УДК 004.9

Л.Э. Чалая, А.В. Чижевский, Ю.Ю. Шевякова

МЕТОД ФОРМИРОВАНИЯ ЗАПРОСОВ В СИСТЕМАХ ПОИСКА МУЛЬТИЯЗЫЧНОЙ ИНФОРМАЦИИ

Аннотация. В статье рассмотрен подход к построению систем автоматизированного поиска мультиязычной информации, особенностью которого является возможность существенного уменьшения неоднозначности терминов на основе применения количественных показателей подобия текстов. В практических приложениях данный подход может быть эффективно использован в задачах формирования дайджестов и тезаурусов.

Ключевые слова: система поиска информации, формирование запросов, оценка подобия текстов, мультиязычный текст, многозначность терминов.

Введение

В настоящее время, работы, осуществляемые в области создания систем поиска мультиязычной информации (СПМИ), основываются главным образом на переводе запросов [1]. Такие системы должны обеспечивать решение следующих задач: перевод терминов запроса; снятие многозначности терминов, если существуют возможные замены; присвоение весов терминам переведенного запроса.

Первая задача относится к переводу запросов. Принимая во внимание то, что в настоящее время не всегда имеются качественные и полные двуязычные словари для всех пар языков, рассмотрим технику перевода смешанных текстов, где основным является русский язык. Действительно, как правило, можно найти двуязычные словари, пригодные для качественного перевода слов языка L_1 на русский язык или русских слов на язык L_2. Целесообразно обеспечить возможность перевода текстов, представленных на языке L_1, на язык L_2, даже если отсутствуют необходимые для этого словари. Помимо использования словарей как средства перевода, эффективным является применение техники, основанной на выравнивании корпусов. Главная

 ISSN 1562-9945

идея глобального подхода к построению СПМИ состоит в построении тезауруса ассоциаций между терминами различных языков. Этот тезаурус используется впоследствии как средство перевода терминов запроса.

Вторая задача касается снятия многозначности переведенных запросов. Большая часть существующих методов снятия многозначности основывается главным образом на использовании выровненных корпусов [2].

В данной работе предлагается метод для уменьшения неоднозначности запросов в СПМИ, основанный на анализе мер подобия терминов и запросов, в качестве которых используются лексикографический порядок и коэффициент корреляции.

Общий подход к построению СПМИ

Общая схема обработки информации в СПМИ подход к созданию СПИ со смешанными языками представлен схематично на рис. 1. Этот подход к построению СПМИ основан, прежде всего, на специфике перевода запросов.

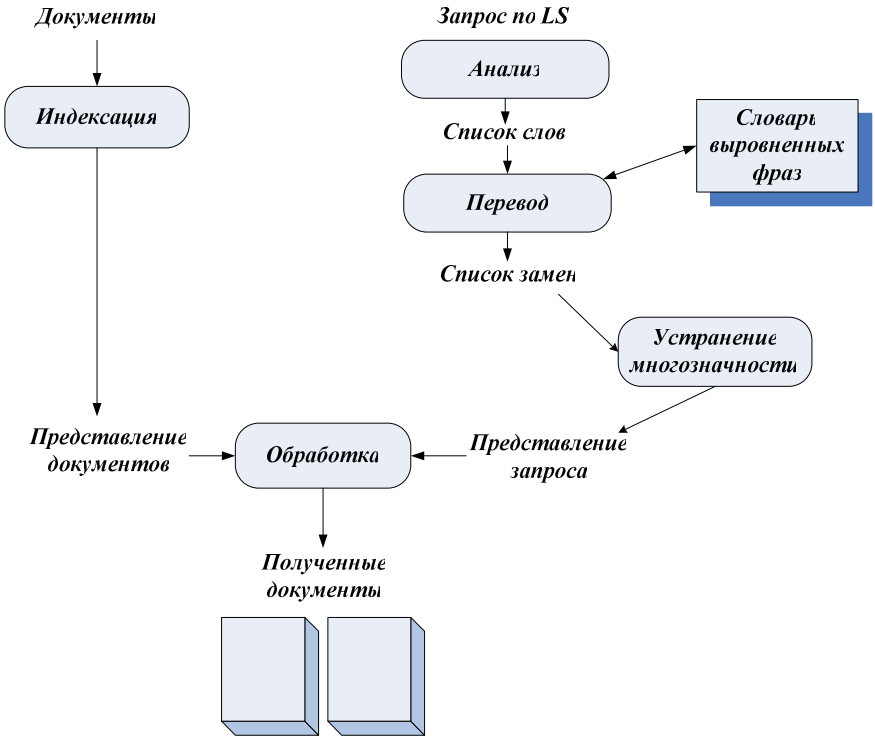

Рисунок 1 - Общая схема обработки информации в СПМИ

Наибольшую сложность в реализации приведенной схемы вызывает решение следующих задач: представление мультиязычной информации, перевод мультиязычных запросов, снятие многозначности переведенных запросов.

