УДК 004.934.11

Федоров Є.Є.

Методика нейромережного аналізу форм слів

Постановка проблеми. На сьогодні актуальною є розробка моделей і алгоритмів інтелектуальних систем керування, що базуються на природно-мовних інтерфейсах. При витягу знань із текстів важливу роль відіграє аналіз слів і їхніх форм.

Аналіз досліджень. Аналіз останніх досягнень і публікацій [1-3], присвячених проблемі аналізу і синтезу форм слів, дозволяє зробити висновок, що ці моделі не враховують морфонологічні перетворення і не використовують кількісні оцінки мовних структур.

Рішення задачі. У даній статті пропонується методика нейромережного аналізу форм слів, для якої розробляються такі правила:

а) формалізації і чисельного дослідження елементів аналізу форм слів;

б) закріплення взаємозв'язків між елементами аналізу форм слів;

в) аналізу форм слів.

Основний матеріал. Перед аналізом форм слів проводиться формалізація таких мовних конструкцій: частин мови H_i; значень Z_m^k морфологічних ознак Z^k і їхніх наборів $\overline{Z1}_j$; букв A_i і фонем Φ_i; слів $\overline{C}_r^s$; основ слів $\overline{\Lambda}_q^s$; форм основ слів $\overline{\tilde{\Lambda}}_q^s$, $\overline{\tilde{\Lambda}}_q^s$, отриманих при лінійних і нелінійних морфонологічних перетвореннях; неосновних частин слів $\overline{V}_u^s$; словозмінних афіксів (суфіксів $\overline{Б}_p^{s1}$, флексій $\overline{Б}_p^{s2}$, постфіксів $\overline{Б}_p^{s3}$) і їхніх наборів $\overline{Б1}_p^s$; буквених і фонемних послідовностей, що усікаються/нарощуються ($\overline{\tilde{\Delta}}_w^{s1}/\overline{\tilde{\Delta}}_w^{s2}$) і чергуються ($\overline{\tilde{\Delta}}_v^{s1}/\overline{\tilde{\Delta}}_v^{s2}$), і їхніх наборів $\overline{\tilde{\Delta}1}_w^s, \overline{\tilde{\Delta}1}_v^s$, що беруть участь у лінійних і нелінійних морфонологічних перетвореннях основ слів; форм слів $\overline{\Omega}_l^s$, при цьому $s=1$ – якщо буква, $s=2$ – якщо фонема.

Цим конструкціям привласнюються кількісні оцінки:

а) ранги – $r(H_i)$, $r(Z_m^k)$, $r(A_i)$, $r(\Phi_i)$;

б) інформаційні міри (що обчислюються для векторів рангів морфологічних ознак, букв і фонем) – $M(\overline{Z1}_j)$, $M(\overline{C}_r^s)$, $M(\overline{\Lambda}_q^s)$, $M(\overline{\tilde{\Lambda}}_q^s)$, $M(\overline{\tilde{\Lambda}}_q^s)$, $M(\overline{V}_u^s)$, $M(\overline{Б}_p^{sk})$, $M(\overline{Б1}_p^s)$, $M(\overline{\tilde{\Delta}}_w^{sk})$, $M(\overline{\tilde{\Delta}}_v^{sk})$, $M(\overline{\tilde{\Delta}1}_w^s)$, $M(\overline{\tilde{\Delta}1}_v^s)$, $M(\overline{\Omega}_l^s)$.

Далі вводяться матриці бінарних відносин $\Gamma(H_i,\overline{\Omega}_l^s),\Gamma(H_i,\overline{Б1}_p^s),\Gamma(\overline{Z1}_j,\overline{\Omega}_l^s),\Gamma(\overline{Z1}_j,\overline{Б1}_p^s),\Gamma(H_i,\overline{Z1}_j),\Gamma(H_i,\overline{C}_r^s),\Gamma(H_i,\overline{V}_u^s)$, що установлюють взаємозвязок між формою слова $\overline{\Omega}_l^s$ і частиною мови H_i, набором словозмінних афіксів $\overline{Б1}_p^s$ і частиною мови H_i, формою слова $\overline{\Omega}_l^s$ і набором морфологічних ознак $\overline{Z1}_j$, набором словозмінних афіксів $\overline{Б1}_p^s$ і набором морфологічних ознак $\overline{Z1}_j$, набором морфологічних ознак $\overline{Z1}_j$ і частиною мови H_i, словом $\overline{C}_r^s$ і частиною мови H_i,

неосновною частиною слова $\overline{V}_u^s$ і частиною мови H_i .

