УДК 519.68+681.513.7+612.8.001.57

Временные и пространственные понятия в текстах на естественном языке и их исследование

Т. В. Батура, Л. В. Ефимова, А. С. Еримбетова,
А. Б. Касекеева, Ф. А. Мурзин[1]

Целью работы является создание базы знаний, содержащей информацию о временных и пространственных понятиях, встречающихся в текстах на естественном языке. Основа базы: наиболее важные понятия, относящиеся к времени и пространству, из толкового словаря С. И. Ожегова; перефразированные варианты предложений; результаты анализа словарных статей (толкований) и примеров использования в художественной литературе соответствующих понятий из словаря С. И. Ожегова с помощью программных систем Link Grammar Parser и ДИАЛИНГ и т.д. Результаты работы могут быть использованы в интеллектуальных системах поиска информации. Диаграммы, полученные на выходе систем Link Grammar Parser и ДИАЛИНГ, представляют собой чрезвычайно интересный материал для дальнейших исследований. Целесообразно исследовать возможности применения в компьютерной лингвистике ряда конструкций и понятий математической логики, таких как конструкция Л. Генкина, реализация и опускание типов, модельная полнота, форсинг, а также ряда неклассических логик.

Ключевые слова: компьютерная лингвистика, семантика, временные и пространственные понятия, Link Grammar Parser, ДИАЛИНГ.

1. Введение

Основная цель состоит в том, чтобы исследовать конструкции на естественном языке, содержащие временные и пространственные понятия, в семантическом плане.

Для этого создается база знаний, содержащая информацию о временных и пространственных понятиях. База знаний содержит наиболее важные понятия, относящиеся к времени и пространству, из толкового словаря С. И. Ожегова; перефразированные варианты ряда предложений; результаты анализа словарных статей (толкований) и примеров использования в художественной литературе соответствующих понятий из словаря С. И. Ожегова с помощью программных систем Link Grammar Parser и ДИАЛИНГ.

Далее предполагается рассмотреть диаграммы, полученные на выходе систем Link Grammar Parser и ДИАЛИНГ средствами математической логики. Речь идет об исследовании возможностей применения в компьютерной лингвистике ряда конструкций и понятий математической логики, таких как конструкция Л. Генкина, реализация и опускание типов, модельная полнота, форсинг, а также ряда неклассических логик.

Предполагается провести анализ средствами математической логики свойств лексических функций И. А. Мельчука в контексте определения смысла текста и теоретико-множественных моделей языка, предложенных С. Маркусом.

[1] Исследования выполнены при финансовой поддержке Министерства образования и науки Республики Казахстан (грант 2018-2020 MES RK № AP05133550), Интеграционного проекта СО РАН (АААА-А18-118022190008-8) и Российского фонда фундаментальных исследований (грант № 19-07-01134).

В будущем работа будет также включать исследование комбинаторных свойств лингвистических определений и конструкций, возможностей применения в компьютерной лингвистике понятий из теоретико-множественной и алгебраической топологии и исследование географических понятий.

Результаты работы могут быть использованы в интеллектуальных системах поиска информации. А именно, для определения релевантности текста поисковому запросу и для определения тем текстов. Она также представляет интерес для ученых-лингвистов.

2. Формальные методы исследования семантики текстов

Семантика – раздел лингвистики, изучающий смысловое значение единиц языка: отдельных слов, словосочетаний, предложений, фрагментов текста. На данный момент существует ряд машинно-ориентированных методов отображения смысла высказываний.

Например, И. А. Мельчук ввел понятие лексической функции, развил понятия синтаксических и семантических валентностей и рассмотрел их в контексте толково-комбинаторного словаря [1]. В. Ш. Рубашкин и Д. Г. Лахути ввели иерархию синтаксических связей для более эффективной работы семантического анализатора. Подход И. А. Мельчука поддержан в программной системе ДИАЛИНГ [2].

