Автоматическое выделение слов и словосочетаний из вьетнамских печатных текстов[1]

Ле Чунг Хьеу Санкт-Петербургский гос. университет
Ле Ань Ву Университет ELTE, Будапешт, Венгрия
Ле Чунг Кьен Гуэ Университет, Вьетнам
vkhhieukien@yahoo.com, leanhvu@inf.elte.hu, hieukien@hotmail.com

---

В работе рассматриваются способы применения методов статистической обработки для построения списка вьетнамских слов и словосочетаний. На основе полученных результатов могут быть подготовлены входные данные для автоматической обработки текстов (Natural Language Processing) на естественном вьетнамском языке.

Ключевые слова: распознавание слов и словосочетаний на вьетнамском языке, методы статистической обработки, адаптация.

1. Введение

За последние 20 лет область автоматической обработки текстов (Natural Language Processing, NLP) достигла значительных успехов в сфере разнообразного лексико-грамматического анализа, выявления темы, или поиска информации [1–4]. Тем не менее, большинство работ были проведены для западных языков индоевропейской группы, и не могут применятся к другим языковым группам. Таким образом, сегодня возникает необходимость в разработке инструментов для остальных языковых групп.

Вьетнамские ученные только недавно начали заниматься NLP [5–10]. Авторам неизвестны публикации ни по формальной лингвистики, ни по распознаванию формальных правил для определения границ вьетнамских слов и словосочетаний. Основные задачи обработки текстов на вьетнамском языке, такие как лексико-грамматический анализ, синтаксический анализ и т. п., очень сложны для вычислительной лингвистики.

Построение списка вьетнамских слов и разделение фраз по словам является важной проблемой не только для лингвистики, но

[1]©Ч. Х. Ле, А. В. Ле, Ч. К. Ле, 2008

имеет важное применение для автоматической обработки текстов на естественном вьетнамском языке. Эти проблемы достаточно трудны, из-за отсутствия конкретных определений, и не могут быть решены до конца. Раньше вьетнамские специалисты решали эти проблемы только пользуясь словарями [11–13].

Целью нашего проекта является построение общей лингвистической базы данных, которая могла бы свободно и легко использоваться для автоматической обработки вьетнамского языка.

В европейских языках слова в предложении разделяются пробелами. Поэтому начало и конец слова довольно просто определяются как человеком, так и компьютером. Во вьетнамском языке слова не разделяются пробелами. Слово может являться одним слогом, либо комбинацией слогов, расположенных вместе в предложении. К сожалению, нет определенных правил как эти комбинации строятся. Поэтому проблема определения границ вьетнамских слов в печатных текстах не проста. Более того, структура и смысл одного вьетнамского слова в одном конкретном предложении зависят не только от способов связи между слогами, но и от слогов, которые расположены рядом. Эти трудности восприятия иллюстрируются следующим примером. В предложении “hoc sinh hoc sinh hoc”, которое по-русски означает “школьник учит биологию”, все комбинации “hoc sinh”, “sinh hoc”, “sinh”, “hoc” являются вьетнамскими словами:

“hoc sinh” — школьник,

“sinh hoc” — биология,

“hoc” — учиться,

“sinh” — родиться.

Для разделения предложения на слова важно содержание предложения. Исходя из смыслы предложения определяется, какие комбинации слогов являются словами. Поэтому в рассматриваемом примере, учитывая его основную мысль, правильная расстановка границ такова: “hoc sinh|hoc|sinh hoc”.

Проблема разделения фраз по словам во вьетнамском языке так же трудна, как и проблема построения вьетнамских слов и словосочетаний.

На протяжении долгого времени вьетнамские, а также иностранные специалисты, разрешали эту проблему вручную. Однако построение списка слов вручную требует колоссальных усилий и все же не обеспечивает полноты словаря вьетнамских слов. Одна из причин этого — широкое использование вьетнамского языка в различных сферах со специальными словами, которые редко используется. Другая — в различных регионах могут использоваться 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 internet, iphone и т. п. требуется 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.

