А. С. БЕРЛІН, Е. В. МАЛАФЄЄВ

ПРОГРАМУВАННЯ ЕОМ В СИСТЕМАХ АВТОМАТИЧНОГО ПРОГРАМУВАННЯ СКЛАДАННЯ

Технічний прогрес в області складальних процесів розвивається зараз у напрямі вдосконалення як буквовідливних, рядковідливних і фотоскладальних методів, так і методів підготовки керуючих ними програм. Особливо велику увагу за останні 3—4 роки приділяють способам підготовки програм керування рядковідливними складальними автоматами.

Програми керування можуть бути підготовлені системами ручного, напівавтоматичного та автоматичного програмування складання (АПС). Системи ручного та напівавтоматичного програмування складання потребують в процесі виготовлення програми роботи оператора. Застосовуючи системи АПС, можна повністю автоматизувати цю операцію.

Робота з системою АПС проходить так. При передрукуванні оригінала, наприклад у видавництві, паралельно виготовляється первинна перфолента, в якій містяться тільки коди тексту і деяких команд, але без поділу тексту на рядки, тобто в такій перфоленті немає кодів кінця рядка. Одержану в такому вигляді перфоленту, яку називають неповнокодовою, вводять у зчитуючий пристрій системи. Електронна обчислювальна машина (ЕОМ), яка входить до складу системи за спеціально розробленою програмою здійснює розбивку тексту на рядки заданого формату з одночасним виготовленням повнокодової ленти за неповнокодовою. Програма роботи ЕОМ враховує місця можливих переносів слів мови, яка обробляється, формат набору, гарнітуру та кегль шрифту, тип клина та ін.

Створення систем АПС йде двома шляхами. Перший передбачає використання універсальних ЕОМ, а другий — спеціалізованих. Обидва шляхи мають свої переваги й недоліки.

До переваг універсальних ЕОМ порівняно з спеціалізованими можна віднести такі:

1) високу продуктивність;

2) відсутність необхідності в розробці, оскільки використовується готова техніка;

3) можливість заміни програм більш досконалими.

До недоліків універсальних ЕОМ слід віднести необхідність розробки програм. Проте цей недолік умовний, тому що трудомісткість їх складання непомірно менша трудомісткості розробки спеціалізованої ЕОМ.

Першим етапом програмування роботи ЕОМ є складання алгоритмів, згідно з якими повинен здійснюватися процес обробки тексту. Алгоритми роботи ЕОМ можна поділити на три групи:

а) алгоритми переносу слів;

б) алгоритми розрахунку виключки рядка;

в) алгоритми виконання деяких правил складання.

Алгоритми переносу слів — збір правил та способів, за якими можна здійснити «машинний» перенос слів.

В наш час відомі три основні методи складання алгоритмів для переносу слів.

Перший метод створення алгоритмів переносу слів (логічний) передбачає перенос слів по складах, а поділ слова на склади проводиться шляхом аналізу приголосних та голосних букв слова. Для виключення неправильних варіантів переносу слів у машину повинен бути введений словник закінчень, префіксів та суфіксів.

Середній час, необхідний для здійснення переносу слів англійської мови, за цим методом дорівнює 15 мілісекундам, а точність розділу слів — близько 92%.

Другий метод (імовірносний) побудований на використанні статично-імовірносних способів вирішення питання переносу слів. Результати підрахунків показують, що цим методом можна здійснювати правильний перенос до 80% слів, що зустрічаються у тексті. При роботі за імовірносним методом проводиться аналіз можливих переносів слів даної мови. Для цього здійснюється попередня обробка слів мови з метою отримання даних про можливість поділу двох букв слова, причому враховуються дві попередні та дві наступні букви. Результати обробки вносяться у таблицю можливих поділів. Для поділу слова проводиться порівняння комбінації букв, які входять до слів, з одержаною таблицею можливих поділів. Слово поділяється у тому місці, де імовірність переносу максимальна.

Системи, що працюють за таким методом, також повинні мати словник закінчень, суфіксів та префіксів.

В основі третього (словниковий) методу переносу слів (у застосуванні до англійської мови) покладено правило «3, 5, 7, 9». За цим правилом перенос слова може здійснюватись після 3, 5, 7 або 9 букви. Застосування цього правила для англійської мови дає точність поділу до 90%.

Слова, перенос яких за цим правилом неможливий, вносяться у пам'ять машини. Кількість слів, занесених у словник, може коливатись у широких межах — від 30 до 500 тисяч. При проведенні аналізу слово порівнюється з словником. У випадку, якщо немає збігу, слово поділяється за правилом «3, 5, 7, 9».