Появилось понятие универсального языка представления знаний. Он может быть удобным инструментом для осуществления вывода новых знаний из уже имеющихся. Вполне возможно, что именно в направлении создания подобных семантических языков будут развиваться исследования в будущем. Например, в настоящее время система Knowledge Vault содержит 1.6 миллиарда фактов. Система NELL, разрабатываемая в рамках проекта ReadTheWeb университетом Карнеги – Меллона, содержит более 50 миллионов утверждений, дополнительно характеризующихся различными степенями доверия.

Еще один подход – это использование синтаксического анализатора Link Grammar Parser [3], разработанного в университете Карнеги – Меллона, базирующегося на некоторой специальной теории синтаксиса. Отметим, что данная теория, вообще говоря, отличается от классической теории синтаксиса. Получив предложение, система приписывает к нему синтаксическую структуру, которая состоит из множества помеченных связей (коннекторов), соединяющих пары слов.

Получаемые диаграммы, по сути, являются аналогами так называемых деревьев подчинения предложений. В деревьях подчинения от главного слова в предложении можно задать вопрос к второстепенному. Таким образом, слова выстраиваются в древовидную структуру. Главной причиной, по которой анализатор называют семантической системой, можно считать уникальный по полноте набор связей. Имеется около 100 основных связей, при этом некоторые из них дополнительно имеют 3–4 варианта.

Наши исследования проводятся в соответствии со следующими пунктами.

1. Выборка наиболее важных понятий, относящихся к времени и пространству, из толкового словаря С. И. Ожегова.

2. Создание массива перефразированных вариантов различных предложений и методов оценивания их похожести.

3. Анализ словарных статей (толкований) и примеров использования в художественной литературе соответствующих понятий из словаря С. И. Ожегова с помощью программных систем Link Grammar Parser и ДИАЛИНГ.

4. Анализ диаграмм, полученных на выходе систем Link Grammar Parser и ДИАЛИНГ средствами математической логики. Исследование возможностей применения к географическим понятиям.

5. Интеграция результатов исследований в базу знаний.

3. Программная система Link Grammar Parser

Link Grammar Parser – это синтаксический анализатор английского языка, разработанный в университете Карнеги – Меллона. Как уже было отмечено выше, она базируется на неклассической теории синтаксиса. Link Grammar Parser имеет словари, включающие около 60 000 словарных форм. Он позволяет анализировать большое число синтаксических конструкций, включая многочисленные редкие выражения и идиомы. Анализатор довольно устойчив; может пропустить часть предложения, непонятную ему, и определить структуру оставшейся части предложения. Он способен делать разумные предположения о синтаксической категории неизвестных ему слов (т.е. слов, которые отсутствуют в словарях) из контекста и написания. У анализатора есть данные об именах собственных, о числовых выражениях и разнообразных знаках препинания.

Для каждого слова в словаре записывается, какими коннекторами оно может быть связано с другими словами предложения. Коннектор состоит из имени типа связи, в которую может вступать рассматриваемая единица анализа. Например, пометка S соответствует связи между субъектом и предикатом, O – между объектом и предикатом. Только основных, наиболее важных связей, имеется более ста. Для обозначения направления связи справа к коннектору присоединяется знак «+», слева – знак «–». Левонаправленный и правонаправленный коннекторы одного типа образуют связь (link).

Например, если слову W1 приписан коннектор A+, а слову W2 – коннектор A–, то в синтаксической структуре предложения, состоящего из двух слов W1 W2, будет проведена связь A между словами W1 и W2. Предложение же W2 W1 не получит никакой интерпретации, поскольку W2 приписан коннектор A–, который образует связь только влево, а слову W1 приписан A+, который образует связь только вправо. Отметим, что может быть несколько вариантов разбора одного и того же предложения. На рис. 1 приведены 4 варианта разбора предложения «Вблизи ветер дул сильнее».

