УДК 004.9:371.261

Informatics and Mathematical Methods in Simulation
Vol. 6 (2016), No. 2, pp. 173-183

УЧЕТ МЕЖФРАЗОВЫХ СВЯЗЕЙ ПРИ АВТОМАТИЗИРОВАННОМ ПОСТРОЕНИИ ТОЛКОВОГО СЛОВАРЯ ПРЕДМЕТНОЙ ОБЛАСТИ

А.Б. Кунгурцев, А.И. Гаврилова, А.С. Леонгард, Я.В. Поточняк

Одесский национальный политехнический университет,
просп. Шевченко, 1, Одесса, 65044, Украина; e-mail: abkun@te.net.ua

Предложены алгоритмы и информационная технология выявления межфразовых связей в процессе построения словарей предметной области. Разработанные решения позволяют повысить качество толкового словаря и сократить время на его формирование, что имеет большое значение для создания и развития информационных систем.

Ключевые слова: межфразовые связи, анафора, термин, словарь предметной области

Введение

В соответствии с [1] под толковым словарем предметной области (СПО) будем понимать специализированный словарь, дающий определение множеству понятий, связанных с деятельностью некоторой организационной структуры. Толковый словарь оказывается очень полезным при решении различных задач, связанных с созданием и развитием информационных систем, подготовкой кадров, созданием новой документации, четким распределением обязанностей между сотрудниками.

Каждая запись в словаре представляет одно слово или устойчивое для данной предметной области словосочетание («термин»), для которого приведено толкование, специфическое для данной предметной области, а также список синонимов.

Анализ литературных данных и формулировка задач исследования

Выбор терминов для толкового словаря происходит на основании частоты их появления в анализируемом тексте. При этом, в известных на данный момент работах [1-3] не учитывались межфразовые связи (МС). В качестве примера рассмотрим два предложения: Жесткий диск является основным запоминающим устройством для большинства персональных компьютеров. Обычно он содержит несколько пластин, покрытых слоем ферромагнитного материала, которые вращаются со скоростью в несколько тысяч оборотов в минуту. Здесь словосочетание «Жесткий диск» может быть определено как термин в рассматриваемой предметной области. Однако, без учета МС анализатор текста отметит только одно появление данного термина, тогда как фактически их два, поскольку во втором предложении местоимение «он» замещает термин «жесткий диск». Это может привести к ошибкам при решении вопроса о включении выявленного термина в словарь.

Подробное описание, подходы и особенности систем для разрешения анафоры на английском языке можно найти в работе [4], однако задача разрешения анафоры недостаточно исследована для русского языка, а также имеет свои особенности, ввиду

чего существует лишь небольшое количество систем, с различным успехом справляющимися с данной задачей.

В работе [5], сделана попытка построения алгоритма выделения МС в некотором тексте, однако представленный в работе подход нецелесообразен в рамках СПО, где требуется осуществить поиск анафоры при известном термине. Предложенная в [6] программа является коммерческой разработкой, которая в качестве синтаксического анализатора также использует коммерческий продукт. Поэтому отсутствует информация как об алгоритме выделения МС, так и об оценке эффективности программы. В работе [7] в методике разрешения кореференции рассмотрены только 3 группы местоимений – личные, возвратные и относительное, тогда как на практике часто используются также уточняющие прилагательные и порядковые числительные.

Система построения СПО выдвигает требования, которые ограничивают, и в то же время дополняют задачу учета МС, связанную с восстановлением частоты появления термина в тексте, что отличает данную систему разрешения кореференции от других ей аналогичных, делает ее более мобильной и быстрой.

На основании изложенного можно сформулировать следующие задачи исследования:

- анализ возможных типов МС и выделение тех, которые должны быть обнаружены при построении СПО;
- определение частот появления МС выделенных типов в текстах различного вида с целью обоснования целесообразности использования механизма выявления МС при построении СПО;
- создание формализованного описания процесса определения МС в текстах;
- разработка алгоритмов и программного обеспечения для выявления МС и соответствующей корректировки СПО.