В общем случае, документы текстовых наборов индексируются согласно их языку, индекс также строится на этом языке. Обработка запроса осуществляется следующим образом. Сначала первоначальный запрос анализируется для извлечения простых слов. Каждое слово переводится затем на выходной язык (с использованием одного из возможных методов перевода). Список переведенных терминов индексируется с целью устранения пустых слов и подвергается обработке для обеспечения совместимости его представления с представлением разыскиваемых документов.

Рассмотрим особенности процедур устранения многозначности переведенных запросов в СПМИ.

Метод устранения неоднозначности запросов в СПМИ

Главная проблема методов перевода простых слов, основанного на использовании словарей, часто состоит в том, что словари предлагают несколько возможных вариантов перевода для заданного термина. Эти методы обычно комбинируются со стратегиями устранения многозначности для того, чтобы облегчить проблему однозначности переводов. В сущности, эта проблема состоит в том, чтобы выделить термин, который лучше всего соответствует контексту запроса.

Существуют три стратегии устранения многозначности: по контексту запроса; по мерам подобия; с применением бинаправленного перевода. Рассмотрим подробнее стратегию устранения многозначности, основанную на применении мер подобия.

Техника устранения многозначности, основанная на анализе подобия (рис.2), состоит в осуществлении моноязычного поиска для каждого термина входного запроса и для каждого из этих переводов на язык основного корпуса. Таким образом, каждый термин входного запроса дополняет список документов, начиная с части корпуса, представленной на входном языке. Также список документов на выходном языке дополняется каждым из переводов этого термина. Каждый из этих списков сравнивается со входным списком. Это сравнение основывается на таких мерах подобия, как лексикографический порядок и коэффициент соответствия Спирмена, которые позволяют обнаруживать список документов, наиболее близкий к входному списку. Это способствует выбору наилучшего перевода среди всех полученных.

