Семантика текста как модель ситуации

Харламов А.А.,
доктор технических наук старший научный сотрудник Института высшей нервной деятельности и нейрофизиологии РАН, Москва, профессор Департамента программной инженерии НИУ ВШЭ, Москва, профессор Кафедры прикладной и экспериментальной лингвистики МГЛУ, Москва

Аннотация

Как правило, отдельный текст описывает отдельную ситуацию. Поэтому можно с достаточной степень уверенности соотнести описание ситуации с самой ситуацией. Далее, конкретный текст может быть представлен несколькими способами. Одним из них является формирование семантической сети текста, которая, теперь уже, будет представлять упомянутую ситуацию. Для того чтобы такое представление оказалось эффективным, необходимо автоматизировать процесс построения семантической сети. В работе представлен способ описания текста с помощью однородной семантической сети. Для начала покажем, что текст может быть представлен неоднородной семантической сетью. Не вызывает сомнений, что отдельные предложения текста могут быть представлены в графическом виде расширенными предикатными структурами. Если мы соберем все расширенные предикатные структуры текста воедино, отождествив соответствующие вершины, мы получим граф, представляющий смысл этого текста. Убрав разметку дуг неоднородной семантической сети, мы получим однородную (ассоциативную) семантическую сеть. Семантические сети представляют статику ситуации. Если же мы выйдем за рамки собственно семантической сети, то легко смоделируем и динамику ситуации тоже. Для этого нам надо построить не одну, а несколько семантических сетей, каждая на своем корпусе текстов, описывающем ситуацию позавчера, вчера, сегодня. Соединив их одноименные вершины, мы получим полное представление о динамике ситуации. Однородная семантическая сеть, сформированная автоматически, таким образом, может быть удобным инструментом для анализа ситуаций.

Ключевые слова: scемантическая сеть текста как модель ситуации, текст как модель ситуации, Расширенная предикатная структура предложения, семантическая структура текста, однородная (ассоциативная) семантическая сеть, тематическая структура текста, анализ динамики ситуации, инструментарий для автоматического анализа текста.

ВВЕДЕНИЕ

Как правило, отдельный текст описывает отдельную ситуацию. Исключения в виде учебников и монографий таковыми не являются: отдельные главы

представляют ровно такие описания ситуаций. Поэтому можно с достаточной степень уверенности соотнести описание ситуации с самой ситуацией [1].

Далее, конкретный текст может быть представлен несколькими способами [2]. Одним из них является формирование семантической сети текста [3], которая, теперь уже, будет представлять упомянутую ситуацию.

Для того чтобы такое представление оказалось эффективным, необходимо автоматизировать процесс построения семантической сети. В работе представлен способ описания текста с помощью однородной семантической сети, которая автоматически формируется из текста.

1. Неоднородная семантическая сеть

Для начала покажем, что текст может быть представлен неоднородной семантической сетью. Не вызывает сомнений, что отдельные предложения текста могут быть представлены в графическом виде расширенными предикатными структурами [4]:

$$P_i \cong < S, O_1, < O_{i_{i\neq 1}} >, < A_i \gg, \quad (1)$$

где S — субъект, O_1 — главный объект, $O_{i_{i\neq 1}}$ — второстепенные объекты, A_i — атрибуты. Предикат Р связан с субъектом предикативной связью, остальные актанты — валентными связями.

