Библиографические ссылки

1. Havas G., Wall G., Wamsley J. The two generated Burnside group of exponent five // Bull. Austral. Math. Soc. 1974. № 10. P. 459–470.
2. Кузнецов А. А., Кузнецова А. С. Компьютерное моделирование конечных двупорожденных групп периода 5 // Вестник СибГАУ. 2012. № 1 (45). С. 59–62.
3. Филиппов К. А. О диаметре Кэли одной подгруппы группы $B_0(2,5)$ // Вестник СибГАУ. 2012. № 1 (41). С. 234–236.
4. Kuznetsov A. A., Shlepkin A. K. Comparative analysis of the Burnside groups $B(2,5)$ and $B_0(2,5)$ // Proceedings of the Steklov Institute of Mathematics. 2010. № 2 (15). P. 111–117.
5. Holt D., Eick B., O'Brien E. Handbook of computational group theory. Boca Raton: Chapman & Hall/CRC Press. 2005. 514 p.
6. Sims C. Computation with Finitely Presented Groups. Cambridge: Cambridge University Press, 1994. 628 p.
7. Кузнецов А. А., Кузнецова А. С. Быстрое умножение элементов в конечных двупорожденных группах периода пять // Прикладная дискретная математика. 2013. № 1 (18). С. 110–116.

References

1. Havas G., Wall G., Wamsley J. The two generated Burnside group of exponent five // Bull. Austral. Math. Soc., 1974, no. 10, p. 459–470.
2. Kuznetsov A. A., Kuznetsova A. S. *Vestnik SibGAU*, 2012, no. 1 (45), p. 59–62.
3. Philippov K. A. *Vestnik SibGAU*, 2012, no. 1 (41), p. 234–236.
4. Kuznetsov A. A., Shlepkin A. K. Comparative analysis of the Burnside groups $B(2,5)$ and $B_0(2,5)$. Proceedings of the Steklov Institute of Mathematics, 2010. no. 2 (15), p. 111–117.
5. Holt D., Eick B., O'Brien E. Handbook of computational group theory. Boca Raton: Chapman & Hall/CRC Press, 2005, 514 p.
6. Sims C. Computation with Finitely Presented Groups. Cambridge: Cambridge University Press, 1994. 628 p.
7. Kuznetsov A. A., Kuznetsova A. S. *Prikl. Diskr. Mat.*, 2013, no. 1 (18), p. 110–116.

УДК 004.9

ОСУЩЕСТВЛЕНИЕ ГЕНЕРАЦИИ СИНОНИМИЧНЫХ ПОИСКОВЫХ ЗАПРОСОВ НА ОСНОВЕ СЕМАНТИЧЕСКОЙ КЛАССИФИКАЦИИ*

Д. В. Личаргин[1], К. В. Сафонов[2], О. И. Егорушкин[2], И. В. Колбасина[2], Е. Д. Старовойт[2]

[1]Сибирский федеральный университет
Россия, 660074, Красноярск, ул. Академика Киренского, 26
[2]Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Россия, 660014, Красноярск, просп. им. газ. «Красноярский рабочий», 31
E-mail: lichdv@hotmail.ru

В работе рассматривается проблема усовершенствования семантических естественно-языковых запросов к поисковым системам. Предлагается модель представления предложений естественного языка в качестве функций пространства с заданными семантизированными координатами. Предложен способ представления процесса порождения синонимичных предложений естественного языка на основе «таблиц синонимизации», модель может позволить порождать классы синонимических предложений на основе базовых предложений из многомерного пространства предложений естественного языка. Затрагивается вопрос об организации интерфейса расширенных синонимизированных семантических запросов к поисковым системам на основе иерархии общих и частных предложений с выделением темы и ремы. Делается вывод о необходимости продолжения данного исследования на основе апробации систем генерации осмысленных предложений.

Ключевые слова. Поисковые системы, компьютерная лингвистика, семантическая классификация, синонимичные запросы к поисковым системам».

