N E W S
OF THE NATIONAL ACADEMY OF SCIENCES OF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 5, Number 297 (2014), 25 – 32

MODELS OF DETERMINATION OF RELEVANCE OF THE TEXT TO THE GIVEN THEME, GRAPHS ASSOCIATED WITH TEXTS AND A PROBLEM OF SUMMARIZING

T. V. Batura[1], F. A. Murzin[1], D. O. Speransky[2], B. S. Baizhanov[3], M. V. Nemchenko[3]

[1]A. P. Ershov Institute of Informatics Systems, Russian Academy of Sciences, Siberian Branch,
[2]Novosibirsk National Research State University,
[3]Institute of Mathematics, Informatics and Mechanics,
Committee of Science of the Ministry of Education and Science of Republic Kazakhstan.
E-mails: tatiana.v.batura@gmail.com, murzin@iis.nsk.su, speranskydanil@gmail.com,
baizhanov@hotmail.com, nemchenko.imim@mail.ru

Key words: processing of texts in a natural language, syntactic analysis, Link Grammar Parser, summarization, relevance.

Abstract. Article is devoted to algorithms of the summarization. The problem consists of that to allocate from the text fragments corresponding to the given theme. The theme is understood as a set of the sentences concerning to one concept, the phenomenon, sequence of events, etc. Fragments not necessarily follow one after another. In the text between them, other themes can be included. Further the allocated fragments can be united in the resume on the given theme. In the work, some generalization of the method described in the work of Niraj Kumar, etc. [1] is offered. The considered approach allows us to take into account syntactic relations, which can be received on output of program system Grammar Parser.

УДК 519.68; 681.513.7;
316.472.45; 007.51/.52

МОДЕЛИ ОПРЕДЕЛЕНИЯ РЕЛЕВАНТНОСТИ ТЕКСТА ЗАДАННОЙ ТЕМЕ, ГРАФЫ АССОЦИИРОВАННЫЕ С ТЕКСТАМИ И ЗАДАЧА РЕФЕРИРОВАНИЯ

Т. В. Батура[1], Ф. А. Мурзин[1], Д. О. Сперанский[2], Б. С. Байжанов[3], М. В. Немченко[3]

[1]Институт систем информатики им. А. П. Ершова СО РАН,
[2]Новосибирский национальный исследовательский государственный университет,
[3]Институт математики, информатики и механики КН МОН РК

Ключевые слова: обработка текстов на естественном языке, синтаксический анализ, Link Grammar Parser, резюмирование, релевантность.

Аннотация. Статья посвящена алгоритмам реферирования. Задача состоит в том, чтобы из текста выделить фрагменты, соответствующие заданной теме. Под темой понимается набор предложений, относящихся к одному понятию, явлению, последовательности событий и др. Фрагменты не обязательно следуют друг за другом. В тексте между ними могут присутствовать вставки на другие темы. Далее выделенные фрагменты могут быть объединены в резюме по данной теме. В работе предложено некоторое обобщение метода, описанного в работе Нираджа Кумара и др. [1]. Рассматриваемый подход позволяет учесть синтаксические отношения, которые могут быть получены на выходе программной системы Link Grammar Parser.

1. Введение. Основой для написания данной статьи послужила работа [1], в которой авторы рассматривают процесс автоматического резюмирования. В научной литературе как синоним используется также слово «реферирование». Суть работы состоит в том, что из текста могут выделяться фрагменты, соответствующие заданной теме. Под темой понимается набор предложений, относящихся к одному понятию, явлению, последовательности событий и др. Фрагменты не обязательно следуют друг за другом. В тексте между ними могут присутствовать вставки на другие темы. Далее выделенные фрагменты могут быть объединены в резюме по данной теме. В общем случае из текста могут быть выделены несколько таких резюме, соответствующих различным темам.

Одна из возникающих проблем состоит в том, что перестановка слов в предложении может существенно менять его смысл, что приводит к некорректной работе алгоритмов, оперирующих с отдельными ключевыми словами, их частотами и т. д. Сравните, например, такие выражения, как «кровь с молоком» и «молоко с кровью». В упомянутой выше работе предложен метод, позволяющий учесть порядок слов и показавший свою эффективность.

