Розробка моделей відкритого тексту на основі стохастичних контекстно-вільних граматик в нормальній формі Грейбах

Грубіян Євген Олександрович

Керівник: к.ф.-м.н. Фесенко Андрій В'ячеславович

Актуальність

Мови та граматики моделюють складні внутрішні структури багатьох об'єктів із різних галузей науки.

- Формалізація природної мови - одна із найважливіших задач штучного інтелекту[Jurafsky, 2009].
- В криптоаналізі вдала модель відкритого тексту дає криптоаналітику додаткову інформацію для зламу[Shannon, 1948; Яглом, 1973].
- Структури послідовностей нуклеотидів в ДНК та РНК утворюються за граматичними правилами[Sakakibara, 1994]

...Актуальність

Jabberwocky, Льюїс Керрол

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

"Beware the Jabberwock, my son!
The jaws that bite, the claws that catch!
Beware the Jubjub bird, and shun
The frumious Bandersnatch!"

Мета, об'єкт та предмет дослідження

Мета

Побудова нової моделі природних мов, що дасть змогу краще використовувати структуру мов для методів криптоаналізу

Об'єкт

Формальна модель природних мов

Предмет

Побудова поліпшеної моделі природних мов на основі стохастичної контекстно-вільної граматики в нормальній формі Грейбах

Завдання

1. Провести огляд опублікованих робіт за тематикою дослідження
2. Розробити модель стохастичних регулярних граматик на основі прихованих моделей Маркова
3. Побудувати модель стохастичних контекстно-вільних(КВ) граматик на основі узагальнення прихованих моделей Маркова та дослідити властивості цієї моделі
4. Розробити програмну реалізацію моделей
5. Перевірити узгодженість моделей емпіричним шляхом

Моделі відкритого тексту

- Класичні моделі M0, M1, M2, M3
- n-грамна модель
- Позиційні моделі
- Моделі на основі нейронних мереж
- Граматичні моделі

Визначення 1.7

Контекстно-вільною граматикою G називається граматика $G = \langle N, \Sigma, S, R \rangle$, всі правила виводу R якої мають вигляд:

$$A \to \gamma, \ A \in N, \ \gamma \in (N \cup \Sigma)^*$$

Модель стохастичної регулярної граматики

Теорема 2.1

Для кожної стохастичної регулярної граматики існує прихована модель Маркова другого порядку за спостереженнями, що допускає таку саму стохастичну мову.

Твердження 2.1

Алгоритм перетворення стохастичної регулярної граматики до прихованої моделі Маркова другого порядку за спостереженнями має складність $\mathcal{O}(n)$, де n - кількість правил виводу в граматиці.

Модель стохастичної КВ граматики

Визначення 2.4

Прихованою моделлю Маркова другого порядку за спостереженнями зі стеком з простором латентних(прихованих) станів $E = \{1, \ldots, n\}$, простором спостережень $O = \{1, \ldots, m\}$ та стеком називається стохастичний процес $\{(X_t, S_t, Y_t),\ t \in \mathbb{N}\},\quad X_t \in E,\ Y_t \in O$, S_t - стек над множиною E, переходи в якому здійснюються за алгоритмом:

- Ініціалізація стеку: $S_1 = \emptyset,\ S_t = S_{t-1}$

Модель стохастичної КВ граматики

- Перехід між латентними станами:

$$\mathrm{P}(\mathrm{X_t = i},\ \mathrm{S_t.Push(k)} | \mathrm{X_{1:t-1}}) = \mathrm{P}(\mathrm{X_t = i},\ \mathrm{S_t.Push(k)} | \mathrm{X_{t-1}}),$$
$$\mathrm{P}(\mathrm{X_t = i},\ \mathrm{S_t.Push(k)} | \mathrm{X_{t-1} = j}) = \mathrm{p_{jik}} \quad (1)$$

- Спостереження:

$$\mathrm{P}(\mathrm{Y_t = i} | \mathrm{X_{1:t}}, \mathrm{S_t.Top}) = \mathrm{P}(\mathrm{Y_t = i} | \mathrm{X_t}, \mathrm{X_{t-1}}, \mathrm{S_t.Top}),$$
$$\mathrm{P}(\mathrm{Y_t = i} | \mathrm{X_t = j}, \mathrm{X_{t-1} = k}, \mathrm{S_t.Top = l}) = \mathrm{q_{lkji}} \quad (2)$$

- Якщо $\mathrm{X_t} = \epsilon$ та $\mathrm{S_t.Top} \neq \epsilon$, то $\mathrm{X_t} := \mathrm{S_t.Top},\ \mathrm{S_t.Pop}$.
  Якщо $\mathrm{X_t} = \epsilon$ та $\mathrm{S_t.Top} = \epsilon$, то ланцюг завершується.

Модель стохастичної КВ граматики

Теорема 2.2

Для кожної стохастичної контекстно-вільної граматики(SCFG) існує прихована модель Маркова другого порядку за спостереженнями зі стеком(HMM2S), що допускає таку саму стохастичну мову.

Твердження 2.2

Алгоритм перетворення стохастичної граматики в 2-нормальній формі Грейбах до прихованої моделі Маркова другого порядку за спостереженнями зі стеком має складність $\mathcal{O}(n)$, де n - кількість правил виводу в граматиці.

Узгодженість моделей

Було перевірено емпіричним шляхом узгодженість запропонованої моделі і стохастичних КВ граматик за допомогою розробленого програмного пакету:
`https://github.com/juja256/hmm2s`

Тестова граматика

$$(0.2)\mathrm{A}_1 \rightarrow \mathrm{a}\mathrm{A}_3\mathrm{A}_2$$
$$(0.8)\mathrm{A}_1 \rightarrow \mathrm{c}\mathrm{A}_2$$
$$(0.5)\mathrm{A}_2 \rightarrow \mathrm{a}$$
$$(0.5)\mathrm{A}_2 \rightarrow \mathrm{b}$$

$$(0.3)\mathrm{A}_3 \rightarrow \mathrm{d}\mathrm{A}_2$$
$$(0.3)\mathrm{A}_3 \rightarrow \mathrm{c}\mathrm{A}_2$$
$$(0.4)\mathrm{A}_3 \rightarrow \mathrm{a}$$

Узгодженість моделей

Слово	Частота / SCFG	Частота / HMM2S
a a a	5	1
a a b	1	2
a c a a	1	1
a c a b	3	0
a c b a	2	0
a c b b	3	2
a d a a	0	2
a d a b	0	2
a d b a	1	1
a d b b	1	3
c a	45	45
c b	38	41

Висновки

- Розроблено нову модель природної мови(відкритого тексту) на основі запропонованої прихованої моделі Маркова другого порядку за спостереженнями зі стеком
- Доведено еквівалентність запропонованої моделі та стохастичних КВ граматик та запропоновано поліноміальний алгоритм зведення стохастичної КВ граматики до запропонованої моделі
- Емпірично підтверджено узгодженість моделі і стохастичної КВ граматики
- Нова модель дозволяє розробку ефективніших алгоритмів навчання стохастичних КВ граматик