В данной статье предлагается новая система для построения списка вьетнамских слов, основанная на статистическом методе и информационной энтропии [14–15], а также двух новых алгоритмах. Первый — алгоритм разделения фраз, использующий статистические значения. Он разделяет фразы, упрощая их дальнейшую обработку. Второй — алгоритм адаптации. Он выполняет циклический процесс статистической обработки и разделения фраз для получения статистических значений для исходных данных.

Предлагаемая система может быть использована при решении задач распознавания образов или классификации. Система в состоянии обучиться разделять на классы элементы предъявляемой ей последовательности векторов, представляющих некоторые объекты, в условиях, когда информация о том, к какому классу принадлежит каждый из используемых в процессе обучения векторов, отсутствует. Система сама “собирает” выборочные векторы в некоторые компактные группы.

Статья построена следующим образом. Сначала дается краткое описание характеристик вьетнамского языка и определяются используемые в дальнейшем основные статистические функции. Далее предлагаются две функции: связи и сравнения, характеризую-

щие способности связи между слогами. В следующем разделе обсуждается модель системы, включающая алгоритм разделения фраз и алгоритм адаптации. В пятом разделе описывается применение предлагаемой системы к обработке набора текстовых данных, состоящего из 11 060 000 вьетнамских фраз. В заключение подводятся итоги и формулируются некоторые задачи для дальнейших исследований.

2. Известные подходы

Вьетнамские Буквы.

Буква является мельчайшей единицей слова. Определенный набор букв языка называется алфавитом. Вьетнамский алфавит состоит из 29 букв.

Вьетнамский Слог.

Слог — специальная единица, которая соответствует в одно и то же время слогу в фонологии, морфеме в морфологии, а также слову с точки зрения синтаксиса. Слог является последовательностью букв. Слоги разделяется друг от друга пробелами и знаками в текстах.

Пусть $\mathcal{L}$ — множество всех вьетнамских букв. Обозначим

$$\mathcal{S} = \{s = l_1 l_2 \dots l_k | l_1, l_2, \dots, l_k \in \mathcal{L}\}$$

множество вьетнамских слогов. $\mathcal{S} \subset \mathcal{L}^*$. Для всех $s = l_1 l_2 \dots l_k \in \mathcal{S}$ по определению положим $l(s) = k$ — длина слога s.

Вьетнамское Слово и Словосочетание.

Вьетнамское слово и словосочетание является либо слогом, либо комбинацией слогов.

Обозначим

$$\mathcal{W} = \{w = s_1 s_2 \dots s_k | s_1, s_2, \dots, s_k \in \mathcal{S}\}$$

множество вьетнамских слов. $\mathcal{W} \subset \mathcal{S}^*$. Для всех $w = s_1 s_2 \dots s_k \in \mathcal{W}$ определим $l(w) = k$ — длина слова w.

Вьетнамская Фраза.

Фраза является последовательностью слогов. Пусть

$$\mathcal{C} = \{c = s_1 s_2 \dots s_k | s_1, s_2, \dots, s_k \in \mathcal{S}\}$$

множество вьетнамских фраз в исследуемом наборе данных, $\mathcal{C} \subset \mathcal{S}^*$. Для всех $c = s_1 s_2 \dots s_k \in \mathcal{C}$ определим $l(c) = k$ — длина фразы c.

Перечислим некоторые свойства вьетнамских слогов.

Частота появления.

Для всех слогов $s \in \mathcal{S}$ обозначим $N(s)$ частоту появления слога s во вьетнамских текстах.

Для всех слогов $r, s \in \mathcal{S}$ обозначим $N(r, s)$ частоту появления во вьетнамских текстах слога r перед слогом s.

Распределение вероятности.

Для всех слогов $s \in \mathcal{S}$ определим

$$\mathbb{P}(s) = \frac{N(s)}{N1},$$

где $N(s)$ — частота появления слога s, $N1 = \sum_s N(s)$ — сумма частот появления всех слогов во вьетнамских текстах.