В данной работе авторы предлагают некоторое обобщение данного метода, которое позволяет учесть дополнительно также синтаксические отношения, которые могут быть получены на выходе программной системы Link Grammar Parser. Речь идет об анализе текстов на английском языке.

2. Базовый алгоритм резюмирования. Базовый алгоритм резюмирования описан в вышеупомянутой работе [1] и состоит из нескольких этапов. Ниже дано краткое описание алгоритма.

1. Проводится предварительная обработка статьи, могут удаляться отдельные элементы, специальные обозначения, неподдерживаемые символы.

2. Вычисляются веса слов.

3. Выполняется разбиение статьи на отдельные фрагменты (топики). Вычисляются веса топиков. Под топиком будем понимать набор предложений, идущих не обязательно последовательно и охватывающих некоторую самостоятельную подтему в исходной статье. Отметим, что для статьи достаточно большого объема таких подтем может быть довольно много.

4. Вычисляется оценка релевантности (соответствия) топиков эталонному сообщению по заданной теме. Множество релевантных фрагментов объединяются в одно резюме.

На рисунке 1 схематически показано, что исходная статья разбивается на топики, каждый из которых сравнивается с эталонным сообщением Summary.

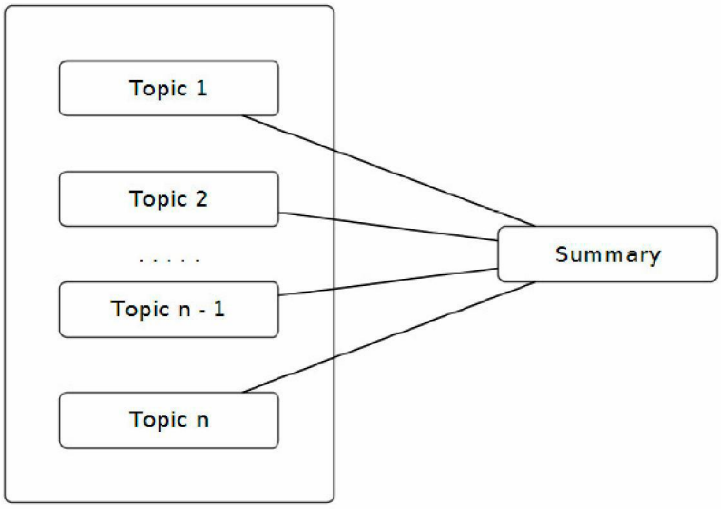
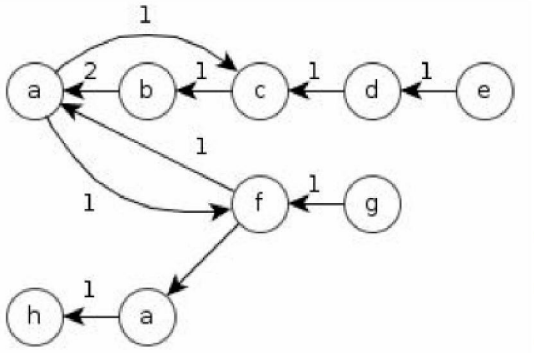


Рисунок 1 – Общая схема алгоритма

Считаем, что произвольный текст (статьи в целом или топика) представляет собой последовательность предложений, каждое из которых состоит из слов. Тогда текст можно представить в виде графа, вершины которых слова, а направленные ребра показывают очередность следования слов. Ребра направлены от последующих слов к предыдущим. Пример построения графа приведен на рисунке 2. Веса ребер определяются естественным образом. Это количества вхождений соответствующих пар слов в весь текст. Заметим, что метод может применяться как к исходному тексту, так и предварительно «очищенному» от предлогов, союзов и других слов.



$$S_1 =< a, b, c, d, e >$$
$$S_2 =< c, a, f, g >$$
$$S_3 =< h, i, f, a, b >$$

Рисунок 2 – Граф слов, ассоциированный с набором предложений S_1, S_2, S_3

Далее для вычисления весов слов рассматриваются именно такие графы. Веса слов можно вычислять различными способами. В частности, в работе [1] предлагается вычислять их по формуле:

$$W(v_i) = \frac{1-\lambda}{N} + \lambda \sum_{v_j \in In(v_i)} \frac{W(v_j)}{|Out(v_j)|},$$

где N – количество вершин в графе; $In(v_k)$ – множество вершин, соединенных входящими в v_k связями; $Out(v_k)$ – множество вершин, соединенных исходящими из v_k связями; λ – коэффициент затухания.

