САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ

На правах рукописи

Ле Чунг Хьеу

МАТЕМАТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ МЕТОДОВ РАСПОЗНАВАНИЯ ОБРАЗОВ ПРИ ОБРАБОТКЕ ТЕКСТОВ НА ВЬЕТНАМСКОМ ЯЗЫКЕ

05.13.11 — Математическое и программное обеспечение
вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Санкт-Петербург
2011

Работа выполнена на кафедре системного программирования Математико-механического факультета Санкт-Петербургского государственного университета.

Научный руководитель:	доктор физико-математических наук, профессор ГРАНИЧИН Олег Николаевич
Официальные оппоненты:	доктор технических наук, профессор ТИМОФЕЕВ Адиль Васильевич, (Санкт-Петербургский институт информатики и автоматизации РАН)
	доктор физико-математических наук, доцент КРИВУЛИН Николай Кимович (Санкт-Петербургский государственный университет)
Ведущая организация:	Санкт-Петербургский государственный университет информационных технологий, механики и оптики.

Защита состоится "____" _____ 2011 года в ____ часов на заседании совета Д 212.232.51 по защите докторских и кандидатских диссертаций при Санкт-Петербургском государственном университете по адресу 198504, Санкт-Петербург, Петродворец, Университетский пр.,28, Математико-механический факультет.

С диссертацией можно ознакомиться в Научной библиотеке им.М.Горького Санкт-Петербургского государственного университета по адресу: 199034, Санкт-Петербург, Университетская наб. 7/9.

Автореферат разослан "____" _____ 2010 г.

Ученый секретарь
диссертационного совета Даугавет И. К.

Общая характеристика работы

Актуальность темы. В последние десятилетия методы распознавания образов находят приложения в самых разнообразных областях. Многие из них активно используются при автоматической обработке текстов (АОТ). АОТ достигла значительных успехов в лексико-грамматическом анализе, выявлении темы, в поиске информации и т. п. Большинство работ по АОТ были проведены для языков индоевропейской группы. Их результаты не могут быть непосредственно применены к вьетнамскому языку, который, являясь разговорным языком (как китайский, японский и др.), оперирует слогами, а не словами. Границы слова определяются контекстом. Для построения новых лексических единиц или слов используются сочетания различных слогов. Роль приставок и суффиксов также выполняют слоги, что еще более запутывает процесс анализа текста. Похожие проблемы характерны и для распознавания текстов на других восточных языках. Но, например, для китайского они решаются за счет большого объема уже сформированных и подготовленных аннотированных корпусов текстов.

Проблемы распознавания образов слов и словосочетаний во вьетнамских текстах исследовались в современных работах Д. Дьена, Х.Н. Као, Х.П. Ле, К.Т. Нгуена, Х. Нгуена, Л.А. Ха и др. Основные задачи обработки текстов на вьетнамском языке (лексико-грамматический анализ, синтаксический анализ и т. п.) сложны для вычислительной лингвистики в первую очередь из-за нерешенности проблемы делимитации слова, так как слово во вьетнамском языке не является единицей, которую можно было бы всегда четко выделить по каким-либо формальным признакам. При автоматической обработке вьетнамского языка методы распознавания образов целесообразно использовать не только в традиционных сферах приложений по распознаванию символов и звуков, но и неожиданной с точки зрения обработки индоевропейских текстов сфере — распознавании образов слов и фраз.

Для вьетнамского языка серьезной проблемой для автоматической обработки является отсутствие достаточно полных словарей вьетнамских слов и вьетнамскоязычных корпусов текстов. На протяжении долгого времени вьетнамские, а также иностранные специалисты, решали эту проблему вручную. Однако построение списка слов вручную требует колоссальных усилий и все же не обеспечивает полноты словаря. Одна из причин этого — широкое использование вьетнамского языка в различных сферах со специальными словами, которые редко используется. Другая

— в различных регионах используются разные диалекты и словосочетания. Кроме этого, условия жизни быстро меняются. С развитием новых технологий и увеличением объемов информации постоянно расширяется лексикон вьетнамского языка. Например, новые слова: Интернет, айфон и т. п. надо включать в словари как новые понятия. Все эти причины делают процесс построения списка вьетнамских слов вручную трудновыполнимым. По последним данным самый полный вьетнамский словарь содержит только 75 000 слов, но в реальности по оценке специалистов количество вьетнамских слов насчитывает уже более 200 000. Это означает, что более половины вьетнамских слов нигде не сохранены.