Для всех слогов $r, s \in \mathcal{S}$ определим

$$\mathbb{P}(r, s) = \frac{N(r, s)}{N2},$$

где $N(r, s)$ — частота появления слога r расположенного перед слогом s в текстах, $N2 = \sum_{r,s} N(r, s)$ — сумма частот появления всех пар слогов, расположеных вместе во вьетнамских текстах.

Энтропия.

Будем считать, что набор вьетнамских текстов расчленен на m частей: $B_1, B_2, \dots, B_m$. Для всех слогов $s \in \mathcal{S}$ считаем, что $N_i(s)$ —

частота появления s в подмножестве B_i. Распределение вероятности s в B_i задается следующим образом:

$$\mathbb{P}(s \in B_i) = \frac{N_i(s)}{N(s)}$$

Энтропия слога s определяется следующим образом:

$$\mathbb{H}(s) = -\sum_{i=1}^{n} \mathbb{P}(s \in B_i) \log[\mathbb{P}(s \in B_i)].$$

Пусть для всех слогов $r, s \in \mathcal{S}$ $N_i(r, s)$ — частота появления слога r расположенного перед s в B_i. Вероятность слога r, расположенного перед слогом s в B_i, задается так:

$$\mathbb{P}(r, s \in B_i) = \frac{N_i(r, s)}{N(r, s)}.$$

Энтропия пары слогов (r, s) определяется формулой

$$\mathbb{H}(r, s) = -\sum_{i=1}^{n} \mathbb{P}(r, s \in B_i) \log[\mathbb{P}(r, s \in B_i)].$$

Известно, что $\mathbb{P}(r, s) \neq \mathbb{P}(s, r)$ и $\mathbb{H}(r, s) \neq \mathbb{H}(s, r)$. Значение $\mathbb{P}(\cdot)$ характеризует частоту повторения элементов, $\mathbb{H}(\cdot)$ — дисперсию.

Посл- и Пред-энтропия:

Для слога $s \in \mathcal{S}$ определим

$$N_n(s) = \sum_r N(s, r) \qquad ; \qquad \mathbb{P}_n^s(r) = \frac{N(s, r)}{N_n(s)}$$

и

$$N_p(s) = \sum_r N(r, s) \qquad ; \qquad \mathbb{P}_p^s(r) = \frac{N(r, s)}{N_p(s)}$$

Посл-энтропия слога s:

$$\mathbb{H}_n = -\sum_r \mathbb{P}_n^{(s)}(r) \log[\mathbb{P}_n^{(s)}(r)].$$

Пред-энтропия слога s:

$$\mathbb{H}_p = -\sum_r \mathbb{P}_p^{(s)}(r) \log[\mathbb{P}_p^{(s)}(r)].$$

3. Функции Связи и Сравнения

Как отмечалось в предыдущем разделе, вьетнамские слова являются либо одним слогом, либо комбинацией слогов, у которых есть связь между собой. К сожалению, эти комбинации строятся не по определенным правилам. Для того, чтобы разрешить проблему выделения вьетнамских слов в тексте, необходимо ответить на два вопроса: “Где граница слова?” и “Какая комбинация слогов является словом?”. В этом разделе рассматриваются два статистических свойства, характеризующие распознаваемое слово. Первое предназначено для оценивания связи между слогами в этом слове, а второе — для оценивания отношения между этим словом и другими словами в одной фразе.

3.1. Функция Связи

Функция связи характеризует способность связи между слогами. Она основана только на статистике, строится по следующим функциям: появления, комбинации и условной вероятности. $F(s)$ определена для всех слогов $s \in \mathcal{S}$, $F(r, s)$ для пар слогов $r, s \in \mathcal{S}$. Вычисляется по частоте появления $N(\cdot)$, распределению вероятности $\mathbb{P}(\cdot)$ и Посл- и Пред- энтропиями $\mathbb{H}_n(\cdot), \mathbb{H}_p(\cdot)$ слогов.

