OSTIS-2013
(Open Semantic Technologies for Intelligent Systems)

УДК 004.82

ИДЕНТИФИКАЦИЯ ПРОСТРАНСТВЕННО-ВРЕМЕННЫХ СВЯЗЕЙ МЕЖДУ ВЫСКАЗЫВАНИЯМИ В ЗАДАЧАХ СЕМАНТИЧЕСКОГО АНАЛИЗА ТЕКСТА

Дмитриев А.С., Заболеева-Зотова А.В.

Волгоградский государственный технический университет,
г. Волгоград, Российская Федерация

dmitriealeksan@yandex.ru

zabzot@gmail.com

В работе представлена методика интеллектуального анализа, в основе которой лежит обработка пространственной и темпоральной информации в тексте на естественном языке. Описывается применение пространственно-темпоральной логики для идентификации связи между событиями в тексте на естественном языке. Идентификация связи между событиями в тексте производится при помощи методов машинного обучения, в частности логико-марковских сетей.

Ключевые слова: семантический анализ; коммуникативная грамматика; пространственные отношения; темпоральные отношения; комбинированная логика; машинное обучение.

Введение

Общеизвестно, что в последнее время существует большое количество попыток (и иногда довольно успешных) по созданию современных интеллектуальных систем семантического анализа текстов и речи на естественных языках. Существующие системы с разным успехом справляются с поставленными перед ними задачами, но их общий и главный недостаток состоит в том, что семантика текста строится в основном по конечным высказываниям и предложениям. При этом смысл всего текста рассматривается крайне редко. Такой подход чреват сильным искажением описанной информации в тексте, если, например, какой-либо факт расписан в нескольких последовательных или разбросанных по тексту предложениях.

Естественно, для решения данной проблемы необходимо использовать методику построения сюжетных линий, или по-другому, цепочки событий. Именно события в тексте играют в связке различных предложений ключевую роль. Как правило, события описываются глаголами или отглагольными образованиями. Определив все события текста, используя метод анализа текста на основе коммуникативной грамматики, можно попытаться выстроить между ними причинно-следственную связь, которая позволит достаточно точно описать семантику всего текста, а не отдельно взятых предложений.

Исходя из вышесказанного, было принято решение о разработке системы, способной распознавать взаимосвязи между событиями в тексте, используя пространственно-временные отношениями между данными событиями для повышения семантической связности естественно-языкового (ЕЯ) текста.

1. Способы выражения пространственных и темпоральных отношений в тексте на ЕЯ

1.1. Выражение пространственных отношений

Согласно работе Всеволодовой [Всеволодова и др., 2008], определение пространственных отношений в тексте производится на основе так называемых оппозиций. На первом уровне разбиения всех существующих в русском языке именных локативных групп наиболее важным является отношение локализуемого предмета к локуму, которое основано на нахождении предмета в какой-либо момент времени в пределах локума. В том случае, если предмет в любой момент времени есть, был или будет в пределах локума, то можно говорить о семе сопространственности: «в поля, в поле, из поля, полем, через поле, по полю». Иначе,

если локализуемый предмет не находится, или не будет находиться в пределах локума, то можно говорить об отношении несопространственности: «около поля, у поля, в двух километрах от поля, к полю, мимо поля».

Остальные уровни описания пространственных отношений в русском языке представлены на рисунках 1 и 2.

Рис. 1. Описание отношения сопространственности

Рис. 2. Описание отношения несопространственности в русском языке.

1.2. Выражение темпоральных отношений

В построении временного порядка используется множество грамматических категорий. К ним относятся видо-временные формы глаголов, наречия времени, лексико-семантическая информация и представление о познаваемом мире.

Большинство событий описываются через конструкцию глаголов. Известно, что видо-временные формы глаголов налагают ограничения на временной порядок событий (прошедшее, настоящее, будущее, совершенный и несовершенный вид).

Для событий, описывающихся не через конструкцию глаголов, т.е. с помощью имени существительного, эти параметры не учитываются.