Перейдем от предикатоцентрической структуры (1) к субъектоцентрической [4], то есть предикат мы убираем из рассмотрения, актанты предиката мы присоединяем к субъекту: главный объект — предикативной связью, второстепенные объекты и атрибуты их валентными связями:

$$S_i \cong < O_1, < O_{i_{i\neq 1}} >, < A_i \gg, \quad (2)$$

Если мы объединим расширенные предикатные структуры всех предложений текста воедино, отождествив соответствующие вершины, мы получим граф, представляющий смысл этого текста. Перецентрированная расширенная предикатная структура является направленным графом (связи идут от субъекта к объектам и атрибутам), поэтому полученная отождествлением вершин неоднородная семантическая сеть N также будет направленным графом:

$$N \cong \{S_i\} = \{< c_i r_{1_i} c_{1_i}, < c_i r_{j_{i\neq 1}} c_{j_{i\neq 1}} \gg\}, \quad (3)$$

где S_i — расширенная предикатная структура, приведенная к субъектоцентрическому представлению, c_i — субъект i-й предикатной структуры, r_{1_i} предикативное отношение, c_{1_i} — главный объект, $r_{j_{i\neq 1}}$ — валентные отношения, $c_{j_{i\neq 1}}$ — второстепенные объекты и атрибуты.

2. Разметка связей сети

Разметка отношений в графе расширенной предикатной структуры осуществляется на основе лингвистических правил. Это достаточно громоздкая процедура, и сильно зависит от конкретного языка. В частности, для этого необходимо наличие словаря

валентностей глаголов. В нескольких словах эта процедура, которая включает в себя последовательность морфологического и синтаксического анализа, а также семантического анализа предложения, сводится к следующим шагам [5].

На уровне синтаксического анализа строится дерево зависимостей, для чего предложение текста фрагментируется на начальные сегменты, формируются их синтаксические интерпретации в случае наличия омонимии, выявляются синтаксические группы для каждой интерпретации, устанавливается иерархия между сегментами.

Последний шаг связан с выявлением расширенной предикатной структуры предложения. Для этого сформированные ранее начальные сегменты объединяются в простые предложения. Множество типов простых предложений русского языка может быть задано перечнем минимальных структурных схем предложений, описывающих предикативный минимум предложения [6]. Для этого используются правила, по которым, на основе морфологической информации о словах, входящих в предложение, можно определить тип минимальной структурной схемы. Результатом является выявленная предикатная пара «предикат-субъект». Далее на основе словаря валентностей глаголов оцениваются и вставляются в расширенную предикатную структуру оставшиеся члены предложения.

К сожалению, упомянутая процедура не дает стопроцентного результата анализа всех предложений текста: для разных языков процент правильно распознанных предикатных структур колеблется от 30 до 90 процентов [7]. Дело в том, что и человек не в состоянии точно решить эту задачу. Тем не менее, даже частичная разметка отношений однородной семантической сети дает дополнительную информацию, которая значительно улучшает формальное описание ситуации, представленной в тексте.

Поэтому для простоты построенная как в разделе 1 неоднородная семантическая сеть может быть упрощена до однородного представления заменой всех типов отношений между вершинами сети только одним типом — ассоциативным. При этом сеть упрощается: все одинаковые пары вершин, которые были связаны разными типами отношений, собираются вместе.

Такая процедура сильно сказывается на робастности вычислений. Процедура переранжирования весов вершин однородной сети резко улучшает качество ранжирования.

3. Однородная семантическая сеть

Убрав разметку дуг r_{j_i} неоднородной семантической сети, мы получим однородную (ассоциативную) семантическую сеть:

$$N \cong \left\{< c_i c_{1_i}, < c_i c_{j_{i\neq 1}} \gg\right\} \equiv \{< c_i c_j >\}, \quad (4)$$

Иначе эта сеть может быть переописана как множество звездочек:

$$N \cong \{s_i\} \equiv \{< c_i < c_j >>\}, \quad (5)$$

где c_i — первое слово звездочки (пар слов, составляющих звездочку), а $<c_j>$ — вторые слова этих пар — ближайшее ассоциативное окружение первого слова — его семантические признаки.

Если неоднородная семантическая сеть не может быть автоматически получена из текста, то однородная семантическая сеть может быть автоматически получена из текста [3, 6, 8]. Для этого сначала строится частотная сеть, весовые характеристики вершин которой — частота встречаемости z_i слов в тексте, и частота попарной встречаемости z_{ij} слов в предложениях текста — потом переранжируются для выявления степени их важности в анализируемом тексте.