Определение межфразовых связей, которые должны учитываться при построении словаря

Приведем определение некоторых понятий.

Межфразовая связь – это связь между предложениями, абзацами, главами и другими частями текста, организующая его смысловое и структурное единство. Для реализации межфразовой связи используются синтаксические повторы, синонимы, анафорические местоимения, слова с временным и пространственным значением и т.д.

Анафора – лингвистическое явление, когда интерпретация некоторого выражения (анафорическое выражение) зависит от другого выражения, которое, встречалось в тексте ранее анафорического выражения (антецедент), либо находится после него (постцедент).

Анафора считается частным случаем такого явления, как «дейксис». Согласно [8], дейктическим является такой элемент, который выражает идентификацию объекта – предмета, места, момента времени, свойства, ситуации – через его отношение к речевому акту, его участникам или контексту. В работе [9] говорится: «дейксис является базисным явлением, нежели анафора, а анафора в некотором смысле производна от дейксиса». Данное утверждение объясняется тем, что в то время, как дейксису характерно изменение фокуса внимания, анафоре характерно его сохранение.

Рассмотрим общую классификацию анафор для анализа возможности их использования при построении СПО, приведенную в работе [10].

1. По корреляционному расположению. Если в предыдущем предложении или ранее в рассматриваемом предложении найден 1 и более антецедентов, согласованных в роде, числе и падеже с анафорой, значит, анафора – ретроспективная.

Если антецедент найден в том же предложении, что и анафора, и следует после нее, значит, анафора – проспективная (катафора).

2. По позиции относительно предложения. Если в одном и том же предложении найден и антецедент, и анафора, значит, анафора – внутрисентенциальная. Если в предложении найдена только анафора, а антецедент – в предшествующих предложениях, значит, анафора – дискурсивная.

3. По формальной выраженности. Если в предложении анафора выражена в виде слова, значит, анафора имеет полную форму. Если в предложении нет слова, которое представляет анафору, и предложение – неполное (напр. отсутствует подлежащее, есть только сказуемое), а в предыдущем предложении есть антецедент (подлежащее), согласованный в роде и числе со сказуемым в последнем предложении, то перед нами нулевая анафора.

4. По наличию лексического компонента. Если анафора состоит только из анафорического местоимения, анафора – простая (я, это, они). Если анафора состоит из именной группы, содержащей местоимения в приименном употреблении (эта комната, такие сооружения и т.п.), анафора – составная.

5. По лексико-грамматическому классу. Если анафора выражена личным или указательным местоимением, значит, анафора – местоименная. Если анафора выражена именной группой или существительным, синонимичным по значению с антецедентом, значит, анафора – именная. Если анафора явно отсутствует, но исходя из контекста можно заключить, что она присутствует, и это выражается в виде неполного предложения, где отсутствует сказуемое, такая анафора – глагольная.

6. По вербальному описанию антецедента. Если антецедент явно указан в тексте, значит, анафора – эксплицитная. Если вербальное описание антецедента в тексте отсутствует или анафора представляет собой довольно значительные фрагменты текста различного синтаксического статуса, значит, анафора – имплицитная.

7. По степени полноты. Полная анафора означает прямой повтор наименования референта и служит для выражения отношения тождества между референтами двух наименований.

При частичной анафоре второе обозначение отличается от первого, представляя собой, например, согласованное в роде и числе местоимение, его синоним, или гипероним, соотносящийся с гипонимом.

Ассоциативная анафора предполагает наиболее широкие возможности выбора анафорического обозначения, поскольку логико-смысловые отношения, лежащие в её основе (метонимические, метафорические), могут быть чрезвычайно разнообразными. В следующем примере слова металлу и пластику являются отсылочными компонентами, ассоциирующимися с материалами изготовления антецедента жесткий диск.

Научные сотрудники института высшей математики и программирования Израиля заявили, что через 2 года жесткий диск превратится в мягкий накопитель информации. Носители будущего будут изготавливаться из материалов, противоположных металлу и пластику.

