УДК 004.021

В. О. Лесков

ФОРМИРОВАНИЕ ЛЕКСИЧЕСКИ СВЯЗАННЫХ КОМПОНЕНТОВ ИНФОРМАЦИОННО-ТЕРМИНОЛОГИЧЕСКОГО БАЗИСА

Рассмотрена проблема информационной поддержки обучения иностранной лексике на основе внутриязыковых ассоциативных зависимостей. Сформулирована задача построения информационно терминологического базиса. В качестве ее решения предлагается нисходящий алгоритм формирования лексически связанных компонентов.

Ключевые слова: МЛ-технология, лексически связанный, ИТБ, ЛСК-методика, ЛС-компонент.

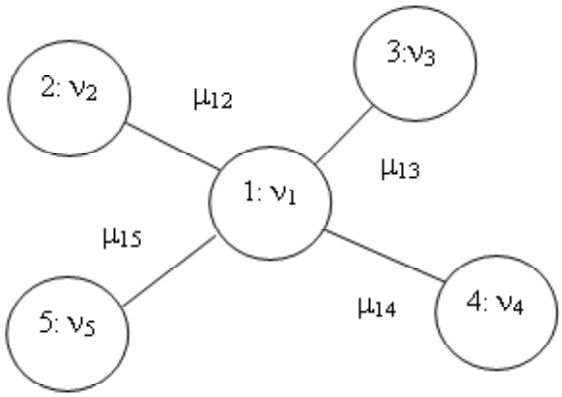
Обучение иностранной лексике подразумевает установление однозначных ассоциативных связей между терминами различных языков, обозначающими одно и то же понятие. Такие связи являются наиболее простыми и устанавливаются непосредственно в процессе обучения. Построение более сложных ассоциативных связей происходит естественным путем в процессе использования усвоенных знаний, в том числе это касается и установления ассоциативных связей внутри изучаемого языка. Однако современные средства обработки и анализа текстов позволяют уже на этапе обучения искусственно воссоздавать такие ассоциативные зависимости.

Так, например, назначение методики обучения на основе лексически связанных компонентов (ЛСК-методики) состоит в построении систем внутриязыковых ассоциативных связей с относительно жесткой структурой непосредственно в процессе обучения [1].

Для понимания сути этой методики обратимся к ее информационному обеспечению, т. е. к информационно-терминологическому базису (ИТБ). Дело в том, что ИТБ ЛСК-методики качественно отличается от классического ИТБ [2]. Различие состоит в информационных базисных компонентах: если в классическом ИТБ таким компонентом является термин, его качества и языковые аналоги, то в случае ЛСК-методики речь идет о более сложной структуре, а именно о лексически связанном компоненте (ЛС-компоненте).

Лексически связанный компонент. ЛС-компонент имеет следующую структуру (см. рисунок).

Лексему, связанную со всеми без исключения лексемами ЛС-компонента ИТБ, принято называть *основной лексемой*, а лексемы, имеющие только одну связь, – *связанными лексемами*.



Лексически связанный компонент ИТБ

Методика обучения строится на совместном применении двух алгоритмов:

– адаптивного алгоритма обучения, рассмотренного в [3], с тем исключением, что в качестве элементов обучающего материала выступают не лексемы, а ЛС-компоненты;

– алгоритма построения внутриязыковых ассоциативных полей.

Алгоритм построения внутриязыковых ассоциативных полей состоит в последовательной подаче к изучению элементов ЛС-компонента. Порядок выполнения этого алгоритма следующий:

– основная лексема – перевод, подсказка на иностранном языке (в рамках мультилингвистической адаптивно-обучающей технологии (МЛ-технологии));

– связанная лексема – перевод, подсказка на иностранном языке (в рамках МЛ-технологии);

– лексическое сочетание основной и связанной лексем – перевод сочетания, подсказка на иностранном языке (языковой аналог лексического сочетания, а не лексем по отдельности);

– переход к следующей связанной лексеме;

– переход к следующему лексически связанному компоненту.

Таким образом, на этапе обучения усваивается не только информация о языковых аналогах терминов, но и информация о лексических связях этих аналогов, что в свою очередь ведет к построению систем внутриязыковых ассоциативных связей.

