Модель морфологического анализа текстов вьетнамского языка[1]

Ле Чунг Хьеу
Санкт-Петербургский государственный университет
vkhhieukien@yahoo.com

В работе рассматривается новый подход к морфологическому анализу текстов в системе обработки вьетнамского языка. Используя вероятностные модели, набор грамматических правил, а также "ручной" с участием лингвистов процесс морфологической разметки, он позволяет провести: (i) графематический анализ — выделение различных нестандартных элементов текста, присваивание им соответствующих графематических дескрипторов, сегментация предложений текста на слова и фразы; (ii) на его основе далее осуществляется поиск по морфологическому словарю для каждого слова в предложениях; (iii) автоматический морфологический анализ вьетнамских текстов на основе Марковской модели и набора грамматических правил.

Основные результаты — формализованный набор грамматических правил и аннотированный корпус вьетнамских текстов — могут быть использованы в развитии средств обработки вьетнамского языка.

Ключевые слова: морфологическая разметка корпуса, морфологический анализ, обработка вьетнамского языка, Марковская модель, алгоритм Витерби, гибридный метод.

1. Введение

Автоматизированная обработка печатных документов — одна из основных задач формирующихся систем искусственного интеллекта [1]. Стадия морфологического анализа является наиболее проработанным лингвистическим этапом процесса обработки естественного текста [2]. Цель морфологического анализа заключается в определении морфологических признаков слов для использования их на последующих этапах обработки текста. За последние два десятилетия создано, по крайней мере, несколько десятков алгоритмов для разных языков, например, английского, германского, русского и т. п. В некоторых восточных языках, таких как вьетнамский, китайский и японский языки, проблемы морфологического анализа становятся более сложными из-за нерешенности проблемы делимитации слова. Слово в этих языках не является единицей, которую можно было бы всегда четко выделить по каким-либо формальным признакам.

Морфологический анализ текстов вьетнамского языка *затруднен* следующими двумя причинами:

Не существует официального определения слова и алгоритма, который сегментирует вьетнамское предложение на слова точно в соответствии с его смыслом. Вьетнамский язык является разговорным языком, в котором самый важный элемент является слогом, не словом. Граница слова может измениться от человека человеку и не влияет на коммуникационный процесс. Кроме того, сочетание различных слогов является единственным способом для построения новых лексических единиц или слов во вьетнамском языке. Во вьетнамском языке не существуют приставки и суффиксы, используются только слоги, что запутывает исследователей. Например, слоги "*học*", "*sinh*" и их сочетание "*học sinh*" также являются словами в вьетнамском языке. Система определения частей речи во вьетнамском языке (также как и в китайском или японском) не является очень четкой, это приводит к трудностям определения границы слова.

Не существует полного словаря вьетнамского языка и хороших вьетнамскоязычных корпусов текстов. На протяжении долгого времени вьетнамские, а также иностранные специалисты, разрешали эту проблему вручную. Однако построение словаря или хороших аннотированных корпусов текстов вручную требует колоссальных усилий и все же не обеспечивает полноты словаря вьетнамских слов [3–5]. Согласно [6] крупнейшие вьетнамские словари содержат менее чем 33 000 слов, но во втором издании Оксфордского словаря английского языка содержится более 250 000 слов. Кроме того, по сведениям автора, нет полного вьетнамского словаря собственных имен и названий мест и организаций.

В этой работе представляется новый подход, который преодолевает эти трудности с использованием гибридных методов. На основе преимуществ символических и вероятностных методов, во-первых, минимизируются человеческие усилия при обработке типовых лингвистических конструкций и, во-вторых, максимизируются гибкость и эффективность обработки данных.

Во вьетнамском языке проблема делимитации слова не разрешена из-за некоторых классов слов, которые не могут быть однозначно определены не только по формальным признакам, но и по семантическим. Но вместе с такими словами существуют многие классы слов (более 80%), которые могут быть легко выделено в тексте.

