РОЗПІЗНАВАННЯ ПРИРОДНОЇ МОВИ НА ГРАМАТИЧНИХ МАРКОВСЬКИХ МЕРЕЖАХ

Т.В. Грищук
Вінницький національний технічний університет
Україна, м. Вінниця, вул. Хмельницьке шосе, 95,
e-mail: thryshuk@cms.com.ua

Abstract
The mathematical approach to the speech recognition systems modeling is offered. Its advantage is in joining two processes: generating and recognizing the phrases of the discourse. A new theory - the theory of the Grammar Markov Nets - is based on the theory of the Hidden Markov Models and on the theory of the context-free grammars.

Вступ

Розробкою систем автоматичного розпізнавання природної мови займаються наукові колективи в усьому світі. Особливо актуальною дана проблема стала в останні десятиліття в зв'язку з розповсюдженням мовних інтерфейсів. Через велику складність природної мови створення систем розпізнавання мови (СРМ) є спільною задачею математиків, лінгвістів та програмістів.

Основними етапами, з яких складається процес розпізнавання довільного мовного сигналу, є параметрична обробка сигналу, синтаксичний аналіз та класифікація. Дана стаття присвячена огляду двох останніх етапів.

Постановка задачі дослідження

В загальному вигляді модель розпізнавання мови містить такі складові елементи:

- дискурс, що формальним чином описується за допомогою граматики; кожна фраза дискурсу відповідає виключно одному з класів мовних образів $\{w_1,...,w_N\}$, що розпізнаються;
- механізм виводу фраз граматики;
- механізм класифікації мовних образів.

Основою системи розпізнавання мови є граматика. Граматикою ми називаємо четвірку

$$G = (V_N, V_T, P, S),$$

де V_N - множина нетермінальних символів;

V_T - множина термінальних символів;

P - множина граматичних правил або правил підстановки;

S - початковий або кореневий символ.

Кожному нетермінальному символу з множини V_N ставиться у відповідність автоматичний класифікатор, який визначає ступінь належності мовного сигналу до даної лексичної одиниці. Найбільш поширеними типами класифікаторів мовних сигналів на сьогодні є приховані марковські моделі [1, 2] та штучні нейронні мережі [3].

В процесі розпізнавання на основі граматики G породжуються фрази дискурсу $L(G)$. Основною проблемою обробки природної мови є мовна неоднозначність. Існують різні типи неоднозначності: синтаксична (структурна), змістовна, відмінкова, референційна та літераційна [4]. В даній статті основна увага приділяється задачі усунення синтаксичної

неоднозначності. Така неоднозначність виникає у випадках, коли одне й те саме слово може мати різне значення в залежності від порядку розташування в реченні, наприклад, термінал "color" в таких словосполученнях: "color object" та "this color". В першому випадку дане слово є прикметником, а в другому – іменником.

Після завершення процесу генерації фраз виконується процес конкатенації класифікаторів відповідних термінальних символів. В результаті ми отримуємо класифікатори для визначення належності мовного сигналу, що розпізнається, до одного з класів мовних образів.

Недоліком представленого підходу до розпізнавання мовних образів є те, що процеси граматичного виводу і класифікації виконуються на основі використання різних математичних апаратів та розділені в часі [1].

В даній статті описується математична модель генерації і розпізнавання фраз системи розпізнавання мови, що дозволяє об'єднати процес граматичного виводу фраз граматики та процес розрахунку ймовірності належності фрази, яку вимовляє користувач, до дискурсу СРМ.

Модифікований синтаксис запису формальних граматик

Одним з найпоширеніших способів запису граматик є форми Бекуса-Наура або БНФ. Згідно даного синтаксису, нетермінальні символи прийнято позначати заголовними літерами в дужках (<A>, <B>, …), а термінальні символи – маленькими літерами (a, b, c, …) [5, 6]. Граматичні правила складаються з двох частин: правої і лівої, які розділяються символом :=. Загальний вигляд правила:

$$< A >:= \alpha ,$$

де α - граматичний ланцюжок, що складається з термінальних та нетермінальних символів.