Принципы формирования ЛС-компонентов. Пусть имеется ИТБ с данными о лексических связях. Для применения ЛСК-методики необходимо изменить структуру ИТБ, выделив основные лексемы и построить на их

лексемы:
1 – основная лексема
2,3,4,5 – связанные лексемы
лексические связи:
1-2, 1-3, 1-4, 1-5
количественные характеристики:
ν_i – абсолютная частота i-ой лексемы;
μ_{ik} – абсолютная частота сочетания i–ой и k–ой лексем

основе ЛС-компоненты. Рассмотрим пути решения этой задачи.

Для выделения основных лексем нужно выработать критерий на основе данных о частоте использования лексем и их лексических связях.

Процесс обучения и, соответственно, процесс формирования порций обучающей информации (ОИ) подчинены адаптивному алгоритму обучения. Согласно данному алгоритму, скорость забывания лексем изменяется по закону

$$\alpha_i^n = \frac{b_i^n}{\hat{h}k \dfrac{\sum\limits_k (1-p_{ik})\mu_{ik}}{\sum\limits_k \mu_{ik}} + 1}, \qquad (1)$$

где b_i^n – скорость забывания i-го элемента ОИ на n-м сеансе; $(1 - p_{ik})$ – вероятность знания k-го элемента ОИ, который лексически связан, т. е. порождает ассоциацию, с i-м элементом ОИ; μ_{ik} – относительная частота сочетания i-й и k-й лексем, отражающая силу ассоциативной связи; k – количество связанных лексем в компоненте; $\hat{h}$ – параметр, характеризующий активность ассоциативных связей, $0 < \hat{h} < 1$.

Вероятность незнания лексемы через функцию от времени может быть представлена в виде

$$p_i^n = p_i(t_i^n) = 1 - e^{-\alpha_i^n t_i^n}, \qquad (2)$$

где α_i^n – скорость забывания i-го элемента ОИ на n-м сеансе с учетом его связи с элементами ранее изучавшихся иностранных терминологий; t_i^n – время с момента последнего заучивания i-го элемента ОИ.

Согласно адаптивному алгоритму обучения, порции обучающей информации формируются с учетом критерия

$$Q_n = \sum_{i=1}^{N} p_i(t_i^n) q_i \to \min, \qquad (3)$$

где $p_i^n(t_i^n)$ – вероятность незнания i-го элемента ОИ в момент времени t_i^n; q_i – относительная частота, выражающая долю лексической единицы в тексте, подвергшемся статистической обработке при составлении частотного словаря, $0 < q_i < 1$:

$$q_i = \frac{q_i^{\max}}{V}, \qquad (4)$$

здесь $q_i^{\max} = \max q\{q_{i1}, q_{i2}, \ldots, q_{in}\}$ – абсолютная частота появления лексической единицы в тексте; $q_{i1}, q_{i2}, \ldots, q_{in}$ – частоты из мультилингвистического словаря, если речь идет о мультилингвистической адаптивно-обучающей технологии [4].

Таким образом, для обеспечения оптимального значения Q_n к концу n-го сеанса обучения необходимо найти M_n максимальных членов суммы в критерии, индексы которых и определят очередную порцию ОИ, выдаваемую обучаемому для запоминания. Процедура поиска индексов для терминов записывается следующим образом:

$$\begin{gathered} u_1 = \arg\max p_i(t_i^n) q_i,\ 1 \le i \le N, \\ u_2 = \arg\max p_i(t_i^n) q_i,\ 1 \le i \le N, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ u_{M_n} = \arg\max p_i(t_i^n) q_i,\ 1 \le i \le N, \\ i \ne u_j,\ j = 1, 2, \ldots, M_n - 1, \end{gathered} \qquad (5)$$

где $\arg\max \{a_i\} = i^* $ – индекс $i^* \in U$ максимального значения a_i, т. е. $a_i^* = \max a_i$; $\{u_1, \ldots, u_{Mn}\} = U^*$ – та порция ОИ, которая выдается для заучивания на n-м сеансе.

С учетом применения методики обучения на основе лексически связанных компонентов за u_i берется не термин, а лексическое сочетание основной и связанной лексемы, если наибольший вес приобрела связанная лексема, или лексически связанный компонент, если наибольший вес приобрела основная лексема.