Конкретная анафора может относится одновременно к различным видам приведенной классификации. Рассмотрим два предложения.

Устройство позиционирования головок (жарг. актуатор) представляет из себя малоинерционный соленоидный двигатель. Он состоит из неподвижной пары сильных неодимовых постоянных магнитов, а также катушки (соленоид) на подвижном кронштейне блока головок.

Анафора он является ретроспективной (1), дискурсивной (2), эксплицитной (6), простой (4), местоименной (5), частичной выраженности (7) и имеет полную форму 0 (3).

Анализ различных видов МС с точки зрения их применимости при построении СПО предусматривает определение следующих характеристик МС:

- соответствие решаемой задаче;

- возможность формализации процесса определения анафоры в тексте;
- частота встречаемости в тексте, влияющая на качество СПО.

При определении соответствия МС решаемой задаче были удалены из дальнейшего рассмотрения следующие виды анафор:

- полная (предусматривает повторение ранее приведенного определения (возможно, термина из СПО); её учет не влияет на частоту использования термина);
- глагольная (термин в СПО непосредственно действие не обозначает);
- именная (в анафоре используется именная группа или существительное, которые могут быть ранее определенным термином или его синонимом);
- составная (то же, что и именная анафора, но с возможными сопутствующими ей указательными местоимениями);
- ассоциативная (при данном типе анафоры присутствует семантическая связь между словами и отсутствуют местоимения);
- проспективная (редко встречаемый вид, используется преимущественно в художественной литературе);

При определении возможности формализации следует исходить из следующего:

- имеется ли формальная возможность определить предложение, которое содержит анафору;
- имеется ли формальная возможность определить предложение, которое содержит антецедент или постцедент;
- имеется ли возможность связать анафору с термином путем использования морфологических атрибутов (часть речи, род, число, падеж) и синтаксических связей.

Используя указанные критерии из дальнейшего рассмотрения были удалены следующие типы анафоры:

- нулевая (нельзя точно выделить предложение, содержащее анафору);
- имплицитная (нельзя точно выделить предложение, содержащее антецедент или постцедент).

Таким образом, остались следующие типы МС, которые будут учтены при построении словаря предметной области:

1. Ретроспективная. Накопитель на жестком диске (HDD) относится к наиболее совершенным и сложным устройствам современных систем хранения цифровой информации, характеризующийся значимым объемом хранимой информации при низкой себестоимости. Однако, исходя из исследований доктора Бианки Шредер и Google, в силу своих конструктивных особенностей и элементов количество от- казов данного устройства после 3-го года работы стабильно увеличивается.
2. Полной формы. Операция дефрагментации должна быть отключена, так как она практически никак не влияет на производительность SSD-носителя и лишь дополнительно изнашивает его.
3. Местоименная (простая). Наследование — механизм языка, позволяющий описать новый класс на основе уже существующего (родительского, базового) класса или интерфейса. Оно является одним из основных принципов объектно-ориентированного программирования.
4. Эксплицитная. Все штрихкоды можно разделить на два типа: линейные и двухмерные. Первый – это код, который читается в одном направлении, характеризуется простой эксплуатацией и низкой себестоимостью.
5. Внутрисентенциальная. Функциональное программирование предполагает обходиться вычислением результатов функций от исходных данных и результатов других функций, и оно не предполагает явного хранения состояния программы.
6. Дискурсивная. Функциональное программирование предполагает обходиться вычислением результатов функций от исходных данных и результатов других функций, и не предполагает явного хранения состояния программы. Соответственно, не предполагает оно и изменяемость этого состояния.

Оценка частоты появления МС в текстах из различных предметных областей

Для определения влияния МС на качество СПО было проведено исследования 50 научно-популярных, публицистических и научных текстов объемом от 200 до 1500 слов на предмет выявления частоты появления МС.

В результате исследования установлено:

- различные виды анафор встречаются в текстах различной тематики в среднем 54,3 раз на 1000 слов;
- в публицистических и научно-популярных текстах – 54,1 раз на 1000 слов;
- в технических текстах – 35,6 раз на 1000 слов;
- в научно-технических текстах – 33,1 раз на 1000 слов.