Такие слова помогают высокоэффективно сегментировать предложения на слова и фразы с хорошей точностью. Высокая эффективность означает, что многие предложения сегментируются на слова, а не на фразы. В работах [3–6] представлены и обсуждены методы и модели распознавания слов и сегментации предложений на слова, а также описана модель графематического анализа текстов вьетнамского языка, которая выполняла две функции: (i) выделение различных нестандартных элементов текста и присваивание им соответствующие графематических дескрипторов; (ii) сегментация предложений текста на слова и фразы.

Процесс сегментации предложений на слова и фразы помогает значительно снизить человеческие усилия при морфологической разметке текстов. Для преодоления трудностей делимитации неопределенных слов по формальным признакам, а также построения лингвистических конструкций вьетнамских предложений или фраз выполняется морфологическая разметка наборов слов и фраз, которые сегментируются из корпуса. Процесс морфологической разметки производится вручную с участием лингвистов.

Как и в проблеме делимитации слова, при морфологической разметке вручную некоторые слова и фразы также не могут быть однозначно определены из-за того, что анализ проводится только в узком контексте. Однако важной особенностью вьетнамского языка является то, что для обозначения грамматических связей слова не изменяют свои формы, а соединяются между собой путем примыкания с учетом различных оценок лексических значений отдельного слова или путем добавления служебных слов. Порядок слов в предложении играет особую роль во вьетнамском языке и является основным средством для выражения синтаксических правил. Иными словами, на наш взгляд, с использованием хорошего набора грамматических правил на основе характеристик служебных слов и учета порядка слов в фразах возможно достаточно подробно описание структуры вьетнамских предложений.

Далее в этой работе рассматривается новый подход к морфологическому анализу текстов в системе обработки вьетнамского языка. В частности, используя вероятностные модели, набор грамматических правил, а также “ручной” с участием лингвистов процесс морфологической разметки вручную. Он позволяет решать три основные проблемы морфологического анализа текстов: (i) графематический анализ — выделение различных нестандартных элемен-

тов текста, присваивание им соответствующие графематических дескрипторов, сегментация предложений текста на слова и фразы; (ii) на основе сегментации и морфологической разметки поиск каждого слова в предложениях по морфологическому словарю; (iii) выполнение автоматического морфологического анализа вьетнамских текстов на основе Марковской модели и набора грамматических правил.

Основные результаты являются набором грамматических правил и аннотированным корпусом вьетнамских текстов, которые могут быть использованы в развитии средств обработки вьетнамского языка.

Дальнейший текст статьи построен следующим образом. Во втором разделе дается обзор исследований проблемы морфологического анализа вьетнамского текста. Новый подход автора описан в третьем разделе. В четвертом — представляются основные понятия, постановка задачи и описание модели автоматического морфологического анализа. В пятом — рассматривается система морфологических признаков вьетнамского языка. В шестом — описываются эксперименты. В заключении подводятся итоги.

2. Другие исследования

Вьетнамские ученые только недавно начали изучать область автоматической обработки текстов (Natural Language Processing, NLP). Насколько известно автору, его группа является первой, которая изучает проблему распознавания вьетнамских слов на основе статистических методов. Проблема сегментации вьетнамского предложения на слова исследовалась в [8–13], где выделялись два основных подхода: *подход на основе множества* [8–10] и *подход на основе обучения без учителя* [11, 12].

Первая группа методов относится к теории контролируемого обучения (обучение с учителем). Для реализации Динь [8] использовал модели WFST и нейронных сетей, Нгуен [10] — модели CRF (условных случайных полей) и SVM (метод главных векторов), Ле [9] — гибридные алгоритмы с *методом максимального сопоставления*. Эти методы основаны или на использовании словаря (из 34 000 слов), или аннотированных документов (около 1 400). При таких способах обучения используется ограниченное число различных слов. Авторы утверждают, что точность их методов более 90%,

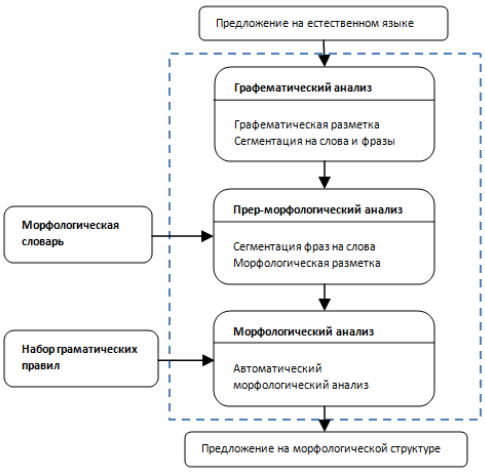

Рис. 1: Модель автоматического морфологического анализа.