Функция появления характеризует частоту появления элементов:

$$F_{ap}(s) = \log(N(s)),$$

где $N(s)$ — частота появления слога s в $\mathcal{C}$,

$$F_{ap}(r, s) = \log(N(r, s)),$$

где $N(r, s)$ — частота появления слога r, расположенного перед слогом s в $\mathcal{C}$.

Функция условной вероятности характеризует способность комбинации между элементами

$$F_{cd}(r, s) = \frac{\mathbb{P}(r, s)^2}{\mathbb{P}(r)\mathbb{P}(s)},$$

где $\mathbb{P}(\cdot)$ и $\mathbb{P}(\cdot, \cdot)$ — распределение вероятности.

Функция комбинации характеризует способность комбинироваться с другими слогами

$$F_{cb}(s) = \frac{\mathbb{H}_n(s) + \mathbb{H}_p(s)}{2},$$

где $\mathbb{H}_n(s)$ и $\mathbb{H}_p(s)$ — Посл- и Пред-энтропия слога s в $\mathcal{C}$,

$$F_{cb}(r, s) = \frac{\mathbb{H}_n(r) + \mathbb{H}_p(s)}{2},$$

где $\mathbb{H}_n(r)$ и $\mathbb{H}_p(s)$ — Посл- и Пред-энтропия слогов r и s соответственно.

Функция Связи.

О п р е д е л е н и е 1.

$$F(s) = F_{cd}(s), \tag{1}$$

$$F(r, s) = F_{cd}(r, s), \tag{2}$$

где $F_{cd}(\cdot)$ и $F_{cd}(\cdot, \cdot)$ — функция условной вероятности.

О п р е д е л е н и е 2.

$$F(s) = \frac{F_{ap}(s) * F_{cd}(s)}{F_{cb}(s)}. \tag{3}$$

$$F(r, s) = \frac{F_{ap}(r, s) * F_{cd}(r, s)}{F_{cb}(r, s)}, \tag{4}$$

где $F_{ap}(\cdot)$ и $F_{ap}(\cdot, \cdot)$ — функция появления, $F_{cb}(\cdot)$ и $F_{cb}(\cdot, \cdot)$ — функция комбинации, $F_{cd}(\cdot)$ и $F_{cd}(\cdot, \cdot)$ — функция условной вероятности.

Функцию связи по первым формулам построить просто. В процессе выполнения алгоритма она сходится устойчиво, но медленно. Вторую построить сложнее, но она выводит слова точнее. Сходится быстро, но не устойчиво (см. п. 4).

3.2. Функция Сравнения

В предложении слоги образуют слова, некоторые комбинации рядом стоящих слогов могут встречаться чаще, чем другие. Функция сравнения служит для сравнения строгости связи между слогами в предложении.

Функция сравнения использует функцию связи для сравнения каждой пары стоящих рядом слогов. Она строится с помощью пяти функций: логической, сравнения-1, сравнения-2, сравнения-3, сравнения-4.

Для фразы $c = s_1 s_2 \dots s_l \in \mathcal{C}$ определим.

Логическая функция.

$$
\begin{aligned}
B(c,0) &= 1, \qquad B(c,l) = 1, \\
B(c,k) &= \begin{cases} 1 & \text{если между } s_k \text{ и } s_{k+1} \text{ разделитель} \\ 0 & \text{иначе} \end{cases}, .0 < k < l.
\end{aligned}
$$

Функция Сравнения-1.

$$
\begin{aligned}
C_1(c,0,\theta) &= 0, \qquad C_1(c,l-1,\theta) = 0, \qquad C_1(c,l,\theta) = 0, \\
C_1(c,k,\theta) &= \begin{cases} 1 & \text{если } B(c,k-1) = 0 \text{ и } \Delta > \theta \\ 0 & \text{иначе} \end{cases}, .0 < k < l-1,
\end{aligned}
$$

где $\theta > 0,\ \Delta = F(s_{k-1}, s_k) + F(s_{k+1}, s_{k+2}) - 2F(s_k, s_{k+1})$.

