

7. Ревзин, И. И. Основы общего и машинного перевода / И. И. Ревзин, В. Ю. Розенцвейг. М.: Высш. шк., 1964.
8. Сепир, Э. Градуирование // Новое в зарубежной лингвистике. М.: Прогресс, 1985. Вып. XVI. С. 43–78.

9. Федоров, А. В. Основы общей теории перевода. М.: Высш. шк., 1968.
10. Швейцер, А. Д. Теория перевода: Статус, проблемы, аспекты. М.: Наука, 1988.
11. Levý, J. Umění překlada. Praha: Panorama, 1983.

Вестник Челябинского государственного университета. 2011. № 24 (239). Филология. Искусствоведение. Вып. 57. С. 69–71.

В. Ю. Гудков, Е. Ф. Гудкова

N-ГРАММЫ В ЛИНГВИСТИКЕ

В статье анализируется содержание и применение N -грамм как средства фиксации языковых реалий. Показывается отношение моделей N -грамм к формальной грамматике, предлагается рассматривать их как инструмент автоматического анализа печатных текстов и непрерывной речи человека.

Ключевые слова: N -грамма, порождающая грамматика Хомского, вероятностная модель речи, автоматический анализ текстов.

Модель N -граммы в лингвистике. Пусть задан некоторый конечный алфавит $V = \{w_i\}$, где w_i — символ. Языком $L(V)$ называют множество цепочек конечной длины из символов w_i . Высказыванием называют цепочку из языка. N -граммой на алфавите V называют произвольную цепочку длиной N , например последовательность из N букв русского языка одного слова, одной фразы, одного текста или, в более интересном случае, последовательность из грамматически допустимых описаний N подряд стоящих слов [1]. Грамматически корректные N -граммы могут нести разную смысловую нагрузку — во фразах «Она разинула пасть» и «Она решила пасть» слово «пасть» имеет разные значения.

N -граммы для понимания естественного языка стали применять сравнительно недавно. Предложена вероятностная модель речи на основе теории цепей Маркова, различающая разных авторов и даже фольклор. Значение N -грамм исчерпывается их прикладной направленностью: они являются эффективным инструментом решения важной задачи — отбраковки вариантов, а их использование сводится к наложению допустимых N -грамм на имеющиеся данные [1; 2].

Пусть $C(w | w = w_1, w_2, \dots, w_n)$ — число вхождений строки w в генеральную совокупность Ω текстов языка. Вероятность $p(w)$ появления N -граммы w находят в виде

$$p(w) = \frac{C(w)}{\sum_{x \in \Omega} C(x)}.$$

Подобно определяют вероятность $p(w_i)$ униграммы как вырожденного случая N -граммы [3]. Если вероятности появления символов в любой позиции цепочки независимы и одинаково распределены, то

$$p(w) = \prod_{i=1}^n p(w_i).$$

Таким образом, перестановки символов $w_i \in w$ имеют одну и ту же вероятность. Например, в языке вероятность встретить выражения «красно-коричневый» та же, что и выражение «к-рснкрчнваооиеый». Для разрешения указанного недоразумения вводят условные вероятности [3]. Тогда вероятность очередного символа строки задается в зависимости от предшествующих ему символов в виде

$$p(w) = p(w_n | w_1, w_2, \dots, w_{n-1}) p(w_1, w_2, \dots, w_{n-1}),$$

а модель N -граммы — марковской цепью $(N-1)$ -го порядка. Задача оценивания статистических параметров N -граммы сводится к задачам по марковским цепям, а оценкой вероятности N -граммы служит частота ее встречаемости:

$$\begin{aligned} \hat{p}(w) &= f(w_n | w_1, w_2, \dots, w_{n-1}) = \\ &= \frac{C(w_1, w_2, \dots, w_n | L)}{C(w_1, w_2, \dots, w_{n-1} | L)}. \end{aligned} \quad (1)$$

Формула (1) для условных вероятностей триграмм использовалась в системе распознавания речи, разработанной IBM. Эксперименты пока-

зали, что в обучающей выборке отсутствовало значительное число триграмм, обнаруженное при проверке системы. Вероятность таких триграмм по (1) равна нулю, поэтому расчет $\hat{p}(w)$ модифицируют [4].