но только на небольших наборах аннотированных документов.

При другом подходе Ха [11] применял модель “три-граммы” над большим набором документов, Тхань [13] использовал формулы взаимной информации (MI) и модель N-граммы с генетическим алгоритмом. Различия между идеями этой работы и работами других исследователей состоят в том, что, во-первых, у них не было процесса обучения для повышения точности статистической информации, во-вторых, способность связи между слогами в одном слове характеризуется различными формулами, а не только формулой MI [11] или максимальной вероятностью N-граммы [13], в-третьих, авторы использовали сравнительно небольшие наборы документов. Точности их алгоритмов составляли 50% [11] и 80% [13], что существенно ниже, чем у метода рассматриваемого в этой статье (95%).

3. Описание нового подхода

Для решении задачи морфологического анализа текстов вьетнамского языка рассмотрим новый подход, состоящий из трех этапов, представленных на рис. 1.

Графематический анализ представляет собой начальный этап процесса морфологического анализа, в ходе которого определяются различные нестандартные элементы текста (цифры и числа, даты в цифровых форматах, буквенно-цифровые комплексы, цифровознаковые комплексы и т. д.) и выполняются первые предварительные действия над текстом — выделение элементов текста и присваивание им соответствующие графематических дескрипторов. Далее проводится сегментация предложений текста на слова, фразы и графематические дескрипторы [3–6]. На выходе — подготовленный корпус, в котором все предложения сегментированы на слова и фразы, содержащие только слоги и графематические дескрипторы.

Пред-морфологический анализ производят вручную с участием лингвистов и использованием *морфологического словаря*.

Морфологический словарь — набор слов и фраз, которые содержатся в подготовленном корпусе, и соответствующие им возможные морфологические признаки. Процесс построения морфологического словаря сделан вручную с участием лингвистов, в котором соответствующие слова в подготовленном корпусе сохраняют все возможные морфологические признаки в квадратных скобках (части речи или грамматические категории). Также с участием лингвистов соответствующие фразы были сегментированы на слова, парсер анализирует каждое слово в узком контексте фразы и присваивает им соответствующие морфологические признаки. Возможно, что у слова будет найдено несколько частей речи. В этом случае программа указывает их всех через разделитель “запятая”. Например,

1. *quạt* [Noun, Verb].
2. *cái*[Affix] *quạt*[Noun].

В морфологическом словаре гарантируется, что соответствующие фразы имеют только одну сегментацию.

На вход для пред-морфологического анализа поступает подготовленный текст. Далее на основе поиска по морфологическому словарю выполняются две задачи:

- сегментация фраз на слова и установка графематических дескрипторов;
- морфологическая разметка всех возможных частей речи для каждого слова в предложениях.

На этом этапе морфологический анализ как таковой отсутствует, только сохраняются все возможные части речи соответствующего слова в предложениях. На выходе — пред-морфологический корпус, в котором все предложения сегментированы на слова, вставлены графематические дескрипторы и каждому слову приписываются в квадратных скобках соответствующие все возможные морфологические признаки.

Морфологический анализ производит последующую обработку текста. На входе — проаннализированное предложение — последовательность $w_1[T_1] \dots w_n[T_n]$, где w_i — слово, и соответствующий T_i — набор возможных морфологических признаков этого слова в этом предложении. На основе вероятностной модели определяются правильные последовательности морфологических признаков $t_1 t_2 \dots t_n$ этого предложения, где $t_i \in T_i$.

В модели вероятности оцениваются на основе специальных функций, полученных или из лингвистических знаний (набор грамматических правил), или в результате обучения на наборах данных. В результате обеспечивается удаление неправильных вариантов. Например, рассмотрим фразу "*cái quạt*". Слово "*quạt*" может быть существительным или глаголом, но по правилу грамматики: "Аффикс "*cái*" всегда ставится непосредственно перед существительным, поэтому модель удаляет вариант с частью речи "глагол" "*cái*[Affix] *quạt*[Verb]". Правильной морфологической разметкой в этом случае остается только "*cái*[Affix] *quạt*[Noun]".