*Исследование выполнено при поддержке Министерства образования и науки Российской Федерации, соглашение 14.B37.21.1010.

IMPLEMENTATION OF THE GENERATION OF SYNONYMIC SEARCH QUERIES BASED ON SEMANTIC CLASSIFICATION

D. V. Lichargin[1], K. V. Safonov[2], O. I. Egorushkin[2], I. V. Kolbasina[2], E. D. Starovoyt[2]

[1]Siberian Federal University
26 Kirenskiy str., Krasnoyarsk, 660074, Russia
[1]Siberian State Airspace University named after academician M. F. Reshetnev
31 "Krasnoyarskiy Rabochiy" prosp., Krasnoyarsk, 660014, Russia. E-mail: lichdv@hotmail.ru

The article considers the problem of up-grade of the semantic natural language queries to the search engines. The authors offer a model for presentation of sentences of the natural language as functions in a notional space with given semantically assigned coordinates, along with a method of presenting the process of generating synonymous sentences of the natural language based on the "table synonymization" taken from the multidimensional space of a sentence of the natural language. The problem of interface design within expanded synonymized queries to the search engines based on the hierarchy of general and partial sentences within theme- and rheme- parsing is discussed. The conclusion about the necessity to continue the investigation on the basis of testing the systems of meaningful sentences generation is made.

Keywords: search engines, computational linguistics, semantic classification, synonymic requests for search engines.

В работе рассматривается проблема усовершенствования семантических естественно-языковых запросов к поисковым системам с точки зрения перехода от строки запроса к выбору иерархических структур и фрагментов семантической классификации.

На сегодняшний день широко распространены и разрабатываются разнообразные поисковые и метапоисковые системы. Проблема усовершенствования поисковых систем является актуальной в связи с необходимостью получать более системную и упорядоченную информацию в качестве ответов на поисковые запросы продвинутым пользователем.

Проблема генерации осмысленных и, в частности, синонимичных фраз языка решается на стыке таких наук, как лингвистика, компьютерная лингвистика, искусственный интеллект, логика, математика, психология и философия.

Проблема создания программ синонимизаторов и другие проблемы [1–5] давно и широко решаются различными авторами, в частности на основе подстановок слов из списка синонимов.

Однако вопрос о переборе всех трансформаций замкнутых на множестве синонимичных, и расширяемых до множества частично синонимичных фраз, и нахождения множества фраз более абстрактного вида требует дополнительных исследований в рамках теории лексико-грамматических трансформаций.

Цель данной работы состоит в формулировке принципа решения задачи генерации синонимических фраз для генерации синонимических поисковых запросов.

Задачи данной работы заключаются в:

1. Привлечении к решению указанной проблемы семантического словаря, описанного в работе [2].

2. Построении модели множеств линейных соответствий с полным или частичным сохранением не только осмысленности фразы в целом, но и специфики ее базового содержания.

3. Построения табличного интерфейса в целях генерации синонимических фраз языка.

Основная идея работы состоит в построении модели последовательного умножения дерева разбора (парсинга) строки с поисковым запросом на лексический ряд с прописанными синонимическими заменами.

Новизна работы состоит в использовании модели умножения на «синонимизирующие» ряды слов, основанной на словаре, описанном в работе [2] в применении к созданию генератора поисковых запросов на английском и других языках, например, русском [3] в рамках общей модели естественного языка.

В работе рассматривается проблема представления информации, а именно текста на естественном языке для запросов поисковых систем. Объектом исследования является мультииерархическая система естественного языка (ЕЯ).

Классификация слов и понятий естественного языка. Рассмотрим, многомерное грамматическое пространство единиц естественного языка: слов и предложений.

Такое пространство слов позволяет генерировать грамматически, но не семантически осмысленные фразы естественного языка. Так, фраза «очень он» является грамматически бессмысленной, фраза «карта выключает чашку» – грамматически осмыслена, но семантически бессмысленна, а фраза «девочка выключает компьютер» – грамматически и семантически осмыслена.