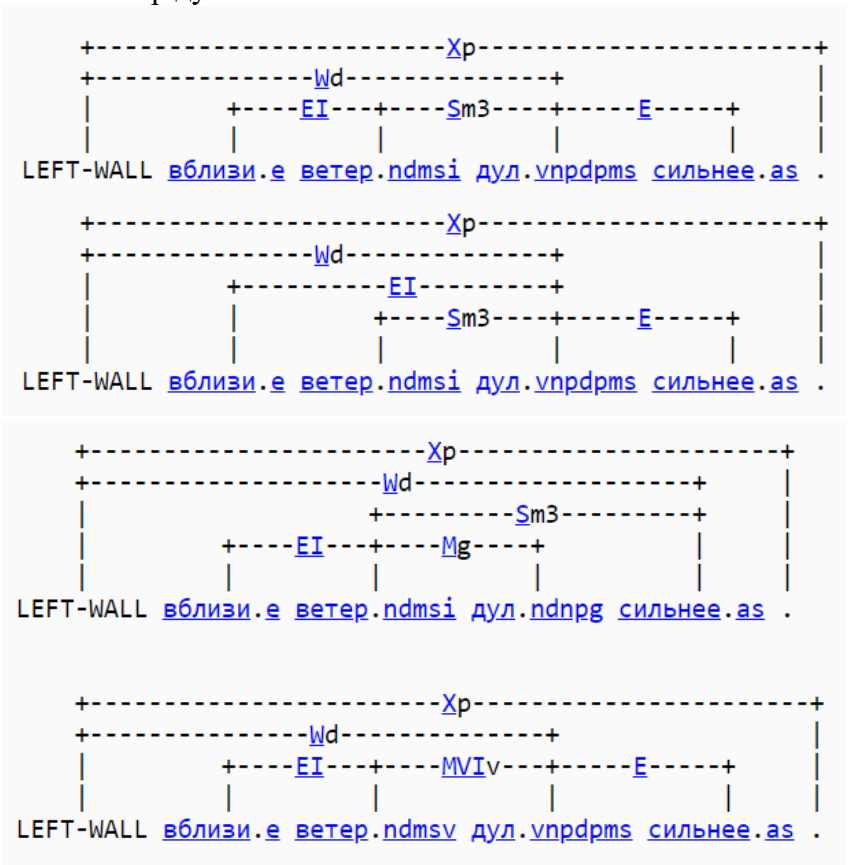


Рис. 1. Варианты разбора предложения при помощи Link Grammar Parser

Здесь приняты следующие обозначения:

Xp – связь для соединения начала предложения с точкой (в конце предложения).

Wd – связь указывает на вершину предложения.

EI – связь с наречием, обозначающим пространственное понятие.

Sm3 – соединяет подлежащее с глаголом (строчные буквы дополняют информацию о роде и лице).

E – связь между глаголом и наречием, модифицирующим его.

Mg – связь между существительным и причастием.

MVIv – соединяет сказуемое с дополнением.

Мы видим, что связь Xp присутствует во всех схемах. Можно сказать, что она инвариантна. Слово *вблизи*, относящееся к пространственным понятиям, считаем константой в нашей сигнатуре. Относительно других не инвариантных связей с помощью логических формул можно записать имеющиеся эквивалентности.

$$x_1 = \textit{ветер},$$
$$x_2 = \textit{дул},$$
$$\varphi_1 = EI(\textit{вблизи}, x_1) \wedge Sm3(x_1, x_2),$$
$$\varphi_2 = EI(\textit{вблизи}, x_2) \wedge Sm3(x_1, x_2),$$
$$\varphi_3 = EI(\textit{вблизи}, x_1) \wedge Mg(x_1, x_2),$$
$$\varphi_4 = EI(\textit{вблизи}, x_1) \wedge MVIv(x_1, x_2),$$
$$\varphi_1 \leftrightarrow \varphi_2 \leftrightarrow \varphi_3 \leftrightarrow \varphi_4.$$

4. Перефразирования предложений

Известно, что естественные языки обладают многообразием способов выражения смысла, одну и ту же мысль можно передать разными словами. Эта особенность значительно затрудняет анализ текстов на естественном языке, так как задача обнаружения похожего смысла в разных высказываниях является трудно формализуемой. Формализация анализа временных и пространственных понятий в текстах позволяет хотя бы частично приблизиться к решению этой задачи. В данном разделе формально описан процесс обнаружения перефразированных предложений, содержащих некоторые пространственные и временные понятия.