4. Модель автоматического морфологического анализа

4.1. Основные понятия

Слово — оригинальный слог ("*của*", "*đã*") или соединение слогов ("*công_ việc*", "*chúng_ tôi*", "*thành_ công*").

Тэг — код одной группы частей речи или грамматических категорий слов, например, [Noun] — существительное, [Verb] — глагол. $\mathcal{T} = \{t_1, t_2, \dots, t_m\}$ — конечный набор тэгов.

Помеченное слово — слово с некоторыми своими помеченными тэгами, представляется в форме: $w[T]$, где $T = \{t\} \subseteq \mathcal{T}$ является множеством тэгов слова w. Например, в помеченном слове "*quạt* [Noun, Verb]", слово w является "*quạt*" и соответственно тэгами слова w являются [Noun] и [Verb].

Предложение является последовательностью слов: $s = w_1 w_2 \dots w_k$. Например, предложение "*Công_ việc của chúng_ tôi đã thành_ công*" записывается как $s = w_1 w_2 \dots w_5$, где w_1="*công_ việc*", w_2="*của*", ... и w_5="*thành_ công*".

Помеченное предложение является последовательностью помеченных слов.

Помеченное предложение $s' = w'_1[T_1] w'_2[T_2] \dots w'_k[T_k]$ называется *набором разметки* предложения $s = w_1 w_2 \dots w_l$ если удовлетворятся условиями: $l = k$ и $w_i = w'_i \quad \forall i = 1, \dots, k$. Например, помеченное предложение "*Công_ việc[Noun] của[Prepos] chúng_ tôi [Pronoun] đã[Particle] thành_ công[Verb]*" является одним из некоторых наборов разметки предложения "*Công_ việc của chúng_ tôi đã thành_ công*".

Набор предложений $\mathcal{C}$=$\{s_1, s_2, \dots, s_n\}$ является конечной совокупностью предложений, и соответствующие конечные совокупности помеченных предложений $\mathcal{C}' = \{s'_1, s'_2, \dots, s'_n\}$, в которых каждое помеченное предложение $s'_i \in \mathcal{C}'$ является набором разметки предложения $s_i \quad \forall i = \overline{1, n}$, называются *наборами помеченных предложений.*

Пусть Σ_C — множество всех слов в наборе $\mathcal{C}$, $w_\alpha, w_\beta \in \Sigma_C$ — слова. Обозначим $N(w_\alpha)$ — число появлений w_α в $\mathcal{C}$, $N(w_\alpha, w_\beta)$ число появлений $w_\alpha w_\beta$ в предложениях, принадлежащих $\mathcal{C}$.

Определим $N_1 = \sum_{w_\alpha \in \Sigma_C} N(w_\alpha)$ — число появлений слова w_α в $\mathcal{C}$, и вероятность появления w_α в $\mathcal{C}$ в качестве независимого слова

$$P(w_\alpha) = \frac{N(w_\alpha)}{N_1}.$$

Аналогично определим $N_2 = \sum_{w_\alpha, w_\beta \in \Sigma_C} N(w_\alpha, w_\beta)$ и вероятность появления пары $w_\alpha w_\beta$ в $\mathcal{C}$ в качестве независимого слова

$$P(w_\alpha, w_\beta) = \frac{N(w_\alpha, w_\beta)}{N_2}.$$

Пусть $t_\alpha \in \mathcal{T}$ — тэги. Также определим $N(w_\alpha, t_\alpha)$ — число появлений помеченного слова $w_\alpha[T_\alpha](t_\alpha \in T_\alpha)$ в $\mathcal{C}'$, и вероятность появления слова w_α с тэгом t_α

$$P(w_\alpha, t_\alpha) = \frac{N(w_\alpha, t_\alpha)}{N(w_\alpha)}.$$

4.2. Функция оценивания и функция контекста

Основная проблема морфологического анализа — определение тэгов каждого слова в одном предложении.