Для выработки критерия отбора основных лексем рассмотрим ИТБ в момент времени t_{n+1}, когда прошло уже некоторое время с тех пор как базис был пройден. Допустим, что ИТБ был построен как совокупность ЛС-компонентов, тогда основные лексемы, согласно приведенному выше адаптивно-обучающего алгоритму, будут заучены намного лучше, чем их связанные лексемы:

$$p_i^{n+1} = p_i(t_i^{n+1}) = 1 - e^{-\alpha_i^{n+1} t_i^{n+1}} \to \max. \qquad (6)$$

Значения $p_{\text{зн}i}^{n+1}$ позволяют определить, какие из лексем ИТБ более других подходят на роль основных, но это возможно только к концу обучения при условии, что базис уже сформирован как совокупность ЛС-компонентов.

Перед нами стоит обратная задача – сформировать ИТБ до начала обучения. Однако по частотным характеристикам лексем и лексических сочетаний можно попытаться оценить значение $p_{\text{зн}i}^{n+1}$ для всех лексем базиса и, исходя из этого набора значений, построить критерий для выделения основных лексем.

Предположим, что нам удалось вычислить оценку $p_{\text{зн}i}^{n+1}$ для каждой лексемы, тогда, учтя относительную частоту лексем q_i, построим искомый критерий:

$$L_i = \hat{p}_{\text{зн i}}^{n+1} q_i \to \max \qquad (7)$$

Теперь найдем значение $\hat{p}_{\text{зн i}}^{n+1}$. Подставив выражение (1) в (6), получим

$$\hat{p}_{\text{зн}i}^{n+1} = e^{-\frac{b_i^{n+1}}{\hat{h}k \frac{\sum\limits_k p_{\text{зн}ik}^{n+1} \mu_{ik}}{\sum\limits_k \mu_{ik}} + 1} t_i^{n+1}}, \qquad (8)$$

где t_i^{n+1} – время с последнего заучивания i-й лексемы, о котором нам ничего не известно; b_i^{n+1} – скорость забывания i-й лексемы к моменту времени t_i^{n+1}, о которой нам также ничего не известно и которая вычисляется итеративно в процессе обучения; $p_{\text{зн}ik}^{n+1}$ – вероятность знания k-й лексемы, которая лексически связана, т. е. порождает ассоциацию, с i-й лексемой.

Здесь нужно заметить, что задача выработки критерия для выделения основных лексем сводится к тому, чтобы для каждой лексемы было рассчитано значение (7) и это значение было тем больше, чем больше эта лексема подходит на роль основной. Это обстоятельство позволяет нам во многом смириться с частичным незнанием элементов выражения (8).

Так, например, для разбивки ИТБ на ЛС-компоненты совершенно не важны значения t_i^{n+1} и b_i^{n+1}, которые мы заменим соответственно на условную безразмерную единицу времени и на b_0, получаемое из условия $p = 0{,}5$, $t = 1$ для всех лексем ИТБ. Также примем за единицу оценку параметра h.

Что касается вероятности $p_{\text{зн}ik}^{n+1}$, то ее мы можем оценить через относительную частоту k-й лексемы, так как к

концу обучения, согласно адаптивно-обучающему алгоритму, вероятности знания всех лексем будут прямо пропорциональны их относительным частотам. Тогда выражение для оценки $p_{\text{зн}i}^{n+1}$ будет иметь вид

$$\hat{p}_{\text{зн}i}^{n+1} = e^{-\frac{b_0}{k\frac{\sum_k q_k \mu_{ik}}{\sum_k \mu_{ik}}+1}}, \qquad (9)$$

где $b_0 \approx 0,7$ (получено из условия $p = 0,5$, $t = 1$); q_k и μ_{ik} содержатся в ИТБ.

Подставив выражение (9) в (7), получим критерий для выделения основных лексем:

$$L_i = e^{-\frac{0,7}{k\frac{\sum_k q_k \mu_{ik}}{\sum_k \mu_{ik}}+1}} q_i \to \max. \qquad (10)$$

Таким образом, мы получили значение L_i для каждой лексемы ИТБ. Теперь нам необходимо определить количество основных лексем, и задачу о формировании ИТБ как совокупности ЛС-компонентов можно будет считать решенной.