Час розшуку будь-якого слова в словнику винятків дорівнює в існуючій системі 1 *сек* (при об'ємі словника винятків 28 000 слів). Точність поділу слів для переносу наближається до 100%.

Слід зауважити, що існує модифікація словникового методу, яка не передбачає використання правила «3, 5, 7, 9» або аналогічного. В цьому випадку в словник заносяться всі слова даної мови, поділені в місцях можливих переносів.

Системи, побудовані за логічним методом та методом імовірності, будуть більш швидкодіючі, ніж системи, побудовані за словниковим методом. Це викликано тим, що використання словникового методу передбачає наявність словника винятків великого об'єму (до 500 000 слів), який не може бути введений в оперативну пам'ять машини. Використання ж зовнішньої пам'яті вимагає значних затрат часу на вибір потрібного слова з словника. В цей же час словники суфіксів та префіксів, необхідні при роботі цими двома методами, мають значно менший об'єм, що дозволить заносити їх в оперативну пам'ять ЕОМ. При цьому навіть при використанні зовнішньої пам'яті час виборки буде незначним.

З 1965 р. УНДІПП запланував розробки по створенню систем АПС з використанням універсальної ЕОМ. У зв'язку з цим виникла не-

обхідність розробки алгоритмів поділу слів переносом для російської мови.

Дослідження можливих переносів слів російської мови для створення системи АПС, яка працювала б за методом імовірності, вимагає виконання великого об'єму статистичних робіт, а точність поділу слів за цими методами (для англійської мови) складає всього 80%, в той час, коли логічний метод дає точність поділу близько 92%.

Недоліки словникового методу були вказані вище. Тому ми скористалися логічним методом, який забезпечує більшу швидкість обробки текстового матеріалу і дає досить високу точність поділу слів. При роботі за цим методом система алгоритмів повинна забезпечити поділ

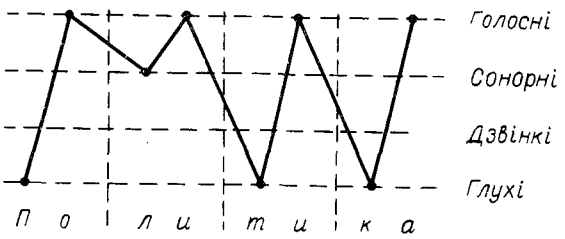

Рис. 1. Поділ слова на склади за дзвінкістю звуків.

слова на склади, який може проводитись як на основі аналізу дзвінкості звуків, що входять до нього, так і шляхом проведення аналізу комбінацій голосних та приголосних звуків слова.

Розглянемо принципи побудови цих систем алгоритмів переносу.

Алгоритми поділу слова на склади за дзвінкістю звуків. Правильний поділ слова на склади може бути здійснений після аналізу дзвінкості складових його звуків. Більшість складів у російській мові побудована за принципом зростаючої дзвінкості, кінець складу знаходиться в місці спадаючої або однакової дзвінкості.

За дзвінкістю звуки російської мови поділяються на такі групи (в порядку падаючої дзвінкості):

1) голосні: а, е, и, о, у, ы, э, ю, я;
2) сонорні: л, м, н, р;
3) дзвінкі: б, г, в, д, ж, з;
4) глухі: к, м, с, т, ф, х, ц, ч, ш, щ.

Для спрощення аналізу при поділі слова на склади слід будувати графіки дзвінкості слова. Графічно її можна представити ломаною лінією, зображеною над даним словом.

Проаналізуємо, наприклад, слово «политика» (рис. 1). З графіка видно, що поділ слова на склади, проведений у місцях падаючої дзвінкості, буде таким: по-ли-ти-ка.

Отже, основний алгоритм виразиться так: поділ слів для переносу неможливий в місцях зростаючої дзвінкості.

Проте подібний спосіб може давати помилки поділу:

а) з початку і наприкінці слова — за рахунок наявності в ньому суфіксів та префіксів;

б) у складних словах.

В обох випадках порушення складоутворення спостерігаються в основному тоді, коли в слові зустрічаються підряд дві або більше голосних або приголосних букви. В результаті статистичної обробки російської мови були складені такі правила поділу, які зменшують кількість помилок:

1. Якщо в слові зустрічаються підряд дві або три голосні букви, їх можна розділити переносом.

2. Якщо в слові зустрічаються дві і більше приголосних підряд, перенос перед першою приголосною заборонений. У випадку, коли групу приголосних не можна розбити переносом (зростаюча дзвінкість), перенос необхідно робити перед першим приголосним. Останні алгоритми, хоч і зменшують кількість помилок при поділі слів з приставками та складних слів, проте не виключають повністю можливості появи помилок при їх поділі.

