Список литературы

1. Лебедев Н. Н., Скальский И. П., Уфлянд Я. С. Уравнения математической физики. М., 1959.
2. Прусов В. И. // Вестн. Белорусского ун-та. Сер. 1: Физ. Мат. Мех. 1986. № 3. С. 52.
3. Уфлянд Я. С. Интегральные преобразования в задачах теории упругости. М., 1963.
4. Мусхелишвили Н. И. Некоторые основные задачи математической теории упругости. М., 1966.

Поступила в редакцию 03.06.87.

УДК 801.73:681.3

Н. К. РУБАШКО, И. В. СОВПЕЛЬ

АВТОМАТИЧЕСКИЙ КОНТРОЛЬ ТЕКСТОВ ЕСТЕСТВЕННЫХ ЯЗЫКОВ. I

Одним из важнейших условий эффективного функционирования систем обработки естественно-языковой (ЕЯ) информации как средства описания действительности и коммуникации с вычислительной системой является наличие в них подсистемы автоматического контроля данных (АКД). Исчерпывающий АКД затрагивает не только морфологический, но и синтаксический и семантический уровни ЕЯ, а если говорить о его структурных уровнях, то уровень слов, конфигураций, фраз, предложений, дискурсов, текстов и включает корректировку как орфографии (будет рассматриваться именно этот случай), так и стиля.

Ошибки в сообщениях возникают по нескольким причинам. Очевидно, что обобщенная схема линии ЕЯ коммуникации l в плане этапности передачи информации может быть представлена следующим образом: $a_0 \quad a_1 \quad a_2 \quad a_{k-1} \; a_k$
$|\longrightarrow|\longrightarrow \ldots |\longrightarrow|$, где a_0 — источник информации; $a_1, \ldots, a_{k-1}$ — так называемые релейные станции (одновременно приемники и источники информации); a_k — конечный пользователь (приемник информации); каждой дуге (a_i, a_{i+1}), $i=\overline{0, k-1}$ соответствует определенный «технический» тип коммуникации (радиоволны, провода и т. п.).

Первый источник ошибок может быть связан с техническими аспектами коммуникации, второй — с тем, что именно человек посылает сообщения (речевой ввод, ввод оригинальных или отперфорированных текстов и т. п.) как в случае a_0, так и в случае $a_1, a_2, \ldots, a_{k-1}$. Специальные инструкции на формат сообщений не всегда соответствуют как формальному синтаксису, так и заранее определенному словарю. Следует отметить, что в этом случае существуют некоторые закономерности «производства» ошибок, которые могут и должны учитываться при построении алгоритмов их распознавания и корректировки.

При проектировании средств АКД необходимо иметь в виду некоторые важные характеристики систем ЕЯ коммуникации [1].

Будучи, как правило, системами реального времени, они требуют минимальных временных затрат на корректировку, интерпретацию и, если необходимо, доступ к соответствующей базе данных, а также аккуратности и надежности; системы должны быть в состоянии обработать ограниченный ЕЯ и быстро идентифицировать случаи, невозможные для автоматической обработки.

Принято различать две стратегии автоматического контроля ЕЯ информации: исчерпывающий контроль и контроль переменной глубины, который способен на тщательную обработку отдельных, наиболее важных или наиболее вероятных с точки зрения возникновения ошибок частей сообщения.

Предпосылкой для успешного решения проблемы АКД является 4-кратная избыточность ЕЯ, а также возможность формализации описания основных типов ошибок.

Предварительно введем следующие определения. Обозначим через Σ алфавит: непустое конечное множество символов, включая «пустой» символ (пробел), обозначаемый $\otimes$.

Цепочкой в алфавите Σ назовем конечную последовательность элементов $x_1, x_2, \ldots, x_n$, где $x_i \in \Sigma$, $i = \overline{1, n}$ и запишем $X = x_1 x_2 \ldots x_n$. Число n назовем длиной цепочки X, обозначим $n = |X|$.

Тогда естественный язык L можно определить как четверку $L = (A, M, S_1, S_2)$, где A — алфавит ЕЯ, M, S_1, S_2 — соответственно множества морфологических, синтаксических и семантических правил образования цепочек из A с помощью операции конкатенации, а любой текст $T_i \in L$ — как конечную цепочку $x_1^{(i)} x_2^{(i)} \ldots x_m^{(i)}$, $x_p^{(i)} \in A$, $p = \overline{1, m}$, образованную определенными подмножествами $M^{(i)}$, $S_1^{(i)}$, $S_2^{(i)}$ множеств M, S_1, S_2 соответственно.

Пусть T_i, $T_j \in L$. Будем говорить, что текст T_i равен тексту T_j, запишем $T_i = T_j$, если $|T_i| = |T_j|$ и $x_p^{(i)} = x_p^{(j)}$ для всех p от 1 до $m = |T_i|$. В противном случае — T_i не равен T_j ($T_i \neq T_j$). Сообщение, поступающее в линию ЕЯ коммуникации, назовем начальным текстом для l, обозначим T_0 ($T_0 \in L$), а сообщение, поступающее в приемник информации a_k, — конечным текстом $T_k = l(T_0)$; l назовем корректной, если $T_k \in L$ и $T_0 = T_k$, а в случае нарушения хотя бы одного из этих условий — некорректной.

Очевидно, что для рассматриваемого аспекта ЕЯ коммуникации количество этапов в соответствующей линии не имеет принципиального значения, поэтому положим $k = 1$.

Поскольку в a_1 T_0 неизвестно, а известно только T_1, то, очевидно, решение вопроса корректности l в общем случае является приближенным и сводится к решению вопроса принадлежности T_1 языку L, который прежде всего и в основном рассматривается на уровне слов текста.