Задачу о нахождении оптимального количества основных лексем решим с помощью перебора. Для этого введем функцию качества ИТБ как совокупности ЛС-компонентов от числа основных лексем:

$$L(n) = \sum_i q_i e^{-\frac{0,7}{k\frac{\sum_k q_k \mu_{ik}}{\sum_k \mu_{ik}}+1}} \to \max, \qquad (11)$$

где $L(n)$ показывает сумму взвешенных вероятностей знания лексем по всему базису. Чем больше эта сумма, тем более удачно построен базис. Отсюда, максимизируя $L(n)$, получаем оптимальное значение количества основных лексем.

Нисходящий алгоритм формирования ЛС-компонентов. Алгоритм формирования ЛС-компонентов состоит из следующих фаз:

1. Подготовка ИТБ.

1.1. Для каждой лексемы ИТБ вычисляется значение L_i (10).

1.2. ИТБ упорядочивается по убыванию значения L_i таким образом, что чем меньше будет порядковый номер лексемы, тем выше вероятность образования на ее основе ЛС-компонента.

1.3. Данные о лексических связях упорядочиваются по убыванию значения $qk\mu_{ik}$. Тем самым увеличивается вероятность попадания в ЛС-компонент тех связанных лексем, которые более всего могут улучшить качество ИТБ (11).

2. Поиск оптимального количества основных лексем.

2.1. Осуществляется перебор возможного количества основных лексем k от 1 до объема ИТБ (возможно сужение разработчиком интервала поиска).

2.2. Для текущего значения k определяются основные лексемы (k первых лексем ИТБ).

2.3. Для выбранных основных лексем определяются связанные лексемы (как правило, задается максимум количества связанных лексем).

2.4. Подсчитывается значение функции качества (11).

2.5. Если перебор окончен, то переход к п. 2.6, если иначе, то возврат к п. 2.1.

2.6. Определяется максимум функции качества (оптимальное число основных лексем $k_{\max}$).

3. Формирование ИТБ как совокупности ЛС-компонентов. Искомый ИТБ получается при выполнении пп. 2.2 и 2.3 для $k_{\max}$ основных лексем.

Алгоритмы представленного выше вида будем называть *нисходящими* (Н-алгоритмы), так как определение связанных лексем (см. п. 2.3) происходит непосредственно из свойств текущей основной лексемы (см. п. 2.2), т. е. сверху вниз.

Приведем пример работы такого алгоритма.

Настраиваемые параметры базиса:

– объем базиса в терминах (1 000);

– максимальное количество связей на одну лексему (10);

– максимальное значение абсолютной частоты лексем (100/50 000);

– максимальное значение частоты сочетаний лексем (20/50 000);

– объем материала, по которому произведен частотный анализ (50 000);

– коэффициент связанности лексем (1).

Тестирование Н-алгоритма дает следующие результаты:

– максимум $L(n)$, равный 0,496 118 565 143 325, приходится на 188 основных лексем;

– количество итераций – 4 981 096;

– время исполнения – 3,41 с;

– количество элементов, не задействованных в ЛС-компонентах, – 320.

Таким образом, в данной статье рассмотрена проблема информационной поддержки ЛСК-методики: поставлена и решена задача формирования ИТБ как совокупности ЛС-компонентов; выработаны критерий для выделения основных лексем ИТБ и функция качества построения ИТБ; разработаны общие принципы решения поставленной задачи, реализованные в нисходящем алгоритме формирования ЛС-компонентов, применение которого к ИТБ является ее частным решением.

Библиографический список

1. Ковалев, И. В. Внутриязыковые ассоциативные поля в мультилингвистической адаптивно-обучающей технологии / И. В. Ковалев, В. О. Лесков, М. В. Карасева // Системы упр. и информ. технологии. 2008. № 3.1 (33). С. 157–160.

2. Карасева, М. В. Модель архитектуры мультилингвистической адаптивно-обучающей технологии / М. В. Карасева, И. В. Ковалев, Е. А. Суздалева // Новые информационные технологии в университетском образовании : тез. докл. Междунар. науч.-метод. конф. / Кемер. гос. ун-т. Кемерово, 2002. С. 204–205.