При аналізі форми слова відповідно до частини мови і набором морфологічних ознак виробляється визначення слова (у початковій формі) і набору словозмінних афіксів.

При використанні рангів й інформаційних мір мовних конструкцій, а також матриць бінарних відносин здійснюється нейромережний аналіз форм слів.

На рис.1 наведена п'ятишарова повнозв'язна з прямими і зворотними зв'язками нейронна мережа аналізу форм слів.

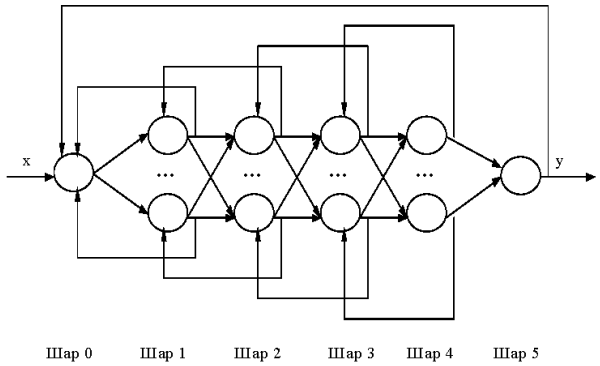

Рис. 1. Нейронна мережа аналізу форм слів

Нульовий (*вхідний*) *шар* містить один нейрон і здійснює зіставлення інформаційної міри форми слова x_1 з інформаційною мірою форми слова $M(\overline{\Omega}_l^s)$ з бази даних у вигляді

$$\rho(x_1, M(\overline{\Omega}_l^s)) = | x_1 - M(\overline{\Omega}_l^s) | \ , 77$$

зіставлення рангу частини мови слова і його форм x_2 з рангом частини мови $r(H_i)$ з бази даних у вигляді:

$$\rho(x_2, r(H_i)) = | x_2 - r(H_i) | \ ,$$

зіставлення інформаційної міри набору значень морфологічних ознак форми слова x_3 з інформаційною мірою набору значень $M(\overline{Z1}_j)$ з бази даних у вигляді:

$$\rho(x_3, M(\overline{Z1}_j)) = | x_3 - M(\overline{Z1}_j) | \ .$$

Якщо $\exists M(\overline{\Omega}_l^s) \ \ \rho(x_1, M(\overline{\Omega}_l^s)) = 0$ й $\exists r(H_i) \ \ \rho(x_2, r(H_i)) = 0$ і $\exists M(\overline{Z1}_j) \ \ \rho(x_3, M(\overline{Z1}_j)) = 0$, то встановлюється відповідність $x_1 \leftrightarrow M(\overline{\Omega}_l^s)$, $x_2 \leftrightarrow r(H_i)$, $x_3 \leftrightarrow M(\overline{Z1}_j)$ і по прямому зв'язку здійснюється перехід до першого шару.

Перший шар складається з $\eta(\overline{\mathbf{Б1}}^s)$ нейронів, кожний з яких відповідає вектору словозмінних афіксів (суфікс, флексія, постфікс) $\mu(\overline{Б1}_p^s) = (M(б_{p1}^s), M(б_{p2}^s), M(б_{p3}^s))$, $M(б_{pk}^s) \in M(\overline{\mathbf{Б}}^{sk})$ з інформаційною мірою $M(\overline{Б1}_p^s)$. На цьому шарі виконується виділення інформаційної міри форм основ (з нарощенням/усіканням і чергуванням букв/фонем) шляхом усікання інформаційної міри форми слова $M(\overline{\Omega}_l^s)$, що надійшла із вхідного шару, за допомогою інформаційних мір суфікса $M(б_{p1}^s)$, флексії $M(б_{p2}^s)$ і постфікса $M(б_{p3}^s)$ в такому вигляді:

$$M(\zeta^s) = M(\overline{\Omega}_l^s) \circ (M(б_{p1}^s) \Diamond M(б_{p2}^s) \Diamond M(б_{p3}^s)) .$$