Вказані вище групи поділу букв російського алфавіту також були дещо змінені. Так [3, § 119, п. 1 і 2], букви **ъ, ь, й** не повинні відриватись від попередніх; це привело до необхідності виділення ще одної групи дзвінкості. Більш глибоке вивчення правил переносу та обробки словника російської мови показали, що можливе об'єднання в одну групу

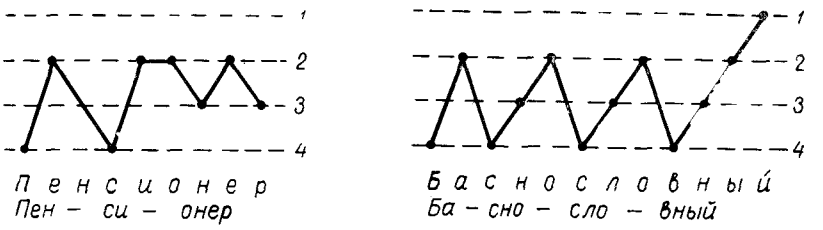

Рис. 2. Приклади поділу слів для переносу.

глухих і дзвінких приголосних без появи помилкових переносів. У зв'язку з вищевказаним був прийнятий такий розподіл букв російського алфавіту на групи звучності:

1. й, ь, ъ;
2. а, е, и, о, у, ы, э, ю, я;
3. л, м, н, р;
4. б, г, в, д, ж, з, к, н, с, т, ф, х, ц, ч, ш, щ.

Розглянемо декілька прикладів поділу слів для переносу за цими алгоритмами (рис. 2).

З цих прикладів видно, що поділ слів для переносу за цими алгоритмами приводить до втрати переносів. Так, у поданих вище прикладах втрачені переноси: пе-нсионер, пенсио-нер; бас-нословний, баснослов-ний.

Застосування цих алгоритмів для аналізу складних слів та слів з префіксами не ліквідувало б помилок поділу. Наприклад: нас-коблить, небос-клон, нес-частный, разос-паться та ін. Це викликає необхідність введення словника префіксів, а також основ перших частин складних слів, що найчастіше зустрічаються.

Алгоритми поділу слова на склади на основі аналізу голосних та приголосних звуків слова. Для проведення аналізу всі букви алфавіту розділено на групи:

1. Приголосні;
2. Голосні;
3. ь, ъ, й.

Поділ слова для переносу здійснюється після аналізу комбінації різних букв у слові.

Так, до складу обов'язково повинні входити як приголосна, так і голосна букви, а тому частини слів, які не вміщують приголосних або голосних, вважаються не підлягаючими переносу.

Переважна більшість складів у словах російської мови починається з приголосної букви, а тому основний алгоритм переносу записується так: перенос можливий після голосного перед приголосним.

Наприклад: ре-бя-та, па-стух, лю-бовь і т. п.

Якщо в слові зустрічається декілька приголосних підряд, поділ на склади в цьому місці вільний [3, § 118 та 119], тобто перенос може бути поставлений як перед групою приголосних, так і в будь-якому місці всередині групи. Наприклад: де-рзкий, дер-зкий, дерз-кий; се-стра, сестра, сест-ра; ца-пля, цап-ля. Проте останній голосний не може бути відірваний від наступного голосного.

Аналіз та статистична обробка слів російської мови показали, що у випадку, коли в слові зустрічаються дві або три голосні підряд, їх можна розділити переносом, причому перенос може стояти і після останнього голосного.

Згідно з правилами переносу не можна розділяти слово перед знаками 3-ї групи (**ь, ъ, й**), що забороняється спеціальним алгоритмом.

Необхідно відмітити декілька другорядних алгоритмів, таких як дозвіл на перенос після знаків дефісу і тире, заборона переносу після знаку №, заборона поділу цифр та ін.

Згідно з наявністю слів, які не можна розділити, немає необхідності проводити аналіз переносу кожного слова. Непереносними вважаються слова, які мають менш ніж два голоси, або менше чотирьох знаків.

Система алгоритмів на основі аналізу голосних і приголосних також приводить до появи помилок при аналізі слів з префіксами і складних слів.

Поділ складних слів і слів з префіксами. Обробка словника російської мови показала, що обидві системи алгоритмів дають приблизно однакову кількість похибок. Але при обробці слів за першою системою алгоритмів губиться деякий процент можливих переносів.