Пусть $s = w_1[T_1]w_2[T_2]\dots w_n[T_n] \in \mathcal{C}'$ — помеченное предложение, $t_k \in T_k, 1 \le k \le n$ — один из возможных тэгов слова w_k. Для решения проблемы морфологического анализа надо дать ответ на вопрос: "*Какова вероятность того, что слову w_k приписывается тэг t_k в предложении s*?".

В предложении определение тэгов каждого слова зависит от соседних слов и тэгов. В этом разделе опишем вероятностные функции над контекстами слова для определения тэгов.

Контекст описывает структуру текста с определенным порядком тэгов: $H = \langle t_1, t_2, \dots, t_l \rangle$, где тэг $t_i \in \mathcal{T}$ — элемент контекста, $1 \le i \le l$.

Пусть $\mathcal{H} = \{H_1, H_2, \dots, H_m\}$ — конечный набор контекстов.

О п р е д е л е н и е 1. *Функция контекста $F_c : \mathcal{H} \times \mathcal{T} \mapsto [0,1]$ — вероятностная функция, которая оценивает какова вероятности того, что данный тэг $t \in \mathcal{T}$ находится в контексте $H \in \mathcal{H}$.*

Например, для контекста $H = \langle$ [*Aff*],$t\rangle$ определим функцию контекста, оценивающую вероятности тэга t в контексте H:

$$R_1(H,t) = \begin{cases} 1 & \text{if } t\text{=[Noun];} \\ \\ 0 & \text{otherwise.} \end{cases}$$

В этом примере функция контекста $R_1(H,t)$ задает правило, что в условии контекста H: если тэг "[*Aff*]" находится перед тэгом t, то вероятность того, что тэг t является [Noun] равна 1 и иначе равна 0.

Одна функция контекста определяется на конкретном контексте каждого слова в предложении. Известно, что в одном предложении одно слово имеет разные контексты, завися от длины предложения. Соответственно разные контексты одного слова в предложении могут определять и разные функции контекста. Кроме того, над одним контекстом можно определить другие функции контекста. Например, над тем же контекстом H и его элементом (тэгом t), определим другую функцию контекста как

$$R_2(H,t) = P(t|[Aff]),$$

где $P(t|[Aff])$ — это условная вероятность, т. е. вероятность тэга t при условии, что предыдущий тэг в контексте является "$[Aff]$". Правило $R_2(H,t)$ получено из процесса обработки данных.

Функции контекста могут быть получены из лингвистических знаний или в процессе обработки данных.

Пусть задано помеченное предложение $s = w_1[T_1]w_2[T_2]\dots w_n[T_n] \in \mathcal{C}'$, $t_k \in T_k, 1 \le k \le n$ — один из возможных тэгов слова w_k, и $\mathcal{R}_{w_k}$ — набор всех функций контекста, определенных над разными контекстами слова w_k.

Определение 2. *Функция оценивания $F : \mathcal{C}' \times \mathcal{T} \mapsto [0,1]$, оценивающая какова вероятность того, что слово w_k приписывает тэг t в помеченном предложении s, определяется следующим образом:*

$$F(s,t) = \prod_{R(H,t)\in\mathcal{R}_{w_k}} R(H,t).$$

Использование функций оценивания повышает гибкость, эффективность применения лингвистических знаний и процессов обработки данных в построении Марковской модели.

4.3. Постановка задачи морфологического анализа

Задачу морфологического анализа можно сформулировать следующим образом. Пусть задано помеченное предложение $s = w_1[T_1]$ $w_2[T_2]\dots w_n[T_n]$. Необходимо найти разметку t^*, которая доставляет максимум

$$t^* = arg\max_{t_1^n} P(t_1^n|w_1^n) = arg\max_{t_1^n} \prod_{i=1}^{n} P(w_i|t_i)F(s,t_i), \qquad (1)$$

где $t_1^n = \langle t_1 \dots t_n \rangle$ — набор тэгов помеченного предложения s и для каждого тэги t_i слова w_i: $t_i \in T_i$; $w_1^n = w_1 w_2 \dots w_n$, $F(s,t_i)$ — функция оценивания тэга t_i, соответствующего слова w_i в помеченном предложении s.