Отриманий результат зіставляється з інформаційною мірою форми основи (з нарощенням/усіканням і чергуванням букв/фонем) $M(\overline{\widetilde{\Lambda}}_q^s)$ з бази даних у вигляді:

$$\rho(M(\zeta^s),M(\overline{\widetilde{\Lambda}}_q^s))=|M(\zeta^s)-M(\overline{\widetilde{\Lambda}}_q^s)|\ ,$$

Якщо $\forall M(\overline{\widetilde{\Lambda}}_q^s)$ $\rho(M(\zeta^s),M(\overline{\widetilde{\Lambda}}_q^s))>0$ чи $\neg(\Gamma(H_i,\overline{\Omega}_l^s)=1\wedge\Gamma(H_i,\overline{B1}_p^s)=1\wedge\Gamma(\overline{Z1}_j^s,\overline{\Omega}_l^s)=1\wedge\Gamma(\overline{Z1}_j^s,\overline{B1}_p^s)=1\wedge\Gamma(H_i,\overline{Z1}_j^s)=1)$, то по зворотному зв'язку здійснюється повернення до вхідного шару.

Якщо $\exists M(\overline{\widetilde{\Lambda}}_q^s)$ $\rho(M(\zeta^s),M(\overline{\widetilde{\Lambda}}_q^s))=0$ і $\Gamma(H_i,\overline{\Omega}_l^s)=1\wedge\Gamma(H_i,\overline{B1}_p^s)=1\wedge\Gamma(\overline{Z1}_j^s,\overline{\Omega}_l^s)=1\wedge\Gamma(\overline{Z1}_j^s,\overline{B1}_p^s)=1\wedge\Gamma(H_i,\overline{Z1}_j^s)=1$, то установлюється відповідність $M(\zeta^s)\leftrightarrow M(\overline{\widetilde{\Lambda}}_q^s)$ і по прямого зв'язку здійснюється перехід до другого шару.

Другий шар складається з $\eta(\overline{\widetilde{\Delta}\mathbf{1}}^s)$ нейронів, кожний з який відповідає вектору буквених/фонемних послідовностей $\mu(\overline{\widetilde{\Delta}1}_v^s)=(M(\widetilde{\delta}_{v1}^s),M(\widetilde{\delta}_{v2}^s))$, $M(\widetilde{\delta}_{vk}^s)\in M(\overline{\widetilde{\mathbf{\Delta}}}^{sk})$, що чергуються, з інформаційною мірою $M(\overline{\widetilde{\Delta}1}_v^s)$. На цьому шарі здійснюється формування форми основи слова, що породжує (з нарощенням/усіканням букв/фонем, без чергування) шляхом перетворення інформаційних мір форми слова (з чергуванням і нарощенням/усіканням букв/фонем$M(\overline{\widetilde{\Lambda}}_q^s)$), отриманої на першому шарі за допомогою інформаційних мір що $M(\widetilde{\delta}_{v1}^s)$чергуються $M(\widetilde{\delta}_{v2}^s)$, буквених/фонемних послідовностей у такому вигляді:

а) якщо $M(\widetilde{\delta}_{v2}^s)=0\wedge M(\widetilde{\delta}_{v1}^s)\neq 0$, тобто в одержуваній формі основи слова з`являється гласна між приголосними, що йдуть наприкінці основи, використовується правило