Цель и задачи работы. 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 диссертации через решение следующих задач:

- разработка и обоснование математических статистических моделей распознавания образов вьетнамских слов и словосочетаний, создание на их основе математического обеспечения для сегментации предложений на слова и фразы, использующего методы теории вероятностей и математической статистики, а также алгоритмы обучения без учителя;

- разработка обеспечения методов графематического анализа вьетнамских текстов, основанных на статистических моделях распознавания образов вьетнамских слов, словосочетаний и фраз и на сопоставлении образцов в большом текстовом массиве данных, позволяющих эффективно выполнять процесс выделения различных лексем вьетнамского текста и присваивать им соответствующие графематические дескрипторы;

- разработка и обоснование теоретико-вероятностной модели, использующей метод скрытых марковских моделей, для выполнения процесса морфологического анализа вьетнамских текстов;

- создание программной системы для автоматической обработки вьетнамских текстов и построение с ее помощью значительных наборов данных, включающих графематический, морфологический и статистический словари, а также

аннотированный корпус вьетнамских текстов.

Методы исследования. В диссертации применяются методы распознавания образов, машинного обучения без учителя, теории вероятностей и математической статистики, имитационного моделирования и системного программирования.

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.

Научная новизна. Все основные научные результаты диссертации являются новыми.

Теоретическая ценность и практическая значимость. Теоретическая ценность работы состоит в разработке, обосновании и реализации новых алгоритмов распознавания образов слов, сегментации предложений на слова и фразы, а также автоматической морфологической разметки вьетнамских текстов.

Предложенные новые алгоритмы могут быть эффективно использованы при решении практических задачах обработки текстов на вьетнамском языке, а также на ряде других (китайском, японском, корейском и т. п.). Созданный программный

комплекс для автоматической обработки вьетнамских текстов может быть использовать лингвистами для дальнейшего изучения языка. Результаты автоматической обработки текстов, получаемые с помощью разработанной системы, могут использоваться как лингвистами, так и в других системах поиска и автоматического перевода.

Апробация работы. Материалы диссертации докладывались на семинарах кафедры системного программирования математико-механического факультета СПбГУ и на международной конференции: The 2nd Asian Conference on Intelligent Information and Database Systems (Hue, Vietname, March 24—26, 2010).

Результаты диссертации были частично использованы в работе по НИР из средств бюджета "Математическая модель распознавания и процессинга текстов на восточных языках на основе сегментации релевантных составляющих", выполняемой в СПбГУ.

Публикации. Основные результаты диссертации опубликованы в шести работах. Из них две публикации [1, 2] в изданиях из перечня ВАК. Работы [1,2,3] написаны в соавторстве. В работе [1] Граничину О.Н. принадлежит общая постановка задачи, а Ле Ч.Х. реализации и обоснования описываемых методов, создание демонстрационных примеров и программных средств. В работах [2, 3] Ле Ч.Х. предложил новые статистические методы распознавания образов и теоретико-вероятностную модель для автоматической сегментации предложений на вьетнамском языке, а его соавторы участвовали в подготовке наборов текстовых данных для апробации новых методов и выполнили часть работы по созданию и доработке нового словаря вьетнамских слов.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав, заключения, списка литературы, включающего 105 источников. Текст занимает 102 страницы, содержит 10 рисунков и 11 таблиц.

Содержание работы

Во **введении** обосновывается актуальность тематики диссертационной работы и кратко излагаются ее основные результаты.

В **первой главе** "*Особенности обработки вьетнамских текстов*" анализируются общие проблемы автоматической естественных языков, а также представлены

лингвистические характеристики и атрибуты вьетнамского языка.

Одной из серьезных проблем организации человеко-машинного взаимодействия является лингвистический анализ предложения на естественном языке с последующим переводом его на машинный язык вычислительной системы. Общие подходы к решению этой задачи рассматриваются в п. 1.1. Обычно текст подвергается последовательной обработке графематическим, морфологическим, синтаксическим и семантическим анализаторами.