Кроме этих параметров, события могут быть классифицированы различными аспектуальными классами (из которых основными являются состояния, процессы, моментальные события, и события в развитии), способами глагольного действия.

Также они могут быть классифицированы модальностями (т.е. возможность, вероятность или обязанность выполнения действия) – «На этой неделе можно ожидать подъема котировок основных акций».

Полярностями (т.е. положительность или отрицательность выполнения действия) – «Пока не будет остановлен огонь, ни о каком перемирии речи идти не может».

В таблице 1 представлены 4 способа временного поведения глагола в предложениях.

Для связи двух событий используются временные союзы. Они часто появляется в сложных предложениях, и описывают отношения между частями. Например: перед тем как, после, во время, с тех пор, когда, пока, …

Кроме временных союзов, существуют ещё другие лексические единицы, которые влияют на временной порядок событий. Это временные предлоги, наречие, местоимение и частицы (таблица 2).

Таблица 1 – Способы глагольного действия

Действия	Значение	Приставки или суффиксы к глаголам
Начинательный	Обозначает начало некоторого действия	по- (пойти, полететь, поскакать)
Пердуративный	Обозначает действие, которое заполняет некоторый промежуток времени	про- (проработал, проговорил, просидеть, пролежать, простоять,…)
Финитивный (завершенный)	Обозначает прекращение некоторой деятельности	от- (Отговорил; Отцвели)
Кумулятивный	Обозначает «накопление результата» действия	на- (накупить, натворить, накопить ,…)

Таблица 2 – Временные индикаторы

Предлог	с, до, по, перед, под, при, спустя,…
Наречие	сначала, раньше, отныне, скоро, потом
Местоимение	который, какой, кто, что, чей
Частицы	еще, уже, и, а, …
Другие	как только, когда-то, чтобы,…

2. Коммуникативная грамматика с использованием категорий пространства и времени

В настоящей работе для обработки текстов на естественном языке с целью выявления и извлечения пространственных и темпоральных данных используется коммуникативная грамматика русского языка, разработанная Золотовой Г.А.

Основная идея данной грамматики состоит в том, что синтаксис и семантика тесно взаимосвязаны в рамках анализа смысла предложений.

Анализ текста с использованием коммуникативной грамматики русского языка состоит из следующих этапов:

- Лемматизация
- Морфологический анализ
- Синтаксический анализ
- Семантический анализ (с использованием семантического словаря и лингвистических правил)

Главная характерная особенность синтаксического анализа состоит в том, что для дальнейшего семантического анализа составляются все допустимые деревья синтаксических зависимостей рассматриваемого высказывания. В дальнейшем на этапе семантического анализа производится отбор наиболее подходящего по смыслу дерева зависимостей.

Основным термином коммуникативной грамматики является синтаксема. Она представляет собой слово или словосочетание, значение которого определяется в зависимости от категориального значения слова и морфологической формы, которые в свою очередь реализуются в определенной синтаксической позиции. Смысл предложения (высказывания) определяется совокупностью значений входящих в него синтаксем и отношений между ними.

Для поиска значений синтаксем и отношений между ними используется реляционно-ситуационный анализ. В основе его работы лежит поиск предикатных слов (глаголов и отглагольных образований) на основе семантического словаря, после чего производится поиск значений синтаксем путем сопоставления предикатного слова и составленных синтаксическим анализатором деревьям зависимости, с применением семантического словаря со словарными статьями по синтаксемам и лингвистических правил построения синтаксем. Лингвистические правила и семантический словарь составляются экспертами-лингвистами.

Для реализации реляционно-ситуационного анализа используется предложенная Осиповым Г.С. [Осипов и др., 2008] и др. в интеллектуальной поисковой системе Exactus неоднородная семантическая сеть с расширенным семейством отношений. Вершинами этой сети являются значения синтаксем, ребрами – отношения между синтаксемами.

После установки значения синтаксем, определяются отношения, в которые входят те или иные пары синтаксем. При этом происходит замыкание – отношения в семантической сети дополняются новыми связями.