При этом наиболее часто встречается местоименная анафора:

- в публицистических и научно-популярных текстах – 23,2 раз;

В научно-технических текстах местоименная анафора встречается 19,4 раза; в технических – 20,6 раз.

Учитывая, что при построении СПО слова и словосочетания, встречающиеся в анализируемых текстах в количестве 25 раз на 1000 слов, обычно принимаются в качестве терминов СПО, можно сделать вывод, что учет МС заметно изменит частотные характеристики терминов в СПО.

Результаты анализа видов МС и текстов приведены в табл. 1.

Таблица 1.

Оценка возможности использования МС при построении СПО

Классификация	Вид	Подлежит формализации	Высокая частота встречаемости	Соответствует решаемой задаче
По корреляционному расположению	Ретроспективная	+	+	+
	Проспективная	+	-	+
По позиции относительно предложения	Внутрисентен-циальная	+	+	+
	Дискурсивная	+	+	+
По формальной выраженности	Полная форма	+	+	+
	Нулевая	+	-	+
По наличию лексического компонента	Простая	+	+	+
	Составная	+	+	-
По лексико-грамматическому классу	Местоименная	+	+	+
	Именная	+	+	-
	Глагольная	-	-	-
По вербальному описанию антецедента	Эксплицитная	+	+	+
	Имплицитная	-	-	-
По степени полноты	Полная	+	+	-
	Частичная	+	+	+
	Ассоциативная	-	-	-

Формализация описания и выявления МС

Представим анализируемый текст S в виде множества предложений

$$S = \{S_i\} i = 1, n, \quad (1)$$

а каждое предложение – в виде последовательности элементов (слов и знаков препинания)

$$e_1, ..., e_l, .., e_m. \quad (2)$$

Каждый элемент будет характеризоваться текстом N и множеством атрибутов A

$e =< N, A >$

Определим некоторые из них. Пусть А1 представляет часть речи, А2 – число, А3 – род, А4 – лицо, А5 – падеж, А6 – время, А7 – залог, А8 – одушевленность.

Будем считать, что к началу процесса выявления МС завершился первый этап построения СПО. Представим каждую запись СПО в виде $d =< N, q >$,

где N – термин (слово или словосочетание); q - количество появления термина в тексте S.

В общем случае $N = \{e\}$.

Тогда СПО можно представить множеством записей:

$$D = \{d_j\} j = 1, k, \quad (3)$$

где k – количество терминов, обнаруженных в S (до процесса редактирования словаря).

Ведем определение принадлежности термина предложению $d_j \in_d S_i$.

Поскольку анализаторы текстов обычно не дают сведений о типе местоимения, то предложено ввести множество личных местоимений третьего лица, указательных и притяжательных местоимений, которые могут играть роль анафоры.

$M\,\mathrm{Pr} = \{он, она, оно, они, тот, этот, такой, таков, столько, свой, его, её, их\}$

Также необходимо ввести множество уточняющих прилагательных $MAdj$, порядковых числительных $MNum$

$$MAdj - \{указанный, данный, последний, ...\},$$

$$MNum = \{первый, второй, третий\},$$

В ходе анализа относительных местоимений, их изменяемости по введенным выше атрибутам, их было решено разделить на две группы, таким образом введем множество относительных местоимений, для которых условия проверок отличаются от условий проверок принадлежности $M\,\mathrm{Pr} - M\,\mathrm{Re}l\,\mathrm{Pr}$

$$M\,\mathrm{Re}l\,\mathrm{Pr} = \{кто, что, сколько\}.$$

Также дополним множество $M\,\mathrm{Pr}$ относительными местоимениями, проверки по атрибутам, для которых аналогичны элементам этого множества:

$M\,\mathrm{Pr} = \{он, она, оно, они, тот, этот, такой, таков, столько, свой, его, её, их,$

$который, какой, чей\}$

Рассмотрим ряд возможных ситуаций.