В п. 1.2 описываются лингвистические характеристики вьетнамского языка: фонетика, слог, морфология, синтаксис, лексические категории и предложения. Особенностью вьетнамского языка является то, что он разговорный, и в нем самым важным элементом является слог, а не слово. Автоматический анализ текстов на вьетнамском языке затруднен нерешенностью проблемы делимитации слова. Границы слов могут меняться в зависимости от контекста, что приводит к трудностям их определения. Эти трудности восприятия иллюстрируются следующим примером. В предложении "*học sinh học sinh học*", которое по-русски означает "*школьник учит биологию*", все комбинации "*học sinh*", "*sinh học*", "*sinh*", "*học*" являются вьетнамскими словами:

- "*học sinh*" — школьник,
- "*sinh học*" — биология,
- "*học*" — учиться,
- "*sinh*" — родиться.

Для разделения предложения на слова важно содержание предложения. Исходя из смысла предложения определяются те комбинации слогов, которые являются словами. В рассматриваемом примере, учитывая его основную мысль, правильная расстановка границ такова: "*học sinh* / *học* / *sinh học*".

Другой трудной проблемой является отсутствие достаточно полных вьетнамских словарей и корпусов текстов.

В п. 1.3 представлены классификация и краткие описания программных продуктов, связанных с анализом текстов и вычислительной лингвистикой, которые исследовались при разработке в ходе работы над диссертацией новой программной системы по автоматическому распознаванию вьетнамских текстов.

Во **второй главе** "*Методы распознавания образов при графематическом анализе*" описываются применения методов распознавания, основанные алгоритмах на обучении без учителя и поиска образцов, при выделении лексем во вьетнамских текстах.

Графематический анализ представляет собой начальный этап обработки текста, представленного в виде цепочки ASCII символов, подготавливающий информацию, необходимую для дальнейшей обработки морфологическим и синтаксическим процессорами. При графематическом анализе вьетнамского языка решаются две основные задачи: первая — выделение различных нестандартных элементов текста и присваивание им соответствующих графематических дескрипторов (например, знаков пунктуации, цифровых комплексов, собственных имен, сокращений и т. д.); вторая — распознавание слов и словосочетаний, сегментация предложений на слова.

Для выявления в текстах графематических дескрипторов необходимо иметь правила формирования структуры текстовых сегментов (шаблоны, образцы) и правила извлечения. Первые выявляют лингвистические свойства структуры текстов, тогда как вторые, используют эти свойства для распознавания текстовых фактов. Формирование таких правил в существующих разработках производится вручную, что является причиной сложности настройки системы графематического анализа.

В п. 2.1 предлагается и исследуется модель первичного графематического анализа вьетнамских текстов, основанного на сопоставлении образцов в большом текстовом массиве данных, позволяющая эффективно выполнять процесс выделения различных лексем вьетнамского текста и присваивания им соответствующих графематических дескрипторов. Модель базируется на исследовании и выявлении статистических характеристик образования лексем во вьетнамских текстах и построении набора соответствующих графематических правил.

Пусть $\Sigma = \{\sigma_i\}$ — *алфавит* (конечное упорядоченное множество символов). $\mathcal{L} \subseteq \Sigma^* = \{w = \langle\sigma_i\rangle | \sigma_i \in \Sigma, |w| \geq 0\}$ — *некоторый язык*, заданный над этим алфавитом.

$\mathcal{C}_\mathcal{L} = \{s_1, s_2, \ldots, s_N\}$ — *конечный набор всех текстовых сегментов*, где текстовый сегмент $s = \sigma_1\sigma_2\ldots\sigma_n$ является последовательностью символов алфавитов языка.

$\mathcal{A} = \{A_i | A_i \subseteq \mathcal{C}_\mathcal{L}, A_i \neq \emptyset\}$, — *конечная совокупность элементарных атрибутов*. Говорят, что текстовый сегмент $s \in \mathcal{C}_\mathcal{L}$ имеет атрибут A_j если $s \in A_j$.

Образец $P = \langle A_1, A_2, \ldots, A_k\rangle$ представляют собой шаблон фразы, состоящий из

элементарных атрибутов. $\mathcal{C}_\mathcal{A} = \{P = \langle A_i \rangle | A_i \in \mathcal{A}\}$ — *конечный набор образцов.*

$\mathcal{T} = \{T_i | T_i - \textit{дескриптор}\}$ — *конечное множество дескрипторов. Графематический дескриптор* $M_i = (P_i, T_i)$ — особый образец, состоящий из шаблона и дескриптора. $\mathcal{C}_\mathcal{T} = \{M_i = (P_i, T_i)\}$ — *конечное множество классов графематических дескрипторов.*

$\mathcal{R} = \{R : (P_c, M_o = (P_o, T_o)) \to T_o\}$ — множество правил извлечения, где P_c — образец, $M_o = (P_o, T_o)$ — графематический дескриптор.

