальнейшие проблемы.

Основные понятия и определения

Опишем ряд математических объектов, исследуемых в настоящей работе. Следующие определения введены в работах [1, 2].

О п р е д е л е н и е. Лингвистические единицы — это буквы, морфемы, звуки и т. д., передаваемые по каналу связи (речь, сообщение, составленное из символов некоторого языка (кода), и т. д.).

З а м е ч а н и е. В языке (коде) имеют место ограничения не только на вероятность появления лингвистических единиц в сообщениях, но и на их сочетаемость.

В [1] установлено, что для определения ограничений, регулирующих появление лингвистических единиц в сообщениях, требуется учитывать их соотношение с такими величинами, как количество информации, энтропии и избыточности. В работе [2] введены соответствующие характеристики.

В силу того, что в области синтаксиса и лексики число элементарных единиц не ограничено, а в области фонетики трудно выделить и разграничить дискретные едини-

цы, ограничим дальнейшее изложение письменной речью, где буквенный код включает конечное число дискретных единиц.

О п р е д е л е н и е. Пусть $p_1, \dots, p_N$ — распределение вероятностей букв алфавита мощности N. Тогда величина $H = -\sum_{i=1}^{N} p_i \log p_i$ называется энтропией этого распределения, где N — мощность алфавита.

Обзор имеющихся решений

1. Модель марковского источника сообщений.

О п р е д е л е н и е. Дискретный стационарный источник называется марковским источником порядка m, если для $\forall l \geqslant m$ и $\forall c_l = (a_{i_1}, \dots, a_{i_l})$ выполняется равенство $p(a_{i_l}/a_{i_{l-1}}, \dots, a_{i_1}) = p(a_{i_l}/a_{i_{l-1}}, \dots, a_{i_{l-m+1}})$, $a_{i_j} \in A$, где A — алфавит источника.

О п р е д е л е н и е. Величина $H_\infty = -\lim_{k\to\infty} \sum_{c_k} p(a_{i_1}, \dots, a_{i_k}) \log p(a_{i_k}/a_{i_{k-1}}, \dots, a_{i_1})$ называется энтропией марковского источника.

2. Метод Шеннона. Метод К. Шеннона по оценке энтропии и избыточности, описанный в работе [1], основан на изучении возможности предсказания исследуемого текста, а именно, точности предсказания очередной буквы по известным буквам текста.

Для данного языка вычисляются вероятности $p_{i_1,\dots,i_n}(j)$ того, что после n — граммы $(i_1, \dots, i_n)$ будет стоять буква j. Рассматривается преобразование множества букв алфавита мощности N во множество чисел от 1 до N, при котором в 1 отображается буква алфавита с наибольшей частотой встречаемости (т. е. с наибольшей вероятностью), в 2 — буква со второй по величине частотой и т. д.

$q_1^N = \sum_{(i_1,\dots,i_n)} p\{i_1, \dots, i_n, 1\}$ — вероятность появления наиболее частой буквы,

$q_2^N = \sum_{(} i_1, \dots, i_n) p\{i_1, \dots, i_n, 2\}$ — вероятность появления второй по частоте буквы, и т. д.

Тогда условная энтропия H_n при наличии n-граммы может быть оценена следующим образом:

$$\sum_{i=1}^{N} i(q_i^n - q_i^{n+1}) \log q_i^n \leqslant H_n \leqslant \sum_{i=1}^{N} q_i^n \log q_i^n.$$

3. Метод Пиотровского. Метод подсчета энтропии Р. Г. Пиотровского, основанный на условных вероятностях, описан в работе [4].

Пусть $p_1, \dots, p_N$ —распределение вероятностей букв алфавита.

Пронумеруем все цепочки символов $l_1, \dots, l_{n-1}$. Обозначим через b_i^{n-1} событие, заключающееся в появлении i-й цепочки длины $n-1$.

Обозначим через $p(\frac{j_{i,k}}{b_i^{n-1}})$ условную вероятность того, что l_n примет значение j_k при условии b_i^{n-1}. То есть вероятность того, что в следующей после b_i^{n-1} позиции стоит именно j_k.

