Математическое обеспечение методов распознавания образов при обработке текстов на Вьетнамском языке

Ле Чунг Хьеу

Научный руководитель: доктор физ.-мат. наук, проф.
Граничин Олег Николаевич

Математическо-механический факультет
Санкт-Петербургский государственный университет

17/2/2011

1. Актуальность темы

Методы распознавания образов и автоматическая обработка текстов (АОТ).

Основные трудности проблемы обработки вьетнамских текстов

1. проблема делимитации слова
   - "*học sinh học sinh học*" — "*школьник учит биологию*".
     - "*học sinh*" — школьник,
     - "*sinh học*" — биология,
     - "*học*" — учиться,
     - "*sinh*" — родиться.
   - "*học sinh | học | sinh học*".
2. вьетнамские словари и вьетнамскоязычные корпусы текстов
   - 75 000 слов — 200 000 слов.

2. Цель исследования и задачи

Цель исследования:

- Создание математического обеспечения, реализующего методы распознавания образов для автоматической разметки текстов на вьетнамском языке, результаты применения которого могут быть использованы для дальнейшей обработки лингвистами или другими программными системами поиска и автоматического перевода.

Цель достигается в работе через решение следующих задач:

- разработка и обоснование математических статистических моделей распознавания образов вьетнамских слов и словосочетаний;
- разработка обеспечения методов графематического анализа вьетнамских текстов;
- разработка и обоснование теоретико-вероятностной модели, для выполнения процесса морфологического анализа вьетнамских текстов;
- создание программной системы для автоматической обработки вьетнамских текстов и построение с ее помощью значительных наборов данных.

3. Структура работы

1. Введение.
2. Особенности обработки вьетнамских текстов.
   - Обработка естественного языка.
   - Особенности вьетнамских текстов.
   - Программные продукты для обработки естественного языка.
3. Методы распознавания образов при графематическом анализе.
   - Извлечение графематических дескрипторов.
   - Распознавание слов и сегментация предложений.
4. Оптимизация параметров скрытых марковских моделей при распознавании морфологической структуры.
   - Морфологическая разметка корпусов текстов.
   - Автоматический морфологический анализ.
5. Система автоматической обработки вьетнамских текстов.
   - Описание функционирования программ.
   - Состав программного обеспечения.
   - Результаты экспериментов.
6. Заключение

4. Описание задачи

Пусть $\Sigma = \{\sigma_i\}$ — алфавит.

$\mathcal{L} \subseteq \Sigma^* = \{c = \langle\sigma_i\rangle | \sigma_i \in \Sigma, |\omega| \geq 0\}$.

$\mathcal{C} = \{s_1, s_2, \ldots, s_N\}$.

$\mathcal{T} = \{t_1, t_2, \ldots, t_M\}$.

Проблема разметки текстов *состоит в том, чтобы по предложению* $s \in \mathcal{C}, \quad s = c_1 c_2 \ldots c_k$ *— последовательность текстовых сегментов, сформировать* $s' = w_1[T_1]w_2[T_2]\ldots w_l[T_l]$ *помеченное предложение, где* $w_i = c_i^1 \ldots c_i^{l_i}$ *— последовательность вьетнамских текстовых элементов (слово, знак пунктуации, цифровый комплекс и т.п.), а* $[T_i]$ *— множество возможных графематических дескрипторов или морфологических признаков текстового элемента в этом предложении.*

Задача автоматической разметки текстов на вьетнамском языке

- первичный графематический анализ;
- распознавание слов и словосочетаний, сегментация предложения на слова;
- морфологическая разметка корпуса вьетнамских текстов;
- морфологический анализ вьетнамского предложения.

5. Извлечение графематических дескрипторов

Задача первичного извлечения графематических дескрипторов — выделение различных нестандартных элементов текста и присваивание им соответствующих графематических дескрипторов.

- Элементы графематического дескриптора.
  - структурные элементы текста.
  - различные элементы текста, не являющиеся слогами.
  - собственные имена, названия, аббревиатуры.
  - иностранные лексемы.
- Основные трудности задачи.
  - распознавание случаев использования знаков препинания в иных целях, а не в качестве разделителя фразы.
  - выделение фразы построено на анализе правого и левого окружения знаков препинания.
  - распознавание собственных имен, сокращений, технических терминов, аббревиатур и иностранных лексем.