У описанной методики анализа естественно-языкового текста, несмотря на то, что анализ на основе коммуникативной грамматики считается одним из передовых в семантическом анализе, есть существенный недостаток. Этот анализ способен выявить семантику в рамках одного предложения или высказывания. Провести анализ и составить семантическую модель всего текста данная методика не может.

Исходя из этого, было принято решение расширить анализ естественных языков на основе коммуникативной грамматики дополнительными категориями событийности, которые выражаются на всем протяжении текста. Это позволит создавать более полные семантические модели текстов, где раскрывается смысл не каждого отдельно взятого предложения, а, по возможности, большей части анализируемого текста.

Для установления связи между предложениями используются пространственно-темпоральные отношения между предикатными словами и синтаксемами в предложениях. Структурно это представляет собой расширение неоднородной семантической сети.

Например: «Вася опаздывал в школу. Он срезал путь через дворы» (Рисунок 3).

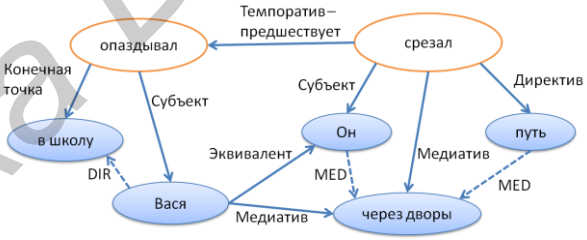

Рис. 3. Пример пространственно-темпоральной связи между семантическими моделями двух предложений

На рисунке сплошными линиями показаны значения синтаксем, а штриховыми отношения между синтаксемами. Для установки семантической взаимосвязи между двумя предложениями используется категория событийности. В тексте повествование идет, как правило, или последовательно, или колеблется относительно определенной сюжетной линии. Соответственно, в тексте прослеживается цепочка событий, попеременно сменяющих друг друга. Для выявления событийности используется изменение пространственных и временных данных, связанных с указанными событиями.

В отличие от построения семантики одного высказывания определение взаимосвязи событий между разными предложениями является более сложной задачей. Воспользуемся для решения данной проблемы механизмом машинного обучения на основе логико-марковских сетей. Для составления событийной цепочки необходимо выявить темпоральные и пространственные характеристики из текста, которые извлекаются с помощью описанного выше метода на основе коммуникативной грамматики, и затем, оперируя

разработанной в рамках данной работы пространственно-темпоральной логикой и методами классификации на основе логико-марковских сетей, выявить принадлежность события к той или иной цепочке событий.

3. Комбинированная пространственно-темпоральная логика

Для определения смысла конечных предложений используется неоднородная семантическая сеть, выявляющая события в рамках одного предложения. Получив семантическую сеть, мы получаем большой набор логических правил, среди которых содержатся правила, описывающие пространственную и временную информацию. Но эта информация довольно слабо связана между собой, поэтому целесообразно использовать комбинированную логику темпоральных и пространственных отношений, для того, чтобы описываемые в предложениях события представляли собой максимально четкую модель в пространстве и времени.

Кроме того, использование комбинированной пространственно-временной (или по-другому топо-темпоральной) логики снижает потребление вычислительных ресурсов. Использование каждой логики в отдельности с дальнейшими попытками связать результат их работы с событиями требует существенных временных затрат. Так или иначе, задача построения эффективной и достаточно полной топо-темпоральной логики требует тщательного исследования, поскольку разные ее комбинации относятся к принципиально разным классам сложности.

3.1. Пространственная логика RCC-8

За основу анализа пространственных отношений взята топологическая логика. Логика топологических пространств в настоящий момент является одним из самых успешных подходов в описании пространственных отношений в искусственном интеллекте. Но до сих пор не было составлено эффективной модели взаимодействия пространственных и временных отношений, поскольку обычное сложение пространственных и временных логик не дает желаемого результата. Связано это в первую очередь с проблемами достижимости и сохранения условий динамических систем. Наиболее полное исследование, посвященное комбинированию пространственных и темпоральных логик, проведено в исследованиях David Gabelaia, Roman Kontchakov, Agi Kurucz, Frank Wolter, Michael Zakharyaschev и др.