Тогда средняя условная энтропия для l_n имеет вид:

$$H_n = -\sum_{b_i^{n-1}} p(b_i^{n-1}) \sum_{k=1}^{s} p\Big(\frac{j_{i,k}}{b_i^{n-1}}\Big) \log p\Big(\frac{j_{i,k}}{b_i^{n-1}}\Big).$$

4. Метод Аматова. Метод А. М. Аматова основан на определении информационной энтропии по Шеннону (см. [5]). В методе используется понятие «показателя уровня неопределенности языкового знака». Этот показатель равен отношению суммы планов содержания (число объектов $C_i^{(k)}$, описываемых одним словом x_k) к сумме планов выражения (число различных значений $F_j^{(k)}$ рассматриваемого слова x_k), зафиксированных в языке на тот или иной момент времени. Для некоторой подсистемы

языка из n элементов, имеющих m, значений имеет место равенство:

$$U_k = \frac{\sum_{i=1}^{m} C_i^{(k)}}{\sum_{j=1}^{n} F_j^{(k)}}, \quad k = \overline{1, N}, \quad H = -\sum_{k=1}^{N} U_i \log U_i.$$

5. Метод Олешкова. Метод подсчета энтропии текста М. Ю. Олешкова [9] основан на методах лингвистики. Определяются следующие параметры:

— объем текста V на пропозициональном уровне (равен количеству пропозиций, где пропозиция может быть определена, например, как самая маленькая единица знания);

— количество точек бифуркации N, в случае возникновения флуктуаций определяемое либо значениями n_1 (когда текст уходит с основной магистрали), либо n_2 (когда текст возвращается на нее);

— количество «коммуникативных дефектов» m и количество «коммуникативных дефектов», вызывающих флуктуацию p.

В итоге, формула расчета энтропии текста имеет вид:

$$H = \frac{m + N}{V} + N\frac{n_2 - n_1 + p}{V}.$$

В ы в о д ы. Проведенные экспериментальные расчеты показали, что среди представленных методов наиболее точные оценки энтропии удается получить методом, основанным на модели марковского источника. Поэтому дальнейшие исследования будут направлены на построение модификаций именно этого подхода.

СПИСОК ЛИТЕРАТУРЫ

1. *Некипелова И. М.* Организация, самоорганизация и дезорганизация языковой системы: Механизмы оптимизации языка и речи. Серия Философия. М.: ИКИ, 2014.
2. *Шеннон К.* Сборник статей «Теория информации», М.: ИЛ, 1963.
3. *Савчук А. П.* Об оценках энтропии языка по Шеннону. — Теория вероятн. и ее примен., 1964, т. 9, в. 1, с. 154–157.
4. *Пиотровский Р. Г.* Информационные измерения печатного текста. — В сб.: «Энтропия текста и статистика речи». Л.: Наука, 1968.
5. *Аматов А. М.* Информационная энтропия как фактор конвергенции синтаксических структур в языках разных типов (на примере русского и английского языков). — Вестник Волгоградского гос. ун-та. Серия 2, Языкознание», 2008, в. 2.
6. *Кокин А. С., Чепьюк О. Р.* Уравнение стоимости бизнеса: энтропия, как мера стоимости. — Вестник Нижегородского ун-та им. Н. И. Лобачевского, 2007, т. 6, с. 201–207.
7. *Прангишвили И. В.* Энтропийные и другие системные закономерности: Вопросы управления сложными системами, Ин-т проблем управления им. В. А. Трапезникова. М.: Наука, 2003.
8. *Хорошилов А. А.* Теоретические основы и методы построения систем фразеологического машинного перевода. Дисс. на соискание уч. ст. д.т.н., М.: 2006.
9. *Олешков М. Ю.* Энтропия текста: опыт дискурсивного анализа. — Вестник Пятигорского гос. лингвистического ун-та, 2005, т. 3, в. 4.
10. *Кузнецова А. В.* Информация в СМИ и энтропия: лингвистический аспект. — Изв. ЮФУ, «Филологические науки», 2008, № 2, с. 104–113.
11. *Кузнецова А. В.* Проблемы информации и энтропии в медиатексте. Дисс. на соискание уч. ст. канд. филологических наук, ФГАОУ ВПО ЮФУ, 2012.