Подход и основные понятия

Метод, основанный на сопоставлении образцов.

- *Образец* — объект, который представляет собой шаблон фразы.
- *Графематический дескриптор* — особый образец.
- *Правила* — "образец → действие".

P-Date {/ ngày 12/07/1982 */*
P: word Day/Month/Year;
C_1: IsVietnameseWord(word) & word = "ngày";
C_2: Length(Day) = 2 & IsNumeric(Day);
C_3: Length(Month) = 2 & IsNumeric(Month);
C_4: Length(Year) = 4 & IsNumeric(Year);
};

M-Date {/ 12/07/1982 [Date]*/*
M: "[Date]";
P: Day/Month/Year;
C_1: Length(Day) = 2 & IsNumeric(Day);
C_2: Length(Month) = 2 & IsNumeric(Month);
C_3: Length(Year) = 4 & IsNumeric(Year);
};

R-Date {/ ngày 12/07/1982 → ngày "[Date]" */*
O: P-Date;
A: M-Date.P → M-Date.M;
};

Модель извлечения

Пусть $\Sigma = \{\sigma_i\}$ — *алфавит*.

$\mathcal{L} \subseteq \Sigma^* = \{w = \langle\sigma_i\rangle | \sigma_i \in \Sigma, |w| \geq 0\}$.

$\mathcal{C}_{\mathcal{L}} = \{s_1, s_2, \ldots, s_N\}$.

$\mathcal{A} = \{A_i | A_i \subseteq \mathcal{C}_{\mathcal{L}}, A_i \neq \emptyset\}$.

$\mathcal{C}_{\mathcal{A}} = \{P = \langle A_i\rangle | A_i \in \mathcal{A}\}$.

$\mathcal{T} = \{T_i | T_i - \textit{дескриптор}\}$.

$\mathcal{C}_{\mathcal{T}} = \{M_i = (P_i, T_i)\}$.

$\mathcal{R} = \{R : (P_c, M_o = (P_o, T_o)) \to T_o\}$

Пусть задан кортеж $\mathcal{M} = \langle\Sigma, \mathcal{C}_{\mathcal{L}}, \mathcal{C}_{\mathcal{A}}, \mathcal{C}_{\mathcal{T}}, \mathcal{R}\rangle$. Тогда основной задачей распознавания графематических дескрипторов в условиях $\mathcal{M}$ будем называть задачу построения для произвольного текстового сегмента $s \in \mathcal{C}_{\mathcal{L}}$ набора графематических дескрипторов M_s.

6. Распознавание слов и сегментация предложений

При решении задачи распознавания слов рассматриваются две проблемы:

- распознавание слов с вероятностной точки зрения;
- построение по большому набору предложений с помощью процесса обучения без учителя адекватной вероятностной модели.

Вероятностная модель

$\mathcal{M} = (\mathcal{C}, \Sigma_{\mathcal{C}}, \mathcal{F}_{\mathcal{C}})$ — *вероятностная модель.*

$\mathcal{C}=\{s_1, s_2, \ldots, s_n\}$ — *набор предложений.*

$\Sigma_{\mathcal{C}}$ — множество слогов.

$\mathcal{F}_{\mathcal{C}}$ — *множество вероятностных функций.*

Definition

Пусть $c \in \mathbb{R}$ является константой. *Функция достоверности* $f_{c,\mathcal{M}}(\alpha, \beta) : \Sigma_{\mathcal{C}}^2 \mapsto \mathbb{R}$ над вероятностной моделью $\mathcal{M}$ и набором $\mathcal{C}$ определяется следующим образом:

$$f_{c,\mathcal{M}}(\alpha, \beta) = c\frac{P(\alpha\beta)^2}{P'(\alpha\beta)}.$$

Функция распознавания слов и локальная максимальная достоверная последовательность

$m_{sup}, M_{sup}, m_{con}, M_{con} \in \mathcal{F}_{\mathcal{C}}, \quad f_c \in \mathcal{F}_{\mathcal{C}}$