В ряде случаев в базе данных целесообразно иметь перефразированные варианты предложений и методы оценивания их близости или, иначе говоря, похожести. Это особенно важно для предложений, содержащих глаголы каузации движения и изменения положения в пространстве, ввиду сложности предложений, их содержащих. В этот класс входят глаголы, выражающие такие понятия, как: перемещать, приближать, удалять, нести, давать, брать, класть, поднимать, опускать, бросать, ловить, посылать и др. Определенный интерес представляют предложения, содержащие некоторые наречия: впереди, позади, сбоку, раньше, позже, еще, уже; именно с точки зрения их перефразирования.

В [4] дано описание, каким образом можно сравнивать перефразированные предложения, для случая использования анализатора Link Grammar Parser. Предположим, что L – множество слов некоторого естественного языка. Для любого слова $x \in L$ обозначим $Norm(x)$ его нормализованную форму. Запись $Syn(x, y)$ обозначает, что x, y – синонимы.

Возникают два вида эквивалентностей:

1) $x_1 \approx x_2 \leftrightarrow x_1 = x_2 \vee Syn(x_1, x_2)$;

2) $x_1 \equiv x_2 \leftrightarrow Norm(x_1) = Norm(x_2)$.

Предложение рассматриваем как вектор с компонентами из слов $\bar{x} =< x_1,\ldots,x_n >$. Функция $Norm$ может быть естественно распространена на предложения $Norm(\bar{x}) =< Norm(x_1),\ldots,Norm(x_n) >$. Текст $T = < \bar{x}_1,\ldots,\bar{x}_n >$ есть последовательность предложений.

Пусть запись $\bar{x} \models P(x_i, x_j)$ обозначает, что в схеме разбора предложения $\bar{x} =< x_1,\ldots,x_n >$ посредством анализатора Link Grammar Parser имеется коннектор типа P, идущий от слова x_i к слову x_j. Знак $\models$ означает, что фактически мы имеем дело с моделью. Основным множеством модели является множество пар $\{<1, x_1 >,\ldots,< n, x_n >\}$. Так как одно и то же слово может входить в предложение два и более раз, то это приводит к необходимости рассмотрения именно пар, а не отдельных слов. В силу сказанного выше корректным является даже обозначение $\bar{x} \models \varphi$, где φ – формула, например, исчисления предикатов первого порядка. Фактически $\bar{x}$ одновременно является обозначением и для вектора, и для модели.

Предположим, что даны два предложения $\bar{x} = < x_1,\ldots,x_n >$, $\bar{y} =< y_1,\ldots,y_m >$. Интерес представляют функции f, такие, что $dom(f) \subseteq \{1,\ldots,n\}$, $range(f) \subseteq \{1,\ldots,m\}$ с дополнительными свойствами типа: $f(i) = j \to x_i \approx y_j$, $f(i) = j \to x_i \equiv y_j$ и другие подобные им.

При сопоставлении двух предложений, точнее при анализе их на близость, осуществляется проверка ряда логических свойств. Например, пусть $f(i_1) = j_1$, $f(i_2) = j_2$. Теперь приведем примеры такого рода свойств.

Инвариантность коннектора:

$\bar{x} \models P(x_{i_1}, x_{i_2}) \to \bar{y} \models P(y_{j_1}, y_{j_2})$.

Замена коннектора на дизъюнкцию других:

$\bar{x} \models P(x_{i_1}, x_{i_2}) \to \bar{y} \models \bigvee_t Q_t(y_{j_1}, y_{j_2})$.

Расщепление коннектора на два коннектора:

$\bar{x} \models P(x_{i_1}, x_{i_2}) \to \exists k\, (\bar{y} \models Q(y_{j_1}, y_k) \wedge R(y_k, y_{j_2}))$.

Расщепление коннектора на два коннектора с инверсией:

$\bar{x} \models P(x_{i_1}, x_{i_2}) \to \exists k\, (\bar{y} \models Q(y_{j_2}, y_k) \wedge R(y_k, y_{j_1}))$.