Поясним последнюю формулу. Вероятность $P(t_1^n|w_1^n)$ соответствующих последовательностей слов w_1^n и тэгов t_1^n может быть преобразована по правилом Байеса как:

$$P(t_1^n|w_1^n) = \frac{P(w_1^n|t_1^n)P(t_1^n)}{P(w_1^n)}.$$

В предложении слова являются независимыми друг от друга, поэтому

$$P(w_1^n|t_1^n) = \prod_{i=1}^{n} P(w_i|t_i).$$

Так как в модели считаем, что вероятность появления тэга в предложении зависит от его контекстов в предложении, то

$$P(t_1^n) \approx \prod_{i=1}^{n} F(s, t_i).$$

Отсюда следует формула: $P(t_1^n|w_1^n) = \prod_{i=1}^{n} P(w_i|t_i)F(s, t_i)$.

4.4. Алгоритм Витерби

В этом разделе описывается использование алгоритма Витерби для оценивания значений уравнения (1) для всех возможных тэгов предложения $s = w_1[T_1]w_2[T_2]\dots w_n[T_n]$.

Алгоритм Витерби состоит из трех шагов: (i) инициализация; (ii) индукция; (iii) завершение и считывание.

Определим две функции:

1. функция $\delta_i(t)$, которая дает вероятность того, что слову w_i соответствует тэг t;

2. функция $\psi_{i+1}(t)$, которая дает наиболее вероятный тэг слова w_i при условии, что слову w_{i+1} соответствует тэг t.

Шаг инициализации определяется вероятность соответствующего тэга [BoS] (Begin of sentence) равная 1.0

$$\delta_1(t) = 0.0 \quad \forall t \in \mathcal{T}, \quad t \neq [BoS], \quad \delta_1([BoS]) = 1.0.$$

Шаг индукции основан на уравнении (1):

$$\delta_{i+1}(t^j) = \max_{t \in T_i}[\delta_i(t) \times P(w_{i+1}|t^j) \times F(t^j|t)], \quad 1 \leq j \leq \mathcal{T},$$
$$\psi_{i+1}(t^j) = arg\max_{t \in T_i}[\delta_i(t) \times P(w_{i+1}|t^j) \times F(t^j|t)], \quad 1 \leq j \leq \mathcal{T},$$

где $F(t^j|t)$ — функция оценивания над тэгом t^j и помеченным предложением $s = w_i[t]w_{i+1}[t^j]$.

Шаг завершения и считывания определяет набор тэгов $t^* = \langle t_1^*, t_2^*, \ldots, t_n^* \rangle$ соответствующего помеченного предложения $s = w_1[T_1]$ $w_2[T_2] \ldots w_n[T_n]$ следующим образом:

$$\begin{aligned} t_n^* &= arg \max_{t \in T_n} \delta_n(t), \\ t_i^* &= \psi_{i+1}(t_{i+1}^*), \qquad 1 \leq i \leq n, \\ P(t_1^*, \ldots, t_n^*) &= \max_{t \in \mathcal{T}} \delta_{n+1}(t). \end{aligned}$$

Алгоритм тэггинга:

1.**comment:** Given: a annotated sentence of length n
2.**comment:** Initialization
3. $\delta_1([BoS]) = 1.0$
4. $\delta_1(t) = 0.0$ for $t \neq [BoS]$
5.**comment:** Induction
6. **for i:= 1 to n step 1 do**
7. **for all tags** t^j **do**
8. $\delta_{i+1}(t^j) = \max_{t \in T_i}[\delta_i(t) \times P(w_{i+1}|t^j) \times F(t^j|t)]$
9. $\psi_{i+1}(t^j) = arg \max_{t \in T_i}[\delta_i(t) \times P(w_{i+1}|t^j) \times F(t^j|t)]$
10. **end**
11. **end**
12.**comment:** Termination and path-readout
13. $t_{n+1}^* = arg \max_{t \in \mathcal{T}} \delta_{n+1}(t)$
14. **for j:=n to 1 step -1 do**
15. $t_j^* = \psi_{j+1}(t_{j+1}^*)$
16. **end**
17. $P(t_1^*, \ldots, t_n^*) = \max_{t \in \mathcal{T}} \delta_{n+1}(t)$

---

5. Система морфологических тэгов вьетнамского языка

Вьетнамский язык характеризуется отсутствием *словоизменения* и наличием *аналитических форм*. В отличие от английского и русского слова во вьетнамском языке не изменяются по падежам и по числам. При морфологическом исследовании вьетнамского слова интересуются только частью речи, грамматическими категориями и множеством грамматических признаков слова.

