УДК 621.391

В.С. Чернега, доцент, канд. техн. наук
Севастопольский национальный технический университет
ул. Университетская 33, г. Севастополь, Украина, 99053
E-mail: vs_chernega@mail.ru

СЖАТИЕ ТЕКСТОВЫХ СООБЩЕНИЙ НА ОСНОВЕ АДАПТИВНОГО КОНТЕКСТНО-МОРФОЛОГИЧЕСКОГО МОДЕЛИРОВАНИЯ

Описан способ повышения эффективности сжатия текстовых сообщений на основе контекстного моделирования, особенностью которого является то, что при оценке вероятности ожидаемого символа учитывается не только предыдущие символы, но и морфологическая часть языка, которой принадлежит контекст.

Ключевые слова: *сжатие информации, текстовые сообщения, контекст, моделирование.*

При динамическом сжатии текстовых сообщений необходимы сведения о вероятностях появления текущих символов в сжимаемом потоке. Для получения этих вероятностей нужно построить модель информационного источника, генерирующего сообщение, которое подлежит сжатию. Одним из распространенных и эффективных способов получения вероятностей текущих символов является контекстное моделирование, основанное на процедуре предсказания вероятности символа на основе предыдущих символов, называемых контекстом [1, 2]. Повышение точности предсказания вероятности символов приводит к повышению степени сжатия текстового сообщения. Недостатком известных способов контекстного моделирования является то, что они не учитывают морфологические свойства сжимаемого текста. Вероятность предсказания очередного символа текстового сообщения можно повысить за счет дополнительной информации о контексте, в частности учета его грамматических свойств.

Целью статьи является описание способа контекстно-морфологического моделирования и оценка вероятности появления очередного символа, определяемой путем контекстно-морфологического моделирования, т.е. с учетом морфологической части речи, которой принадлежит анализируемый контекст.

Основной морфологической единицей текстового сообщения, как известно, является слово. Морфема есть наименьшая значащая часть слова, которую можно выделить в его структуре с помощью сопоставления с другими словами [3]. К морфемам в русском и украинском языках относятся префикс (приставка), корень, суффикс, окончание. Морфема обозначает определенное лексическое или грамматическое значение и выполняет важнейшую функцию при словообразовании. Для украинского и русского языков существуют следующие морфологические способы словообразования: суффиксный, префиксный, префиксно-суффиксный, безафиксный, сложение основ и слов.

С целью определении возможности использования морфологических свойств языка при контекстном моделировании для оценки вероятности появления очередного символа были проведены статистические исследования морфологического состава текстов различных стилей на русском и украинском языках. В результате этих исследований было установлено следующее [4].

1) Наиболее часто слова украинского языка начинаются с приставок «*в-*», «*з-*», «*с-*», после них в порядке убывания вероятностей располагаются приставки «*о-*», «*на-*», «*по-*». Для русского языка наиболее частыми приставками в предложениях являются «*по-*», «*не-*», «*на-*» и «*за-*», «*от-*», «*при*». Следует отметить, что вероятности первых трех приставок в два раза превышают значения вероятностей следующих за ними морфем.

2) Частота встречаемости приставок в текстах различных стилей заметно отличается. Так в научной и художественной украинской литературе вероятность появления приставки «*с-*» составляет 12—13 %, а для официально-делового стиля — 8 %.

3) Имеется существенное различие вероятности появления в текстах украинского языка приставки «*пре-*» для художественного, научного и официально-делового стилей; вероятность появления приставки в первом случае составляет 0,07 %, а для двух других — 0,5 %.

4) Наименее вероятными украинскими приставками являются «*понад-*» и «*анти-*», их суммарная вероятность составляет 0,02 %, русскими – «*сверх-*», «*анти-*».

5) Некоторые украинские приставки встречаются примерно с одинаковой вероятностью для всех стилей: «*з-*», «*за-*», «*під-*», «*ні-*», «*пере-*». Для текстов одного стиля, но различных тематик приставки используются практически с одинаковой частотой.

6) Наиболее часто для образования украинских слов используются суффиксы «*-в-*», «*-ість-*», «*-н-*», «*-ок-*», «*-ов-*», «*-ер-*», «*-ова-*», «*-ик-*», «*-ин-*», «*-ник-*», их суммарная вероятность составляет 87 %. В русском языке такой группой суффиксов являются «*-ость-*», «*-ность-*», «*-ик-*», «*-ова-*»,«*-ник-*», «*-тель-*», «*-ств-*», суммарная вероятность которых равна 85 %.