Принимая во внимание, что $\bar{y}$ является обозначением для соответствующей модели, формула из третьего пункта может быть переписана в виде $\bar{x} \models P(x_{i_1}, x_{i_2}) \to \bar{y} \models \exists y Q(y_{j_1}, y) \wedge R(y, y_{j_2})$. По аналогии может быть записана формула из четвертого пункта.

Резюмируя, можно сказать, что в нашем распоряжении имеются правила вида

$R_i : \bar{x} \models \varphi_i(x_1, x_2) \to \bar{y} \models \psi_i(y_1, y_2)$.

Отметим, что для английского языка такого рода правил нами зафиксировано более тридцати. Для русского и других языков этот вопрос менее изучен. Для системы ДИАЛИНГ тоже могут быть сформулированы аналогичные правила, но это более сложный вопрос.

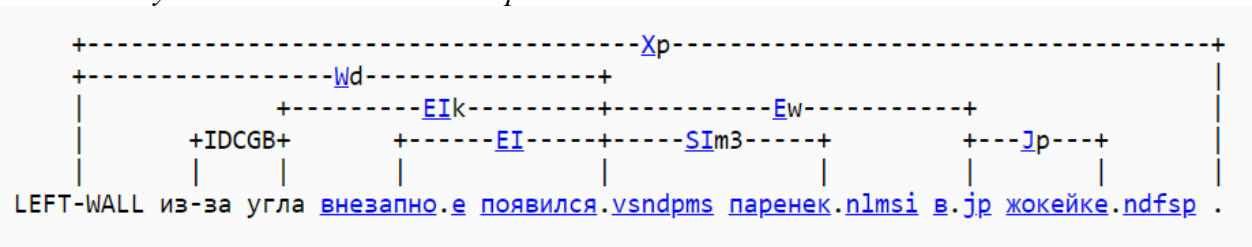
Далее строится функция f и проводится анализ по поводу выяснения того, встречаются ли индексы i_1, i_2, $j_1 = f(i_1)$, $j_2 = f(i_2)$, такие, что на конкретных словах из предложений $\bar{x}, \bar{y}$ выполнено правило R_i, т.е. $\bar{x} \models \varphi_i(x_{i_1}, x_{i_2}) \to \bar{y} \models \psi_i(y_{j_1}, y_{j_2})$. Для простоты можно говорить, что правило выполняется на паре $< i_1, i_2 >$.

Теперь рассмотрим множество всех таких пар $< i_1, i_2 >$, на которых выполнено одно из правил. Обозначим это множество I, и пусть его мощность $|I| = n$. Отметим, что анализатор Link Grammar Parser допускает между двумя словами наличие только одного коннектора. Поэтому будет выполняться не более чем одно правило.

Далее пусть n_1, n_2 – количество коннекторов, получающихся в результате анализа предложений $\bar{x}, \bar{y}$ соответственно. В качестве меры похожести двух предложений можно ввести $\mu_0(\bar{x}, \bar{y}) = n / \max(n_1, n_2)$ или $\mu_1(\bar{x}, \bar{y}) = 2n/(n_1 + n_2)$.

Известно, что предлог – это служебная часть речи, выражающая синтаксические отношения между именем существительным, местоимением, числительным и словами других частей речи, а также между существительными. Часть предлогов совмещают ряд значений. Так, предлоги *за, под, из, от, в, на* совмещают причинные, пространственные и временные значения. Предлог *через*, выражая пространственные (*через горы*) и временные (*через века*) отношения, в просторечии встречается при выражении причинных отношений (*через тебя я лишился семьи*). Другие предлоги совмещают причинные значения со значениями цели (например, *для, по*). На рис. 2 приведен пример перефразирования довольно простого предложения, содержащего предлог *из-за*.

1. *Из-за угла внезапно появился паренёк в жокейке*.



2. *Внезапно появившийся из-за угла паренёк был в жокейке*.

