

УДК 510.62

З. В. ДИМА, А. Ф. ОСЫКА, канд. техн. наук

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ СПРЯЖЕНИЯ НЕВОЗВРАТНЫХ ГЛАГОЛОВ РУССКОГО ЯЗЫКА

Математическое описание способности человека владеть языком — сложная и обширная задача. Ее полное решение требует усилий многих специалистов, работающих в области искусственного интеллекта.

В настоящей статье описывается моделирование русского языка лишь на морфологическом уровне, т. е. на уровне обработки отдельных слов. А это в свою очередь также является сложной задачей, поэтому ограничим ее следующим образом. Будем строить математическую модель словоизменения русского языка. Такие явления, как чередования в основе слова, беглости, постановка ударения и некоторые другие, модель не охватывают. Нас будет интересовать, как в процессе словоизменения меняется окончание той или иной словоформы, т. е. по-

пытаемся описать фрагментное отношение $L(X, Y)$, отражающее связь между смыслом X и переменным фрагментом текста Y . В качестве переменного фрагмента текста из всех морфем слова выберем окончание. В данной статье ограничимся рассмотрением окончаний глагольных словоформ. Под смыслом текста X понимаем некоторый набор признаков, заменяющий действие отброшенной части текста.

Статья посвящена выбору подходящей структуры многокомпонентных векторов $X = (x_1, x_2, \dots, x_n)$ (n — число компонентов вектора смысла) и $Y = (y_1, y_2, \dots, y_m)$ (m — число компонентов вектора фрагмента текста). В отношении вектора Y все относительно просто. В качестве фрагмента текста мы приняли окончание. Полагаем, что глагольные окончания состоят не более чем из трех букв, поэтому формально представим окончание в виде трехкомпонентного вектора $Y = (y_1, y_2, y_3)$, где y_1, y_2, y_3 — буквы, стоящие на первом, втором, третьем местах фрагмента (окончания) соответственно. Нумерация букв фрагмента производится слева направо. Переменные y_1, y_2, y_3 имеют области изменения, задаваемые следующими уравнениями алгебры конечных предикатов [3]:

$$\bar{y}_1 \vee y_1^0 \vee y_1^1 \vee y_1^2 \vee y_1^3 \vee y_1^4 \vee y_1^5 \vee y_1^6 \vee y_1^7 \vee y_1^8 \vee y_1^9 = 1;$$

$$y_2^0 \vee y_2^1 \vee y_2^2 \vee y_2^3 \vee y_2^4 \vee y_2^5 \vee y_2^6 \vee y_2^7 \vee y_2^8 \vee y_2^9 = 1;$$

$$y_3^0 \vee y_3^1 \vee y_3^2 \vee y_3^3 \vee y_3^4 \vee y_3^5 \vee y_3^6 \vee y_3^7 \vee y_3^8 \vee y_3^9 = 1. \quad (1)$$

Выбор структуры вектора X зависит от ограничений, накладываемых на модель спряжения глаголов, поэтому сначала обсудим эти ограничения.

Построение математической модели будем производить для невозвратных глаголов. Соответствующие формы возвратных глаголов, если они вообще имеются, отличаются лишь наличием конечного постфикса *-ся* (*-сь*), который присоединяется к словоформе посредством несложной формальной процедуры. В глагольную парадигму будем включать лишь личные формы (наиболее важные компоненты глагола как части речи). Причастия и деепричастия рассматриваются как самостоятельные части речи.

В качестве исходной формы для всей системы глагола примем неличную форму — инфинитив, которая не имеет форм словоизменения. Глагольная парадигма включает два залога — действительный и страдательный. В действительном залоге формы противопоставляются по наклонениям, временам, числам, лицам и (в прошедшем времени) по родам.

Модель будем строить для грамматического разряда с максимальной парадигмой, т. е. для глаголов переходных несовершенного вида (не многократных и не безличных), например: *читать* [1]. Остальные грамматические разряды глаголов имеют

сокращенные парадигмы, получение которых выходит за рамки нашей задачи.

Математическая модель не учитывает сослагательное наклонение, сложное будущее время, одну из форм первого лица множественного числа повелительного наклонения (форму обращения ко многим лицам), например, *пойдемте*, так как в данной форме четырехбуквенное окончание (а у нас с целью экономии аппаратурных средств выбран трехбуквенный регистр); присоединение частиц *-ся(-сь)* к страдательным формам и к окончаниям возвратных глаголов.