Пусть x_p — некоторый непустой символ из T (T — начальный либо конечный текст). Максимально возможную по длине подцепочку из T, содержащую x_p и не содержащую пустого символа, назовем словом текста T. Очевидно, что T содержит конечное множество слов текста, в общем случае не являющихся попарно различными.

Любое слово языка L можно определить как конечную цепочку элементов из A, образованную соответствующими подмножествами множеств M, S_1, S_2. Условимся любое конечное множество слов языка L называть эталонным словарем, обозначим D, а любое $X \in D$ — значимым словом.

Пусть $X \in T_0$, $Y = l(X) \in T_1$. Очевидно, что понятие равенства для текстов распространяется и для слов. Если $X \neq Y$, то будем говорить, что Y является ошибочным словом.

Положим $X \in D$, $|X| = m$, $Y \in T_1$, $|Y| = n$. Будем говорить, что X и Y удовлетворяют условию (*), если выполняется одно из следующих условий:

при $|m - n| = 1$

(1) $x_i = y_j$, $i = j = \overline{1, m}$, $m < n$;

(2) $x_i = y_j$, $j = i + 1$, $i = \overline{1, m}$, $m < n$;

(3) $x_i = y_j$, $i = j = \overline{1, n}$, $m > n$;

(4) $x_i = y_j$, $i = j + 1$, $j = \overline{1, n}$, $m > n$;

(5) если $m < n$, то существует $1 < i < n$ такое, что $x_j = y_j$ для $j = \overline{1, i-1}$ и $x_j = y_{j+1}$ для $j = \overline{i, n-1}$;

(6) если $m > n$, то существует $1 < i < m$ такое, что $x_j = y_j$ для $j = \overline{1, i-1}$ и $x_{j+1} = y_j$ для $j = \overline{i, m-1}$;

(7) в T_1 существует два последовательных слова Y', $|Y'| = n'$ и Y'', $|Y''| = n''$ (полагаем, что Y есть $Y' \otimes Y''$, $n = n' + n'' + 1$) таких, что $m =$

$= n' + n''$, $x_i = y'_i$ для $i = \overline{1, n'}$ и $x_{n'+i} = y''_i$ для $i = \overline{1, n''}$;
при $|m - n| = 0$

(8) существует $1 \leqslant i \leqslant m$ такое, что $x_i \neq y_i$ и для $j = \overline{1, m}$, кроме $j = i$, $x_j = y_j$;

(9) существует $1 \leqslant i < m$ такое, что $x_i = y_{i+1}$, $x_{i+1} = y_i$ и для $j = \overline{1, m}$, кроме $j = i$, $j = i + 1$ $x_j = y_j$;

(10) существуют $X' \in D$, $|X'| = m'$, $X'' \in D$, $|X''| = m''$ (полагаем, что X есть совокупность X' и X'', $m = m' + m''$) такие, что $m' + m'' = n$, $x'_i = y_i$ для $i = \overline{1, m'}$ и $x'_i = y_{m'+i}$ для $i = \overline{1, m''}$.

Очевидно, что (*) формально определяет X как слово (в случае (10) — пару слов) языка L в T_0, в котором, возможно, в процессе ЕЯ коммуникации возникла ошибка следующего типа.

Замена: один символ в X заменен на другой, отличный от пробела.

Вставка: один символ, отличный от пробела, вставлен в X, возможно, в качестве первого или последнего.

Перестановка: два соседних символа в X обменены местами;

Пропуск: один символ, возможно, первый или последний, удален из X.

Разбиение: в X вставлен пробел, но не в качестве первого или последнего символа;

Слияние: между двумя последовательными словами удален пробел.

Y определяется как $Y = l(X)$ либо $Y = l(X' \otimes X'')$.

Приведенные типы ошибок выделены на основании статистических данных, полученных при анализе большого объема текстов, подготовленных с помощью УПДЛ ЕС9003. В частности, доля ошибок типа замена составляет 35 % их общего числа, пропуск — 26, вставка — 19, слияние — 5, перестановка — 3, разбиение — 2 %. Возможны ошибки и других типов, но их процент значительно ниже указанных. Эти результаты в целом согласуются с данными [2].

Формальное определение основных типов ошибок служит основой для разработки простых, но достаточно эффективных алгоритмов АКД. При этом мы исходим из следующих положений.

В системах, работающих с ЕЯ информацией, как правило, заранее заданы в виде декларативных знаний определенные лингвистические данные и прежде всего словарь соответствующего ЕЯ.

Эффективность средств АКД находится в прямой зависимости от объема эталонного словаря и полноты покрытия им данной предметной области.

Задача АКД в общем случае имеет неоднозначное решение на морфологическом уровне ЕЯ.

По сравнению с существующими, например [3, 4], предложенный критерий в виде условия (*) достаточно прост, исключает, в частности, морфологический анализ текста, определяет все основные (в соответствии с приведенными статистическими данными) типы ошибок и с этой точки зрения полностью решает задачу АКД на морфологическом уровне, создавая предпосылки для ее решения на синтаксическом и семантическом уровнях языка.

Список литературы

1. Rosenberg J. // Applications of Artificial Intelligence II. 1985. V. 548. P. 233.
2. Pollock J. J., Zamora A. // Journ. Amer. Soc. Inform. Sci. 1983. V. 34. № 1. P. 51.
3. Белоногов Г. Г., Дуганова И. С. и др. // НТИ. Сер. 2. 1982. № 6. С. 29.
4. Матвеев С. А., Сотникова Р. А. // Программирование. 1984. № 5. С. 68.

Поступила в редакцию 04.06.87.