Нехай ми маємо граматику $G1$:

0. <S>:= <M>;
1. <M>:= <A>;
2. <M>:= <B>;
3. <A>:= x y;
4. <A>:= x;
5. <B>:= y;

В процесі синтаксичного виводу виконується послідовність підстановок один правил в інші. Вивід починається з кореневого символу <S>. Так фразу x можна отримати шляхом послідовних підстановок правила 1 в правило 0 та правила 4 в правило 1:

$$\begin{array}{ccccccc} < S > & \Rightarrow & < M > & \Rightarrow & < A > & \Rightarrow & x \\ \uparrow & & \uparrow & & \uparrow & & \\ 0 & & 1 & & 4 & & \end{array}$$

Зрозуміло, що описана вище схема не є ефективною, так як вимагає повного перебору граматичних правил та постійної "прив'язки" механізму породження фраз до правил граматики.

Автором пропонується модель виводу фраз граматики, яка базується на модифікованому синтаксисі запису контекстно-вільних граматик.

Розглянемо довільне правило граматики:

$$< M >:= x_1 \, x_2 \ldots x_n , \qquad (1)$$

де $< M >$ - нетермінальний символ;

x_i - термінальний чи нетермінальний символ граматики.

В модифікованому синтаксисі кожний нетермінальний символ $< X >$ розбиваємо на два символи: $< X$ і $X >$, які будемо називати, відповідно, лівою і правою дужкою нетерміналу. Тобто ми розширюємо множину нетермінальних символів:

$$V_N = V_{N_{left}} \cup V_{N_{right}},$$

де $V_{N_{left}}$ - символи лівих дужок нетерміналів ($< nonterm$);

$V_{N_{right}}$ - символи правих дужок нетерміналів ($nonterm >$).

Надалі будемо називати новий спосіб запису граматик дужковим.

Ліву і праву дужки нетерміналу $< M >$ з (1) помістимо, відповідно, на початку та на закінченні продукції.

Розглянемо таку продукцію:

$$< M >:= x < A > y.$$

В дужковому записі отримуємо продукцію:

$$< M \ x < A \ A > \ y \ M >.$$

Граматика $G1$ в дужковому синтаксисі має такий вигляд:

<S <M M> S>;
<M <A A> M>;
<M <B B> M>;
<A x y A>;
<A x A>;
<B y B>;

Вивід фраз граматики

Наступним етапом є розробка методу генерування фраз дискурсу СРМ.

Введемо поняття граматичної мережі.

Граматична мережа (ГМ) – це зважений граф, вершинами якого є символи дужкового синтаксису граматики G, а дуги R проводяться згідно правил P.

$$G - net = (V_T \cup V_N, R).$$

Навантаження дуг графу представлені не в чисельному вигляді, а в вигляді вихідних конструкцій з правил граматики.

Навантаження – це правило граматики в модифікованому синтаксисі без першого входу в нетермінал і останнього виходу з нетерміналу.

Таким чином, ГМ містить $2n + k$ вершин, де n - кількість нетермінальних символів граматики, а k - кількість термінальних символів граматики.

Правило побудови ребер графу граматики:

> Проглядаємо зліва направо правило в модифікованому синтаксисі, з'єднуючи вершини. При цьому необхідно пам'ятати, що вершину типу $< nonterm$ не з'єднуємо з вершиною $nonterm >$.

Нехай продукція в дужковому синтаксисі має вигляд:

$$x_1 ... x_i x_{i+1} ... x_n. \qquad (2)$$

Пара $x_i \ x_{i+1}$ задає ребро тільки в тому випадку, якщо x_i і x_{i+1} не є, відповідно, лівою і правою дужкою одного терміналу (рис. 1).

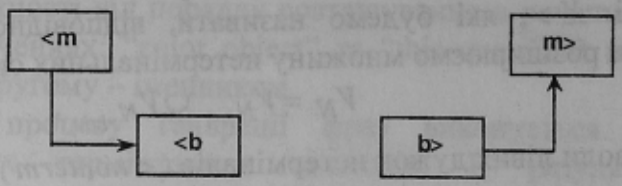

Рис. 1. Фрагмент ГМ, що відповідає продукції <m <b b> m>

Нехай в дужковому синтаксисі деякої продукції (2) пара $x_i\ x_{i+1}$ є ребром, а x_i - ліва дужка деякого нетермінала. Тоді, ребро $x_i\ x_{i+1}$ навантажується ланцюжком $x_2 \ldots x_i\ x_{i+1} \ldots x_{n-1}$.

Для граматики $G1$ граматична мережа має вигляд, який представлено на рис.2.