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.

В п. 2.2 анализируются задачи распознавания слов и словосочетаний и сегментации предложений на слова, словосочетания и фразы. Рассматриваются две проблемы:

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

Предлагаемый в работе подход к решению первой проблемы заключается в том, что по изучению большого множества последовательностей слогов выделяются *пары слогов,* (α, β)*, являющиеся словами или частями слов.*

Вероятностная модель $\mathcal{P}$ определяется как тройка $(\mathcal{C}, \Sigma_C, \mathcal{F}_\mathcal{C})$, в которой *набор предложений* $\mathcal{C}$=$\{s_1, s_2, \ldots, s_n\}$ является конечной совокупностью предложений, $\Sigma_\mathcal{C}$ — множество слогов, которые являются частями некоторых предложений s_i из набора $\mathcal{C}$, $\mathcal{F}_\mathcal{C}$ — *множество вероятностных функций* $F_c \in \mathcal{F}_\mathcal{C} : \Sigma_\mathcal{C}^* \mapsto \mathbb{R}$.

Функции достоверности — вероятностные функции пары слогов, которые оценивают какова вероятность того, что эта упорядоченная пара слогов являются словом или частью слова.

Определение 1: *Функция достоверности* $f_{c,\mathcal{M}}(\alpha, \beta) : \Sigma^2 \mapsto \mathbb{R}$ над вероятностной моделью $\mathcal{M}$ и набором $\mathcal{C}$ определяется следующим образом:

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

где $c \in \mathbb{R}$ некоторая константа (например, $c = 1$).

На основе функций достоверности строятся функции распознавания, которые для пар слогов дают вероятности того, что они могут быть частью слова.

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

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

где $f_c \in \mathcal{F}_{\mathcal{C}}$ — функция достоверности, $\mathcal{P}$ — вероятностная модель, $N(\alpha\beta)$ — вероятность появления пары $(\alpha\beta)$, $m_{sup}, M_{sup}, m_{con}, M_{con} \in \mathcal{F}_{\mathcal{C}}$ — некоторые постоянные функции: $0 < m_{sup} \leq M_{sup}$ и $0 < m_{con} \leq M_{con}$.

Пусть $f_c, f_R^* \in \mathcal{F}_{\mathcal{C}}$ — функции достоверности и универсальная распознавания, $D_{con} \in \mathcal{F}_{\mathcal{C}}$ — положительная постоянная; $s = \alpha_1\alpha_2 \ldots \alpha_k \in \mathcal{C}$ — предложение в наборе, и $w = \alpha_l\alpha_{l+1} \ldots \alpha_{l+m}$ является частью предложения s $(1 \leq l < k, 0 < m \leq k - l)$.

Определение 3: Пусть $s = \alpha_1\alpha_2 \ldots \alpha_k \in \mathcal{C}$ — предложение в наборе $\mathcal{C}$. Часть предложения $w = \alpha_l\alpha_{l+1} \ldots \alpha_{l+m}$ $(1 \leq l < k, 0 < m \leq k - l)$ называется *локальной максимально 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 $f_R^*(\alpha_{l-1}, \alpha_l) = 0$ and $f_c(\alpha_l, \alpha_{l+1}) > f_c(\alpha_{l-1}, \alpha_l) + D_{con}$
(iii) если $l + m < k : f_R^*(\alpha_{l+m}, \alpha_{l+m+1}) = -1$ или
$f_R^*(\alpha_{l+m}, \alpha_{l+m+1}) = 0$ and $f_c(\alpha_{l+m-1}, \alpha_{l+m}) > f_c(\alpha_{l+m}, \alpha_{l+m+1}) + D_{con}$,
где $f_c, f_R^* \in \mathcal{F}_{\mathcal{C}}$ — функции достоверности и универсальная распознавания, $D_{con} \in \mathcal{F}_{\mathcal{C}}$ — положительная постоянная.