Опустив подробности, сразу же отметим, что данная формула корректна, т.е. веса всегда однозначно вычисляются. Также отметим, что приведенное определение напоминает понятие ранга формулы в математической логике.

Чтобы разбить исходную статью на топики в [1] используется метод кластеризации, например, Group-Average Agglomerative Clustering. Простейший метод – это топиками считать абзацы статьи. Некоторые топики могут быть выделены вручную.

Переходим непосредственно к процессу резюмирования. Итак, на данной стадии имеется некоторый заранее заготовленный текстовый фрагмент Set_1, точно относящийся к определенной теме. Фрагмент Set_2 – анализируемый фрагмент. В действительности, анализируемых фрагментов много. Он сопоставляется с фрагментом Set_1 на предмет релевантности, т.е. соответствия той же теме. Напомним, что в нашем случае он представляет собой фрагмент, выделенный из исходного (большого) текста посредством кластеризации или другим способом, как было указано выше. Процесс сбора всех фрагментов, удовлетворяющих определенному критерию соответствия, и является, по сути дела, процессом резюмирования.

Перейдем теперь к более детальному описанию алгоритма. Фрагменту Set_1 соответствует ориентированный граф $G = (V, E)$, описанный выше. Пусть далее w_{ij} – вес ребра $(V_i, V_j) \in E$, если таковое имеется, и $w_{ij} = 0$ в противном случае. Рассмотрим неориентированный граф, ассоциированный с графом G естественным образом. Множество вершин то же самое. Ребра возникают посредством «забывания» об ориентации. Вес ребра в новом графе будет равен $Link\_Strength_{ij} = 2 \times \min(w_{ij}, w_{ji})$. Большое значение данной величины означает, что данные два слова многократно встречаются рядом в двух вариантах. А именно, первое слово предшествует второму или наоборот второе слово предшествует первому.

Далее вводим величину $Path\_Length_{ij} = 1/Link\_Length_{ij}$. То есть, осуществлен переход к обратным величинам. Таким образом, получаем, что чем меньше $Path\_Length_{ij}$, тем более выражено упомянутое выше свойство, иначе говоря, данные слова «очень связаны» между собой внутри данной темы.

Центральность по близости (Closeness centrality) – это понятие обычно используется при изучении социальных сетей и является показателем, насколько быстро распространяется информация в

сети от одного участника к остальным. В качестве меры расстояния между двумя участниками используется кратчайший путь в графе (геодезическое расстояние). Так, непосредственные друзья участника находятся на расстоянии 1, друзья друзей – на расстоянии 2, друзья друзей друзей – на расстоянии 3 и т.д. Далее берется сумма всех расстояний и нормируется. Полученная величина называется удаленностью вершины V_i от других вершин. Близость определяется как величина, обратная удаленности.

$$C_C(V_i) = \frac{N-1}{\sum_{t \in V \setminus v} d_G(V_i, t)},$$

где $d_G(V_i, t)$ – кратчайший путь от вершины V_i до вершины t.

Другими словами, центральность по близости позволяет понять, насколько близок рассматриваемый участник ко всем остальным участникам сети. Таким образом, важно не только наличие непосредственных друзей, но и чтобы у самих этих друзей тоже были друзья.

В нашем случае мы имеем дело с текстами, а не с социальными сетями. Но в обоих случаях используются теоретико-графовые конструкции. Поэтому, аналогично, можно вычислить центральность по близости для любой вершины в рассматриваемом неориентированном графе. При вычислении геодезического расстояния используются веса ребер $Path\_Length_{ij}$.