Розглянемо ГМ як недетермінований пристрій для синтаксичного виводу.

Вивід фраз граматики необхідно починати з асоціації з початковою вершиною ГМ (це ліва дужка початкового нетермінала) ланцюжка:

$$< S\ \ S >$$
$$\uparrow$$

Опишемо побудову деякого шляху з вершини <S у вершину S>.

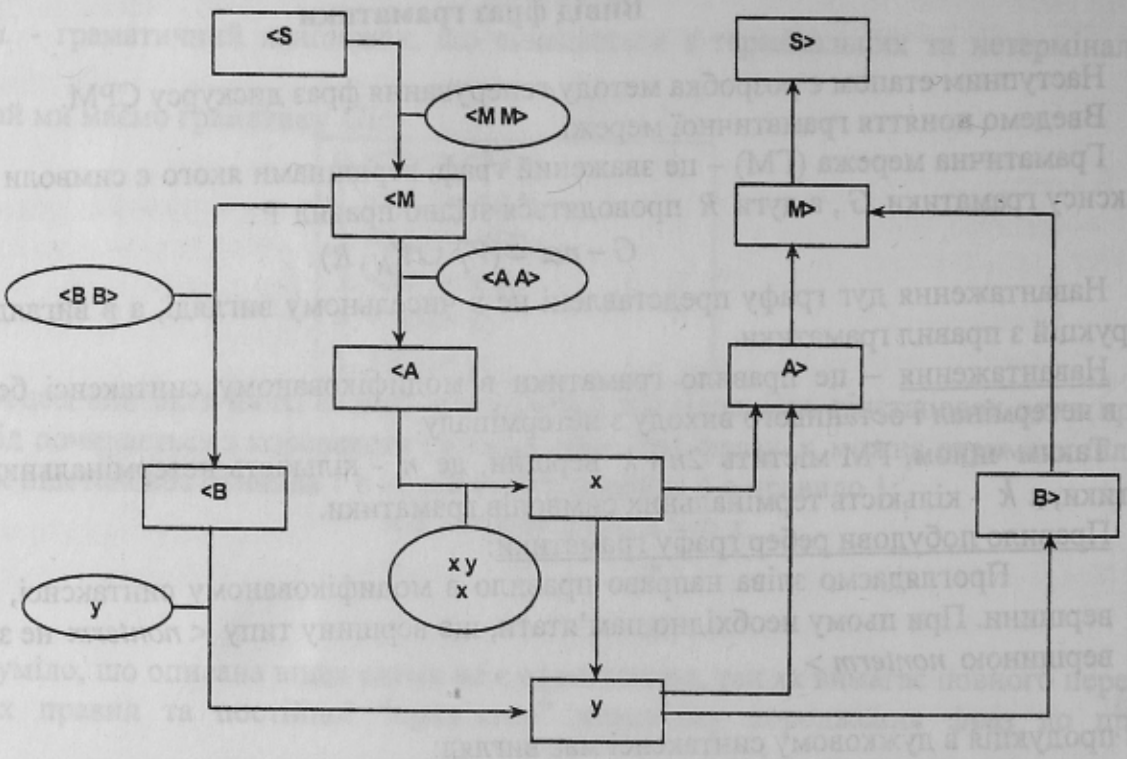

Рис. 2. Приклад ГМ

Розглянемо поточну вершину даного шляху і асоційований з нею ланцюжок:

$$x_1 \ldots x_k\ \ x_{k+1} \ldots x_p$$
$$\uparrow$$

Стрілка направлена на поточну вершину x_k.

Існує 2 правила переходу по ребрам ГМ.

Правило 1. Вершина x_k - ліва дужка нетермінала (тобто має вигляд $< X$). В даному випадку здійснюється перехід по довільному ребру. При цьому необхідно замінити символ x_k на довільне навантаження цього ребра. Стрілка направлена на вершину, до якої відбувається перехід.

Правило 2. Вершина x_k не є вершиною виду $< X$. В цьому випадку відбувається перехід по ребру до вершини, яка безпосередньо вказана за символом, на який направлена стрілка. Стрілка при цьому зміщується на символ вправо. Відзначимо, що якщо стрілка ланцюжка направлена на символ виду $X >$, то при переході до наступної вершини цей символ вилучається з ланцюжка.

Фінальним ланцюжком буде одна з фраз граматики.