Частотная сеть — это множество направленных пар слов, выявленных в предложениях текста. Выстроенные друг за другом (первое слово последующей пары объединяется со вторым словом предыдущей пары) пары слов текста соединяются в направленный граф с ветвлениями и вхождениями (тот же результат может быть получен проведением аналогичной процедуры со звездочками). Сама по себе однородная сеть не очень интересна. Она становится мощным инструментом для анализа текста, как только ее вершины приобретают численные характеристики, соответствующие рангам понятий в тексте. В частотной сети эти характеристики — частоты встречаемости слов и пар слов в тексте. Частотную сеть можно использовать для анализа больших текстов [7], но при анализе сравнительно небольших текстов требуется дополнительная переоценка ранга вершин сети с точки зрения их связанности в тексте (в рамках n-граммной модели текста, где n существенно больше 3).

Для переранжирования весов вершин частотной сети используется итеративная процедура [3], позволяющая учесть на n шагов влияние вторых слов звездочек (5) на вес первых слов. Вес первого слова вычисляется с учетом весов вторых слов звездочки и весов связей первого и вторых слов:

$$w_i(t+1) = \sum_{\substack{j \\ j \neq i}} w_j(t) w_{ij}, \quad (6)$$

где $w_i(0) = z_i$, $w_{ij} = z_{ij}/z_i$. Эти суммы нормируются на каждой итерации некоторым образом, например, с учетом суммарной энергии сети [3].

4. Динамика ситуации

Семантические сети обвиняют в неспособности (слабой способности) представлять динамику ситуации. Однако динамика не является особенностью семантической сети, которая по сути своей формирует статику ситуации. Если же мы выйдем за рамки собственно семантической сети, то легко смоделируем и динамику ситуации. Для этого нам надо построить не одну, а несколько семантических сетей, каждая на своем корпусе текстов, описывающем ситуацию позавчера, вчера, сегодня [10]:

$$\ldots \rightarrow N(t_{i_1}) \rightarrow N(t_{i_2}) \rightarrow N(t_{i_3}) \rightarrow \ldots . \quad (7)$$

Соединив их одноименные вершины, мы получим полное представление о динамике ситуации.

5. Тематическая структура текста

Минимальный древовидный подграф, извлеченный из однородной семантической сети, также как и сама сеть, представляет и текст (корпус текстов), и ситуацию. Но он проще, и работать с ним проще. А, главное, он фактически является оглавлением текста (см. рис. 1).

Тематическое дерево формируется как минимальный древовидный подграф однородной семантической сети, с корневой вершиной — наиболее значимой вершиной сети. Он получается разрывом слабых обратных связей на сети и ранжированием вершин по степени их важности. Корневая вершина тематического дерева представляет основную тему текста, дочерние вершины, в зависимости от их уровня, представляют подтемы, подпод— и подподпод— и т.д. темы соответственно.

6. Результаты анализа конкретного текста

Рассмотрим пример анализа текста с использованием представленных выше соображений.

6.1. Тематическое дерево

Для примера проанализируем текст этой работы и представим тематическое дерево, полученное с помощью технологии для автоматического смыслового анализа текстов TextAnalyst, разработанной на основе упомянутой теории (см. рис. 1).