При використанні будь-якої з описаних вище систем алгоритмів для поділу слів з префіксами або складних слів можуть виникати помилкові переноси, які протирічать правилам. Найчастіше виникають помилки в таких випадках:

1. Якщо префікс закінчується на приголосну і знову починається з приголосної, наприклад: по-дбить, ра-змах.

2. Якщо слово, яке слідує за префіксом, починається з декількох приголосних, наприклад: прис-лать, отс-транить.

3. Якщо в складних словах друга основа починається з декількох приголосних, наприклад: пятиг-раммовый, красног-вардеец.

Для правильного поділу слів в таких випадках необхідно встановити наявність префікса в слові. Це здійснюється порівнянням початкової частини слова з словником префіксів, закладеним в пам'ять ЕОМ. Проте немає необхідності аналізувати на наявність префікса кожне слово, тому що навіть однобуквений префікс не зустрічається в словах, які мають менше чотирьох букв.

Односкладовий префікс не може бути поділений переносом, якщо за ним іде приголосний, при ньому не може залишатись початкова частина кореня, тобто в такому випадку префікс повинний поділятись переносом. Але якщо за префіксом на приголосний іде далі голосний (крім **ы**), його поділ переносом можливий, наприклад: бе-зумный. У випадку, якщо за префіксом на приголосний іде **ы**, переносити частину слова, яка починається з **ы** не дозволяється, наприклад, ра-зыскать [3, § 118, 119]. Те ж саме можна сказати про найбільш часто вживані основи перших частин складних слів. Винятки становлять основи, що закінчуються на **-х**, які не можуть бути розділені переносом. До таких основ відносяться частини слів: двух-, трех- та ін. Отже, якщо виявлений в слові префікс закінчується на приголосний (крім **-х**), після якого слідує голосний, аналіз слова на можливість переносу необхідно здійснювати як і для слів, що не вміщують префіксу. Решта префіксів відділяється, а частина слова після префікса аналізується окремо.

Складання словника префіксів. Використання словника префіксів значно зменшує процент похибок, які виникають при машинному переносі. Складати словник префіксів можна так. Слова опрацьовуваного словника досліджуються з метою встановлення можливих переносів. Слова, в яких наявні похибки, підраховуються. Після того кожне з цих слів аналізується для визначення префікса, внесення якого в словник виправило б невірний перенос.

Необхідно відмітити, що введення кожного окремого префікса в словник приводить до виправлення даного числа неправильних переносів. Тому від кількості введених у словник префіксів буде залежати і похибка поділу слів переносами. На рис. 3 зображена залежність похибки переносів від кількості префіксів у словнику.

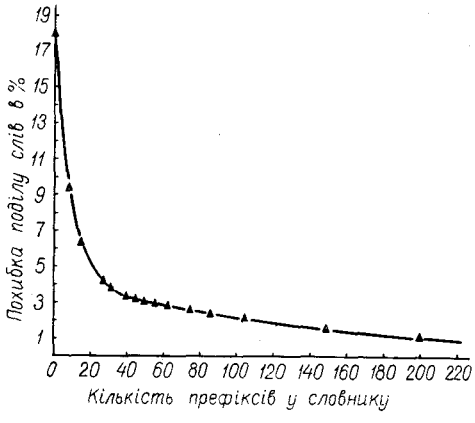

Рис. 3. Залежність похибки поділу слів від кількості префіксів у «словнику» ЕОМ.

Величина точності поділу слів переносами визначає число неправильно набраних рядків у тексті. При недостатній точності поділу процес коректури може стати настільки громіздким, що вся система стане неефективною. За даними зарубіжної літератури точність поділу слів, що дорівнює 85%, вважається незадовільною. Це дозволяє зробити висновок про те, що програма, складена без словника префіксів, безперспективна. Зараз зарубіжні програмісти намагаються добитись точності поділу слів не менше 99%. Спроби використання ЕОМ при складанні крупноформатних книжкових видань показали, що рядки з переносами зустрічаються у процесі складання з частотою від одного рядка на 100 до одного рядка на 1000. При точності поділу 99% це відповідає необхідності повторного складання 2—4 рядків з 10 000, що вважається задовільним показником. Проте приведені вище співвідношення між точністю поділу та кількістю повторно складених рядків не враховують того факту, що навіть у неправильно перенесених словах є правильні переноси. Імовірність того, що при поділі кінцевого слова рядок закінчиться саме в місці неправильного переносу, далеко не стопроцентна. Крім того, частота появи рядків з переносами залежить від формату набору (зростає при зменшенні формату), гарнітури та кеглю шрифту (збільшується разом з зростанням ємкості шрифту).