Части речи вьетнамского языка представляют собой классы слов, выделяемые на основании сходства их синтаксических, морфологических и логико-семантических свойств. Каждой части речи свойствен свой набор грамматических категорий, причем этим набором охватывается абсолютное большинство слов данной части речи [2].

Часть речи или грамматическая категория — категория слов языка, определяемая морфологическими и синтаксическими признаками, имеющих:

1. одно и то же обобщенное лексическое значение;
2. одно и то же обобщенное грамматическое значение, или одинаковый набор морфологических признаков;
3. одни и те же синтаксические функции.

Многие слова, относящиеся к одной и той же части речи, могут быть сгруппированы в отдельный флективный класс (ФК), который описывает закон их словообразования, т. к. для характеристики системы окончаний слова-представителя ФК нет необходимости перечислять окончания всех его форм, а достаточно это сделать для нескольких типичных форм. По флективному классу при морфологическом анализе определяют постоянные параметры слова.

На основании этих признаков в морфологической системе вьетнамского языка выделяется девять основных частей речи:

1. [N] — *Существительные* — указывают на отдельные предметы, лица, имена людей, мест или организаций. Например, "*máy tính*", "*xe hơi*", "*bàn ghế*";
2. [Adj] — *Прилагательные* — обозначающие качественные признаки предмета. Например, "*đẹp*", "*xinh xắn*", "*dễ thương*", "*xấu xí*", "*thông minh*";
3. [ProN] — *Местоимения* — лишенные собственного лексического значения и употребляемые вместо имени существительного, прилагательного, имени числительного или наречия, не называя предмет или его характеристику, а лишь указывая на них. Некоторые основные местоимения во вьетнамском языке:
   - [ProN-P] — *Личные местоимения*: "*tôi*", "*chúng tôi*", "*mình*", "*anh*";
   - [ProN-I] — *Указательные местоимения*: "*đây*", "*đấy*", "*đó*", "*kia*";
   - [ProN-D] — *Определительные местоимения*: "*mỗi*", "*cả*", "*tất cả*";

- [ProN-Q] — *Вопросительные местоимения*: “*ai*”, “*gì*”, “*nào*”, “*mấy*”;
- [ProN-nD] — *Неопределенные местоимения*: “*nào đó*”, “*một số*”.

4. [V] — *Глаголы* — обозначают действие, процесс, состояние или качество. Например, “*viết*”, “*đọc*”, “*nghe*”, “*đi dạo*”, “*mua sắm*”;

5. [Adv] — *Наречия* — обозначающие признаки действия, признаки признака предмета. Принято говорить, что слова этого класса отвечают на вопросы “как?”, “где?”, “куда?”, “когда?”, “зачем?”, “с какой целью?”, “в какой степени?”. Некоторые основные наречии во вьетнамском языке:

   - [Adv-Q] — *Качественные наречия*: “*dần dần*”, “*bỗng*”, “*ào ào*”, “*chợt*”;
   - [Adv-T] — *Наречия времени*: “*ban ngày*”, “*ban đêm*”, “*buổi sáng*”;
   - [Adv-P] — *Наречия места*: “*đây*”, “*đấy*”, “*đó*”, “*ở đây*”, “*ở đây*”, “*ở đó*”;
   - [Adv-I] — *Указательные наречия*: “*thế này*”, “*như thế*”, “*như vậy*”;
   - [Adv-R] — *Наречие степени*: “*rất*”.