В процессе исследований также установлено, что вероятности появления ряда символов в определенном контексте зависят от принадлежности анализируемой буквенной комбинации к определенной морфеме (префиксу, суффиксу или окончанию слова). Так, например, для текстов украинского языка количество появлений буквы "з" после буквосочетания (контекста) "бе" в 7,5 раз выше в начале слова, чем в другой его части, а появление буквы "о" после "п" в 3,9 раза чаще в начале слова, чем в конце и т.п. Аналогичные свойства присущи и другим языкам, однако комбинации букв в них, естественно, другие. Учитывая это явление можно дополнительно повысить точность оценки вероятности появления символов входного сообщения по контексту, если ввести процедуру распознавания морфологической конструкции, к которой относится текущий контекст и осуществлять коррекцию оценки итоговой вероятности на величину коэффициента аффиксальных морфем или морфологических частей речи.

Еще более ярко выражена статистико-грамматическая связь в проблемно-ориентированных текстах: данных гидрометеорологических и гидрохимических наблюдений, банковских документах, технических описаниях машин и конструкций и др. Такая связь позволяет при сжатии этого вида сообщений предсказывать не только отдельные символы, но и их сочетания, а в ряде случае и группы слов.

Предлагаемый способ архивации относится к области сжатия текстовых сообщений, представленных на национальных языках и предназначен для использования в компьютерных системах и сетях, WEB-узлах, архиваторах, специальных хранилищах данных.

В известных способах сжатия на основе контекстного моделирования текстовых сообщений [1] для повышения точности оценка вероятности текущих символов осуществляется в зависимости от предыдущих символов, т.е. с учетом контекста. Причем, для определения вероятности текущего символа используют не один, а несколько предыдущих символов (на практике обычно не более четырех). Такая процедура получила название контекстно-ограниченного моделирования. Если вероятность текущего символа оценивается с учетом O предыдущих символов, то такая модель называется моделью O-го порядка. В лучшем, на настоящее время, способе сжатия на основе контекстного моделирования [2], используется контекстно-смешанная оценка вероятности текущего символа Φ алфавита A. Этот способ сжатия текстовых сообщений на базе контекстного моделирования основан на подсчете на каждом шаге сжатия количества появления символов, связанных с текущими контекстами различных порядков, вычислении суммарных оценок вероятностей появления символов в соответствующих контекстах, последующем умножении этих оценок на весовые коэффициенты соответствующих контекстов и присвоения символам текста кодовых комбинаций в соответствии с их вероятностями.

В соответствии с этим способом вычисляются вероятности появления символа Φ при условии предыдущих O символов $P(O,\Phi)$. Каждой вероятности контекста O присваивается весовой коэффициент $w(O) <.1$. Итоговая оценка вероятности символа Φ определяется как взвешенная сумма вероятностей по всем m используемым контекстам:

$$p(\Phi) = \sum_{O=-1}^{m} w(O)\, p(O,\Phi)\,, \qquad (1)$$

где $w(O)$ — вес модели O-го порядка, причем $\sum_{O=-1}^{m} w(O) = 1$.

Вес $O = -1$ означает, что всем символам используемого алфавита присваиваются одинаковые вероятности.

Недостатком этого способа является неполное использование информации о контексте сообщения и, как следствие, не достаточно высокая степень сжатия текстовых сообщений. В частности, в нем не используется тот факт, что вероятности появления отдельных символов зависят не только от того, какими являются предыдущие символы, а и от морфемной и морфологической формы языка (префиксов, суффиксов, союзов и т.д.), в которой встречается символ с данным контекстом.

Для устранения отмеченного недостатка в устройство сжатия заносятся таблицы с аффиксальными морфемами (префиксы, суффиксы, окончания), и наиболее часто используемыми морфологическими частями речи (союзы и частицы) вместе с их весовыми коэффициентами для украинского, русского, английского и др. языков, а также усредненная таблица весовых коэффициентов, аналогичная прототипу. Кроме этого, в устройство сжатия вводится блок определения вида языка и стиля сообщения, а также блок анализа, позволяющий различить префикс, суффикс или другую морфему от аналогичного буквосочетания, не являющегося морфемой.