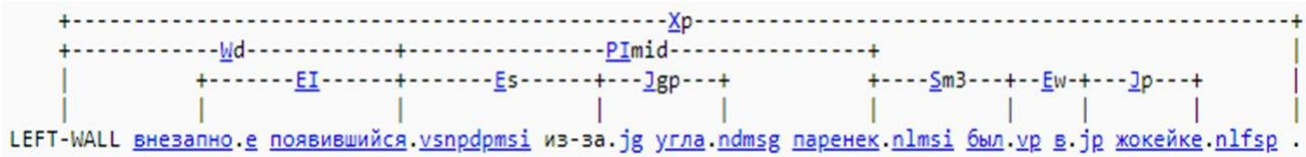


Рис. 2. Пример перефразирования предложения с предлогом *из-за*

Здесь приняты следующие обозначения в дополнение к тем, которые имеются на рис. 1:

Ew – соединяет глагол с предлогом.

J – соединяет предлог с существительным (строчные буквы дополняют информацию о падеже).

PImid – связь существительного с пассивным причастием прошедшего времени.

Очевидно, что здесь имеются инвариантные коннекторы, например: Xp, Jp, EI. Таковым можно считать даже Wd. Слова *появился*. *появившийся* считаем эквивалентными. Далее, слова *из-за, внезапно, появился*, относящиеся к пространственным и временным понятиям, полагаем константами в нашей сигнатуре. Логические формулы, приведенные ниже, позволяют описать перефразирования.

$$x_1 = угол,$$

$$x_2 = паренёк,$$

$$\varphi_{11} = IDBCP(из-за, x_1) \wedge EIk(x_1, появился) \wedge EI(внезапно, появился),$$

$$\varphi_{12} = S\,\mathrm{Im}3(появился, x_2),$$

$$\varphi_1 = \varphi_{11} \wedge \varphi_{12},$$

$$\varphi_{21} = EI(внезапно, появившийся) \wedge Es(появившийся, из-за) \wedge Jgp(из-за, x_1),$$

$$\varphi_{22} = \mathrm{P}\,\mathrm{Im}id(появившийся, x_2),$$

$$\varphi_2 = \varphi_{21} \wedge \varphi_{22},$$

$$\varphi_1 \leftrightarrow \varphi_2.$$

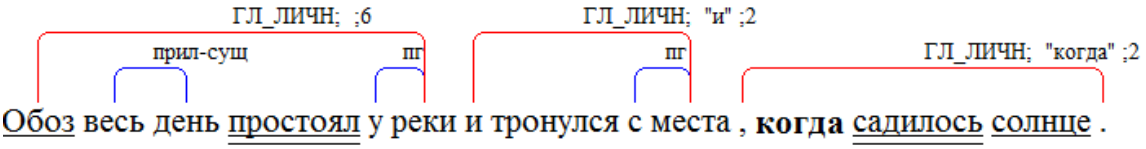
5. Система ДИАЛИНГ

Идеи, рассмотренные в предыдущих разделах, могут быть применены для анализа перефразированных вариантов предложений, а именно, оценивания их близости, также на основе диаграмм, получаемых на выходе системы ДИАЛИНГ. Система ДИАЛИНГ разрабатывалась как система русско-английского перевода с 1999 г. по 2002 г на базе ООО Диалинг (г. Москва) [2]. В разное время в работе над системой принимали участие более 20 специалистов, большинство из которых – известные ученые-лингвисты. Как и все современные системы обработки текста, ДИАЛИНГ включает в себя все основные этапы анализа текста.