$$|\overline{\widetilde{\Lambda}}_q^s|=len\wedge(r(\overline{\widetilde{\Lambda}}_{qj}^s)\in\{r(T_i^s)\}\wedge r(\overline{\widetilde{\Lambda}}_{q,j+1}^s)\in\{r(T_i^s)\})\wedge$$
$$\left((j+1=len)\vee\left(\bigwedge_{k=j+1}^{len-1}\neg(r(\overline{\widetilde{\Lambda}}_{qk}^s)\in\{r(T_i^s)\}\wedge r(\overline{\widetilde{\Lambda}}_{q,k+1}^s)\in\{r(T_i^s)\})\right)\right)\rightarrow$$
$$\left(\bigwedge_{p=1}^{j}r(\zeta_p^s)=r(\overline{\widetilde{\Lambda}}_{qp}^s)\right)\wedge r(\zeta_{j+1}^s)=r(\widetilde{\delta}_{v11}^s)\wedge\left(\bigwedge_{b=j+2}^{len}r(\zeta_b^s)=r(\overline{\widetilde{\Lambda}}_{q,b-1}^s)\right),\ j\in\overline{1,|\overline{\widetilde{\Lambda}}_q^s|-1},$$

де $\{r(T_i^1)\}$ – множина приголосних,

б) якщо $M(\widetilde{\delta}_{v2}^s)\neq 0\wedge M(\widetilde{\delta}_{v1}^s)=0$, тобто в одержуваній формі основи слова зникає гласна між приголосними, що йдуть наприкінці основи, використовується правило

$$|\overline{\widetilde{\Lambda}}_q^s|=len\wedge(r(\overline{\widetilde{\Lambda}}_{q,j-1}^s)\in\{r(T_i^s)\}\wedge r(\overline{\widetilde{\Lambda}}_{qj}^s)=r(\widetilde{\delta}_{v21}^s)\wedge r(\overline{\widetilde{\Lambda}}_{q,j+1}^s)\in\{r(T_i^s)\})\wedge$$
$$((j+1=len)\vee$$
$$\left(\bigwedge_{k=j+1}^{len-1}\neg(r(\overline{\widetilde{\Lambda}}_{q,k-1}^s)\in\{r(T_i^s)\}\wedge r(\overline{\widetilde{\Lambda}}_{qk}^s)=r(\widetilde{\delta}_{v21}^s)\wedge r(\overline{\widetilde{\Lambda}}_{q,k+1}^s)\in\{r(T_i^s)\})\right)\rightarrow$$
$$\left(\bigwedge_{p=1}^{j}r(\zeta_p^s)=r(\overline{\widetilde{\Lambda}}_{qp}^s)\right)\wedge\left(\bigwedge_{b=j+1}^{len-1}r(\zeta_b^s)=r(\overline{\widetilde{\Lambda}}_{q,b+1}^s)\right),\ j\in\overline{2,|\overline{\widetilde{\Lambda}}_q^s|-1};$$

в) якщо $M(\widetilde{\delta}_{v2}^s)\neq 0\wedge M(\widetilde{\delta}_{v1}^s)\neq 0$, тобто в одержуваній формі основи слова заміняються послідовності букв/фонем, що йдуть наприкінці основи, використовується правило

$$|\widetilde{\delta}_{v2}^s|=len1\wedge|\widetilde{\delta}_{v1}^s|=len2\wedge|\overline{\widetilde{\Lambda}}_q^s|=len3\wedge\left(\bigwedge_{m=1}^{len1}r(\overline{\widetilde{\Lambda}}_{q,j+m}^s)=r(\widetilde{\delta}_{v2m}^s)\right)\wedge$$

$$\left((j+len1=len3)\vee\bigwedge_{k=j+1}^{len3}\neg\left(\bigwedge_{n=1}^{len1}(r(\overline{\widetilde{\Lambda}}^{s}_{q,k+n})=r(\widetilde{\delta}^{s}_{v2n}))\right)\right)\rightarrow$$

$$\left(j=0\vee\left(\bigwedge_{p=1}^{j}r(\zeta^{s}_{p})=r(\overline{\widetilde{\Lambda}}^{s}_{qp})\right)\right)\wedge\left(\bigwedge_{z=1}^{len2}r(\zeta^{s}_{(j+1)+z})=r(\widetilde{\delta}^{s}_{v1z})\right)\wedge$$

$$\left(\bigwedge_{b=1}^{len3-(j+len1)}r(\zeta^{s}_{(j+len2+1)+b})=r(\overline{\widetilde{\Lambda}}^{s}_{q,(j+len1+1)+b})\right),\ j\in\overline{0,|\overline{\widetilde{\Lambda}}^{s}_{q}|-1}.$$