Ситуация 1. В предложении S_i обнаружен термин d_j, т. е.

$$\exists d_j \mid d_j \in_d S_i \wedge d_j \in D \tag{4}$$

Тогда следует искать МС сначала в данном предложении, а потом в следующем. Если предложение не содержит местоимений третьего лица, указательных и притяжательных местоимений, уточняющих прилагательных и порядковых числительных, то в нем нет анафор. Для этого осуществляем ряд проверок для каждого элемента предложения (2):

$$\begin{gathered} e_l \to A1 \neq pronoun \vee e_l \to A1 = pronoun \wedge e_l \to N \notin M\,\mathrm{Pr} \wedge N \notin M\,\mathrm{Re}\,l\,\mathrm{Pr} \\ e_l \to A1 \neq adjective \vee e_l \to A1 = adjective \wedge e_l \to N \notin MAdj \\ e_l \to A1 \neq numeral \vee e_l \to A1 = numeral \wedge e_l \to N \notin MNum \end{gathered} \tag{5}$$

Если в результате проверки каждого элемента предложения S_i результатом было только $true$, то делаем вывод, что это предложение S_i не содержит анафор. Аналогичный анализ производим для следующего предложения S_{i+1}. Если в этих предложениях в результате анализа также получен результат $true$, то можно сделать вывод об отсутствии МС между элементами предложений S_i и S_{i+1}. Тогда следует в соответствии с (4) проверить предложение S_{i+1} на вхождение в него термина из СПО.

Ситуация 2. В предложении S_i обнаружен термин (4), который в общем случае соответствует некоторой последовательности элементов предложения $d_j \to N = e_m e_{m+1} .... e_{m+k}$.

Также в этом предложении обнаружен некоторый элемент e_l, который может быть анафорой, т.е. результатом одной из проверок (5) было false.

Рассмотрим случай, когда термин является именной группой, содержащей опорное управляющее слово – имя существительное, которое нужно выявить. Пусть обнаружен элемент e_{nn}, такой что $e_{nn} \to A1 = noun \wedge e_{nn} \in \{e_m, e_{m+1}, \ldots, e_{m+k}\}$

Если антецедент и анафора находятся в одном предложении, то это предложение должно быть сложным. Антецедент будет находиться в одном простом предложении, а анафора в одном из следующих простых предложений. Записываем это условие. Существует $e_h \to punctuation \wedge h > (m+k) \wedge l > h$

Проверяем связь предполагаемых антецедента и анафоры. По результату проверок (8) получили $e_l \to A1 = pronoun \vee e_l \wedge e_l \notin M\,\mathrm{Re}\,l\,\mathrm{Pr} \to A1 = adjective \vee e_l \to A1 = numeral$.

Тогда условием выявления МС является

$$e_{nn} \to A2 = e_l \to A2 \wedge e_{nn} \to A3 = e_l \to A3 \tag{6}$$

Если $e_l \to A1 = pronoun \vee e_l \wedge e_l \in M\,\mathrm{Re}\,l\,\mathrm{Pr}$, тогда условиями для проверок являются

- e_l = 'кто', то для установления МС должно выполняться $e_{nn} \to A8 = animate$;
- e_l = 'что', то для установления МС должно выполняться $e_{nn} \to A8 = inanimate$;
- e_l = '*сколько*', то условие $e_{nn} \to A2 = e_l \to A2$.

(7)

Если термин состоит из одного слова, то это слово является элементом предложения e_{nn}.

Ситуация 3. В предложении S_i обнаружен термин (4) и не обнаружено кандидатов на анафору. В предложении S_{i+1} при отсутствии терминов обнаружены кандидаты на анафору, т.е такие элементы e_l, которые прошли проверки (5), и результат одной из проверок для каждого элемента оказался равным false.

Тогда соответствие между элементом – опорным словом термина e_{nn} и анафорой e_l нужно искать в соответствии с условиями проверок (6), (7). Кандидаты e_l, удовлетворившие данным условиям, считаем анафорами термина.

