УДК 004.912

Фадеев С.Г.,

старший преподаватель кафедры компьютерных технологий

Чувашский государственный университет им. И.Н.Ульянова

Россия, г. Чебоксары

ОПТИМИЗАЦИЯ МАТЕМАТИЧЕСКОЙ МОДЕЛИ ЕСТЕСТВЕННОГО ЯЗЫКА ДЛЯ МОРФЕМНОГО АНАЛИЗА

Аннотация:

В статье рассматривается оптимизация ранее предложенной расширенной математической модели для морфемного анализа словоформ естественного языка, создаваемой на основе статистических данных. Оптимизация опирается на то, что вероятность встречи морфа или комбинации морфов в естественном языке может зависеть от места этого морфа в соответствующей морфемной группе. Оптимизация позволяет уменьшить количество элементов в матрицах вероятностей за счет разбиения матриц и последующего исключения из них нулевых строк. Рассмотрены достоинства и недостатки предложенной оптимизации.

Ключевые слова: морфемный анализ, словоформа, морф, математическая модель, оптимизация.

Fadeev S.G.,

senior lecturer of the Department of computer technologies

Chuvash State University named after I.N. Ulyanov

Russia, Cheboksary

OPTIMIZATION OF THE MATHEMATICAL MODEL OF NATURAL LANGUAGE FOR MORPHEMIC ANALYSIS

Annotation:

The article considers the optimization of the previously proposed extended mathematical model for the morphemic analysis of word forms of natural language, created on the basis of statistical data. Optimization is based on the fact

that the probability of encountering a morph or a combination of morphs in a natural language may depend on the location of this morph in the corresponding morpheme group. Optimization makes it possible to reduce the number of elements in probability matrices by splitting the matrices and then eliminating the zero rows from them. The advantages and disadvantages of the proposed optimization are considered.

Keywords: morphemic analysis, word form, morph, mathematical model, optimization

В статье [2] предложена расширенная математическая модель естественного языка для морфемного разбора на основе статистических данных. Данную модель можно оптимизировать за счет уменьшения размерностей используемых матриц и тем самым ускорить морфемный разбор словоформ.

Расширенная модель делит словоформу на 3 морфемных группы: префиксную, постфиксную и группу корней [3]. Каждая из групп представлена в модели 2-мя матрицами:

• вектор-столбец морфемной группы M_{ext}:

$$\begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_M \\ \beta_1 \\ \beta_2 \\ \dots \\ \beta_R \end{bmatrix}$$

• матрица вероятностей для разбора морфемной группы P_{ext}:

$$\begin{bmatrix} P_1(b_1) & P_2(b_1) & \cdots & P_{\mathrm{k}}(b_1) \\ P_1(b_2) & P_2(b_2) & \cdots & P_{\mathrm{k}}(b_2) \\ \dots & \dots & \dots & \dots \\ P_1(b_M) & P_2(b_M) & \cdots & P_{\mathrm{k}}(b_M) \\ P_1(\beta_1) & P_2(\beta_1) & \cdots & P_{\mathrm{k}}(\beta_1) \\ P_1(\beta_2) & P_2(\beta_2) & \cdots & P_{\mathrm{k}}(\beta_2) \\ \dots & \dots & \dots & \dots \\ P_1(\beta_R) & P_2(\beta_R) & \cdots & P_{\mathrm{k}}(\beta_R) \end{bmatrix}$$

где

b_i - i-ый морф морфемной группы;

β_i - i-ая комбинация морфов морфемной группы;

M – количество морфов в морфемной группе;

R - количество комбинаций морфов в морфемной группе;

$P_j(x)$ – вероятность встречи x на j-ом шаге;

k – максимальное число шагов при анализе данной морфемной группы.

В естественных языках морфы внутри своих группы встречаются неравномерно: некоторые морфы могут не встречаться, например, в начале или конце групп. Следовательно, нет смысла проводить проверки на их наличие в начале или конце группы соответственно. Исключая часть морфов и их комбинаций из проверок на определенных шагах, можно добиться ускорения морфемного анализа.