Переходим теперь к процессу сравнения двух фрагментов Set_1 и Set_2. Продемонстрируем основную идею на примере [1]. Предположим, что каждый из них состоит из трех предложений:

Set_1	Set_2
$A\,B\,C\,D$	$M\,N\,B\,D$
$B\,N\,D$	$C\,O\,D$
$W\,X\,B\,C$	$A\,B\,C$

Оставим только те слова, которые одновременно входят в оба текстовых фрагмента, получаем соответственно:

Set_1'	Set_2'
$A\,B\,C\,D$	$N\,B\,D$
$B\,N\,D$	$C\,D$
$B\,C$	$A\,B\,C$

Каждому из текстовых фрагментам Set_1' и Set_2' соответствует свой неориентированный граф. Заметим, что множество вершин у этих графов будет одно и то же, а именно $\{A, B, C, D, N\}$. Соответственно для всех вершин V_i можно вычислить центральность по близости в первом графе и во втором. Обозначим их $C_C(V_i)_1$ и $C_C(V_i)_2$ соответственно.

Далее подсчитываем величину $Diff(V_i) = (|C_C(V_i)_2 - C_C(V_i)_1| / C_C(V_i)_1) \times 100$. Она характеризует, насколько одинаково или различно употребление данного слова в контекстах первого и второго текстовых фрагментов. Множитель равный 100 позволяет, грубо говоря, выразить эту величину в процентах. Далее устанавливаем порог, который служит критерием «одинаковости» употребления слова. Например, в качестве порога можно взять медиану величин $Diff(V_i)$, и далее этот порог можно использовать при анализе других фрагментов.

На финальной стадии подсчитывается, сколько слов «преодолели» порог по формуле:

$$Score\ (Set_1, Set_2) = (Count_{match}(Set_1, Set_2) / Count\ (Set_1)) \times 100;$$

$Count_{match}(Set_1, Set_2)$ – количество всех слов из Set_1, которые одновременно входят в Set_1 и Set_2, и которые «преодолели» порог, т.е. прошли проверку на идентичность употребления в обоих текстовых фрагментах;

$Count(Set_1)$ – количество всех слов из Set_1, которые подвергались анализу, т.е. просто одновременно входят в Set_1 и Set_2.

Окончательный вывод делается по тому, насколько велико $Score(Set_1, Set_2)$, т.е. в конечном итоге и здесь эксперт выбирает соответствующий порог, позволяющий сделать вывод, что тема фрагмента Set_2 соответствует теме эталонного текстового фрагмента Set_1.

3. Link Grammar Parser и модель определения релевантности. Обобщим теперь данный алгоритм и попытаемся учитывать синтаксическую структуру предложений. Для этого на третьем этапе перед тем, как вычислять веса топиков, воспользуемся результатом работы системы Link Grammar Parser. Это синтаксический анализатор английского языка, базирующийся на оригинальной теории синтаксиса английского языка. Программная система приписывает заданному предложению синтаксическую структуру, состоящую из множества помеченных связей, соединяющих пары слов. Пример синтаксического разбора предложения анализатором приведен на рисунке 3. Подробное описание этого Link Grammar Parser можно найти в [2,3].

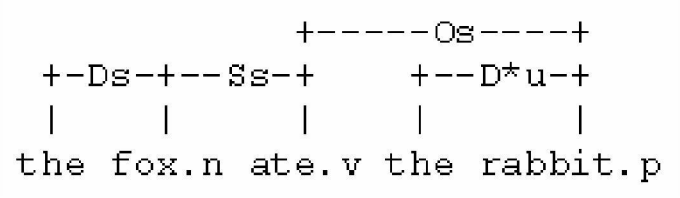


Рисунок 3 – Пример синтаксического разбора предложения

Переходим теперь к собственно описанию обобщенного алгоритма. Как и раньше считаем, что в нашем распоряжении имеется последовательность предложений $S =< S_1, S_2, ..., S_n >$. Далее к каждому из предложений применяем синтаксический анализатор Link Grammar Parser. В результате его работы получаем направленные связи между парами слов предложения. Аналогично, последовательности предложений соответствует ориентированный граф с помеченными ребрами $G = (V, E)$. Но в данном случае меткой ребра является пара $< T, n >$, где T – тип связи (коннектора), n – количество раз, которое данная связь встретилась в последовательности предложений $S =< S_1, S_2, ..., S_n >$.