В работе используется пространственная логика, основанная на пропозициональной логике, в которой унарные предикаты обозначают пространственные объекты, а топологические отношения между ними представляются с помощью внутренних операторов и операторов замыкания, кванторами общности и существования в пространстве и обычными логическими операциями. Данная логика называется модальной логикой и рассматривается как логика топологических пространств. Обозначается как $S4_u$.

Пространственные термы этой логики представляют собой выражения следующего вида:

$$\tau = p_i \mid \bar{\tau} \mid \tau_1 \cap \tau_2 \mid \tau_1 \cup \tau_2 \mid I\tau \mid C\tau\,, \qquad (1)$$

где pi пространственные переменные, I и C – операторы включения и замыкания.

Топологическая модель представляет собой структуру следующего вида:

$$Mod = (P, p_0^{Mod}, p_1^{Mod}, ...)\,, \qquad (2)$$

где P=(U,I) – топологическое пространство, а $p_i^{Mod} \subseteq U$ для любого i.

В работе для описания приведения к одной размерности описываемых пространственных объектов используется понятие регулярных замкнутых множеств (или просто «Регионы»).

Для описания регионов используется язык RCC-8. Синтаксис RCC-8 состоит из переменных, обозначающих регионы r, s, ... и восьми бинарных предикатов:

- DC(r,s) регионы r и s не связаны;
- EC(r,s) r и s внешне связаны;
- и др.

Каждый из этих операторов может использоваться совместно с логическими операциями.

Аргументы RCC-8 предикаты - региональные переменные, интерпретируемые как регулярные замкнутые множества (регионы) топологических пространств. Расширить RCC-8 можно путем представления ее фрагментом логики $S4_u$ (Например):

$$DC(r,s) = (p_r \cap p_s = \varnothing) \qquad (3)$$

и др.

3.2. Темпоральная логика LTL

Временная логика в работе представлена линейной темпоральной логикой LTL. Темпоральная логика является подходом к рассуждению о времени, используя темпоральные связки без явного указания количества времени. Самый популярный вариант этой логики это LTL, которая успешно используется в тестировании и верификации программного обеспечения.

Размеченное течение времени для LTL является любой строгой линейной последовательностью (W,<), с временными точками $w \in W$ и отношением предшествования <. LTL формулы построены из пропозициональных переменных p_0, p_1 используя логические операторы и темпоральный оператор U -

“пока”. Например x*U*y означает, что “x справедливо, пока имеет место быть y”. Другие темпоральные связки Rf - когда-то в будущем, Ff - всегда в будущем Nf - в следующий момент.

LTL-модель представляет собой структуру следующего вида:

$$Mod = (T, p_0^{Mod}, p_1^{Mod}, ...), \quad (4)$$

где $T=(W,<)$, а $p_i^{Mod} \subseteq U$ для любого i.

3.3. Комбинированная топо-темпоральная логика

Далее рассмотрим комбинирование пространственной и темпоральной логики. Передвижение пространственных объектов во времени представляется в виде модели “снимок экрана”, т.е. в каждый момент времени фиксируется текущее положение объекта. Топологико-темпоральная модель это пара Mod=(P,DT), где P=(U, I) - топологическое пространство, а DT это множество пространственных точек p в каждый момент времени $n \in N$.

Язык применяемой логики должен придерживаться нескольких следующих правил:

- Язык должен быть способным выразить изменения во времени значений истинности для простых пространственных суждений;
- Язык должен быть в состоянии выразить изменения или развитие пространственных объектов за несколько фиксированных конечных промежутков времени;
- Язык должен быть в состоянии выразить изменения или развитие пространственных объектов на всем протяжении времени.

В работе в целях повышения выразительности языка используется максимальный подход - комбинация всех 3 правил. Этот подход позволяет неограниченно использовать логические, топологические и темпоральные операторы для построения пространственно-темпоральных термов.