Definition

$f_R : \Sigma_{\mathcal{C}}^2 \mapsto \{-1, 0, 1\}$ является *функцией распознавания слов* над f_c и $\mathcal{M}$ с параметрами $(m_{sup}, M_{sup}, m_{con}, M_{con})$ если:

$$f_R(\alpha, \beta) = \begin{cases} 1 & \text{if } (f_c(\alpha, \beta) \geq M_{con}) \quad \text{and} \quad (N(\alpha\beta) \geq M_{sup}); \\ -1 & \text{if } (f_c(\alpha, \beta) < m_{con}) \quad \text{or} \quad (N(\alpha\beta) < m_{sup}); \\ 0 & \text{otherwise.} \end{cases}$$

$f_c, f_R^* \in \mathcal{F}_{\mathcal{C}} \quad D_{con} \in \mathcal{F}_{\mathcal{C}};$

$s = \alpha_1\alpha_2 \ldots \alpha_k \in \mathcal{C}, \quad w = \alpha_l\alpha_{l+1} \ldots \alpha_{l+m}.$

Definition

w является *локальной максимальной достоверной последовательностью* (ЛМДП) в s над $\mathcal{P}, f_c, f_R^*$ и D_{con}, если удовлетворяются следующие условия:
(i) $\forall i = l, \ldots, m-1 : f_R^*(\alpha_i, \alpha_{i+1}) = 1;$
(ii) если $l > 1$:$(f_R^*(\alpha_{l-1}, \alpha_l) = -1)$ или $(f_R^*(\alpha_{l-1}, \alpha_l) = 0 \quad and \quad f_c(\alpha_l, \alpha_{l+1}) > f_c(\alpha_{l-1}, \alpha_l) + D_{con})$
(iii) если $l + m < k$: $(f_R^*(\alpha_{l+m}, \alpha_{l+m+1}) = -1)$ или
$(f_R^*(\alpha_{l+m}, \alpha_{l+m+1}) = 0 \quad and \quad f_c(\alpha_{l+m-1}, \alpha_{l+m}) > f_c(\alpha_{l+m}, \alpha_{l+m+1}) + D_{con})$

Процесс обучения

Обучающая вероятностная модель.

1. выбор парамертры ($M_{con}, m_{con}, M_{sup}, m_{sup}, D_{con}$);
2. поиск *локальных максимально достовернных*;
3. соединение последовательностей слогов;
4. пересчет всех вероятностных значений;
5. если добавляются новые слоги, то возврат к шагу (2);
6. возврат к шагу (1).

Для обоснования предложенных методов распознавания образов слов, словосочетаний доказаны следующие теоремы:

Theorem

Процесс соединения слогов с определенными параметрами завершается за конечное число итераций.

Theorem

Сложность процесса соединения слогов имеет порядок $O(N_s \times M_s)$.

7. Морфологическая разметка корпусов текстов

$\mathbf{T} = \{t_1, t_2, \ldots, t_m\}$ — конечный набор тэгов ([NN], [VB]).

$s = w_1 w_2 \ldots w_k$ — *предложение*.

$s' = w_1[T_1]w_2[T_2]\ldots w_k[T_k]$ — *помеченное предложение*.

$\mathbf{C}=\{s_1, s_2, \ldots, s_n\}$ — *набор предложений*.

$\mathbf{C}' = \{s'_1, s'_2, \ldots, s'_n\}$ — *набор помеченных предложений*.

Проблема морфологической разметки корпуса *состоит в том, чтобы по набору предложений* $\mathbf{C}=\{s_1, s_2, \ldots, s_n\}$ *создать набор помеченных предложений* $\mathbf{C}' = \{s'_1, s'_2, \ldots, s'_n\}$.

Основные проблемы морфологической разметки корпусов вьетнамских текстов:

- корректная разметка вьетнамских текстов обязательно требует ручной работы;
- полуавтоматической разметки списка фраз;
- морфологической разметки вьетнамских корпусов.

Описание модели разметки

Исходные данные — исходные тексты корпусов.
Результаты работы — аннотированные тексты.

Например,

chúng_ tôi[NN, PP] đã[AD] nghiên_ cứu[VB, NN].

Морфологическая разметка списка фраз.

- *Парсинг.*
- *Фильтрование.*

Морфологическая разметка корпусов текстов.