Отметим, что пара слов и коннектор, их соединяющий, в принципе, могут встретиться многократно. Две вершины могут соединяться однонаправленными параллельными ребрами, имеющими различные пометки, и, конечно, две вершины могут соединяться разнонаправленными ребрами. То есть «ребром с пометкой» назовем четверку $< V_1, V_2, T, n >$. Предполагаем, что каждой связи T_i сопоставлено положительное число α_i, называемое ее весом или важностью.

Далее полагаем:

1) $\mu(T_i, 1) = \alpha_i$,

2) $\mu(T_i, n) = n \cdot \mu(T_i, 1) = n \cdot \alpha_i$,

3) $\mu(T_{i_1}, n_1, \ldots, T_{i_k}, n_k) = \sum_j \mu(T_{i_j}, n_j) = \sum_j n_j \cdot \alpha_j$.

Таким образом, если из вершины V_i в вершину V_j идет множество параллельных ребер с пометками $< T_{i_1}, n_1 >, \ldots, < T_{i_k}, n_k >$, то все они могут быть заменены одним ребром, имеющим вес равный $w_{ij} = \mu(T_{i_1}, n_1, \ldots, T_{i_k}, n_k)$. Далее, как и раньше, можно определить

$Link\_Length_{ij} = 2 \times (w_{ij}, w_{ji})$ и соответственно $Path\_Length_{ij} = 1/(Link\_Length_{ij})$. Отметим, что если для любого i выполнено $\alpha_i \geq 1$, то $Link\_Length_{ij} \geq 2$ и $Path\_Length_{ij} \leq 0.5$.

Соответствующим образом может быть изменена формула вычисления ранга слова. При этом можно использовать два варианта формулы.

Вариант первый:

$$S(V_i) = \frac{1-\lambda}{N} + \lambda \sum_{V_j \in IN(V_i)} \frac{w_{ji} \cdot S(V_j)}{|OUT(V_j)|}.$$

Вариант второй:

$$S(V_i) = \frac{1-\lambda}{N} + \lambda \sum_{V_j \in IN(V_i)} \frac{w_{ji} \cdot S(V_j)}{\left(\sum_{V_k \in OUT(V_j)} w_{jk} \right)}.$$

Приведенные выше формулы могут быть модифицированы на случай размытой логики Заде [4, 5]. А именно, считаем, что $0 \leq \alpha_i \leq 1$. Далее можно, например, положить $\mu(T_i, n) = (1 - 1/2^n)\mu(T_i, 1) = (1 - 1/2^n)\alpha_i$. Формула возникает исходя из следующей идеи. Предполагаем, что если связь входит однократно, то $\mu(T_i, 1) = \frac{1}{2}\alpha_i$, если двукратно, то $\mu(T_i, 2) = \left(\frac{1}{2} + \frac{1}{4}\right)\alpha_i$ и т.д. Соответственно получаем $\mu(T_i, n) = \left(\frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n}\right)\alpha_i = \left(1 - \frac{1}{2^n}\right)\alpha_i$.

Если задействовано несколько связей, то наиболее естественный вариант – это взять их дизъюнкцию $\mu(T_{i_1}, n_1, \ldots, T_{i_k}, n_k) = \mu(T_{i_1}, n_1) \vee \ldots \vee \mu(T_{i_k}, n_k)$. Дизъюнкция в логике Заде означает максимум истинностных значений. Можно также использовать их среднее значение, но этот вариант менее естественный. В случае среднего значения увеличение количества связей может приводить к уменьшению веса ребра в итоговом графе, если добавлять связи с маленькими весами. С другой стороны, вариант, когда используется дизъюнкция, не «чувствует», мало или много связей. Поэтому целесообразно произвести дополнительное уточнение метода.