Ситуация 4. В предложении S_i обнаружено более одного термина (4) и обнаружен кандидат на анафору. Элемент e_l, являющийся кандидатом на анафору, проходит проверки на соответствие с каждым из терминов e_{nn} (6), (7). МС устанавливается с ближайшим e_{nn}, для которого e_l прошел проверку, после чего e_l уже не может быть анафорой для последующих терминов $e_{nn+1}, e_{nn+2}, \ldots$.

Ситуация 5. В предложении S_i обнаружено более одного термина (4) и обнаружено более одного кандидата на анафору. Каждый элемент e_l, являющийся кандидатом на анафору, проходит проверки на соответствие с каждым из терминов e_{nn} (6), (7). МС устанавливается между ближайшими $e_{nn} - e_l$, где e_l прошел проверки для e_{nn}, после чего e_l уже не может быть анафорой для последующих терминов $e_{nn+1}, e_{nn+2}, \ldots$, однако e_{nn} может далее рассматриваться как возможный референт для $e_{l+1}, e_{l+2}, \ldots$.

Ситуация 6. В предложении S_i обнаружено более одного термина (4) и обнаружено один или более одного кандидата на анафору. В предложении S_{i+1} обнаружен термин или более одного теримна и один или более кандидатов на анафору. В таком случае следует произвести проверки (6), (7) для элементов e_l из S_{i+1} – которые не прошли проверки на соответствие с терминами из S_{i+1} – на соответствие с терминами из S_i. МС устанавливается между ближайшими $e_{nn} - e_l$, где e_l прошел проверки для e_{nn}, аналогично ситуации 5.

Технология построения словаря с учетом МС

Технология предусматривает три основных этапа.

На первом этапе производится поиск терминов в некотором тексте. Поскольку число анафор в тексте значительно меньше числа терминов, то параллельно с их поиском определяется вхождение в текст слов из множеств:

$M\,\mathrm{Pr} = \{тот, этот, такой, таков, столько, свой, его, её, их\}$.

$MAdj - \{указанный, данный, последний, \ldots.\}$,

$MNum = \{первый, второй, третий\}$,

$M\,\mathrm{Re}l\,\mathrm{Pr} = \{кто, что, который, какой, чей, сколько\}$,

которые могут оказаться анафорами.

В анализируемый текст, представленный в формате .txt, перед каждой потенциальной анафорой вставляются метасимволы для её индексации.

На втором этапе после предварительного определения множества терминов выделяются предложения, которые предшествуют предложению с потенциальной

анафорой и предложения, которые содержат потенциальную анафору. В работе [5] для каждой анафоры текст просматривается в обратном направлении, в процессе чего составляется множество потенциальных антецедентов. Однако в виду того, что приоритетной задачей нашего алгоритма является не определение любого антецедента, а определения термина-антецедента, при разрешении анафоры в предложениях осуществляется поиск выделенных ранее терминов, и затем осуществляется поиск МС в соответствии с рассмотренными ранее ситуациями. Если анафора найдена, то производится корректировка частоты появления соответствующего термина в тексте.

На третьем этапе экспертом принимается решение об определении нижней частотной границы включения терминов в словарь.

Проведение экспериментов и анализ результатов

Для испытания предложенной технологии определения межфразовых связей была разработана методика определения эффективности выявления МС в текстах из различных предметных областей.

Методика предусматривает выполнение следующей последовательности действий:

- выбор текстов из некоторой предметной области;
- определение терминов (термин, количество вхождений в текст) в выбранных текстах;
- выявление МС с использованием предложенной технологии, определение ранее не учтенных вхождений терминов в текст;
- выявление МС в тех же текстах экспертом без использования предложенной технологии;
- определение ошибок при выявлении МС, которые были учтены в технологии;
- определение МС, которые не были учтены в технологии.

Результаты анализа изменения частотных характеристик терминов после учета МС с использованием предложенной технологии приведены в тал. 2.

Таблица 2.