Функция Сравнения-2.

$$
\begin{aligned}
C_2(c,0,\theta) &= 0, \qquad C_2(c,l,\theta) = 0, \\
C_2(c,k,\theta) &= \begin{cases} 1 & \text{если } B(c,k-1) = 0 \text{ и } \Delta > \theta \\ 0 & \text{иначе} \end{cases}, .0 < k < l,
\end{aligned}
$$

где $\theta > 0,\ \Delta = F(s_{k-1}, s_k) + F(s_{k+1}) - F(s_{k-1}) - F(s_k, sk+1)$.

Функция Сравнения-3.

$$
\begin{aligned}
C_3(c,0,\theta) &= 0, \qquad C_3(c,l-1,\theta) = 0, \qquad C_3(c,l,\theta) = 0, \\
C_3(c,k,\theta) &= \begin{cases} 1 & \text{если } \Delta > \theta \\ 0 & \text{иначе} \end{cases}, .0 < k < l-1,
\end{aligned}
$$

где $\theta > 0,\ \Delta = F(s_k) + F(s_{k+1}, s_{k+2}) - F(s_k, s_{k+1}) - F(s_{k+2})$.

Функция Сравнения-4. Для $\theta > 0$

$$\begin{aligned} C_4(c,0,\theta) &= 0, \qquad C_4(c,l,\theta) = 0, \\ C_4(c,k,\theta) &= \begin{cases} 1 & \text{если } F(s_k, s_{k+1}) < \theta \\ 0 & \text{иначе} \end{cases}, .0 < k < l. \end{aligned} \tag{5}$$

Функция Сравнения. Для $\theta_1 > 0$, $\theta_2 > 0$

$$C(c,k,\theta_1,\theta_2) = \begin{cases} 1 & \text{если } (*) \\ 0 & \text{иначе.} \end{cases} \qquad 0 \le k \le l, \tag{6}$$

$(*) - C_1(c,k,\theta_1) = 1 || C_2(c,k,\theta_1) = 1 || C_3(c,k,\theta_1) = 1 || C_4(c,k,\theta_2) = 1.$

4. Алгоритм

Рассмотрим систему построения списка слов из большого объема данных, состоящего из вьетнамских предложений, в которой используется алгоритм адаптации, представляющий собой циклический процесс статистической обработки для построения функции связи и функции сравнения на текущих данных, и алгоритм разделения фраз. В результате каждой итерации получаются новые данные, более близкие к списку реальных вьетнамских слов и словосочетаний. Используя эти данные, далее можно построить функцию связи, более приближенную к реальным значениям связи между слогами.

4.1. Алгоритм разделения фраз

Пусть задана фраза $c = s_1 s_2 \dots s_l \in \mathcal{C}$. Выберем некоторые коэффициенты $\theta_1 > 0, \quad \theta_2 > 0$.

1. Положим $k = 1$.

2. На k-ом такте вычислим значение функции сравнения

$$C(k, \theta_1, \theta_2),$$

если $C(k, \theta_1, \theta_2) = 1$, то разделяем фразу в месте s_k и s_{k+1} и устанавливаем $B(c, k) = 1$.

3. Увеличиваем номер такта $k = k + 1$.

4. Если $k < l$, то переход к п. 2. Иначе — выход.

В результате получились новые фразы, которые являются частями фразы c.

4.2. Алгоритм адаптации

Процесс сходимости: $Converge(\theta_1, \theta_2, \mathcal{C})$.

Пусть заданы $\mathcal{C}$ — множество начальных фраз, $\theta_1 > 0$, $\theta_2 > 0$, $\varepsilon > 0$.

1. Статистика на $\mathcal{C}$. Рассчитываем значения функции связи $F(\cdot)$ для всех элементов в $\mathcal{C}$. Положим $\mathcal{C}^{'} = \mathcal{C}$.