С учетом изложенных замечаний приведем пример рассматриваемой глагольной парадигмы на примере конкретного глагола: *<делать, делаю, делаешь, делает, делаем, делаете, делают, делай, делаем, делайте, делал, делала, делало, делали>*. Окончания в формах действительного и страдательного залога совпадают с точностью до постфикса *-ся(-сь)*, поэтому на данном этапе признак залога не вводится. Еще два замечания: 1) инфинитив относится к неличным формам глагола, поэтому необходимо учесть противопоставление личных и неличных форм признаком репрезентации со значениями инфинитив, личная форма; 2) глаголы совершенного и несовершенного видов имеют одинаковые окончания в настоящем и будущем времени, поэтому глагольные формы будем противопоставлять только по двум временам — непрошедшему (т. е. настоящему или будущему) и прошедшему.

Таким образом, мы имеем возможность ввести шесть компонентов вектора $X = (x_1, x_2, \dots, x_6)$, т. е. шесть признаков. Значения переменных x_1, x_2, \dots, x_6 являются значениями компонентов смысла. Признаку репрезентации сопоставим переменную x_1 ; признакам рода, числа, наклонения, времени, лица — переменные x_2, x_3, x_4, x_5, x_6 соответственно.

Признак репрезентации имеет одно из двух значений: инфинитив (и) и личная форма (л). Любому слову присуще одно из трех возможных значений рода: мужской (м), женский (ж), средний (с). Глагольные формы противопоставляются по единственному (е) и множественному (м) числу, по изъявительному (и) и повелительному (п) наклонению, по прошедшему (п) и непрошедшему (н) времени, а также по лицам — первому (1), второму (2), третьему (3). Таким образом, области изменения переменных x_1, x_2, \dots, x_6 задаются следующими уравнениями алгебры конечных предикатов:

$$\begin{aligned}x_1^и \vee x_1^л &= 1; & x_2^м \vee x_2^ж \vee x_2^с &= 1; & x_3^е \vee x_3^м &= 1; \\x_4^и \vee x_4^п &= 1; & x_5^н \vee x_5^п &= 1; & x_6^1 \vee x_6^2 \vee x_6^3 &= 1.\end{aligned}\quad (2)$$

Указанных признаков оказывается недостаточно для однозначного определения окончания конкретной словоформы. Например, если задать набор признаков $x_1^л x_3^е x_4^и x_5^н x_6^2$, то возможным

становится любое из трех окончаний *-ешь*, *-ёшь*, *-ишь*; при на боре $x_1^r x_3^e x_4^r x_5^r x_6^2$ возможно любое из двух окончаний *-у*, *-ю*.

Проведенные исследования показали, что дополнительными признаками, обеспечивающими выполнение принципа однозначности, являются следующие: x_7 — признак последней буквы основы с областью значений на множестве букв русского алфавита; x_8 — тип влияния основы со значениями первый (1), второй (2); x_9 — признак спряжения с одним из возможных значений первое (1), второе (2); x_{10} — ударность основы со значениями ударная (у), безударная (б); x_{11} — признак наличия частицы *вы-* со значениями да (д), нет (н); x_{12} — признак предпоследней буквы основы, заданный на множестве букв русского алфавита; x_{13} — признак наличия нормативных ограничений с одним из возможных значений да (д), нет (н). Области изменения переменных x_7, x_8, \dots, x_{13} следующие:

$$\begin{aligned} x_7^2 \vee x_7^6 \vee \dots \vee x_7^r &= 1; & x_8^1 \vee x_8^2 &= 1; & x_9^1 \vee x_9^2 &= 1; \\ x_{10}^y \vee x_{10}^6 &= 1; & x_{11}^d \vee x_{11}^n &= 1; & x_{12}^a \vee x_{12}^6 \vee \dots \vee x_{12}^r &= 1; \\ x_{13}^1 \vee x_{13}^2 &= 1. \end{aligned} \quad (5)$$

Таким образом, вектор смысла имеет тринадцать компонентов $X = (x_1, x_2, \dots, x_{13})$. Нетрудно убедиться в том, что указанный набор признаков не только полон, но и несократим. Исключив из набора, например, признак числа, получим не одно, а два возможных окончания *-ю*, *-ем* (*делаю*, *делаем*); исключив признак ударности основы, вновь получим неоднозначность *-ешь*, *ёшь* (*делаешь*, *делаёшь*).

Впредь фрагментное отношение $L(X, Y)$ будем называть морфологическим в связи с тем, что в качестве фрагмента Y выступает часть слова, а именно — окончание. В силу полноты набора признаков перейдем от морфологического отношения $L(X, Y)$ к морфологической функции $Y = f(X)$.

Весь набор признаков X удобно представить в виде совокупности двух относительно самостоятельных групп. В одну группу войдут признаки, характеризующие влияние основы конкретной словоформы на окончание, т. е. влияние слова на окончание (признаки x_7, x_8, \dots, x_{13}). В другую — признаки (x_1, x_2, \dots, x_6) характеризующие влияние всей остальной части отброшенного текста на окончание. Этот вид влияния назовем просто влиянием текста. Такое разделение признаков позволяет обрабатывать каждую из групп отдельно.