3. Растригин, Л. А. Адаптивное обучение с моделью обучаемого / Л. А. Растригин, М. Х. Эренштейн. Рига : Зинатне, 1988.

4. Ковалев, И. В. Информационно-алгоритмическое обеспечение мультилингвистической технологии обучения / И. В. Ковалев, А. А. Ступина, Е. А. Суздалева // Современное образование: массовость и качество : материалы регион. науч.-метод. конф. Томск, 2001 . С. 98–99.

V. O. Leskov

FORMATION OF LEXICALLY RELATED COMPONETS OF INFORMATION-VOCABULARY BASIS

The problem of information support of foreign vocabulary training based on intralingua associative relationships is considered. The task of the information-vocabulary basis formation is defined and resolved by top-down algorithm of the lexically related components building.

Keywords: ML-technology, lexically related, ITB, LRC-methodology, LR-component.

УДК 681.518.22-25

В. В. Шишов, В. А. Ильин, Н. А. Петрова

ИНТЕРАКТИВНАЯ ИНФОРМАЦИОННАЯ СИСТЕМА ДЕНДРОКЛИМАТИЧЕСКОГО МОНИТОРИНГА[1]

Впервые для РФ спроектирована интерактивная информационная дендроклиматическая система, которая включает обновляемую реляционную базу дендрохронологической и климатической информации на базе SQL-технологий, а также функциональное наполнение, объединяющее разнообразное специальное программное обеспечение по обработке дендроклиматической информации, адаптированное для разных операционных платформ.

Ключевые слова: интерактивная информационная система, TCL-технология.

С начала 90-х гг. XX в. в России формируется единая государственная система экологического мониторинга, в состав которой входит раздел дендрохронологического и дендроклиматического мониторинга, т. е. информационная система слежения, оценки и прогноза изменений годичного прироста деревьев и определяющих этот прирост факторов [1–7].

В настоящее время накоплен уникальный дендрохронологический материал, включающий не только данные по обширной территории России, но и данные по Западной Европе, Южной и Северной Америке, Азии [1–3; 5; 7–15]. По сравнению с другими прямыми и косвенными источниками информации о состоянии биосферы дендрохронологический материал, или временные ряды (ВР) радиального прироста деревьев (древесно-кольцевые хронологии, ДКХ) имеют важные особенности: высокое временное разрешение (1 год); достаточную длительность (ДКХ живых деревьев покрывают интервал до 700...800 лет, с использованием ископаемой древесины – несколько тысячелетий, что позволяет оценивать изменение климата); четкую количественную основу; практически повсеместное распространение на территории суши по земному шару. Система таких пространственно-распределенных ДКХ характеризует основное поведение и динамику одного из важнейших лесных компонентов биосферы Земли. Сами ДКХ, представляющие собой временные ряды или случайные функции (СФ) с дискретным аргументом, несут разнообразную информацию о факторах, определяющих прирост деревьев: внутренних (организменных), фитоценотических и внешних.

Совокупность пространственно-распределенных ДКХ и климатических временных рядов является исходным источником информации для выявления, анализа и моделирования изменений в росте древесных растений, имеющих как специфические (региональные) особенности, так и общие для обширной территории России и Западной Европы характеристики. Такие изменения могут быть ассоциированы с сигналом, характеризующим взаимодействие биосферы с факторами климатической природы.

Выявить, проанализировать и визуализировать подобные закономерности в приросте древесных растений в связи с изменениями климатических факторов можно при помощи разнообразных математико-статистических методов, как новых, так и традиционно используемых в дендроклиматологии, реализованных в виде программного обеспечения с бесплатной лицензией. Но такие задачи являются очень трудоемкими и затратными по времени, так как связаны с обработкой больших массивов данных, а также обобщением и визуализацией пространственно-распределенной информации на огромных территориях.

В связи с этим представляется важной автоматизация решения подобного рода задач на основе соответствующих специально разработанных систем обработки информации, которые позволяли бы решать их в полуавтоматическом режиме с минимальными затратами по времени. Такие системы могут стать информационной основой дендроклиматической системы мониторинга России при выявлении и анализе системных связей и закономерностей в приросте древесных растений в связи с меняющимся климатом на территории страны.

[1] Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект № 09-05-00900-а, проект № 09-04-00803-а).