8. Автоматический морфологический анализ

$\mathbf{T} = \{T_1, T_2, \ldots, T_N\}$.

$\mathbf{W} = \{W_1, W_2, \ldots, W_M\}$.

$\mathbf{A} = \{a_{ij}\} \quad a_{ij} = P(t_{k+1} = T_j | t_k = T_i), \qquad 1 \le i, j \le N,$

$\mathbf{B} = \{b_{ij}\} \quad b_{ij} = P(w_k = W_i | t_k = T_j), \qquad 1 \ge i \ge N, \quad 1 \ge j \ge M,$

$\pi = \{\pi_i\} \quad \pi_i = P(t_1 = T_i), \qquad 1 \le i \le N.$

Модель: $\lambda = \{\mathbf{A}, \mathbf{B}, \pi\}$.

Задача морфологического анализа *состоит в том, чтобы по имеющейся последовательности наблюдений* $W = \{w_1, w_2, \ldots, w_l\}$ *восстановить последовательность состояний (тегов)* $T^* = \{t_1, t_2, \ldots, t_l\}$, *порождающую эти наблюдения с наибольшей вероятностью.*

Определим

$$\delta_k(i, W) = \max_{t_1 \ldots t_k | t_k = T_i} P(W_{[1,k]} | T_{[1,k]}),$$

где $W_{[1,k]} = w_1 w_2 \ldots w_k$ и $T_{[1,k]} = t_1 t_2 \ldots t_k$.

Процесс обучения

Пусть $O = \{w_1, w_2, \ldots, w_l\}$ — последовательность наблюдений.

$$\begin{aligned}
\alpha_{k+1}(i, O) &= \sum_{j=1}^{N} \alpha_k(j, O) a_{ji} b_{iw_k}, \qquad 1 \le k \le l, \\
\beta_k(i, O) &= \sum_{j=1}^{N} a_{ij} b_{jw_{k+1}} \beta_{k+1}(j, O), \qquad 1 \le k \le l^i, \\
\gamma_k(i, O) &= \frac{\alpha_k(i, O)\beta_k(i, O)}{P(O_{[1,l^i]})}, \quad \xi_k(i, j, O) = \frac{\alpha_k(i, O) a_{ij} b_{jw_{k+1}} \beta k + 1(j, O)}{P(O_{[1,l^i]})}.
\end{aligned}$$

$\mathcal{O} = \{O^1, O^2, \ldots, O^n\}$ — максимизировать $P(\mathcal{O}) = \prod_{i=1}^{n} P(O^i)$
$E(\mathcal{O}|\lambda) = -\ln(\mathcal{O}|\lambda)$.
$Q(\mathcal{O}, \lambda^0, \lambda) = -\sum_{i=1}^{n} \sum_{T^i} P(T^i|O^i, \lambda^0) \ln P(O^i, T^i|\lambda)$,
$E(\mathcal{O}, \lambda) \le Q(\mathcal{O}, \lambda^0, \lambda) - Q(\mathcal{O}, \lambda^0, \lambda^0) + E(\mathcal{O}, \lambda^0)$

Theorem

Минимум функции $Q(O, \lambda^0, \cdot)$ достигается в точке λ^* с координатами

$$\pi_j^* = \frac{1}{n} \sum_{i=1}^{n} \gamma_1(j, O^i), \quad a_{jk}^* = \frac{\sum_{i=1}^{n} \sum_{t|w_t^i=k} \gamma_t(j, O^i)}{\sum_{i=1}^{n} \sum_{t=1}^{l^i-1} \gamma_t(j, O^i)}, \quad b_{jk}^* = \frac{\sum_{i=1}^{n} \sum_{t=1}^{l-1} \xi_t(j, k, O^i)}{\sum_{i=1}^{n} \sum_{t=1}^{l^i} \gamma_t(j, O^i)}.$$