Возможно построение многомерного представления данных со следующими координатами вектора понятийного описания.

Таблица 1

Примерный вектор классификации слов и предложений естественного языка

ТЕМА (1)				**ПОЗИЦИЯ (2)**	**ВАРИАНТЫ (3)**	
ОПЕРАЦИЯ	НАД ОБЪЕКТОМ	ЛОКАЛИЗАЦИЯ	СВОЙСТВА		ЛЕКСИЧЕСКАЯ ГРУППА	ЛЕКСИЧЕСКИЕ РЯДЫ
essence **сущность**	consciousness **сознание**	of alive **часть живого**	maximally **максимально**	nominal group … **именная группа …**	***ТОЖЕ, ЧТО В ТЕМЕ (1)***	causation … **причина-следствие …**
essence of something **сущность чего-то**	being **существо**	in alive **в живом**	very **очень**	verbal group… **глагольная группа …**		time … **время …**
property **свойство**	relation **отношение**	on alive **на живом**	rather **достаточно**	bond .. **связь ..**		existing **существующее**
link **связь**	thing **вещь**	at alive **около живого**	a little **несколько**	aspect .. **аспект ..**		non-existing **несуществующее**
action **действие**	information **информация**	of not alive **асть не живого**	little **мало**	property .. **качество ..**		possible **возможный**
connecting **соединение**	idea **идея**	in not alive **в не живом**	minimally **минимально**	system .. **система ..**		necessary **необходимый**
presentation **представление**	place **место**	on not alive **на не живом**	…			start **начало**
changing **обмен**	relation **отношение**	at not alive **около не живого**	positive **позитивный**			stop **остановка**
		of which alive **часть которого живое**	complex **сложный**			continuation **продолжение**
		in which alive **в котором живое**	stable **стабильный**			***И другие***
			И другие			

V_1 = Части речи {«Артикль», «Прилагательное», «Существительное», «Глагол», …};

V_2 = Члены предложения {«Определитель», «Определение», «Подлежащее», «Сказуемое», …};

$V_{3,3,1}$ = Лица {«1-ое», «2-ое», «3-ее», «Не определено»};

$V_{3,3,2}$ = Аспект {«Неопределенный», «Продолженный», «Совершенный», «Совершенный продолженный», «Не определен»};

$V_{3,1,1}$, $V_{3,1,2}$, … – Другие размерности, выраженные грамматическими категориями.

L_1 = Объекты по тематике изучения {идеи {науки, представления, чувства …}, предметы {одежда, еда, части тела, здания, транспорт, …}, существа, …};

L_2 = Порядок слов и члены предложения {субъект, предикат, объект};

L_3 = Варианты подстановок слов в предложение {позитивное{обожать, любить}, негативное {не любить, ненавидеть}} (рис. 1).

Такое многомерное пространство включает комбинаторно сочетающиеся группы слов, например, группа слов {читать, писать, копировать, публиковать, иллюстрировать, …} относится к ячейке многомерного пространства V[СУЩЕСТВИТЕЛЬНОЕ / ИМЕННАЯ ЧАСТЬ / ЧИСЛО-ЕДИНСТВЕННОЕ]-L[ДЕЙСТВИЕ / ПРЕДМЕТ-ИНФОРМАЦИЯ / _].

Грамматические конструкции включаются в ячейки многомерного массива. Пересечение таких координат вектора, как, например V[СУЩЕСТВИТЕЛЬНОЕ / ОПРЕДЕЛЕНИЕ / СОВЕРШЕННЫЙ / …], определяет ячейку многомерного массива с грамматической конструкцией: «have + been + VERB-(e)d». Вектор V[Прилагательное / Сказуемое / 1-ое лицо / Сравнительная степень / Длинное слово / …] = «am more + ПРИЛАГАТЕЛЬНОЕ». Реляционные таблицы как подмножества этого многомерного массива представлены в лингвистике традиционными грамматическими парадигмами.