В [2] предлагается функцию $Y = f(X)$ представить в виде $Y = \varphi(S, t)$, где $S = \xi(X)$, $t = \eta(X)$. Для функции t аргументами являются признаки x_1, x_2, \dots, x_6 , а для функции S — x_7, x_8, \dots, x_{13} . Функции t и S построены методом, изложенным в [2]. В результате проведенного разбиения множеств значений признаков получим пронумерованные произвольным образом

классы смежности разбиения, которые и примем в качестве значений функций

$$t = \eta(x_1, x_2, \dots, x_6) \text{ и } S = \xi(x_7, x_8, \dots, x_{13}).$$

Функция $t = \eta(x_1, x_2, \dots, x_6)$ запишется в виде:

$$\begin{aligned} t^1 &= x_1^n; \quad t^2 = x_1^n x_3^e x_4^n x_5^n x_6^1; \quad t^3 = x_1^n x_3^e x_4^n x_5^n x_6^2; \\ t^4 &= x_1^n x_3^e (x_4^n x_5^n \vee x_4^n) x_6^3; \quad t^5 = x_1^n x_3^m (x_4^n x_5^n \vee x_4^n) x_6^1; \\ t^6 &= x_1^n x_3^m x_4^n x_5^n x_6^2; \quad t^7 = x_1^n x_3^m (x_4^n \vee x_4^n x_5^n) x_6^3; \\ t^8 &= x_1^n x_2^m \wedge x_3^e x_4^n x_5^n; \quad t^9 = x_1^n x_2^m x_3^e x_4^n x_5^n; \quad t^{10} = x_1^n x_2^e x_3^e x_4^n x_5^n; \\ t^{11} &= x_1^n x_3^m x_4^n x_5^n; \quad t^{12} = x_3^e x_5^n x_6^2; \quad t^{13} = x_3^e x_5^n x_6^2. \end{aligned} \quad (4)$$

Каждому значению функции t соответствует определенная группа окончаний. Например, значению $t=2$ соответствуют окончания $-y, -ю$, значению $t=7$ — окончания $-ym, -юm, -am, -ям$. Найденные значения функции t указывают на наличие 13 способов влияния текста на окончание.

В результате построения функции $S = \xi(x_7, x_8, \dots, x_{13})$ оказалось, что слово влияет на окончание 47-ю способами. Произвольно выбранные уравнения для нескольких значений функции S имеют следующий вид:

$$\begin{aligned} S^1 &= (x_7^a \vee x_7^e \vee x_7^e \vee x_7^e \vee x_7^o \vee x_7^y \vee x_7^m \vee x_7^s \vee x_7^o \vee x_7^a) x_8^1 x_9^1 \times \\ &\times x_{10}^y x_{11}^n x_{13}^1; \quad S^{24} = (x_7^m \vee x_7^a \vee x_7^m) x_8^2 x_{10}^y x_{11}^n x_{13}^1 (x_{12}^a \vee x_{12}^e \vee \\ &\vee x_{12}^e \vee x_{12}^m \vee x_{12}^o \vee x_{12}^y \vee x_{12}^m \vee x_{12}^o \vee x_{12}^o \vee x_{12}^a); \\ S^{38} &= (x_7^m \vee x_7^m \vee x_7^m) x_8^1 x_9^2 x_{10}^y x_{13}^1. \end{aligned} \quad (5)$$

Использование функций t и S позволяет компактно записать функцию $Y = f(X)$. Приведем для примера несколько уравнений:

$$\begin{aligned} y_1^a &= t^7 (S^{22} \vee S^{24} \vee S^{38} \vee S^{43}); \quad y_1^a = t^7 (S^7 \vee S^8 \vee S^{12} \vee S^{13} \vee \\ &\vee S^{14} \vee S^{15} \vee S^{23} \vee S^{24} \vee \dots \vee S^{27} \vee S^{39} \vee S^{40} \vee S^{41} \vee \\ &\vee S^{42} \vee S^{44} \vee S^{45} \vee \dots \vee S^{47}); \quad y_2^m = t^3; \quad y_2^m = t^5; \\ y_2^n &= t^9 \vee t^{10} \vee t^{11} \vee t^8 (S^1 \vee S^2 \vee \dots \vee S^{15}); \quad y_3^o = t^{10}. \end{aligned} \quad (6)$$

Ранее проведенные исследования показали, что слово влияет на окончание 160-ю способами, а текст — 19-ю. Большое число способов влияния связано с общей постановкой задачи, которая состояла в формировании полной словоформы глагола с учетом всех изменений, происходящих в ней при спряжении. В данной же статье задача сведена к формированию глагольного окончания. Эта задача органически связана с более обширной задачей формирования окончаний для всех частей речи русского языка.