Обучающая вероятностная модель строится итеративно по процессу соединения слогов. Начальный набор предложений — огромное множество вьетнамских предложений, которое было получено из электронных документов в Интернете. На каждой итерации обучения выполняются следующие шаги: (i) поиск *локальных максимально достоверных* последовательностей слогов в предложениях; (ii) соединение последовательностей слогов, которые являются локальными максимально достоверными; (iii) пересчет всех вероятностных значений нового набора предло-

жений; (iiii) корректировка параметров и возврат к шагу (i).

Для обоснования предложенных в п. 2.2 методов распознавания образов слов, словосочетаний и фраз доказаны следующие теоремы:

Теорема 1. Процесс соединения слогов с определенными параметрами завершается за конечное число итераций.

Теорема 2. Пусть N_s — число всех предложений в наборе и M_s — наибольшее число слогов в предложениях, тогда сложность процесса соединения слогов равна $O(N_s \times M_s)$.

Теорема 3. Процесс обучения завершается за конечное число итераций.

Процесс обучения сам по себе является алгоритмом сегментации слов. Вводные предложении были сегментированы по алгоритму обучения. Он использует статистические значении, которые определяются из формируемой адекватной вероятностной модели.

В **третьей главе** "*Оптимизация параметров скрытых марковских моделей при распознавании морфологической структуры*" предлагается и обосновывается новый метод распознавания для морфологического анализа. Цель морфологического анализа заключается в определении морфологических признаков слов для использования их на последующих этапах обработки текста.

В проблеме морфологического анализа вьетнамских текстов рассматриваются два основные проблемы.

1. морфологическая разметка корпуса вьетнамских текстов;
2. морфологический анализ вьетнамского предложения — снятие морфологической омонимии.

Пусть $\mathbf{C}$=$\{s_1, s_2, \ldots, s_n\}$ — конечная совокупность предложений, $\Sigma_{\mathbf{C}}$ — множество всех слов в наборе $\mathbf{C}$, $\mathbf{T} = \{t_1, t_2, \ldots, t_m\}$ — конечный набор морфологических признаков.

Первая проблема состоит в том, чтобы по предложению $s \in \mathbf{C}, s = c_1 c_2 \ldots c_k$ — последовательность слогов, сформировать $s' = w_1[T_1]w_2[T_2]\ldots w_l[T_l]$ помеченное предложение, где $w_i = c_i^1 \ldots c_i^{l_i}$ — вьетнамские слова, а $[T_i]$ — множество возможных морфологических признаков слова w_i в этом предложении.

В п. 3.1 описан *алгоритм морфологической разметки*, который производит ее полуавтоматически с использованием модели сегментации и списков размеченных ранее фраз. Модель сегментирует каждое предложение на фразы, и размечает их

на основе списков размеченных фраз. Алгоритм выполняется в два этапа: парсинг и фильтрование. *Парсинг* производится автоматически с использованием морфологического словаря. Парсер анализирует каждое слово в узком контексте каждой фразы, и присваивает ему соответствующие морфологические признаки, записывая их в квадратных скобках. *Фильтрование* производится вручную с участием лингвистов, которые проанализированные фразы — последовательности слов и его соответствующие набор возможных морфологических признаков — проверяют и корректируют.

Для решения второй проблемы морфологического анализа вьетнамских текстов в п. 3.2 диссертации предлагается метод автоматического морфологического анализа вьетнамских текстов с использованием скрытых моделей Маркова.

Пусть $\mathbf{T} = \{T_1, T_2, \ldots, T_N\}$ — конечное пространство состояний (тегов), $\mathbf{W} = \{W_1, W_2, \ldots, W_M\}$ — конечное пространство наблюдений (классов слов),
$\mathbf{A} = \{a_{ij}\}$ — матрица вероятностей переходов (или матрица переходов), где

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

$\mathbf{B} = \{b_{ij}\}$ — матрица эмиссии, где

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

$\pi = \{\pi_i\}$ — распределение вероятностей начального состояния, где

$$\pi_i = P(t_1 = T_i), \qquad 1 \leq i \leq N.$$

Моделью называется тройка $\lambda = \{\mathbf{A}, \mathbf{B}, \pi\}$.

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.

Для решения задачи предлагается воспользоваться методом динамического программирования, который в рассматрившаемом контектсе называется алгоритмом Витерби. Определим

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

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

Пусть $O = \{w_1, w_2, \dots, w_l\}$ — последовательность наблюдений и

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

Задача обучения состоит в том, что взять обучающий набор $\mathcal{O} = \{O^1, O^2, \dots, O^n\}$ последовательностей наблюдений и, соответственно, максимизировать правдоподобие наблюдений $P(\mathcal{O}) = \prod_{i=1}^{n} P(O^i)$, варьируя λ.