 ISSN 1562-9945

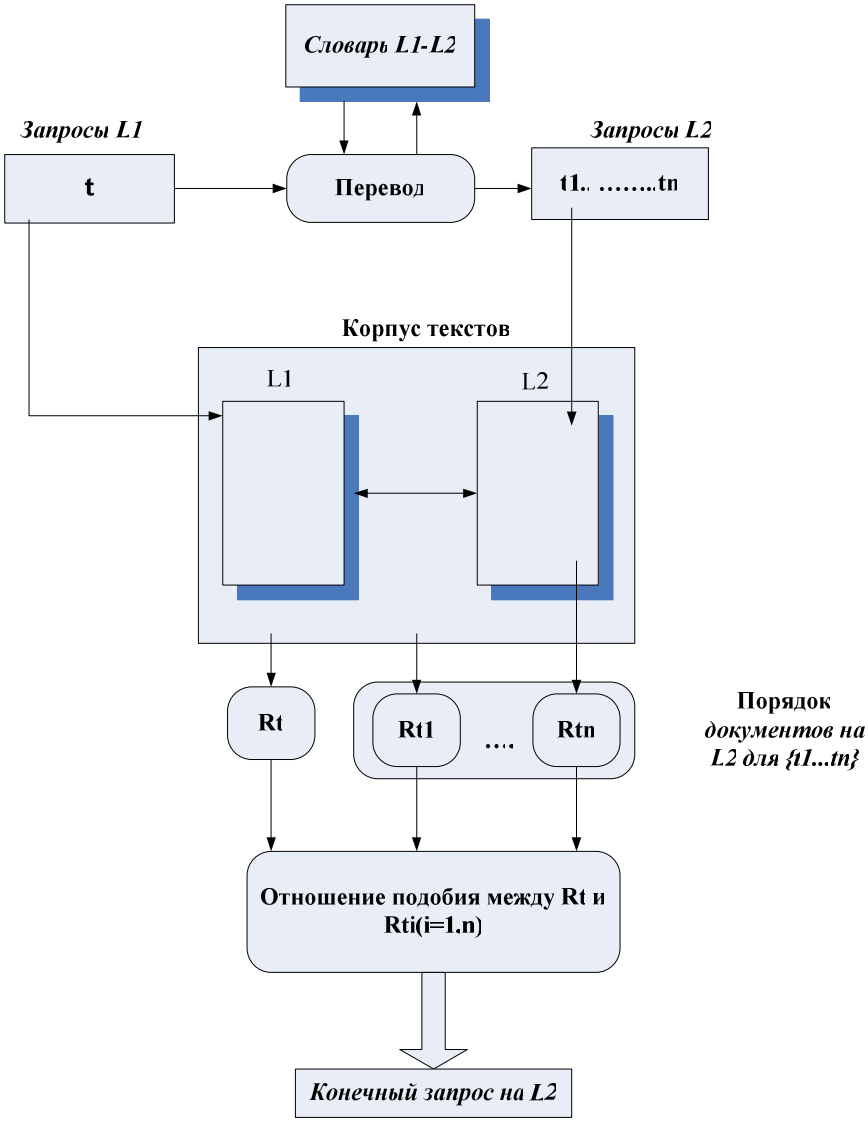

Рисунок 2 - Схема устранения неоднозначности запросов, основанная на анализе подобия

При этом важным является то, что применяемые меры подобия учитывают помимо степени соответствия документов их порядок в списке документов. Рассмотрим пример, иллюстрирующий предлагаемую стратегию.

Пример. Пусть $t_i^{L_1}$ – термин на языке L_1 входного запроса Q^{L_1}.

Пусть $\{t_1^{L_2}, t_2^{L_2}, t_3^{L_2}\}$ – три перевода на языке L_2 термина $t_i^{L_1}$.

Предположим, что у нас есть выровненный корпус, составленный из 20 документов, обозначенный как $D^{L_1} = \{d_1^{L_1}, \ldots, d_{20}^{L_1}\}$ и $D^{L_2} = \{d_1^{L_2}, \ldots, d_{20}^{L_2}\}$, а корпус $D_{t_i^{L_1}}$ (соответственно, корпус $D_{t_j^{L_2}}$) представляет документы, выбранные по термину $t_i^{L_1}$ (соответственно, по

термину $t_j^{L_2}$). Предположим также, что: $D_{t_i^{L_1}} = \{d_2^{L_1}, d_5^{L_1}, d_8^{L_1}, d_{10}^{L_1}, d_{20}^{L_1}\}$ – набор документов, выбранных для $t_i^{L_1}$ на языке L_1 ; $D_{t_1^{L_2}} = \{d_2^{L_2}, d_6^{L_2}, d_7^{L_2}, d_{10}^{L_2}, d_{12}^{L_2}\}$; $D_{t_2^{L_2}} = \{d_2^{L_2}, d_5^{L_2}, d_{13}^{L_2}, d_9^{L_2}, d_{11}^{L_2}\}$; $D_{t_3^{L_2}} = \{d_1^{L_2}, d_8^{L_2}, d_7^{L_2}, d_9^{L_2}, d_{12}^{L_2}\}$.

Эти совокупности уже предполагаются упорядоченными по значимости. Например, в $D_{t_i^{L_1}}$, $d_2^{L_2}$ рассматривается как наиболее существенный документ, а $d_{20}^{L_1}$ как наименее существенный документ (с точки зрения значимости для всей системы).

Отбор наиболее предпочтительных терминов, представляющих окончательный выходной запрос, осуществляется по следующей процедуре:

– изначально каждому документу в выровненном корпусе присваивается свой ранг. Этот ранг определяют, рассматривая степень значимости каждого документа в языке L_1 в соответствии с входным запросом. Наиболее существенному документу присваивают ранг 1, а наименее существенному ранг n. Число n представляет собой порог, который зависит от количества существенных документов в выровненном корпусе. Несущественный документ имеет ранг $K \geq N+1$ (N – число документов в выровненном корпусе). Обозначим как $r(d_i^{L_1})$ ранг документов $d_i^{L_1}$. Для документов на языке L_2 положим просто, что $r(d_i^{L_2}) = r(d_i^{L_1})$.