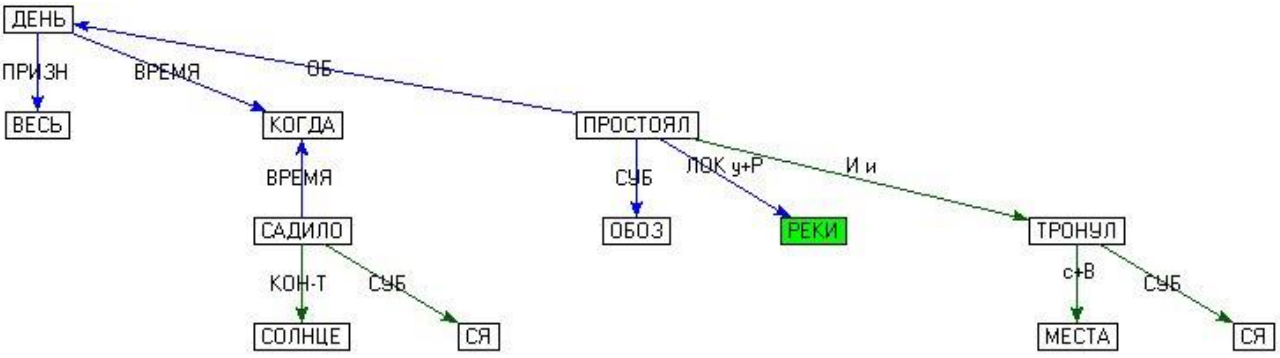
За основу системы автоматического русско-английского перевода ДИАЛИНГ была взята система французско-русского автоматического перевода (ФРАП), разработанная в ВЦП совместно с МГПИИЯ им. М. Тореза в 1976–1986 гг., и система анализа политических текстов на русском языке ПОЛИТЕКСТ, разработанная в центре информационных исследований в 1991–1997 гг. Ниже приведены примеры синтаксического и семантического анализа предложений, являющихся перефразированиями.

1. *Обоз весь день простоял у реки и тронулся с места, когда садилось солнце*.

1.1. Результат синтаксического анализа.



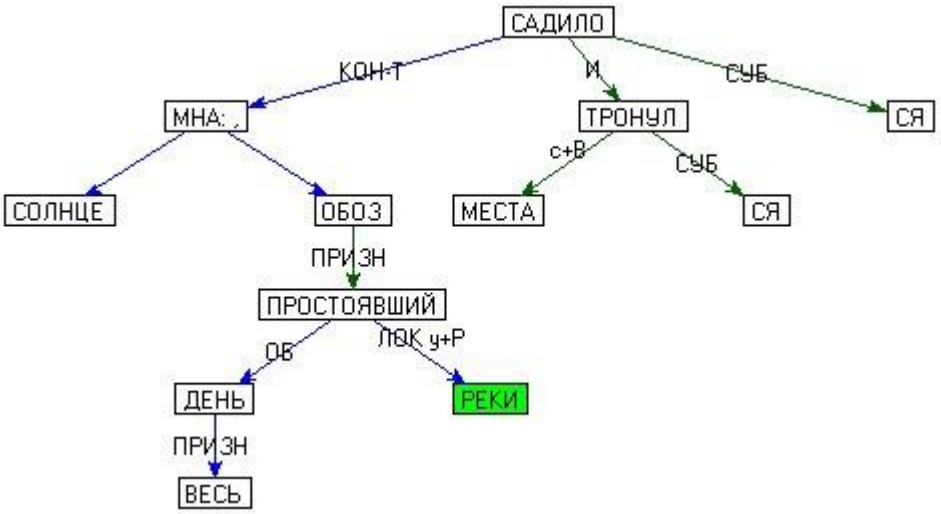
1.2. Результат семантического анализа.



2. *Садилось солнце, обоз, простоявший весь день у реки, тронулся с места*.

2.1. Результат синтаксического анализа.

2.2. Результат семантического анализа.



6. Заключение

В работе описан процесс создания базы знаний, содержащей информацию о временных и пространственных понятиях, встречающихся в текстах на естественном языке. Формально описан процесс обнаружения перефразированных предложений, содержащих некоторые пространственные и временные понятия. В настоящее время ведется работа по наполнению базы знаний. В результате проведенного исследования был сделан вывод о целесообразности применения таких инструментов, как Link Grammar Parser и ДИАЛИНГ, для решения поставленной задачи.

Link Grammar Parser – достаточно необычная система, главной причиной, по которой анализатор называют семантической системой, можно считать уникальный по полноте набор связей. Например, выделяются следующие классы наречий: ситуационные наречия, которые относятся ко всему предложению в целом (clausal adverb); наречия времени (time adverbs); вводные наречия, которые стоят в начале предложения и отделены запятой (openers); наречия, модифицирующие прилагательные, и т.д. Из достоинств системы необходимо отметить, что организация самой процедуры нахождения вариантов синтаксического представления очень эффективна.