Предположим, что $0 \leq \sum_i \alpha \leq 1$, т.е. фактически имеет место неравенство, аналогичное неравенству Крафта, известному в теории информации. Тогда имеем $\mu(T_i, n) = \left(1 - \frac{1}{2^n}\right)\alpha_i \leq \alpha_i$, и в итоге вместо дизъюнкции можно использовать сумму. Соответственно получаем: $w_{ij} = \mu(T_{i_1}, n_1, \ldots, T_{i_k}, n_k) = \sum_j \mu(T_{i_j}, n_j) \leq \sum_j \alpha_j \leq 1$. Очевидно также, что $w_{ij} \geq 0$.

При определении ранга слова можно использовать первый вариант формулы, заменив сумму дизъюнкцией. Тогда значение ранга также будет лежать на отрезке [0,1]. В итоге получаем $S(V_i) = \frac{1-\lambda}{N} + \lambda \cdot \bigvee_{V_j \in IN(V_i)} \frac{w_{ji} \cdot S(V_j)}{|OUT(V_j)|}$. Отметим, что возможны другие варианты данной формулы. Например, дизъюнкция может быть заменена операцией $x \oplus y = x + y - x \cdot y$. Величина $C_C(V_i)$, называемая центральность по близости, может быть вычислена стандартным способом, и далее, для того, чтобы попасть в интервал [0,1], она может быть нормирована. В итоге имеем $\overline{C}_C(V_i) = \frac{C_C(V_i)}{\max_j \{C_C(V_j)\}}$.

4. Заключение. В статье речь идет о резюмировании, но в действительности, задача определения релевантности заданной теме важна сама по себе. Например, эта задача является основной при разработке новых методов интеллектуального поиска информации в сети. Разрабатываемые

методы, в идеале, должны позволять сопоставлять конструкции естественного языка и в ряде случаев отождествлять даже перефразированные варианты предложений, основываясь также на анализе их синтаксических структур. Таким образом, мы можем сопоставить поисковый запрос и текст, взятый из сети интернет или других источников, с целью определения релевантности (соответствия) текста поисковому запросу.

Аналогично, для определения источника распространения информации в социальных сетях можно использовать алгоритм оценки релевантности сообщений, оставляемых в социальных сетях, статьям, публикуемым в интернете. В частности, для того чтобы учесть синтаксическую структуру текста можно использовать результат работы синтаксического анализатора Link Grammar Parser.

Описанная модель представляется перспективной. Тем не менее, настоящая работа еще далека от завершения. Весьма вероятно, предложенный подход потребует доработки. Можно рассматривать множество вариаций обсуждаемого алгоритма: использовать различные методы кластеризации, меры схожести подтем, варьировать формулы для вычисления весов слов и др. Для того чтобы найти наилучшую конфигурацию, при которой алгоритм обеспечит качественные результаты, следует протестировать различные комбинации вариаций этого алгоритма, а также привлечь экспертов для оценки качества работы алгоритма. Это довольно трудоемкий процесс.

В заключение отметим, что ранее были предложены [6, 7] различные способы представления семантико-синтаксических отношений между смысловыми единицами предложения, методы построения этого представления на основе диаграмм Link Grammar Parser, а также способы вычисления степени совпадения естественно-языковых конструкций. Идеи были реализованы в информационно-поисковой системе iNetSearch. В итоге было получено, что методы, учитывающие перефразирования (перефразирование можно считать отдаленным аналогом перестановки слов) позволили улучшить работу системы, но, как показало тестирование, незначительно по сравнению с базовым алгоритмом. С другой стороны, была показана высокая эффективность учета синтаксических отношений, генерируемых системой Link Grammar Parser.

Работа выполнена при поддержке гранта 2581/ГФ3 МОН РК.

ЛИТЕРАТУРА

[1] Kumar N., Srinathan K., Varma V. Using graph based mapping of co-occurring words and closeness centrality score for summarization evaluation // CICLing Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing, 2012. – P. 353-365.

[2] Temperley D., Sleator D., Lafferty J. Link Grammar Documentation [Electronic resource]. – 1998. – Mode of access: http://www.link.cs.cmu.edu/link/dict/index.html (accessed 15 November 2012)

[3] Sleator D., Temperley D. Parsing English with a Link Grammar. – Pittsburgh: School of Computer Science Carnegie Mellon University, 1991. – 93 p.