Изменение частотных характеристик (числа вхождений) терминов

Тип текста и ссылки	Размер текстов (число слов)	Всего обнаружено МС	Термины	
			Всего (к-во)	Изменилось число вхождений на (%)
Электротехника [11], [12]	2500	69	61	11.6
Информатика [13]	2397	47	58	18.9
Энергетика [14]	4700	175	139	26.9
Экономика [15], [16]	5971	197	148	32.8
Юридические науки [17], [18]	3500	109	83	31.3

На основании анализа данных из табл. 2 можно сделать вывод, что учет МС существенно влияет на частотные характеристики терминов и должен быть реализован в технологии создания СПО.

Экспертом был проведен анализ текстов [11-18]. В результате было выявлено, что в автоматическом режиме не учтено 31 МС (учтенное количество МС равно 597), что составляет приблизительно 5% общего количества МС. Причинами ошибок являются синтаксические ошибки, опечатки, широкое использование слов на английском языке, сложная структура предложений с многочисленными вводными конструкциями. Учитывая особенности научных и технических текстов, такое количество ошибок допустимо.

Выводы

Показано, что существующие технологии построения СПО на основе анализа текстов не учитывают МС, что приводит к ошибкам в определении частотных характеристик терминов. Определены типы МС, которые должны быть выявлены при построении СПО. Дано математическое описание процесса определения МС. Разработана технология построения СПО с учетом МС и соответствующее программное обеспечение. Проведены эксперименты, которые подтвердили обоснованность теоретических положений и эффективность разработанной технологии и программного обеспечения.

Список литературы

1. Кунгурцев, А.Б. Метод автоматизированного построения толкового словаря предметной области/ А.Б. Кунгурцев, Я.В. Поточняк, Д.А. Силяев // Технологический аудит и резервы производства –2015. – № 2/2(22). – С. 58 – 63
2. Кунгурцев, А.Б. Застосування мереж фреймів для побудови моделі вилучення фактів з текстів на природній мові / А. Б. Кунгурцев, С. М. Бородавкін // Искусственный интеллект. – 2009. – №4.– С. 202 – 207.
3. Кунгурцев, А. Б. Метод построения словарей предметных областей для извлечения фактов из текстов на естественном языке/ А.Б. Кунгурцев, С.Н. Бородавкин, А.П. Голуб // Восточно-европейский журнал передовых технологий. – 2010. – № 1/4 (43). – С. 32 – 36.
4. Mitkov, R. Anaphora resolution: the state of the art / R. Mitkov // School of Languages and European Studies, University of Wolverhampton. – 1999. – С. 2 – 29
5. Malkovsky, M.G. A method for pronominal anaphora resolution in the course of syntactic analysis / M.G. Malkovsky, A.S. Starostin, I.A. Shilov // Proceedings of Sworld conference. – 2013. – С. 2 – 3
6. Bogdanov, A.V. Anaphora analysis based on ABBYY COMPRENO linguistic technologies / A.V. Bogdanov, S.S. Dzhumaev, D.A. Skorinkin, A.S. Starostin // Becasovo : 20th International Conference on Computational Linguistics "Dialog". – 2014. – С. 1 – 13
7. Ionov, M . The impact of morphology processing quality on automated anaphora resolution for Russian / M. Ionov, A. Kutuzov // Becasovo.: 20th International Conference on Computational Linguistics "Dialog". – 2014. – 3 с.
8. Падучева, Е.В. Семантические исследования/ Е.В. Падучева. // М.: Языки русской литературы. – 1996. – 464 с.
9. Кибрик, А.А. Человеческий фактор в языке: Коммуникация, модальность, дейксис / А.А. Кибрик // М.: Наука. – 1992. – 281 с.
10. Воронкова, А.В. Стратегии когнитивной обработки дискурсивной анафоры пропозитивно-именного типа / А.В. Воронкова. –2009. – С. 27 – 40
11. Зинченко, Е.Е. Методика расчета вентильных индукторно-реактивных двигателей / Е.Е. Зинченко, В.Б. Финкельштейн // Електротехніка і Електромеханіка. – 2009. – №4. – С. 24 – 29.
12. Мурашкин, С.И. Асинхронный частотный электропривод с векторным управление / С.И. Мурашкин // Вестник КрасГАУ. – 2012. – №9 – С. 189 – 196.