Отметим также отрицательные моменты. Практическое тестирование системы показывает, что при анализе сложных предложений, длина которых превышает 25–30 слов, возможен комбинаторный взрыв, и результатом работы анализатора становится «панический» граф – как правило, случайный вариант синтаксической структуры, с лингвистической точки зрения неадекватной. Использование системы затруднено для флективных языков типа русского ввиду значительно возрастающего объема словарей, которые возникают в силу морфологической развитости флективных языков.

Для случая английского языка было проведено тестирование [4]. Основная задача, которая решалась – это определение релевантности текста поисковому запросу. В этом случае был обнаружен минимальный вариант, дававший достаточно хорошие результаты, когда учитывались всего 8 связей. Для английского языка был составлен перечень наиболее важных связей системы Link Grammar Parser. Он содержит 35 связей. Обнаружены связи, которые существенно портили ситуацию в случае неполных предложений, – всего 8 таких связей. Для русского языка такой детальный анализ связей не проведен. Возможности системы ДИАЛИНГ еще менее изучены.

Тем не менее, следует отметить, что диаграммы, полученные на выходе систем Link Grammar Parser и ДИАЛИНГ, представляют собой чрезвычайно интересный материал для

дальнейших исследований, в том числе для исследования временных и пространственных понятий как с точки зрения приложений, так и для теоретической лингвистики.

Литература

1. *Мельчук И. А.* Опыт теории лингвистических моделей типа «Смысл – Текст». М.: Наука, 1974. 315 с.
2. *Сокирко А. В.* Семантические словари в автоматической обработке текста // Канд. дисс., МГПИИЯ, Москва. 2000. 108 с.
3. Link Grammar Documentation, 2015, http://www.abisource.com/projects/link-grammar
4. *Батура Т. В., Бакиева А. М., Еримбетова А. С. И др.* Грамматика связей, релевантность и определение тем текстов: монография. Новосибирск: Изд-во СО РАН. 2018. 91 с.

Статья поступила в редакцию 27.08.2019;
переработанный вариант – 27.09.2019.

Батура Татьяна Викторовна
к.ф.-м.н., с.н.с., Институт систем информатики им. А. П. Ершова СО РАН (630090, Новосибирск, просп. Академика Лаврентьева 6), e-mail: tbatura@iis.nsk.su.

Ефимова Любовь Валерьевна
аспирант, Институт систем информатики им. А. П. Ершова СО РАН.

Еримбетова Айгерим Сембековна
аспирант, Новосибирский государственный университет, e-mail: aigerian@mail.ru.

Касекеева Айсулу Бесеновна
аспирант, Евразийский национальный университет им. Л. Н. Гумилева, Казахстан.

Мурзин Федор Александрович
к.ф.-м.н., зам. директора по научной работе, Институт систем информатики им. А. П. Ершова СО РАН, e-mail: murzin@iis.nsk.su.

Temporal and spatial concepts in natural language texts and their investigation

T. V. Batura, L. V. Efimova, A. S. Yerimbetova, A. B. Kasekeeva, F. A. Murzin

The aim of the work is to create a knowledge base containing information on temporal and spatial concepts found in natural language texts. The content of the base: the most important concepts related to time and space, from the S. I. Ozhegov explanatory dictionary; rephrased sentences; the results of the analysis of vocabulary articles (interpretations) and examples from a literature by means of Link Grammar Parser and DIALING software systems, etc. Results of the work can be used in intelligent information retrieval systems. The diagrams obtained at the output of Link Grammar Parser and DIALING systems are extremely interesting material for further research. It is advisable to investigate the possibility of using a number of constructions and concepts of mathematical logic in computer linguistics, such as: the construction of L. Henkin, realizability and omitting of types, model completeness, forcing, as well as a number of non-classical logics.

Keywords: computational linguistics, semantics, temporal and spatial concepts, Link Grammar Parser, DIALING.