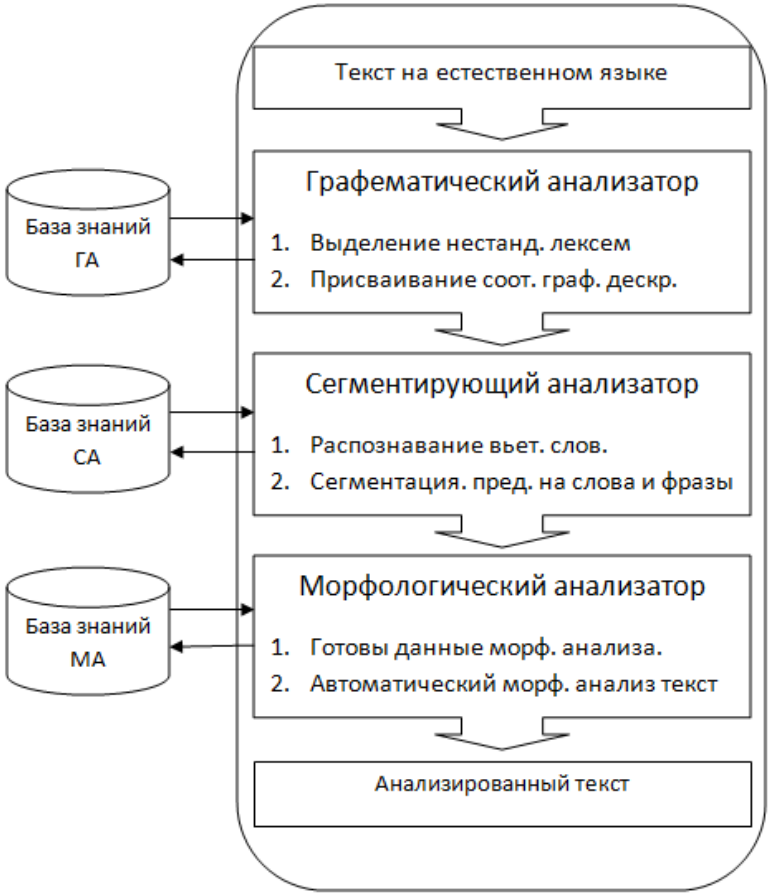
9. Система обработки вьетнамских текстов и основные результаты

Система обработки вьетнамских текстов

- вход работы — текст на естественном языке;
- графематический анализатор;
- сегментирующий анализатор;
- морфологический анализатор;
- результат работы — аннотированный текст.