Алгоритми виключки рядка. Для повного формування рядка необхідно провести розрахунок виключки, тобто дати на повнокодовій ленті код кінця рядка в тому місці, де рядок може бути виключений у даному форматі. З цією метою ЕОМ у процесі розрахунку виключки рядка повинна безперервно провадити аналіз таких нерівностей:

$$\sum_{1}^{n_1} t_i + \sum_{1}^{m_1} t_{\max} \geqslant F, \tag{1}$$

$$\sum_{1}^{n_2} t_i + \sum_{1}^{m_2} t_{\min} \geqslant F, \tag{2}$$

де n — кількість знаків,

m — кількість клинів у даному рядку;

t_{min}, t_{max} — мінімальна та максимальна ширина застосовуваного клина;

F — формат набору.

Момент, коли стає справедливою нерівність (1), називається входом в зону виключки; момент, коли стає справедливою нерівність (2), називається переповненням. Різниця між форматом та довжиною рядка в момент входу в зону виключки називається зоною виключки рядка. Рядок може бути виключений у тому випадку, якщо останній знак рядка знаходиться в зоні виключки.

Враховуючи те, що рядок можна закінчити лише в місці міжсловного пробілу або в місці можливого переносу в слові і враховуючи вимогу правил набору про необхідність формування більш «тугих» рядків, необхідно виключати рядок там, де є останній міжсловний пробіл або останній перенос, які знаходяться в зоні виключки. Але система поділу слів для переносу працює з певною похибкою і для зменшення цієї похибки використовується система, якій властиве намагання зменшити кількість рядків з переносами, так звана система пріоритету. Суть цієї системи зводиться до того, що аналіз слова на можливість переносу здійснюється лише тоді, коли рядок не може бути виключений без переносу.

Виключка рядка при роботі з системою пріоритету проводиться таким чином.

1. Рядок закінчується на останньому міжсловному пробілі, який знаходиться у зоні виключки.

2. Якщо в зоні виключки немає ні одного міжсловного пробілу, пробують здійснити виключку рядка, обмеженого останнім міжсловним пробілом, який лежить перед зоною виключки, добавляючи до кожного клина тонку або півкруглу шпацію.

3. У випадку, якщо рядок не виключається, система виключає рядок з поділом останнього його слова переносом.

Алгоритми виконання правил складання. Програма ЕОМ може реалізувати також алгоритми додержання деяких правил складання. Так, відбивка тире, знаків, №, § і т. п., введення абзацного відступу, формування кінцевих рядків абзацу, збільшення при виділенні в тексті розрядкою пробілів між виділеними розрядкою словами та пробілів, що відділяють ці слова від інших, на задану правилами складання величину, формування рядків, наприкінці яких знаходиться прийменник, що розпочинає наступне речення,— всі ці операції можуть бути запрограмовані.

Необхідно відмітити, що програмування ЕОМ для виконання операції виготовлення повнокодової перфострічки (з кодами кінця рядків) за неповнокодовою є тільки першим етапом автоматизації процесу складання. Наступним етапом є програмування верстки видання, причому для одержання зверстаних полос не можна використати рядкоабо буквовідливних складальних автоматів; в цьому випадку необхідно застосовувати фотоскладальне обладнання.

ЛІТЕРАТУРА

1. "A Means to a Iustified End" by T. V. Higgs (англ.), "The Litho-Printer", 1964, № 7, 33—40.

2. Р. И. Аванесов. *Фонетика современного русского литературного языка.* Изд-во Московского ун-та, 1956.

3. «Правила русской орфографии и пунктуации». Учпедгиз, 1956.

4. С. И. Ожегов. Словарь русского языка, 52000 слов. ГИИНС, 1953.

5. "Computers in '64. Year of Transition From Theory to Practice" (англ.), "Book Production", 1964, March, 44—50.

6. „Der Elektronenrechner als Maschinensetzer", Friedrich Balzer (нім.), „Der Druckspiegel", 1964, № 5, 303—318.

7. „Das GSA — System im halbautomatischen Betrieb", W. Güttinger (нім.), „Graphische Woche", 1965, № 14, 644—646.

A. S. BERLIN, E. V. MALAFEYEV

THE PROGRAMMING FOR ELECTRONIC COMPUTERS IN THE SYSTEMS OF COMPUTERIZED TYPESETTING

Summary

The principles of working out and construction of some russian language algoritms, necessary for programming the computers in the systems of computerized typesetting are described, namely the algoritms of word division, those of line justificating calculation and of some typesetting rules performance.