Семантические же конструкции L[СУЩНОСТЬ-ПРЕДМЕТ-ОДЕЖДА / ДЕЙСТВИЕ / ПОЯВЛЕНИЕ] = «надевать». L[СУЩНОСТЬ-ПРЕДМЕТ-ИНФОРМАЦИЯ / АТРИБУТ / СУЩЕСТВОВАНИЕ] = «содержательный» (документ / книга).

Функции определенного вида (формы) над получаемым векторным пространством образуют осмысленные фразы языка. Необходимо, во-первых, организовать иерархическое множество более общих и более частных по семантике предложений, во вторых задать классы синонимичных и полусинонимичных предложений в рамках классов общности и конкретизации на основе дополнительных признаков.

Примером упорядочения функций осмысленных предложений на понятийном пространстве (см. табл. 1) может служить набор точек:

L[СУЩНОСТЬ-ПРЕДМЕТ-В ЖИВОМ (еда) / ИМЕННАЯ ГРУППА-КАЧЕСТВО (атрибут) / ОЧЕНЬ-ПОЗИТИВ... (отличный)] = {delecious, aromatic} + L[СУЩНОСТЬ-ПРЕДМЕТ-В ЖИВОМ

(еда) / ИМЕННАЯ ГРУППА-СИСТЕМА (именная часть) / СУЩНОСТЬ-СУЩЕСТВО-_-СЛОЖНОЕ (животное)] = {beef, pork, mutton, chicken, …} + L[СВЯЗЬ-ПРЕДМЕТ-НА НЕЖИВОМ (опора) / ГЛАГОЛЬНАЯ ГРУППА-СИСТЕМА (ГЛАГОЛЬНАЯ ЧАСТЬ) / БОЛЬШОЕ] = {is on, lies on, is situated on, is located on, …} + «the» + L[СУЩНОСТЬ-ПРЕДМЕТ-НА КОТОРОМ НЕЖИВОЕ / ИМЕННАЯ ГРУППА-СИСТЕМА (именная часть) / БОЛЬШОЙ-ПЛОСКИЙ] = = {стол, тумбочка, чайный столик}, где «+» означает конкатенацию семантических векторов. В результате могут быть сформированы предложения вида «вкусная курица лежит на тумбочке», «ароматная рыба находится на столе» и другие см. табл. 1. После решения проблемы генерации фраз приведенного вида в первом приближении необходимо перейти к проблеме их усложнения и порождения множеств синонимичных фраз языка.

Подмножества многомерной классификации осмысленных предложений. Для решения проблемы генерации осмысленных синонимичных друг другу предложений предлагается умножать исходный класс предложений C_i на синонимизирующую группу соответствующего уровня G_i, получая класс следующего за ним уровня C_{i+1}, где $C_i * G_i = C_{i+1}$. Здесь под умножением будем подразумевать преобразование определенного вида над семантическими шаблонами, алгоритмическое описание которого является темой отдельной публикации. Синонимизирующие ряды слов имеют следующий вид:

1. Определения: стремление получить, эмоциональная тенденция, направленность в будущее.
2. По частям речи: желание, желать, желательный, желательно.
3. По позиции в предложении: что-то как желание, быть желанием, что-то желания, с учетом желания.
4. Ряды дифинонимов (разные элементы одной речевой ситуации или явления): желающий, объект желания, страстный, вожделенный.
5. Ряд близких синонимов: хотение.
6. Широкий синонимический ряд (с общей 1–3 семами): стремление, нужда, необходимость.
7. Группы слов контекстуальных синонимов (от 1 и более общей семы): важность / значимость / насущность, цель / задача / план, идея / проект / концепция, динамика / направленность / сосредоточение.
8. Метафорические употребления: жажда, алчность, жадность, вожделение.
9. По категориям, например, множественного числа: с желаниями.
10. «Атрибуции»: с множеством желаний, с кучей желаний, с чувством желания, с чувством теплоты / прикосновения / огня желания, с эмоциональным / душевным желанием.
11. Перевод на другой язык: a desire, a wish, to want.