Реализация — C#

- 250 034 вьетнамских Интернет документов.
- http://www.tuoitre.com.vn/



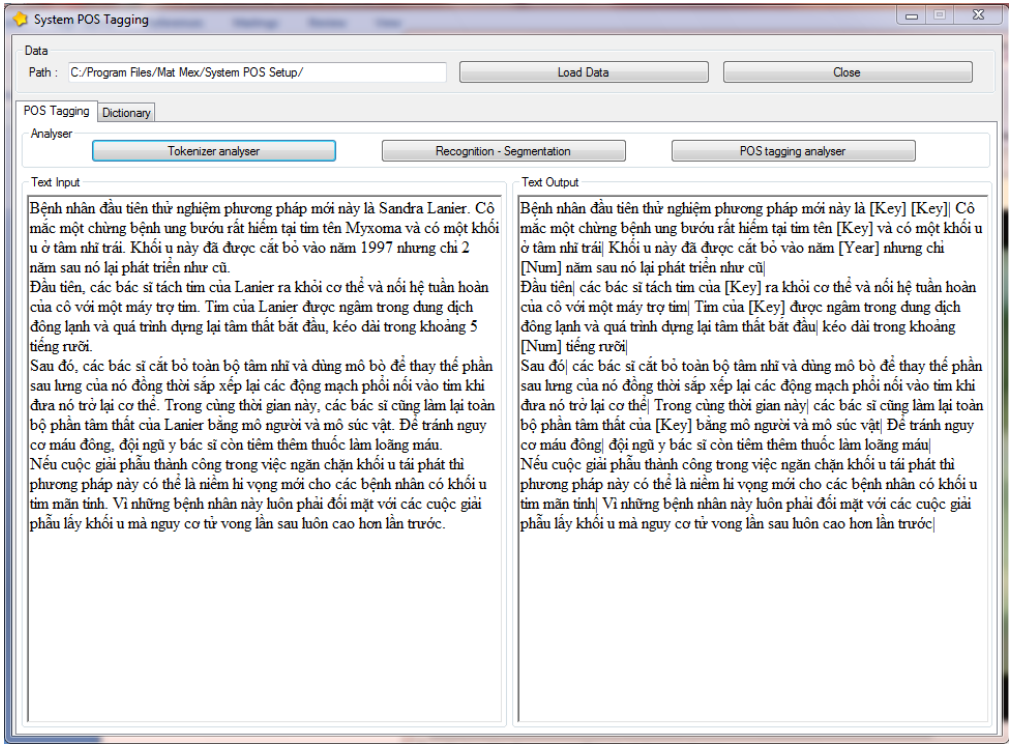
Первичный графематический анализатор

Графематические правила — 279 шт.

Основные результаты:

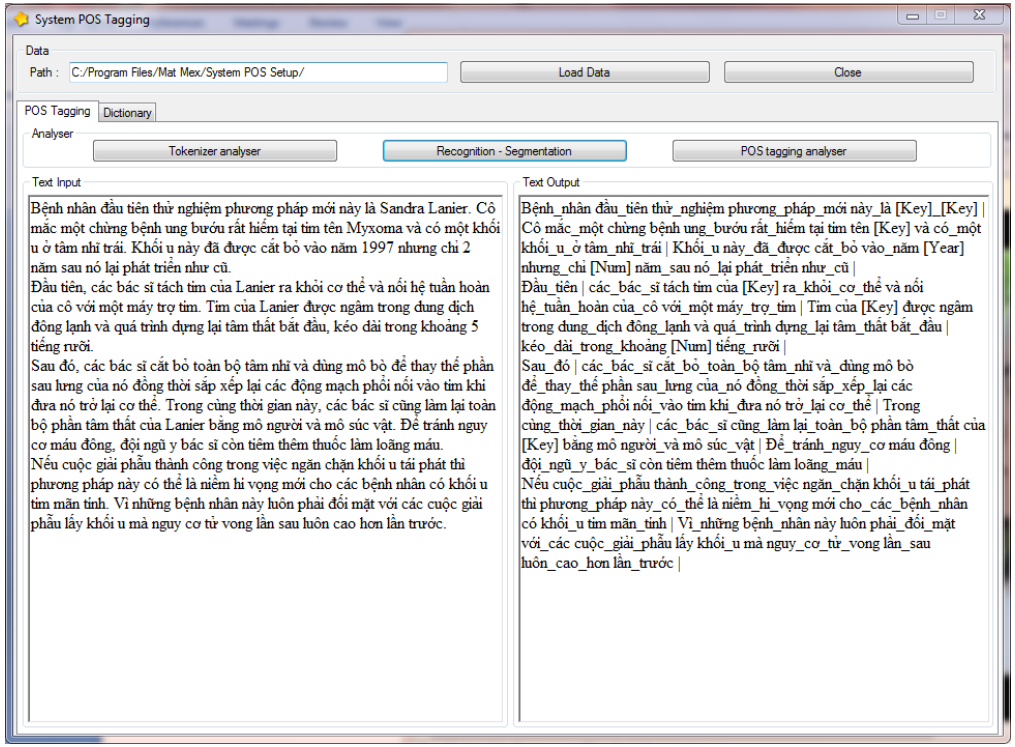
- программа распознавания вьетнамских слогов;
- программа проверки орфографии;
- программа распознавания собственных имен;
- распознавание структурных элементов текста.

Точность получившего словаря составляет 98%.



Сегментирующий анализатор

- соединенные слоги “*bệnh_nhân*”, “*đầu_tiên*”, “*thử_nghiệm*” — вьетнамские слова,
- “*phương_pháp_mới*”, “*này_là*” — вьетнамские фразы.