Граматичний вивід фрази "xy"на ГМ граматики $G1$ символьно ілюструється так:

$$< \underset{\uparrow}{S}\, S > \Rightarrow < \underset{\uparrow}{M}\, M > S > \Rightarrow < \underset{\uparrow}{A}\, A > M > S > \Rightarrow \underset{\uparrow}{x}\, y\, A > M > S > \Rightarrow$$

$$\Rightarrow x\, \underset{\uparrow}{y}\, A > M > S > \Rightarrow x\, y\, \underset{\uparrow}{A} > M > S > \Rightarrow x\, y\, \underset{\uparrow}{M} > S > \Rightarrow x\, y\, \underset{\uparrow}{S} > \Rightarrow x\, y$$

Для порівняння розглянемо класичний вивід цієї фрази:

$$< S > \Rightarrow < M > \Rightarrow < A > \Rightarrow xy$$

$$\uparrow \qquad \uparrow \qquad \uparrow$$

$$0 \qquad 1 \qquad 3$$

Вивід фрази в обох випадках виконується за 3 ітерації, але у випадку породження фрази на ГМ потрібно ще декілька ітерацій на виключення неінформаційних символів. Синтаксична неоднозначність у випадку використання ГМ повністю виключається, так як в процесі виводу чітко зрозуміло, яким символом (символами) повинен закінчуватися граматичний ланцюжок. До того ж, не потрібно весь час звертатись до правил граматики, так як на базі граматичних правил будується граф.

Моделювання дискурсу СРМ

Нехай ГМ – граматична мережа деякої граматики. Поставимо у відповідність кожній термінальній вершині даної мережі її класифікатор - приховану марковську модель (НММ). Отриману таким чином мережу будемо називати граматичною марковською мережею ГММ.

На рис. 3 показаний фрагмент ГММ для правила $< A\ x\ A >$.

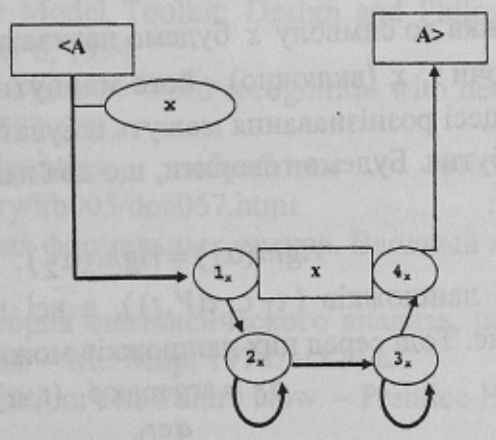

Рис. 3. Приклад ГММ

Вершини марковської мережі термінала x будемо позначати:

$$1_x, 2_x, ..., (N-1)_x, N_x .$$

Частина ребер ГММ – це тонкі ребра (стрілки), а частина – товсті. Це пов'язано з правилами перебігу часу в граматичній марковській мережі. Перехід по тій або іншій стрілці мережі або збільшує значення часу на 1, або залишає значення часу незмінним.

У початковий момент часу t = 0 "ми знаходимося" в вершині $< S$. Перехід по стрілці, що веде з деякої нетермінальної або термінальної вершини в початкову вершину деякої марковської моделі, супроводжується збільшенням часу на 1. Перехід по стрілці з емісійної в емісійну вершину деякої марковської моделі також супроводжується інкрементом часу. В усіх інших випадках час "не тече".

Процедура розпізнавання на граматичних марковських мережах

Математичний апарат граматичних марковських мереж дозволяє розробити новий алгоритм розпізнавання, що є переносом на ГММ класичного алгоритму Вітербі [2, 7].

Визначення:

Нехай $O = o_1, o_2, ... o_T$ - обсервація мовного сигналу, а M - прихована марковська модель.

Тоді величину

$$P_{\max}(O \mid M) = \max_X P(O \mid M)$$

будемо називати оцінкою Вітербі обсервації O в моделі M.

В кожний момент часу t з кожною вершиною V ГММ пов'язується деяка множина граматичних ланцюжків і їх оцінок Вітербі:

$$\tau(V,t) = \{< \alpha, \phi_V(t,\alpha) >\},$$

де α - ланцюжок;

$\phi_V(t,\alpha)$ - оцінка Вітербі граматичного ланцюжка α в вершині V в момент часу t.