2. Для всех $c \in \mathcal{C}^{'}$ и функции связи $F(\cdot)$ выполняем алгоритм разделения. В результате получаем множество фраз $\mathcal{C}^{''}$.

3. Положим $\mathcal{C}^{'} = \mathcal{C}^{''}$.

4. Статистика на $\mathcal{C}^{'}$. Рассчитываем значения функции связи $F^{'}(\cdot)$ для всех элементов в $\mathcal{C}^{'}$.

5. Если $\|F - F^{'}\| < \varepsilon$ — выход, иначе $F = F^{'}$ и переход к п. 2.

Заметим, что с постоянными значениями параметров θ_1 и θ_2 алгоритм будет сходиться. В результате получается множество фраз $\mathcal{C}^*$.

Алгоритм адаптации.

Пусть задано $\mathcal{C}$ — множество начальных фраз. Выберем малый коэффициент $\varepsilon > 0$ и значения $\theta_1^0 > 0$, $\theta_1^n > 0$, $\theta_2^0 > 0$.

1. Положим $k = 1$ и выберем начальные значения $\theta_1 = \theta_1^0$ и $\theta_2 = \theta_2^0$.
2. В начале каждого k-го такта выберем $\theta_1 = \theta_1^0 - k\varepsilon$. Если $\theta_1 < \theta_1^n$ — выход.
3. Выполняем $Converge(\theta_1, \theta_2, \mathcal{C})$, в результате получаем новое множество фраз $\mathcal{C}^*$.
4. Увеличиваем номер такта $k = k + 1$, $\mathcal{C} = \mathcal{C}^*$.
5. Переход к п. 2.

Значения θ_1 и θ_2 зависят от $\mathcal{C}$ и функции связи.

5. Реализация

Данные для эксперимента из 8 808 вьетнамских литературных произведений, разделенных на 40 000 файлов, были получены с веб-сайта www.vnthuquan.net. Для удобства и повышения эффективности статистической обработки данные были нормированы так, чтобы в каждом из файлов было примерно равное количество предложений.

Начальные данные состоят из целых вьетнамских предложений, в которых имеется много знаков, чисел и имен (в том числе иностранных). В лингвистике такие знаки, как числа и имена, помогают разделять предложения на фразы по структуре и содержанию. Наряду с этим, в предложениях есть некоторые слоги, которые написаны с неправильной орфографией. Такие предложения нужно удалить из входных данных. Таким образом, обработка данных включает в себя: разделение предложений по знакам и числам, определение вьетнамских слогов, удаление предложений, в которых присутствуют невьетнамские слоги. В результате были получены данные, содержащие 11 060 000 фраз.

В эксперименте использовалась первая функция связи и были выбраны: малый коэффициент $\varepsilon = 0,2$ и значения $\theta_1^0 = 5,0$; $\theta_1^n = 2,0$; $\theta_2^0 = 1,0$. В результате применения алгоритма был получен список из 143 616 слов и словосочетаний. При сравнении с некоторыми словарями (LacViet dictionary, Vietnamese on-line dictionary:

http://www.saigon.com/ vietdict/,...) этот результат можно считать хорошим, так как 65% выделенных слов действительно являются словами или словосочетаниями, которые можно найти в словарях. В то же время, полученная выборка составляет 85-90% всех слов в словарях. В остальных же 15-20%, многие являются специальными словами. Наряду с этим 26% являются правильными вьетнамскими словами или словосочетаниями, но они не встречаются в словарях. Многие из них являются идиомами, поговорками, словосочетаниями. Кроме того, 4% были написаны с неправильной орфографией. Оставшиеся 5% не являлись вьетнамскими словами или словосочетаниями.

В ходе реализации алгоритма разделения фраз был получен результат, качество которого можно проверить на сайте:

http://edmpc.inf.elte.hu/ vu/index.php .