Синонимизирующие ряды слов. Умножение фрагмента многомерного пространства осмысленных предложений языка на синонимизирующие ряды можно представить следующим образом (табл. 2).

Так, в результате генерации при выборе одного из вариантов синонимизации из каждой колонки таблицы возникают фразы вида:

1. «Городской житель имеет необходимость быть предприимчивым с компьютерной системой с учетом рассрочки».
2. «В пространстве городской среды можно ощутить желание обеспечить выполнение покупки аппаратного обеспечения компьютерной техники в рамках рассрочки».
3. «Заключенный в пространство городской среды, горожанин осуществляет желание получить товар, в частности, компьютерную технику с учетом наличия рассрочки».

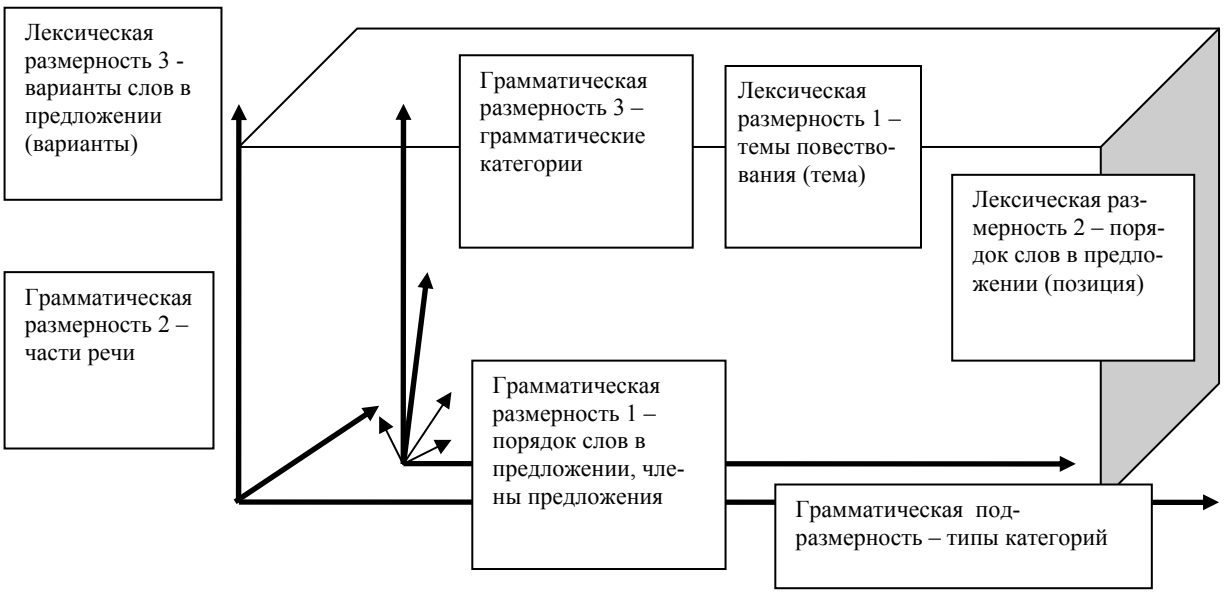

Рис. 1. Координаты многомерного лексико-грамматического пространства

Таблица 2

Схема синонимизации фраз естественного языка

Ключевое слово				Компьютер	
Тема – рема			Продать (ТЕМА)	Компьютер (РЕМА)	
Другие уровни	…	…	…	…	…
Значимый ряд	Горожанин (1)	Хочет (2)	Продать (3)	Компьютер (4)	[Не безвозмездно] (5)
Умножение на лексические ряды	Человек из города	Желание	Продать в рассрочку	Компьютера	(По) акции
	Городской человек	Нетерпение	Продавец	Компьютерной системы	(В) рассрочку
	Городской житель	Намерение	Имеет [такой товар] товар [как]	Вычислительной системы	Дешево
	В городе	Необходимость	Предприимчив в операциях с	Средства вычисления	Дорого
	В городском месте <малоупотребительно>	Требование	Арендовать	Средства осуществления операций над информацией	С доставкой
	В городских местах	Желательно	Арендатор		С учетом дороговизны
	В городской среде	По желанию	Продажа		
		С радостью	Проданный		
Внесение атрибуций	В пространстве городской среды	Ощутить желание	Осуществить покупку	Компьютерного оборудования	Во время проведения акции
	В фокусе городского пространства	Ощутить ясность чувства феномена желания	Выполнить покупку	Компьютерной техники	Во время прохождения акции
		Проявил себя через желание	Обеспечить выполнение покупки	Аппаратного обеспечения компьютерной техники	Во время похождения мероприятия – акции