Слоги	Ш.1	Ш.2	Ш.3	Ш.4
cá nhân	3.58	reg	reg	reg
cá cược	3.60	reg	reg	reg
cá độ	1.86	3.03	reg	reg
cá heo	0.51	1.31	2.01	reg

Слоги	Ш.1	Ш.2	Ш.3	Ш.4
cá cảnh	-2.55	-2.01	-0.66	1.02
cá biển	-2.85	-2.60	-0.8	1.1
cá lóc	0.88	1.5	2.51	reg
cá bỏ	-7.86	-7.49	-6.83	-6.05

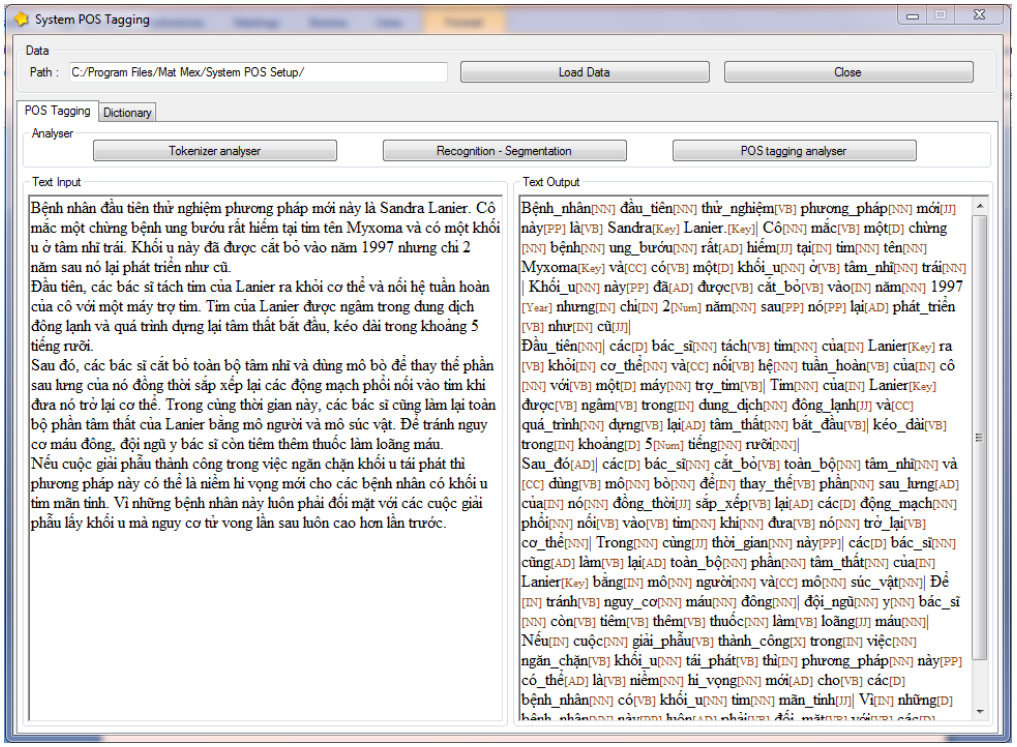
Морфологический анализатор

- слова “*bệnh_nhân*”, “*phương_pháp*” — существительное;
- “*thử_nghiệm*” — глагол;
- “*mới*” — прилагательное.

Набор размеченных фраз — 315 623 вьетнамских фраз.

Точность программы — 96%.