6. Заключение

В статье были предложены два алгоритма: алгоритм разделения фраз и алгоритм адаптации, которые используются для построения словарной основы системы автоматической обработки вьетнамских текстов. Большинство полученных слов были найдены в существующих словарях. Также было выделено много существующих вьетнамских слов, отсутствующих в каких-либо словарях.

В дальнейшем авторами планируется продолжить исследование эффективности системы с другими данными с тем, чтобы выделить более полный список слов и словосочетаний. Предполагается удаление неправильных слов и словосочетаний и включение новых правильных. Планируется запуск системы на специальных данных, состоящих из трудов по научным дисциплинам. Например, математике, информатике, экономике, биологи и т. п. Авторы надеются, что с помощью объемного набора данных и использования методов машинного обучения (например, Reinforcement Learning) можно будет добиться того, что функция связи будет сходиться к реальным значениям связи между вьетнамскими слогами.

При общей постановке задачи об автоматической обработке текстов на вьетнамском языке выделяются следующие подзадачи.

1. Построение списка вьетнамских слов.
2. Построение метода анализа вьетнамских предложений по словам.
3. Поиск ключевых слов и классификация документов.
4. Анализ текста — процесс получения высококачественной информации из текста на естественном языке.

Решения первой и второй подзадач рассмотрены в этой статье. С их помощью становится возможным решить и две оставшиеся. Полученный результат может использоваться в качестве исходных данных для третьей и четвертой подзадач. Третья проблема может решаться алгоритмами DATA MINING для общих языков. Последняя — за счет анализа структур текстов. Она является очень трудной проблемой, решение которой даст возможность автоматически "понимать" тексты.

Список литературы

[1] *Erjavec T., Ide N., Tufis D.* Development and assessment of common lexical specifications for six central and eastern European languages // In Proc. of the First International Conference on Language Resources and Evaluation. Granada, Spain. 1998. PP. 233-240.

[2] *Brill E.* Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging // Computational Linguistics. 1995. PP. 543-565.

[3] *Berger A., Pietra S. D., Ptetra V. D.* A maximum entropy approach to natural language processing // Computat. Ling. 22. 1996. PP. 39-71.

[4] *Covington M. A.* A fundamental algorithm for dependency parsing // In Proc. of the 39th Annual ACM Southeast Conference. ACM. New York, 2001. PP. 95-102.

[5] *Dien D., Hoi P. P., Hung N. Q.* Some lexical issues in electronic Vietnamese dictionary // In PAPILLON-2003 workshop on multilingual lexical databases. Hokaido University, Japan. 2003.

[6] *Dien D., Kiem H.* POS-tagger for English-Vietnamese bilingual corpus // In Workshop: Building and using parallel texts: Data driven machine translation and beyound. Canada: Edmonton. 2003. PP. 88-95.

[7] *Huyen N. T. M., Romary L., Rossignol M., Vu X. L.* A lexicon for Vietnamese language processing // Language Resources and Evaluation. 2007. PP. 291-309.

[8] *Dang V. B., Ho B. Q.* Automatic construction of English-Vietnamese parallel corpus through web mining // RIVF 2007 - Int. Conf. on Research, Innovation and Vision for the Future. March 05-09, 2007. Hanoi, Vietnam. 2007. PP. 261-266.

[9] *Hao C. X.* Vietnamese - Some Questions on Phonetics, Syntax and Semantics. — NXB Giao duc, Hanoi. Vietnam.

[10] *Diep Q. B., Thung V.* Vietnamese Grammar. Vol. 1. — Hanoi, Vietnam: NXB Giao duc. 1999.

[11] *http://www.saigon.com/ vietdict/.*

[12] *http://vdict.com/.*

[13] *http://dict.vietfun.com/.*

[14] *Shannon C. E.* A mathematical theory of communication // Bell System Technical Journal. 27. 1948. PP. 379-423.

[15] *Abramson N. M.* Information Theory and Coding. — McGraw-Hill, New York. 1963.