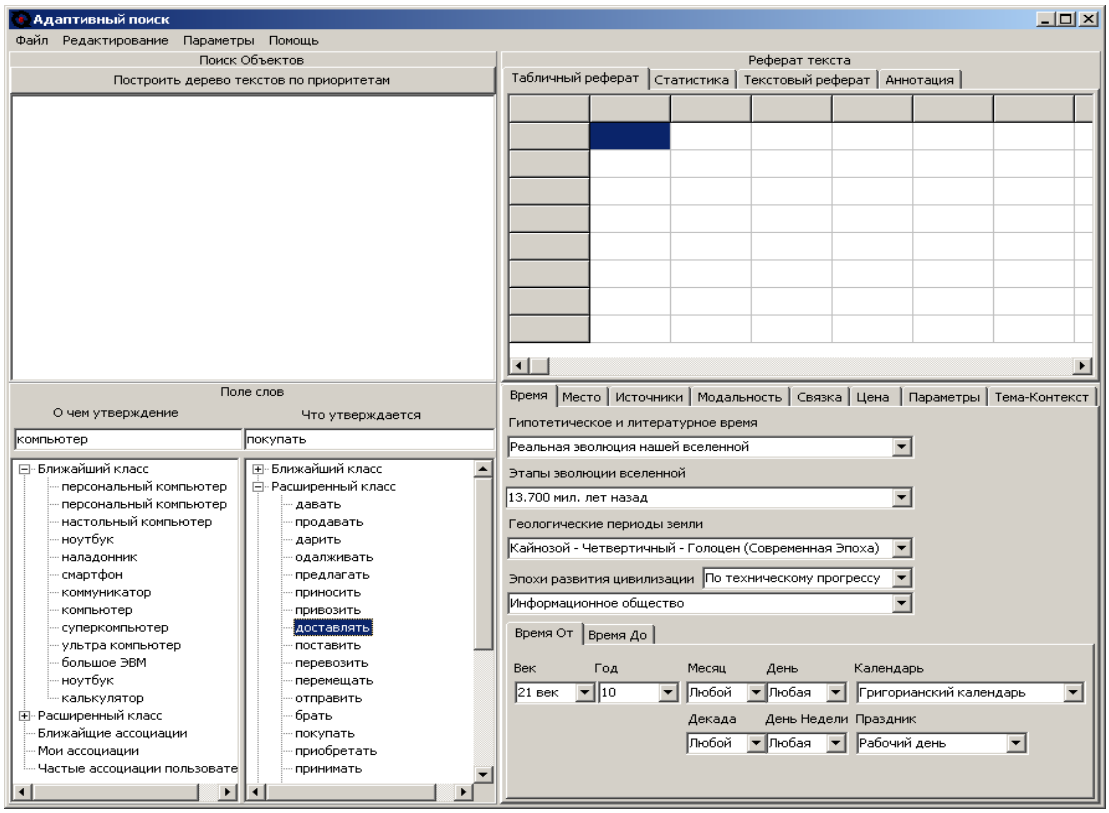

Рис. 2. Интерфейс разрабатываемой программы «Адаптивный Поиск»

Дерево генерации предложений на основе темы и ремы. Традиционно актуальное членение предложений включает в себя деление на тему и рему, рема является ключевым словом в предложении, а тема относится ко всему тексту или его фрагменту.