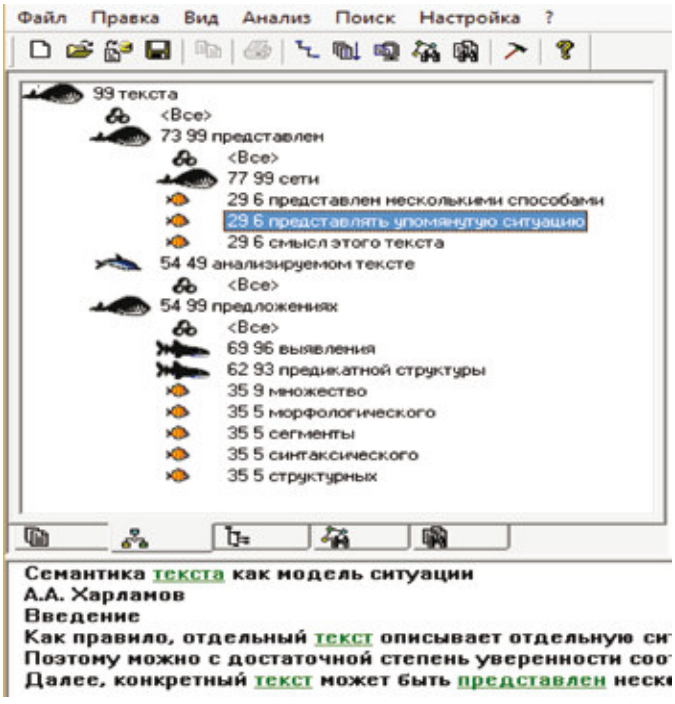

Рис. 1. Тематическое дерево. Здесь корневая вершина имеет наибольший вес, все под- и подподтемы весят меньше.

Упрощенное представление ситуации с помощью минимального древовидного подграфа сети — тематического дерева — дает вполне приемлемую интерпретацию ситуации, описанной в тексте. Просто перечислим концепты, представленные в вершинах тематического дерева, незначительно изменяя их форму с целью согласования. «Текст представлен сетью, несколькими способами, представлена упомянутая ситуация — смысл анализируемого текста. В предложении выявляется предикатная структура с использованием морфологического и синтаксического ...». Оно (тематическое дерево) говорит само за себя.

6.2. Семантическая сеть

Однородная семантическая сеть, также полученная из анализа этого текста, выглядит более громоздко (см. рис. 2). Рассмотрена только одна вершина сети «ситуация», ближайшие ассоцианты которой, в том числе, «текст», «семантическая сеть», «динамика ситуации» — являются семантическими признаками вершины «ситуация», раскрывают ее значение. С этой вершиной посредством гипертекстового механизма ассоциированы предложения текста, в которые входит понятие «ситуация» (в данном случае в паре с понятием «текст»).