6. [PreP] — *Предлоги* — выражающие синтаксическую зависимость имен существительных, местоимений, числительных от других слов в словосочетаниях и предложениях. Например, “*bằng*”, “*cái*”, “*của*”, “*đang*”, “*dưới*”, “*gần*”, “*qua*”, “*theo*”, “*thành*”, “*trong*”, “*trước*”, “*tại*”, “*vào*”, “*với*”, “*rồi*”, “*xong*”;

7. [Conj] — *Союзы* — служебные слова, выражающие смысловые отношения между однородными членами простого предложения или между частями сложного предложения. Например, “*và*”, “*cả ... cả*”, “*không những ... mà còn*”;

   - [Conj-Cr] — *Сочинительный союз*: “*và*”, “*hay*”, “*nhưng*”, “*mà*”, “*song*”;
   - [Conj-Cl] — *Коррелативный союз*: “*và*”;
   - [Conj-S] — *Подчинительный союз*: “*là*”, “*rằng*”;

8. [Part] — *Частицы* — служебная часть речи — вносят различные значения, оттенки в предложение или служит для образования форм слова, например, "*đã*", "*đang*", "*sẽ*";

9. [Aff] — *Аффикс* — морфема, которая присоединяется к корню и служит для образования слов, например, "*cái*", "*sự*", "*ban*";

6. Заключение

Описанные алгоритмы были реализованы автором и апробированы на больших корпусах вьетнамских текстов, взятых из Интернета.

Автоматизация морфологической разметки текстов позволила сократить время подготовки текстов корпуса и уменьшить количество ошибок, совершаемых операторами при ручной обработке текстов. Наиболее "узким местом" в процессе автоматизации является этап ручной обработки, на котором вручную производится отбор и исправление ошибок. За счет улучшения алгоритмов разметки и фильтрования можно полностью ликвидировать этап ручной обработки.

Список литературы

[1] *Граничин О.Н., Кияев В.И.* Информационные технологии в управлении. — М.: Изд-во Бином. 2008. 336 с.

[2] *Евдокимова И.С.* Естественно-языковые системы: Курс лекций. — Улан-Удэ: Издательство ВСГТУ. 2006. - 92с.

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

[4] *Хьеу Л. Ч., Граничин О. Н.* Статистический способ выделения и словосочетаний из вьетнамских печатных текстов // Вестник СПбГУ. 2009. Серия 10. Вып. 3. С. 161–169.

[5] *Le Trung Hieu, Le Anh Vu, Le Trung Kien* An unsupervised learning and statistical approach for Vietnamese word recognition and segmentation // Joining in The 2nd Asian Conf. on Intelligent

Information and Database Systems. Hue City. Viet Nam. March 2010.

[6] *Хьеу Л. Ч.* Обучение без учителя и статистический подход для сегментации и распознавания вьетнамских слов // Стохастическая оптимизация в информатике. 2009. Вып. 5. С. 193–208.

[7] *Thu C.B., Hien P.* Ve mot xu huong moi cua tu dien giai thich. 2007. http://ngonngu.net/index.php?p=319.

[8] *Dien D., Kiem H., Toan N.V.* Vietnamese word segmentation // The Sixth Natural Language Processing Pacific Rim Symposium, Tokyo, Japan. 2001. P. 749–756.

[9] *Ha L.A.* A method for word segmentation in Vietnamese // Proc. of Corpus Linguistics 2003. Lancaster. UK. 2003.

[10] *Le H.P., Nguyen T.M.H., Roussanaly A., Ho T.V.* A hybrid approach to word segmentation of Vietnamese texts // In 2nd Int. Conf. on Language and Automata Theory and Applications. Tarragona. Spain. 2008.

[11] *Nguyen C.T., Nguyen T.K., Phan X.H., Nguyen L M., Ha Q.T.* Vietnamese word segmentationwith CRFs and SVMs: An investigation // In Proc. of the 20th Pacific Asia Conference on Language, Information and Computation (PACLIC 2006). Wuhan. CH. 2006.

[12] *Nguyen H., Nguyen H., Vu T., Tran N., Hoang N.* Internet and genetics algorithm-based text categorization for documents in Vietnamese // Research, Innovation and Vision of the Future. The 3rd International Conference in Computer Science, (RIVF 2005). Can Tho. Vietnam. 2005.

[13] *Nguyen T.V., Tran H.K., Nguyen T.T.T., Nguyen H.* Word segmentation for Vietnamese text categorization: an online corpus approach // Research, Innovation and Vision for the Future. The 4th International Conference on Computer Sciences. 2006.