В рассматриваемом примере мы имеем:

$D_{t_i^{L_1}} = \{d_2^{L_1}, d_5^{L_1}, d_8^{L_1}, d_{10}^{L_1}, d_{20}^{L_1}\}$; $N = 20$;

$r(d_2^{L_1}) = 1, r(d_2^{L2}) = 1;\ r(d_5^{L_1}) = 2, r(d_5^{L_2}) = 2;$

$r(d_8^{L_1}) = 3, r(d_8^{L_2}) = 3;\ r(d_{10}^{L_1}) = 4, r(d_{10}^{L_2}) = 4;\ r(d_{20}^{L_1}) = 5, r(d_{20}^{L_2}) = 5.$

Для $d_i^{L_1} \notin D_{t_i^{L_1}}$ положим $r(d_i^{L_1}) = \alpha$, $\alpha > N$;

– для каждого перевода $t_j^{L_2} \in T^{L_2}(t_i^{L_1})$ выбираем n лучших документов $D_{t_j^{L_2}} = \{d_1^{L_2}, \ldots, d_n^{L_2}\}$ на языке L_2 в выровненном корпусе. Эти документы упорядочиваются от более существенного к менее существенному. Далее переобозначим как $V_{j^{L_2}} = (r(d_1^{L_2}), \ldots, r(d_n^{L_2}))$ результат

 ISSN 1562-9945

замены каждого $d_j^{L_2}$ его рангом $r(d_j^{L_2})$. После этой замены получаем: $V_{t_1^{L_2}} = (1, \alpha, \alpha, 4, \alpha); V_{t_2^{L_2}} = (1, 2, \alpha, \alpha, \alpha);\ V_{t_3^{L_2}} = (\alpha, 3, \alpha, \alpha, \alpha).$;

– на заключительном этапе определяем отношения подобия между каждым вектором $V_{t_j^{L_2}}$ и идеальным вектором $(1, 2, \ldots, N)$, соответствующим начальному запросу. Если $V_{t_j^{L_2}}$ – наиболее близкий вектор, то $t_j^{L_2}$ будет определен как наилучший перевод рассматриваемого термина.

Для вычисления подобия между векторами переведенных документов и документов, полученных по начальному термину, используем две следующие меры подобия: лексикографический порядок (lexicographically ordering (LO)) и коэффициент корреляции Спирмена (rank correlation coefficient (RCC)).

Лексикографический порядок позволяет сравнивать порядки документов $D_{t_j^{L_2}}$, полученных различными переводами. При этом порядок $D_{t_1^{L_2}}$ предпочтительнее порядка $D_{t_2^{L_2}}$, если наиболее важный документ в $D_{t_j^{L_2}}$ имеет порядок меньше, чем наиболее важный документ в $D_{t_2^{L_2}}$. В случае их равенства сравниваются порядки вторых по важности документов и т.д.

Введем формальные определения лексикографического порядка терминов запроса.

Определение 1. Пусть $A = (a_1, \ldots, a_n)$ и $B = (b_1, \ldots, b_n)$ – два входных вектора. A лексикографически предпочтительнее, чем B, если $\exists k \le n$, такой что:

$$a_k < b_k \text{ и } \forall j = 1, \ldots, k-1, a_j = b_j.$$

Определение 2. Термин $t_1^{L_2}$ предпочтительнее термина $t_2^{L_2}$, если $V_{t_1^{L_2}}$ лексикографически предпочтительнее, чем $V_{t_2^{L_2}}$.

В рассматриваемом примере мы имеем: $Q^{L_1} = \{t_i^{L_1}\}$ и $T^{L_2}(t_i^{L_1}) = \{t_1^{L_2}, t_2^{L_2}, t_3^{L_2}\}$. Таким образом, $V_{t_i^{L_2}} = (1, 2, 3, 4, 5).$; $V_{t_1^{L_2}} = (1, \alpha, \alpha, 4, \alpha)$; $V_{t_2^{L_2}} = (1, 2, \alpha, \alpha, \alpha)$; $V_{t_3^{L_2}} = (\alpha, 3, \alpha, \alpha, \alpha)$; $V_{t_3^{L_2}} = (\alpha, 3, \alpha, \alpha, \alpha)$.

Очевидно, что $t_1^{L_2}$ лексикографически предпочтительнее, чем $t_3^{L_2}$, но $t_2^{L_2}$ лексикографически предпочтительнее, чем $t_1^{L_2}$ и $t_3^{L_2}$. Следовательно, конечная заявка будет состоять из $t_2^{L_2}$, $Q^{L_2} = \{t_2^{L_2}\}$.