Формальные грамматики. Порождающей грамматикой G согласно [3] называется четверка $G = \langle N, T, P, S \rangle$, где T — алфавит терминальных, а N — нетерминальных символов; $S \in N$ — начальный символ; P — набор правил порождения (подстановки), имеющих вид $\alpha \rightarrow \beta$, где α — строка, содержащая хотя бы один нетерминальный символ, β — строка, включающая символы из объединенного алфавита $V = N \cup T$. Правила подстановки также называют продукциями, а выражения в их левых частях — посылками. Говорят, что строка $\gamma = w_1\beta w_2$ выводится из $\varphi = w_1\alpha w_2$, если существует правило $\alpha \rightarrow \beta$ (здесь w_1 и w_2 — строки символов из V , возможно, пустые). Запись $\varphi \Rightarrow \gamma$ означает, что существует цепочка выводов, преобразующих строку φ в строку γ . Языком $L(G)$, порождаемым грамматикой G , называют множество всех конечных строк из символов T , выводимых в грамматике G . Множество всех непустых строк из символов алфавита R обозначают R^+ . Очевидно, что $L(G) \subset T^+$.

Наиболее исследован класс контекстно-свободных грамматик (КСГ), в которых правила подстановки имеют вид $A_i \rightarrow \beta$, где $A_i \in N$, а строка $\beta \in V^+$. В частном случае КСГ — автоматные грамматики (АГ) — правила подстановки ограничивают двумя типами: $A \rightarrow \alpha B$ и $A \rightarrow \alpha$, где $A \in N$ и $\alpha \in T$.

Определение стохастической грамматики G_s совпадает с приведенной с той лишь разницей, что все правила $P = \{\alpha_i \rightarrow \alpha_j\}$ снабжают вероятностями p_{ij} при $\sum_j p_{ij} = 1$. Несущей называют

грамматику G , получаемую из G_s выбрасыванием вероятностей. Грамматикой G_s называют согласованной, если в процессе вывода $\lim P(w^k = \{w_i | w_i \in T, i \in 1..n\}) \rightarrow 1$. Рассмотрим стохастическую КСГ (СКСГ) с посылками $\{A_i\} = N$. Для каждого A_i математическое ожидание E_{ij} числа порождаемых нетерминалов (по всем продукциям $A_i \rightarrow A_j$) рассчитывают в виде

$$E_{ij} = E(A_j | A_i) = \sum_{k(i)} p_{ik} N(j, ik),$$

где суммирование производится по всем k продукциям с посылкой A_i ; p_{ik} — вероятность продукции $A_i \rightarrow A_k$; $N(j, k)$ — число вхождений не-

терминала A_j в правую часть продукции $A_i \rightarrow A_k$. Для СКСГ выполняется $\lim_{t \rightarrow \infty} E^t \rightarrow 0$ [Jelinek 1991, Stolcke 1994].

Например, пусть $S \rightarrow A_1 A_2$ с вероятностью 1, $A_1 \rightarrow \beta A_2$ с вероятностью p_1 , $A_1 \rightarrow \eta$ с вероятностью $1 - p_1$, $A_2 \rightarrow A_1 \gamma A_1 A_1$ с вероятностью p_2 , $A_2 \rightarrow \xi$ с вероятностью $1 - p_2$. Здесь $\{A_i\} = N = \{A_0 = S, A_1, A_2\}$. Тогда матрица E имеет вид

$$E = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & p_1 \\ 0 & 3p_2 & 0 \end{bmatrix}.$$

N-граммы и формальные грамматики.

N -граммы как объект теоретического анализа недостаточно изучены. Модель N -грамм не является объяснительной и не входит ни в какую другую объяснительную модель. В качестве носителя для модели N -граммы выступает формальная грамматика. Задача заключается в том, чтобы для формальной грамматики G определить все N -граммы, допустимые в порождаемом ею языке. В вероятностной формулировке задача заключается в том, чтобы для стохастической грамматики G_s определить вероятность каждой N -граммы.

Нормальной формой Хомского (НФХ) называется такая грамматика, в которой правила подстановки имеют вид $X \rightarrow YZ$, $X \rightarrow t$, где $X, Y, Z \in N$, а $t \in T$. К НФХ приводится любая бесконтекстная грамматика [1]. Следуя [4], через $E(w|X)$ с подстроками q и r обозначим сумму $p(X \rightarrow w)$ и сумму по всем подстановкам в виде

$$E(w|X) = p(X \rightarrow w) + \sum_{X \rightarrow YZ} p(X \rightarrow YZ) \times \left(E(\omega|Y) + E(\omega|Z) + \sum_{ab=w} p(Y \Rightarrow qa) p(Z \Rightarrow br) \right).$$

Алгоритм вычисления вероятностей префиксных подстрок для СКСГ приведен в [Stolcke 1994]. Операция, состоящая в замене подстановки $X \rightarrow YZ$ на $X \rightarrow ZY$, не выводит грамматику из класса НФХ. Известны методы, приводящие КСГ к НФХ в виде инвертированной грамматики. Применяв тот же алгоритм, получим вероятности появления хвостовых подстрок для исходной грамматики.