Зафиксируем начальный набор параметров модели λ_0 и будем пытаться увеличить правдоподобие $P(\mathcal{O}|\lambda)$ или, что то же самое, уменьшить $E(\mathcal{O}|\lambda) = -\ln(\mathcal{O}|\lambda)$.

Обозначив

$$Q(\mathcal{O}, \lambda^0, \lambda) = -\sum_{i=1}^{n} \sum_{T^i} P(T^i|O^i, \lambda^0) \ln P(O^i, T^i|\lambda),$$

в диссертации получена оценка вида:

$$E(\mathcal{O}, \lambda) \le Q(\mathcal{O}, \lambda^0, \lambda) - Q(\mathcal{O}, \lambda^0, \lambda^0) + E(\mathcal{O}, \lambda^0),$$

и доказана следующая теорема.

***Теорема 4**: Минимум функции $Q(O, \lambda^0, \cdot)$ достигается в точке λ^* с координатами*

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

В **четвертой главе** — “Система автоматической обработки вьетнамских текстов” — в п. 4.1 описана схема функционирования разработанной автором программной системы, которая представляет собой многоуровневый анализатор: графематический, сегментирующий и морфологический.

На вход системы подается текст в виде последовательности предложений на естественном языке.

Программа первичного графематического анализатора выделяет различные нестандартные лексемы вьетнамского текста и присваивает им соответствующие графематические дескрипторы. Программа включает в себя лексический словарь и набор графематических правил.

Программа сегментирующего анализатора предназначена для распознавания вьетнамских слов и сегментации предложений на слова и фразы. База знаний программы включает в себя словари распознанных сегментов.

Программа морфологического анализатора предназначена для морфологического анализа текстов. База знаний программы — морфологический словарь и набор размеченных фраз. Программа выполняется в два этапа. Сначала — морфологическая разметка вьетнамского текста, потом — автоматический морфологический анализ.

В п. 4.2 описано программное средство, которое обеспечивает загрузку и редактирование анализируемых текстов; анализ текста посредством автоматической системы, составленной из разработанных независимо компонентов: графематического, сегментирующего и морфологического анализаторов.

Система была реализована на языке C# в виде приложения под операционную систему Microsoft Windows. Исследовательский стенд предоставляет функциональность работы с системой со стороны пользователя, реализуя такие возможности, как загрузка, отображение и редактирование текста, запуск анализа текста и отображение результатов анализа. Программируемый конвейер — приложение, реализованное на основе технологии Microsoft Framework 3.5, — предоставляет функциональность работы со стендом со стороны исследователя — разработчика алгоритмов анализа, — реализуя такие возможности, как подключение модулей анализатора к программе, а также связывание их в единый конвейер.

В п. 4.3 анализируются результаты анализа текстов каждым из компонентов. Данные для экспериментов были взяты из 250 034 вьетнамских Интернет-документов с веб-сайта “http://www.tuoitre.com.vn/”. Начальные данные содержали 18 676 877

фраз и 131 318 974 слогов.

Основные В результате проведенных экспериментов были получены: *лексический словарь*, содержащий лексемы вьетнамского текста и соответствующие им графематические дескрипторы и статистические характеристики; *набор графематических правил*; *словари распознанных сегментов*, содержащие слога, пары слогов, соединения слогов вместе с дополнительными характеристиками, включающими количества или вероятности появлений, значения функций достоверности и распознавания; *морфологический словарь*, содержащий вьетнамские слова, словосочетания и соответствующие им части речи; *набор размеченных фраз*, состоящий из вьетнамских фраз и соответствующих морфологических разметок.

В заключении диссертации подведены итоги проведенного и завершенного в рамках поставленных задач исследования.

Работы автора по теме диссертации

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.

Подписано в печать 08.12.2010 г. Формат бумаги 60Х90 1/16. Бумага офсетная.
Печать ризографическая. Объем 1 усл. п. л. Тираж 100 экз. Заказ N ____.
Отпечатано в отделе оперативной полиграфии
химического факультета СПбГУ с оригинал-макета заказчика.
198504, Санкт-Петербург, Старый Петергоф, Университетский пр., 26.