На начальном этапе сжатия текстового сообщения оценки вероятностей символов вычисляются на основе контекстов, аналогично прототипу, с использованием усредненной таблицы весовых коэффициентов. Одновременно, на основе полученных вероятностей символов и выявления морфологических конструкций и аффиксальных морфем, характерных только определенному языку,

определяется принадлежность текстового сообщения конкретному языку (украинскому, русскому, английскому). Морфологические конструкции (союзы и частицы) и афиксальные морфемы (префиксы, суффиксы, окончания) находят путем выделения из входного потока групп символов (не более 5-ти в группе) ограниченных слева, либо справа или с обеих сторон знаками пробела и сравнения их с аналогичными грамматическими конструкциями, размещенными в таблицах типовых морфемных конструкций для каждого из языков, которые используются в компрессоре. Например, обнаружение союзов *і, або, чи, але,* часток *ось, це, лише* и т.п., однозначно свидетельствует, что входное текстовое сообщение относится к украиноязычному.

Затем, при анализе контекстов, проверяется принадлежность буквосочетаний контекста к одной из морфологических конструкций языка, к которому отнесено сжимаемое текстовое сообщение и вычисляется корректирующий компонент оценки вероятности текущего символа. То есть, оценка вероятности символа $Ф$ осуществляется по следующей формуле:

$$p(Ф) = \sum_{O=-1}^{m} w(O)\, p(O, Ф) + w(M_i, L)\, p(M_i, L), \qquad (2)$$

где $w(M_i,L)$ — весовой коэффициент i–й морфологической части речи L-го языка; $p(M_i,L)$ — вероятность принадлежности текущего контекста i-й морфологической части L-го языка. При этом таблица весов $w(O)$ изменяется таким образом, чтобы выполнялось условие

$$\sum_{O=-1}^{m} w(O) + w(M_i, L) = 1 \qquad \text{для всех } i = 1, 2, \ldots, N_m, \qquad (3)$$

где N_m — количество морфологических частей, которые используются в конкретном устройстве сжатия.

Как видно из выражения (2), если текущий контекст не соотносится ни с одной из присущих языку морфологических частей, т.е. при $p(M_i,L) = 0$, то оценка вероятности текущего символа осуществляется традиционным способом. Если же с вероятностью $p(M_i,L)$ определена принадлежность текущего контекста к одной из морфологических частей речи, то вероятность оценки символа увеличивается на величину произведения $w(M_i,L)$ на $p(M_i,L)$, что позволяет уменьшить число битов, затрачиваемых на кодирования данного символа.

В процессе грамматического анализа входного текста осуществляется динамическая процедура присвоения кодовой комбинации текущему кодируемому символу, в зависимости от оценки его вероятности на данном шаге кодирования. То есть, происходит адаптация процедуры сжатия к статистическим и грамматическим свойствам входного сообщения.

Экспериментально установлено, что за счет более точной оценки вероятности текущего символа сжимаемого сообщения и кодирования более вероятного символа меньшим количеством бит сжатое сообщение занимает меньший объем, чем у прототипа в среднем на 13 процентов.

Дальнейшие исследования направлены на экспериментальное определение весовых коэффициентов $w(O)$ и $w(M_i,L)$, оценку снижения точности прогнозирования вероятности очередного символа, а также на решение задачи ускорения процедуры поиска и идентификации в сжимаемом сообщении морфем.

Библиографический список

1. Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео / Д. Ватолин, А. Ратушняк, М. Смирнов, В. Юкин. — М.: ДИАЛОГ-МИФИ, 2002. — 202 с.

2. Bell T. Modeling for Text Compression / T. Bell, I. Witten, J.Cleary // ACM Computing Surveys. — 1989. — V. 21. — No. 4 (Dec.). — P. 557–591.

3. Росінська О.А. Граматика української мови для учнів, абітурієнтів і студентів / О.А. Росінська. — Донецьк: ТОВ ВКФ “БАО”, 2004. — 288 с.

4. Исследование и разработка методов адаптивного сжатия информации для компьютерных сетей: отчет о НИР (заключит.) / Севастоп. нац. техн. ун-т; рук. Чернега В.С.; исполн.: Доронина Ю.В. [и др.]. — Севастополь, 2006. — 185 с. — № ГР 105U002003.

Поступила в редакцию 24.03.2009 г.