Таким образом, N -граммы есть средство фиксации языковой реальности и модель, основанная на грамматике Хомского. Связь модельных N -грамм и формальных грамматик дает эффективный инструмент автоматического анализа

печатных текстов и слитной речи человека независимо от принадлежности языка к языковой группе.

Список литературы

1. Бузикашвили, Н. Е. Задача поиска в неструктурированном тексте и лингвистический анализ / Н. Е. Бузикашвили, Д. В. Самойлов, Л. И. Бродский, А. В. Усков // Интеллектуальные технологии ввода и обработки информации : Труды ИСА РАН. М., 1998. С. 129–141.

2. Звегинцев, В. А. Теоретическая и прикладная лингвистика / В. А. Звегинцев. 2-е изд. М., 2007. 336 с.

3. Jelinek, F. Computation of the probability of initial substring generation by stochastic context free-grammar / F. Jelinek, J. Lafferty // Computational Linguistics. Vol. 17, № 3. 1991. P. 315–323.

4. Stolcke, A. Precise n-gram probabilities from stochastic context-free grammars / A. Stolcke, J. Segal // Proceedings of the 32th Annual Meeting of ACL, 1994. P. 74–79.

Вестник Челябинского государственного университета. 2011. № 24 (239). Филология. Искусствоведение. Вып. 57. С. 71–73.

Е. И. Гуреева

РАЗНОВИДНОСТИ СПЕЦИАЛЬНЫХ ОБОЗНАЧЕНИЙ В СОВРЕМЕННОЙ СПОРТИВНОЙ ТЕРМИНОЛОГИИ

В статье рассматриваются разновидности специальных обозначений в спортивной терминологии современного русского языка. Выделяются собственно термины, номены, термины-эпонимы, а также судейские термины контроля над состязаниями.

Ключевые слова: спортивная терминология, термин, номен, эпоним, судейские термины контроля над состязаниями.

Спортивная терминология является одной из самых активно развивающихся терминологий современного русского языка. Однако системность спортивной терминологии оформилась лишь в последние десятилетия, что связано с осознанием спорта как разновидности профессиональной деятельности (ср. *профессиональный бокс, профессиональный хоккей*), а также с появлением спортивной науки. Как пишет Р. А. Пилюн, «в научной литературе не раз поднимался вопрос о необходимости завершить оформление науки о спорте как самостоятельной системы знаний» [2].

Сегодня к спортивной терминологии применяется научный подход, т. е. признается ее способность концентрировать научное знание. В подтверждение этому приведем несколько цитат из аннотации и предисловия к словарю «Терминология спорта. Толковый словарь-справочник» 2010 года издания (авторы-составители А. Н. Блеер, Ф. П. Сулов, Д. А. Тышлер): «раскрыты <...> более 10 000 терминов, употребляемых в разных видах спорта и *спортивной науке* [здесь и далее выделено нами.—Е. Г.]», «спортивные термины выходят далеко за рамки понятий,

относящихся только к данной отрасли *научных и прикладных знаний*», «в современную *научную* литературу по физической культуре и спорту все шире проникают англоязычные термины <...> и т. д.

Термин служит обозначением некоего ментального объекта, отражающего реальный объект, с которым сталкивается человек в процессе материальной или духовной деятельности. Это слово либо словосочетание, которое выражает специальное понятие, входящее в систему понятий определенной области знания/деятельности. Соответственно, терминология — это система терминов, выражающая систему понятий какой-либо области знания и/или деятельности.

В современной спортивной терминологии русского языка можно выделить несколько разновидностей специальных обозначений.

Прежде всего, это **собственно термины**, единицы, выражающие специальные (общие) понятия в области спорта. Сюда следует отнести: 1) термины с прозрачной внутренней формой (слова, которые могут быть понятны простому обывателю без специального словаря; заметим, что таких терминов в спортивной терминологии