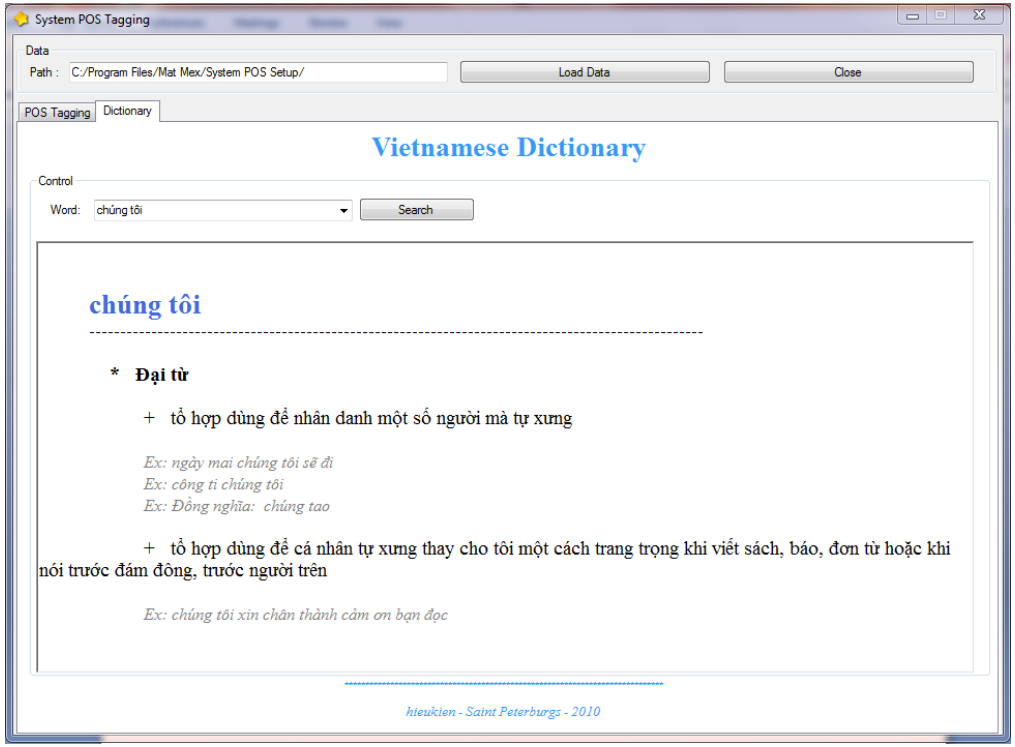
- “*học_sinh[NN] học[VB] sinh_học[NN]*”.
- “*nghiên_cứu_sinh[NN] Lê_Trung_Hiếu[NP]*”.



Морфологический словарь

Морфологический словарь

- 84 555 слов и словосочетаний.
- точность составляет 98%.



Основные результаты

В работе получены следующие основные научные результаты:

1. Предложен, обоснован и реализован метод обучения без учителя для распознавания образов слов, словосочетаний и фраз во вьетнамских текстах, позволяющий производить сегментацию предложений на слова и фразы для последующей автоматической морфологической разметки вьетнамских текстов.
2. Разработано математическое и программное обеспечение, реализующее метод поиска образца, предназначенное для выделения различных лексем вьетнамского текста и присваивания им соответствующих графематических дескрипторов. Исследованы статистические характеристики образования лексем вьетнамского текста.
3. Предложен и реализован метод скрытой марковской модели для распознавания морфологической структуры предложений во вьетнамских текстах, обоснован алгоритм оптимизации его параметров.
4. Разработана новая программная система для автоматической обработки вьетнамских текстов, с помощью которой сформированы графематический, морфологический и статистический словари значительных размеров, а также аннотированный корпус вьетнамских текстов.

Публикации

Статьи в журналах, рекомендованных ВАК:

[1] *Ле Ч. Х., Граничин О. Н.* Статистический способ выделения и словосочетаний из вьетнамских печатных текстов // Вестник СПбГУ. 2009. Серия 10. Вып. 3. С. 161-169.

[2] *Le T. H., Le A. V., Le T. K.* An unsupervised learning and statistical approach for Vietnamese word recognition and segmentation // Lecture Notes in Computer Science "Intelligent Information and Database Systems. Second International Conference, ACIIDS, Hue City, Vietnam, March 24-26, 2010. Proceedings, Part II" / Ngoc Thanh Nguyen, Manh Thanh Le and Jerzy Swiatek editors. Vol. 5991 — Springer, 2010. P. 195–204. [On-line] http://www.springerlink.com/content/7q97147r18158844/

Другие публикации:

[3] *Ле Ч. Х., Ле А. В., Ле Ч. К.* Автоматическое выделение слов и словосочетаний из вьетнамских печатных текстов // Стохастическая оптимизация в информатике. 2008. Т. 4. С. 171-186.

[4] *Ле Ч. Х.* Обучение без учителя и статистический подход для сегментации и распознавания вьетнамских слов // Стохастическая оптимизация в информатике. 2009. Т. 5. С. 193-208.

[5] *Ле Ч. Х.* Модель извлечения графематических дескрипторов в системе обработки вьетнамского языка // Стохастическая оптимизация в информатике. 2010. Т. 6. С. 230–247.

[6] *Ле Ч. Х.* Модель морфологического анализа текстов вьетнамского языка // Стохастическая оптимизация в информатике. 2010. Т. 6. С. 248–263.

Спасибо за внимание!