Таким чином, формується вектор рангів основи

$$\mu(\zeta^{s})=(r(\zeta^{s}_{1}),...,r(\zeta^{s}_{t}),...,r(\zeta^{s}_{M})).$$

Інформаційна міра цього вектора обчислюється у виді

$$M(\zeta^{s})=\left\|\mu(\zeta^{s})\right\|=\sum_{t=1}^{|\zeta^{s}|}\left(r(\zeta^{s}_{t})*2^{(|\zeta^{s}|-t)n}\right)$$

де n – кількість біт, що приходяться на одну букву/фонему.

Отриманий результат зіставляється з інформаційною мірою форми основи слова $M(\overline{\widetilde{\Lambda}}^{s}_{q})$ з бази даних у вигляді:

$$\rho(\zeta^{s},M(\overline{\widetilde{\Lambda}1}^{s}_{q}))=|\zeta^{s}-M(\overline{\widetilde{\Lambda}}^{s}_{q})|$$

Якщо $\forall M(\overline{\widetilde{\Lambda}1}^{s}_{q})\ \ \rho(\zeta^{s},M(\overline{\widetilde{\Lambda}}^{s}_{q}))>0$, то по зворотному зв'язку здійснюється повернення до першого шару.

Якщо $\exists M(\overline{\widetilde{\Lambda}1}^{s}_{q})\ \ \rho(\zeta^{s},M(\overline{\widetilde{\Lambda}}^{s}_{q}))=0$, то встановлюється відповідність $\zeta^{s}\leftrightarrow M(\overline{\widetilde{\Lambda}1}^{s}_{q})$ і по прямому зв'язку здійснюється перехід до третього шару.

Третій шар складається з $\eta(\overline{\widetilde{\Delta}\mathbf{1}}^{s})$ нейронів, кожний з який відповідає вектору буквених/фонемних послідовностей $\mu(\overline{\widetilde{\Delta}1}^{s}_{w})=(M(\widetilde{\delta}^{s}_{w1}),M(\widetilde{\delta}^{s}_{w2}))$, $M(\widetilde{\delta}^{s}_{wk})\in M(\overline{\widetilde{\Delta}}^{sk})$, що нарощуються/усікаються, з інформаційною мірою $M(\overline{\widetilde{\Delta}1}^{s}_{w})$. На цьому шарі здійснюється формування основи слова, що породжує шляхом перетворення інформаційної міри форми слова (з нарощенням/усіканням букв/фонем, без чергування $M(\widehat{\Delta}^{s}_{q})$), отриманої на другому шарі, за допомогою інформаційних мір, що $M(\widetilde{\delta}^{s}_{w1})$ усікаються і $M(\widetilde{\delta}^{s}_{w2})$ нарощуваних буквених/фонемних послідовностей у вигляді:

$$M(\zeta^{s})=(M(\widehat{\Delta}^{s}_{q})\circ M(\widetilde{\delta}^{s}_{w2}))\lozenge M(\widetilde{\delta}^{s}_{w1}).$$

Отриманий результат зіставляється з інформаційною мірою форми основи $M(\overline{\Lambda}^{s}_{q})$ з бази даних у вигляді:

$$\rho(\zeta^{s},M(\overline{\Lambda}^{s}_{q}))=|\zeta^{s}-M(\overline{\Lambda}^{s}_{q})|.$$

Якщо $\forall M(\overline{\Lambda}^{s}_{q})\ \ \rho(M(\zeta^{s}),M(\overline{\Lambda}^{s}_{q})>0$, то по зворотному зв'язку здійснюється повернення до другого шару.

Якщо $\exists M(\overline{\Lambda}^{s}_{q})\ \ \rho(M(\zeta^{s}),M(\overline{\Lambda}^{s}_{q}))=0$, то установлюється відповідність $M(\zeta^{s})\leftrightarrow M(\overline{\Lambda}^{s}_{q})$ і по прямому зв'язку здійснюється перехід до четвертого шару.