Рассмотрим пример морфемной группы для упрощенного естественного языка. Допустим, что в ней всего 3 морфа, 2 комбинации морфов и максимальное число шагов равно 4. Матрица вероятностей P_{ext} в этом случае будет иметь следующий вид:

$$\begin{bmatrix} P_1(b_1) & P_2(b_1) & P_3(b_1) & P_4(b_1) \\ P_1(b_2) & P_2(b_2) & P_3(b_2) & P_4(b_2) \\ P_1(b_3) & P_2(b_3) & P_3(b_3) & P_4(b_3) \\ P_1(b_4) & P_2(b_4) & P_3(b_4) & P_4(b_4) \\ P_1(b_5) & P_2(b_5) & P_3(b_5) & P_4(b_5) \\ P_1(\beta_1) & P_2(\beta_1) & P_3(\beta_1) & P_4(\beta_1) \\ P_1(\beta_2) & P_2(\beta_2) & P_3(\beta_2) & P_4(\beta_2) \\ P_1(\beta_3) & P_2(\beta_3) & P_3(\beta_3) & P_4(\beta_3) \end{bmatrix}$$

где

$b_1 - b_5$ – морфы;

$\beta_1 - \beta_3$ – комбинации морфов;

$P_j(x)$ – вероятность встречи x на j-ом шаге.

Предположим, что морфы b_4, b_5 и комбинация морфов β_3 не встречаются в начале морфемной группы (1 и 2 шаги разбора), а морфы b_1,

b_2 и комбинация морфов β_1 – в конце морфемной группы (3 и 4 шаги разбора). Следовательно, соответствующие им вероятности $P_1(b_4)$, $P_2(b_4)$, $P_1(b_5)$, $P_2(b_5)$, $P_1(\beta_3)$, $P_2(\beta_3)$, $P_3(b_1)$, $P_4(b_1)$, $P_3(b_2)$, $P_4(b_2)$, $P_3(\beta_1)$, $P_4(\beta_1)$ будут равны нулю и матрица вероятностей будет иметь следующий вид:

$$\begin{bmatrix} P_1(b_1) & P_2(b_1) & 0 & 0 \\ P_1(b_2) & P_2(b_2) & 0 & 0 \\ P_1(b_3) & P_2(b_3) & P_3(b_3) & P_4(b_3) \\ 0 & 0 & P_3(b_4) & P_4(b_4) \\ 0 & 0 & P_3(b_5) & P_4(b_5) \\ P_1(\beta_1) & P_2(\beta_1) & 0 & 0 \\ P_1(\beta_2) & P_2(\beta_2) & P_3(\beta_2) & P_4(\beta_2) \\ 0 & 0 & P_3(\beta_3) & P_4(\beta_3) \end{bmatrix}$$

Данную матрицу можно разбить на 2 матрицы – для 1-2 шагов и 3-4 шагов разбора:

• матрица вероятностей P_{12} для 1-2 шагов разбора морфемной группы:

$$\begin{bmatrix} P_1(b_1) & P_2(b_1) \\ P_1(b_2) & P_2(b_2) \\ P_1(b_3) & P_2(b_3) \\ 0 & 0 \\ 0 & 0 \\ P_1(\beta_1) & P_2(\beta_1) \\ P_1(\beta_2) & P_2(\beta_2) \\ 0 & 0 \end{bmatrix}$$

• матрица вероятностей P_{34} для 3-4 шагов разбора морфемной группы:

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ P_3(b_3) & P_4(b_3) \\ P_3(b_4) & P_4(b_4) \\ P_3(b_5) & P_4(b_5) \\ 0 & 0 \\ P_3(\beta_2) & P_4(\beta_2) \\ P_3(\beta_3) & P_4(\beta_3) \end{bmatrix}$$

Можно заметить, что в каждую из этих матриц вошли строки из нулей. Это означает, что соответствующим им морфы и комбинации морфов не встречаются на данных шагах разбора. Следовательно, нет смысла хранить

информацию о них – соответствующие им строки можно удалить из матриц вероятностей.