13. Проскуряков, Н.Е. Анализ и перспективы современных систем хранения цифровых данных / Н.Е. Проскуряков, А.Ю. Ануфриева // Известия Тульского государственного университета. Технические науки. – 2013. – №3 – С. 368 – 377.
14. Ушаков, В.Я. Основные проблемы энергетики и возможные способы их решения / В.Я. Ушаков // Известия Томского политехнического университета. – 2011 – Т. 319. – №4. – С. 5 – 13.
15. Оськина, Ю.Н. Обзор методик анализа финансовых результатов / Ю.Н. Оськина, Е.А. Баева // Социально-экономические явления и процессы. – 2013. - №4(050). – С.126 – 130.
16. Сысо, Т.Н. Оптимизация управления затратами предприятия / Т.Н. Сысо // Вестник Омского университета. Серия «Экономика». – 2011. – № 4. – С. 135 – 143.
17. Мальцева, Л.В. Преступность среди несовершеннолетних и ее предупреждение / Л.В.Мальцева // Общество: политика, экономика, право. – 2011. – № 4. – С.102 – 105.
18. Сафин, З.Ф. Понятие земель общего пользования и их правовой режим / З.Ф. Сафин, Э.Ф. Нигматуллина // Ученые записки Казанского государственного университета. – 2010. – Т. 152. – С. 141 – 148.
19. Kamenskaya, M.A. Data-driven methods for anaphora resolution of Russian texts // M. A. Kamenskaya, I.V. Khramoin, I.V. Smirnov // Becasovo: 20th International Conference on Computational Linguistics "Dialog". – 2014. – 8 с.
20. Protopopova, E.V Anaphoric annotation and corpus-based anaphora resolution: an experiment / E.V. Protopopova, A.A Bodrova, S.A. Volskaya et al. // Becasovo: 20th International Conference on Computational Linguistics "Dialog". – 2014. – 8 с.
21. Toldova, S.J. RU-EVAL-2014: Evaluating anaphora and coreference resolution for Russian / S. J. Toldova, A. Roytberg, A. A. Ladygina, M. D. Vasilyeva, I. L. Azerkovich, M. Kurzukov, G. Sim, D.V. Gorshkov, A. Ivanova, A. Nedoluzhko, Y. Grishina. – 2005. – 6 с.

ОБЛІК МІЖФРАЗОВИХ ЗВ'ЯЗКІВ ПРИ АВТОМАТИЗОВАНІЙ ПОБУДОВІ ТЛУМАЧНОГО СЛОВНИКА ПРЕДМЕТНОЇ ОБЛАСТІ

О.Б. Кунгурцев А.І. Гаврилова, А.С. Леонгард, Я.В. Поточняк

Одеський національний політехнічний університет,
просп. Шевченко, 1, Одеса, 65044, Україна; e-mail: abkun@te.net.ua

Запропоновано алгоритми та інформаційна технологія виявлення міжфразових зв'язків в процесі побудови словників предметної області. Розроблені рішення дозволяють підвищити якість тлумачного словника і скоротити час на його формування, що має велике значення для створення і розвитку інформаційних систем.

Ключові слова: міжфразовий зв'язкок, анафора, термін, словник предметної області.

ACCOUNTING OF INTER-PHRASE CONNECTIONS IN AUTOMATED DEVELOPMENT EXPLANATORY DICTIONARY OF SOME SUBJECT AREA

A. Kungurtsev, A. Gavrilova, A. Leonhard, Ia. Potochniak

Odessa national polytechnic university,
1, Shevchenko Ave., Odessa, 65044, Ukraine; e-mail: abkun@te.net.ua

There was proposed an algorithm and information technology for detection of inter-phrase connections in a domain dictionary building process. Developed solution allows to increase quality of the dictionary and decrease it creation time. It has a big value for creation and development information systems.

Keywords: inter-phrase connections, anaphora, term, dictionary.