Четвертий шар складається з $\eta(\overline{\mathbf{V}}^{s})$ нейронів, кожний з який відповідає інформаційній мірі неосновної частини слова $M(\overline{V}^{s}_{u})$. На цьому шарі виконується формування слова, що породжує, шляхом конкатенації інформаційної міри основи

слова, що $M(\overline{\Lambda}_q^s)$ породжує , отриманої на третьому шарі, за допомогою інформаційної міри неосновної частини $M(\overline{V}_u^s)$ слова в такому вигляді:

$$M(\zeta^s) = M(\overline{\Lambda}_q^s) \lozenge M(\overline{V}_u^s).$$

Отриманий результат зіставляється з інформаційною мірою слова, що $M(\overline{C}_r^s)$ породжує, з бази даних у вигляді

$$\rho(M(\zeta^s), M(\overline{C}_r^s)) = |M(\zeta^s) - M(\overline{C}_r^s)|.$$

Якщо $\forall M(\overline{C}_r^s) \quad \rho(M(\zeta^s), M(\overline{C}_r^s)) > 0$ чи $\neg(\Gamma(H_i, \overline{C}_r^s) = 1 \wedge \Gamma(H_i, \overline{V}_u^s) = 1)$, то по зворотному зв'язку здійснюється повернення до другого шару.

Якщо $\exists M(\overline{C}_r^s) \quad \rho(M(\zeta^s), M(\overline{C}_r^s)) = 0$ і $\Gamma(H_i, \overline{C}_r^s) = 1 \wedge \Gamma(H_i, \overline{V}_u^s) = 1$, то встановлюється відповідність $M(\zeta^s) \leftrightarrow M(\overline{C}_r^s)$ і по прямому зв'язку здійснюється перехід до п'ятого шару.

П'ятий (*вихідний*) *шар* містить один нейрон і здійснює зіставлення сформованої на третьому шарі інформаційної міри слова $M(C_r^s)$, що породжує, отриманого на четвертому шарі, з інформаційною мірою очікуваного на виході слова y_1, що породжує, у вигляді:

$$\rho(M(C_r^s), y_1) = |M(C_r^s) - y_1|.$$

Якщо $\rho(M(C_r^s), y_1) > 0$, то по зворотному зв'язку здійснюється повернення до вхідного шару.

Таким чином встановлюється асоціативний зв'язок $\overline{\Omega}_l^s \to \overline{Б1}_p^s \to \overline{\tilde{\Delta}1}_v^s \to \overline{\hat{\Delta}1}_w^s \to \overline{V}_u^s \to \overline{C}_r^s$

Наукова новизна. У даній роботі розроблені правила лінійних і нелінійних морфонологічних перетворень. Для роботи з мовними конструкціями використовувалися кількісні оцінки – ранги й інформаційні міри.

Практична значимість. Основні положення роботи можуть бути реалізовані в інтелектуальній системі у вигляді алгоритмів, що забезпечують спілкування з користувачем природною мовою.

Література

1. Кисленко Ю.И. Информационный робот // Искусственный интеллект. – 2001. – № 1. – С. 61-68.
2. Cheblakov G.B., Dinenberg F.G., Levin D.Y., Popov I.G., Zagorulko Y.A. Approach to development of a system for speech interaction with an intelligent robot // Perspectives of System Informatics. – Berlin: Springer-Verlag. – 1999. – P. 517-529.
3. Литвинцева Л.В., Ульянов С.В., Танака Т., Охви Д., Ямафудзи К. Интеллектуальное управление мобильным роботом сервисного обслуживания // Труды VIII науч.-техн. конф. "Экстремальная робототехника". – Санкт-Петербург. – 1997. – С. 35-43.
4. Современный русский язык: Учеб. для филол. спец. высших учебных заведений / *В.А.Белошапкова, Е.А.Брызгунова, Е.А.Земская и др.;* Под ред. *В.А.Белошапковой.* – М.: Азбуковник, 1997. – 928 с.
5. Русская грамматика: В 2 т.– М.: Наука, 1980. – т. 1: Фонетика. Словообразование. Морфология. – 784 с.
6. *Криводубский О.А., Федоров Е.Е.* Формальное представление русского языка и речи // Искусственный интеллект. – 2003.– № 4.– С. 402-410.