Для пространственно-темпорального языка LTL* RCC-8 вводится следующие термы и формулы [Gabelaia и др., 2005]:

$$\begin{aligned} \rho &= CIp \mid CI(\rho_1 U \rho_2), \\ \tau &= \rho \mid \bar{\tau} \mid \tau_1 \cap \tau_2 \mid \tau_1 \cup \tau_2 \mid I\rho, \\ \varphi &= Q(\rho_1, \rho_2) \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \varphi_1 \vee \varphi_2 \mid \varphi_1 U \varphi_2, \end{aligned} \quad (5)$$

где ρ - региональный терм, φ - формула комбинированной логики и Q - обозначает 8 предикатов RCC-8.

4. Идентификация пространственных и временных связок с использованием логико-марковских сетей

Идентификация временных и пространственных отношений производится посредством методики машинного обучения. Для этого система на вход получает два типа корпуса текстов. Первый корпус - аннотированный синтаксически размеченный корпус, второй - неразмеченный, по которому вручную расставляются метки с обозначением временных и пространственных отношений. Для обучения системы для выстраивания взаимосвязи между пространственно-временными категориями и событиями в тексте используется механизм логико-марковских сетей, позволяющий оперировать в своих вершинах обычными логическими формулами, представляющими комбинированную пространственно-временную логику, с последующим разбиением сети на атомарные выражения, используемые в обыкновенной Марковской сети. Марковская сеть выполняет задачу классификации по присваиванию тех или иных событий к определенной цепочке событий.

MLNs может ответить на запросы произвольной форме «Какова вероятность того, что формула F_1 верна с учетом того, что формула F_2 верна?» Если F_1 и F_2 являются двумя формулами в логике первого порядка и C является конечным набором констант, включая любые константы, которые появятся в F1 или F2 и L является MLNs, тогда:

$$\begin{aligned} P(F_1 \mid F_2, L, C) &= P(F_1 \mid F_2, M_{L,C}) \\ &= \frac{P(F_1 \wedge F_2 \mid M_{L,C})}{P(F_2 \mid M_{L,C})} \\ &= \frac{\sum_{x \in X_{F_1} \cap X_{F_2}} P(X = x \mid M_{L,C})}{\sum_{x \in X_{F_2}} P(X = x \mid M_{L,C})} \end{aligned} \quad (6)$$

где X_{Fi} является множество «миров» (интерпретаций) где выполняется формула F_i

Подставляя в сеть утверждения, которые необходимо проверить (например, принадлежит ли событие к данной цепочке событий) мы, опираясь на обученную сеть, можем получить с некоторой степенью достоверности ответ на принадлежность события к определенной цепочке событий.

Заключение

Описанная методика призвана повысить качество работы существующих семантических систем обработки текстов, в частности интеллектуального поиска, вопрос-ответных систем и др. Она также может использоваться в задачах семантического поиска, поскольку использование расширенных категорий событийности позволяет извлекать смысл не только из конкретных предложений, но также из связных текстов, в которых прослеживается четкая последовательность событий.

Работа частично поддержана Российским фондом фундаментальных исследований (проекты 12-07-00266, 12-07-00270).

Библиографический список

[Бердник, 2007] Бердник, В.Л. Семантический анализ высказываний идентификации сущности / Бердник В.Л., Заболеева-Зотова А.В. // Известия ВолгГТУ. Серия "Актуальные проблемы управления, вычислительной техники и информатики в технических системах": межвуз. сб. науч. ст. / ВолгГТУ. - Волгоград, 2007. - Вып.3, №9. - С. 43-46.

[Всеволодова и др., 2008] Всеволодова, М.В. Способы выражения пространственных отношений в современном русском языке / М.В Всеволодова, Е.Ю. Владимирский. – М.: Книжный дом «Либроком», 2009. – 288 с.

[Дмитриев, 2011] Дмитриев, А.С. Извлечение пространственно-временных отношений из текста на естественном языке / Дмитриев А.С. // Интегрированные модели и мягкие вычисления в искусственном интеллекте : сб. науч. тр. VI междунар. науч.-практ. конф. (Коломна, 16-19 мая 2011 г.). В 2 т. Т. 2 / Рос. ассоциация искусственного интеллекта [и др.]. - М., 2011. - С. 883-889.