На вершине дерева актуального членения стандартного предложения (без синонимизации, декомпрессии и внесения семантического шума) имеет место ключевое слово (рема); на втором уровне дерева парсинга имеет место тема и рема; на третьем имеет место четверка: тема, связка, рема, модальность; на четвертом уровне добавляются обстоятельства, имеющие важную уточняющую функцию. На пятом уровне имеют место очевидные, понятные из контекста обстоятельства и конкретизация; на шестом – полупустые слова, уточняющие аспекты слов, указанных выше в дереве разбора (назовем их «атрибуциями»).

Так, например, анализ предложения вида «Нельзя не отметить, что образец книги австралийского автора, который я читал вчера, далеко не лишен, того что можно назвать воплощением пошлости» может происходить в несколько этапов развития ключевого слова «пошлая», далее темы и ремы «книга пошлая» и т. д.

0. Тема повествования: «книга»;

1. Ключевое слово: «пошлая»;

2. Тема-Рема: «книга – пошлость» = «книга – пошлая»;

3. Тема-Рема-Связка-Модальность: «плохо, что книга отличается пошлостью»;

4. Важная конкретизация: «книга австралийского автора»;

5. Контекстуальная конкретизация: «книга, которую я читал вчера»;

6. Атрибуции понятий: «образец книги», «воплощение пошлости»;

7. Различные эквивалентные преобразования, например, двойное отрицание: «далеко не лишена пошлости», «нельзя не отметить, что ….».

Генерация классов синонимичных предложений может быть использована при генерации синонимических запросов к поисковым системам.

В работе выполнен анализ проблемы генерации осмысленных фраз синонимичных и частично синонимичных друг другу. Предложена модель генерации синонимичных фраз на основе таблиц синонимизации, и лексических рядов, генерируемых на основе семантического векторизованного словаря [2]. Рассматриваются возможности визуального интерфейса поисковых систем на основе модели членения предложения на тему и рему.

Подчеркивается важность продолжения исследований в области генерации синонимических поисковых запросов.

Библиографические ссылки

1. Avancini H., Lavelli A., Sebastiani F., Zanoli R. Automatic Expansion of Domain-Specific Lexicon by Term Categorization. ACM Translation on Speech and Language Processing. Vol. 3. No. 1. May 2006. P. 1–30.

2. Сафонов К. В., Личаргин Д. В. Elaboration of a vector-based semantic classification over the words and notions of the natural language // Вестник СибГАУ. 2009. № 5 (26). С. 52–56.

3. Личаргин Д. В., Суманеева Я. А., Юрьева Е. В. Метод подстановочных таблиц и его применение в сфере обучения русскому языку для иностранцев // Вестник Сургутского гос. пед. ун-та. 2012. № 6. С. 179–187.

4. Сафонов К. В., Личаргин Д. В. Разработка векторизованной семантической классификации над словами и понятиями естественного языка // Вестник СибГАУ. 2010. № 4. С. 33–37.

5. Сафонов К. В., Личаргин Д. В. Некоторые принципы автоматической генерации учебных материалов на основе баз знаний и лингвистической классификации // Вестник СибГАУ. 2012. № 2 (42). С. 72–77.

References

1. Avancini H., Lavelli A., Sebastiani F., Zanoli R. Automatic Expansion of Domain-Specific Lexicon by Term Categorization. ACM Translation on Speech and Language Processing, Vol. 3, No. 1, May 2006, p. 1–30.

2. Safonov K. V., Lichargin D. V. *Vestnik SibGAU*, 2009, № 5 (26), pp. 52–56.

3. Lichargin D. V., Sumaneeva Ya. A., Yuryeva E. V. *Vestnik of Surgut GPU*, 2012, № 6, p. 179–187.

4. Safonov K. V., Lichargin D. V. *Vestnik SibGAU*, 2010, № 4 (30), p. 33–37.

5. Safonov K. V., Lichargin D. V. *Vestnik SibGAU*, 2012, № 2 (42), p. 72–77.