После удаления строк матрицы вероятностей примут следующий вид:

• матрица вероятностей P_{12}:

$$\begin{bmatrix} P_1(b_1) & P_2(b_1) \\ P_1(b_2) & P_2(b_2) \\ P_1(b_3) & P_2(b_3) \\ P_1(\beta_1) & P_2(\beta_1) \\ P_1(\beta_2) & P_2(\beta_2) \end{bmatrix}$$

• матрица вероятностей P_{34}:

$$\begin{bmatrix} P_3(b_3) & P_4(b_3) \\ P_3(b_4) & P_4(b_4) \\ P_3(b_5) & P_4(b_5) \\ P_3(\beta_2) & P_4(\beta_2) \\ P_3(\beta_3) & P_4(\beta_3) \end{bmatrix}$$

В результате проделанной оптимизации вместо одной матрицы вероятностей P_{ext} из 32 элементов получены 2 матрицы P_{12} и P_{34} с общим количеством элементов, равным 20. Таким образом для данного примера удалось сократить количество элементов более чем в 1.5 раза. Кроме того, будет ускорена и проверка на каждом шаге, т.к. вместо 8-ми проверок (по одной на каждую строку матрицы вероятностей) потребуется выполнять лишь 5 проверок.

Таким образом каждая морфемная группа теперь будет иметь не 2, а 2*n матриц. Но общее количество элементов в матрицах вероятностей будет меньше и количество шагов при разборе морфемной группы тоже уменьшится.

В естественном языке величина оптимизации будет зависеть от того, насколько сильно зависят вероятности появления морфов от их места в морфемной группе. Предполагается, что наибольший выигрыш данная оптимизация принесет для агглютинативных языков, в которых может содержаться много аффиксов в одном слове.

Рассмотрим достоинства и недостатки оптимизированной математической модели в сравнении с ранее предложенной [2].

Достоинства:

1. Позволяет ускорить морфемный разбор за счет исключения проверки тех морфов и их комбинаций, которые не встречаются на данном шаге разбора.

2. Сокращает расходы вычислительных ресурсов (оперативная память и время работы процессоров) для проведения морфемного анализа с помощью вычислительной техники. Особенно это актуально для мобильных приложений, где вычислительная мощность существенно ограничена [1].

Недостатки:

3. Настройка модели усложняется из-за необходимости не только строить матрицы вероятностей, но и выбирать оптимальное разбиение этих матриц для разных шагов разбора.

4. Усложняется сопровождение модели из-за увеличения в ней количества матриц и возможных изменений границ разбиения матрицы вероятностей P_{ext} на отдельные матрицы.

5. При недостаточно накопленной статистике (статистический анализ проводился на ограниченном наборе текстов) может оказаться, что вероятности некоторых морфов на некоторых шагах разбора окажутся равными нулю, хотя на самом деле они отличны от нуля. Это может привести к тому, что после разбиения соответствующие им строки из матриц вероятностей будут исключены. В результате морфемный разбор по этим матрицам может не дать успешного результата. Данная проблема решаема, если в подобных случаях в качестве аварийного варианта продолжить разбор по полному набору морфов данной морфемной группы.

Использованные источники:

1. Мытников А.Н., Мытникова Е.А., Кузнецова Л.Н., Солин С.Ю. Технологии разработки мобильных приложений // Теория и практика современной науки. – 2016. – № 4(10). - С. 504-507.
2. Фадеев С.Г. Расширение математической модели естественного языка для морфемного анализа // Состояние и перспективы развития ИТ-образования: Сборник докладов и научных статей Всероссийской научно-практической конференции (посвящается 50-летию Чувашского государственного университета им. И.Н. Ульянова). (г. Чебоксары, 16-18 ноября 2017 г). 2018. – С. 272-277.
3. Fadeev S.G., Zheltov P.V. Optimization options of word forms morphemic analysis on the basis of statistical knowledge // Russian Linguistic Bulletin. – 2016. – № 3 (7). – с. 15. DOI: 10.18454/RULB.7.33.