Ланцюжки генеруються за допомогою механізму граматичного виводу, а їх оцінки розраховуються за рекурентними формулами [1]. Таким чином на структурі ГММ паралельно виконуються граматична обробка формальної граматики і розрахунок ймовірностей належності мовного сигналу до кожної з фраз дискурсу.

Нехай в процесі генерації ланцюжок α має вигляд:

$$...x...$$
$$\uparrow$$

Тоді частину ланцюжка до символу x будемо називати **минулим** ланцюжка $left(\alpha)$, а частину ланцюжка починаючи з x (включно) – його **майбутнім** $right(\alpha)$.

Зрозуміло, що в процесі розпізнавання можуть існувати ланцюжки, що мають спільне граматична минуле чи майбутнє. Будемо говорити, що два ланцюжка α_1 і α_2 мають спільне граматичне майбутнє, якщо:

$$right(\alpha_1) = right(\alpha_2).$$

Нехай τ_1 - множина ланцюжків ($\tau_1 \subset \tau(V,t)$), а всі ланцюжки множини τ_1 мають спільне граматичне майбутнє. Тоді серед цих ланцюжків можна обрати єдиний:

$$\alpha_0 = \arg\max_{\alpha \in \tau_1} \phi_V(t,\alpha). \qquad (3)$$

Два ланцюжка α_1 і α_2 мають спільне граматичне минуле, якщо:

$$left(\alpha_1) = left(\alpha_2).$$

Нехай τ_2 - множина ланцюжків ($\tau_2 \subset \tau(V,t)$), а всі ланцюжки множини τ_2 мають спільне граматичне минуле. Тоді серед цих ланцюжків можна обрати єдиний:

$$\alpha_0 = \arg\max_{\alpha \in \tau_2} \phi_V(t,\alpha). \quad (4)$$

Таким чином, запропонована математична модель дискурсу у вигляді ГММ дозволяє виконувати неризиковане скорочення варіантів граматичного перебору, що не призводить до втрати потенціальних кандидатів на розпізнавання.

Загальний алгоритм розпізнавання:

1. **В момент часу t виконується асоціювання граматичних ланцюжків з вершинами мережі.**
2. **Розраховуються оцінки Вітербі для кожної вершини.**
3. **Визначаються групи ланцюжків мережі, що мають спільне граматичне майбутнє.**
4. **В кожній групі ланцюжків залишається тільки найкращий кандидат.**
5. **Якщо $t = T$, то переходимо до п. 6, інакше повторюємо п. 1-4.**
6. **За максимумом оцінки Вітербі визначається найкращий кандидат-ланцюжок, який оголошується розпізнаним.**

Таким чином, в даному алгоритмі одночасно виконується синтаксична обробка фраз і їх розпізнавання.

Висновки

Математичний апарат граматичних марковських мереж дозволяє об'єднати процес граматичного виводу фраз граматики та процес розрахунку ймовірності належності фрази, яку вимовляє користувач, до дискурсу системи розпізнавання мови. Основою запропонованої моделі мовного сигналу є модифікований спосіб запису контекстно-вільних граматик, який дозволяє представляти граматику у вигляді навантаженого орієнтованого графа. Описаний в статті алгоритм розпізнавання дозволяє проводити неризиковані скорочення варіантів перебору, що не призводять до втрати потенціальних кандидатів на розпізнавання.

Література

1. Levinson S., Rabiner L., Sondhi M. An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition - Bell Systems Technical Journal, Vol. 62, No.4, p.p. 1035-1074, 1983.
2. The HTKHidden Markov Model Toolkit: Design and Philosophy. S.J. Young, CUED/F-INFNG/TR.152, September 6, 1994.
3. Albesano D., Gemello R, Mana F. Word recognition with neural network// CSELT Techn. Repts. – 1992. – №6. – P.553-559.
4. Попов А.А. Природа обработки естественного языка // http://prof9.narod.ru/library/lib005/doc067.html
5. Рейуорд-Смит Дж. Теория формальных языков. Вводный кур: Пер. с англ. – М.: Радио и связь, 1988. – 129 с.
6. Ахо А., Ульман Дж. Теория синтаксического анализа, перевода и компиляции / т.1 «Синтаксический анализ»/ – М.: Мир, 1978. – 224 с.
7. Koerner S. Speech Recognition: The Future Now. – Prentice-Hall, 1997.–390 p.