[4] Заде Л.А. Понятие лингвистической переменной и его применение к принятию приближенных решений. – М.: Мир, 1976. – 165 с.

[5] Пивкин В.Я., Бакулин Е.П., Кореньков Д.И. Нечеткие множества в системах управления: Учебное пособие. – Новосибирск: НГУ, 1997. – 52 с.

[6] Murzin F., Perfliev A., Shmanina T. Methods of syntactic analysis and comparison of constructions of a natural language oriented to use in search systems // Bull. Nov. Comp. Center, Comp. Science. – 2010. – Iss. 31. – P. 91-109.

[7] Перфильев А.А., Мурзин Ф.А., Шманина Т.В. Методы синтаксического анализа и сопоставления конструкций естественного языка, ориентированные на применение в информационно-поисковых системах // Вестник НГУ. – 2011. – Т. 9, вып. 4. – С. 50-59.

REFERENCES

[1] Kumar N., Srinathan K., Varma V. Using graph based mapping of co-occurring words and closeness centrality score for summarization evaluation. CICLing Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing, 2012. P. 353-365.

[2] Temperley D., Sleator D., Lafferty J. Link Grammar Documentation [Electronic resource]. 1998. Mode of access: http://www.link.cs.cmu.edu/link/dict/index.html (accessed 15 November 2012)

[3] Sleator D., Temperley D. Parsing English with a Link Grammar. Pittsburgh: School of Computer Science Carnegie Mellon University, 1991. 93 p.

[4] Zadeh L.A. The concept of linguistic variable and its application to approximate reasoning. Moscow.: Mir, 1976. 165 p. (in Russian, translated from American. Elsevier Publishing Company, Inc., 1975)

[5] Pivkin V.Ya., Bakulin E.P., Korenkov D.I. Fuzzy sets in control systems: Manual. Novosibirsk State Univ., 1997. 52 p. (in Russian).

[6] Murzin F., Perfliev A., Shmanina T. Methods of syntactic analysis and comparison of constructions of a natural language oriented to use in search systems. Bull. Nov. Comp. Center, Comp. Science. 2010. Iss. 31. P. 91-109.
[7] Murzin F., Perfliev A., Shmanina T. Methods of syntactic analysis and comparison of constructions of a natural language oriented to use in search systems. Vestnik of Novosibirsk State Univ. Ser.: Information Technologies. Novosibirsk, 2012. Vol. 9, is. 4. P. 13-28. (in Russian).

БЕРІЛГЕН ТАҚЫРЫПҚА БАЙЛАНЫСТЫ МӘТІННІҢ РЕЛЕВАНТЫН АНЫҚТАУ ҮЛГІЛЕРІ, МӘТІНМЕН ЖӘНЕ РЕФЕРАТ ОҚУ МІНДЕТІМЕН АССОЦИАЦИЯЛАНГАН ГРАФТАР

Т. В. Батура, Ф. А. Мурзин, Д. О. Сперанский, Б. С. Байжанов, М. В. Немченко

Тірек сөздер: табиғи тілде мәтінді өңдеу, синтаксистік анализ, түйіндеме, Link Grammar Parser, релевanттық.

Аннотация: Мақала реферат оқу алгоритміне арналған. Мақсаты: мәтіннен тақырыпқа байланысты керекті фрагменттерді бөліп алу. Тақырып астында бір түсінікке негізделген, құбылысқа, жүйелі оқиғаға және т.б. байланысты сөйлемдер жиынтығы пайымдалады. Фрагменттер бірінен кейін бірі тұрып реттелуі міндетті емес. Мәтінде олардың арасында басқа тақырыпқа байланысты кірістірмелер болуы мүмкін. Ары қарай белгіленген фрагменттер берілген тақырыпқа байланысты түйіндемеге бірігуі мүмкін. Мақалада Нирадж Кумардың жұмысын сипаттайтын кейбір жалпыламалар ұсынылды. Қарастырылынып отырған әдіс Link Grammar Parser программалық жүйе істен шыққан кезде синтаксистік қателерді ескеруге мүмкіндік береді.

Поступила 01.10.2014 г.