[Кондрашина и др., 1989] Кондрашина, Е.Ю. Представление знаний о времени и пространстве в интеллектуальных системах. / Е.Ю.Кондрашина, Л.В.Литвинцева, Д.А.Поспелов; под ред. Д.А.Поспелова. – М.: Наука, 1989. – 328 с.

[Осипов и др., 2008] Осипов, Г.С. Реляционно-ситуационный метод поиска и анализа текстов и его приложения / Осипов Г.С., Тихомиров И.А., Смирнов И.В. // Искусственный интеллект и принятие решений / Институт системного анализа РАН. – М., 2008. - №2. – С. 3-10.

[Осипов, 1997] Осипов, Г.С. Приобретение знаний интеллектуальными системами/Г.С.Осипов -М., Наука Физматлит, 1997. – 109с.

[Фамхынг и др., 2008] Фамхынг, Д.К. Применение нечёткой нейронной сети к обработке временной информации в тексте на русском языке / Фамхынг Д.К., Захаров С.С. // AIS`08. CAD-2008. Интеллектуальные системы. Интеллектуальные САПР (пос. Дивноморское, 3-10 сент. 2008 г.): тр. междунар. науч.-техн. конференций / ФГОУ ВПО "Юж. федерал. ун-т" [и др.]. - М., 2008. - Т. 3. - С. 16-22.

[Gabelaia и др., 2005] Gabelaia, D. Combining Spatial and Temporal Logics: Expressiveness Vs. Complexity / David Gabelaia, Roman Kontchakov, Agi Kurucz, Frank Wolter, Michael Zakharyaschev // Journal of artificial intelligence research. – 2005. – P.167–243

IDENTIFICATION OF SPACE-TEMPORAL CONNECTIONS BETWEEN STATEMENTS IN PROBLEMS SEMANTIC ANALYSIS TEXT

Dmitriev A.S., Zaboleeva-Zotova A.V.

Volgograd, Russian Federation Volgograd State Technical University, Volgograd, Russia

dmitriealeksan@yandex.ru

zabzot@gmail.com

This paper presents a methodology mining, which is based on processing of spatial and temporal information in natural language text. Describes the use of spatial-temporal logic to identify the connection between the events in natural language text. Identifying relationships between events in the text produced by the methods of machine learning, in particular Markov-logic networks.

It is well known that in recent years there are a large number of times (and sometimes quite successful) to create a modern intelligent systems of semantic text analysis and natural language speech. Existing systems with varying degrees of success to cope with their tasks, but they are common and the main drawback is that the semantics of the text is based mainly on the final statements and proposals. In this case, the meaning of the text is not considered. This approach has strong distortion of information described in the text, for example, if a fact is painted in several successive or scattered through the text sentences.

Naturally, to solve this problem, you must use a method of constructing story lines, or in other words, the chain of events. That event in the text play in a bunch of different offers key role. Typically, the events described by verbs or verbal formations. Defining the events of text using text analysis method based on communicative grammar, you can try to build between cause and effect, which will quite accurately describe the semantics of the text, rather than individual sentences.

Based on the foregoing, it was decided to develop a system that can recognize the relationship between the events in the text, using spatio-temporal relationships between these events to enhance the semantic coherence of text in a natural language.

For this system to the input receives two types of corpus. The first building - annotated syntactically marked up body, the second - unmarked on which labels are placed manually with the designation of temporal and spatial relations. For teaching for building relationships between spatial and temporal categories and events in the text, the mechanism of Markov-logic networks, allowing the use of the tops in their usual logical formulas.

The described technique is designed to improve the performance of existing semantic text processing systems, in particular intelligent search, question-answering systems, etc. This system can be used in the problems of semantic search, as the use of advanced categories eventfulness can extract meaning not only of the specific proposals, but also from coherent texts in which there is a clear sequence of events.

This work was partially supported by the Russian Foundation for Basic Research (projects 12-07-00266, 12-07-00270).