Лексикографический порядок эффективен, если начальный порядок $D_{t_j^{L_1}} = \{d_1^{L_1}, \ldots, d_n^{L_1}\}$ такой, что $d_i^{L_1}$ намного существеннее, чем $d_{i+1}^{L_1}$.

Отметим, что перевод, который позволяет обнаружить $d_1^{L_1}$ в первой позиции, предпочтительнее других переводов, где $d_1^{L_1}$ является наиболее важным документом. Этот подход не очень эффективен, если $d_1^{L_1}$, например, ненамного важнее, чем $d_2^{L_1}$. Чтобы устранить этот недостаток, важно уточнить процедуру классификации документов, имеющих одинаковую степень важности. Для этого нужно предварительно выделить документы, имеющие одинаковый ранг. Эта задача может быть решена с применением коэффициента корреляции Спирмена, используемого обычно в регрессионных методах. В отличие от лексикографического порядка, порядок коэффициентов корреляции использует полный порядок документов и рассматривает не только первый существенный документ.

Определение 3. Пусть $V_{t_j^{L_2}} = \{r(d_1^{L_2}), \ldots, r(d_n^{L_2})\}$ – порядок документов, полученных с помощью $t_j^{L_2}$. Коэффициент корреляции Спирмена между $t_j^{L_2}$ и $t_i^{L_1}$ определится следующим образом:

$$\rho(t_j^{L_2}, t_i^{L_1}) = 1 - \frac{6 * \sum_{i=1}^{n} (r(d_j^{L_2}) - i)^2}{N^3 - N}.$$

Значение RCC повышается, если имеет место существенная зависимость между $V_{t_j^{L_2}}$ и $V_{t_i^{L_1}}$.

Определение 4. $t_1^{L_2}$ предпочтительнее, чем $t_2^{L_2}$ (по оценкам RCC), если $\rho(t_1^{L_2}, t_i^{L_1}) \geq \rho(t_2^{L_2}, t_i^{L_1})$. Конечным запросом при использовании RCC является $Q^{L_2} = \{t_i^{L_2} : t_j^{L_2} \in T^{L_2}(t_i^{L_1}), \exists t_j^{L_2} \in T^{L_2}(t_i^{L_1}), \rho(t_j^{L_2}, t_i^{L_1}) > \rho(t_i^{L_2}, t_i^{L_1})\}$. Отметим, что мера RCC применяется, как правило, для строгого порядка.

 ISSN 1562-9945

Определив в рассматриваемом примере коэффициенты RCC для каждого из переводов $t_1^{L_2} t_2^{L_2} t_3^{L_2}$ и каждого порядка документа, получим: $\rho(t_1^{L_2}, t_i^{L_1}) = -0.3$, $\rho(t_2^{L_2}, t_i^{L_1}) = 0.3$, $\rho(t_3^{L_2}, t_i^{L_1}) = -1.1$.

Так как $\rho(t_2^{L_2}, t_i^{L_1}) > \rho(t_1^{L_2}, t_i^{L_1}) > \rho(t_3^{L_2}, t_i^{L_1})$, то $t_2^{L_2}$ предпочтительнее, чем $t_1^{L_2}$ и $t_3^{L_2}$. Отсюда следует, что $Q^{L_2} = \{t_2^{L_2}\}$.

Выводы и перспективы дальнейших исследований

Проведенные исследования позволяют сделать вывод, что важным этапом построения систем автоматизированного поиска мультиязычной информации является применение процедуры уменьшения неоднозначности запросов, основанный на анализе мер подобия терминов и запросов. Рассмотренные в статье меры подобия (лексикографический порядок и коэффициент корреляции Спирмена) учитывают помимо степени соответствия документов их порядок в списке документов. В практических приложениях это может быть, в частности, использовано.в задачах формирования дайджестов и тезаурусов. При проведении дальнейших исследований целесообразно усовершенствовать предложенный метод, дополнив его анализом контекста запросов.

ЛИТЕРАТУРА

1. [Nassr 2002B] N. Nassr, S. Benferhat, M. Boughanem, C. Chrisment et H. Parde Disambiguation translation in multilingual queries IPMU 2002, The 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based System. 1-6 juillet Annecy pages : 200-207, 2002
2. [Nassr 2001C] M. Boughanem, C. Chrisment et N. Nassr Investigation on disambiguation in CLIR: aligned corpus and bi-directional translation based strategies In CLEF 2001 , lecture Notes in Computer Science 2069, Springer Verlag Darmstadt, 03-04 Septembre Germany. pages 87-92, 2001.