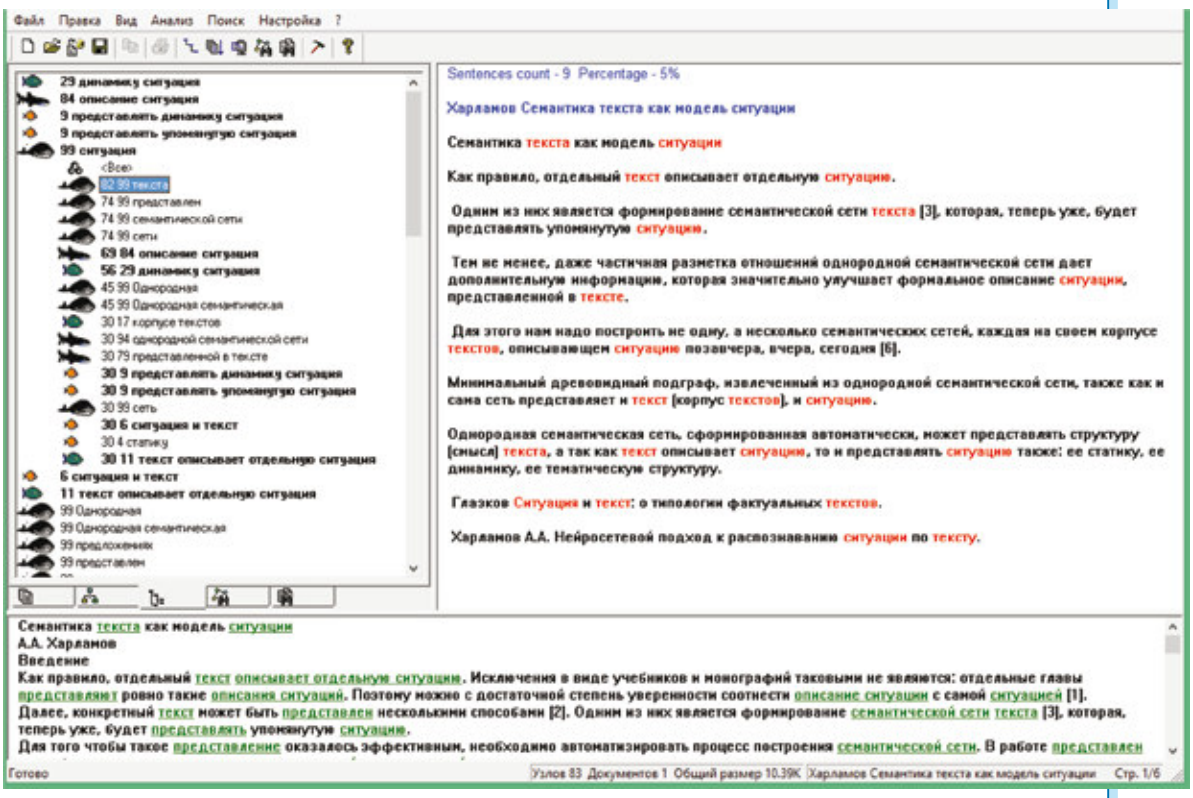

Рис. 2. Семантическая сеть. Она представлена в древовидном виде, но если раскрывать связи вглубь, то можно заметить повторяющиеся концепты, то есть это действительно сеть.

Необходимо сделать одно важное уточнение: автоматический анализ дает объективную структуру текста (а не ту, которую подразумевал автор в процессе его написания), поэтому все несуразности семантической сети, как и тематического дерева, относятся исключительно за счет автора.

Заключение

Однородная семантическая сеть, сформированная автоматически, может представлять структуру (смысл) текста, а так как текст описывает ситуацию, то и представлять ситуацию также: ее статику, ее динамику, ее тематическую структуру. То есть может быть удобным инструментом для анализа ситуаций. Неоднородная семантическая сеть является более точным представлением ситуации, чем его однородная сеть. Автоматическое формирование неоднородной сети сильно зависит от языка текста, а, главное, не дает гарантии стопроцентной разметки типов связей, так как разметка осуществляется исключительно лингвистическими методами.

Литература

1. *Глазков.* Ситуация и текст: о типологии фактуальных текстов. Преподаватель XXI в. 2016. — Выпуск 2 (ч.4). С. 578–588.
2. *Спицын В.Г., Цой Ю.Р.* Прикладные информационные технологии: представление знаний в информационных системах. Томск: Издательство Томского политехнического университета, 2008. 152 с.
3. *Харламов А.А.* Ассоциативная память — среда для формирования пространства знаний. От биологии к приложениям. — Дюссельдорф: Palmarium Academic Publishing, 2017. — 109 с. ISBN 978-3-639-64549-1.
4. *И Мяньчжу.* Позиционная грамматика: теория и приложения (В поисках компьютерной лингвистики). — Харбин: Издательство «Хейяунцзян Женьминь Чубанше», 1999. — 256 с.
5. *Alexander A. Kharlamov, Tatyana V. Yermolenko, Andrey A. Zhonin.* Text Understanding as Interpretation of Predicative Structure Strings of Main Text's Sentences as Result of Pragmatic Analysis (Combination of Linguistic and Statistic Approaches)// Доклад на Международной конференции SPECOM 2013, Пльзень, Чехия, сентябрь 2013г. LNAI 8113 — Pp. 333–339.
6. *Alexander A. Kharlamov.* A neural network approach to the situation recognition based on the texts. (В настоящем сборнике).
7. *Филиппович Ю., Прохоров А.* Семантика информационных технологий: опыты словарно-тезаурусного описания. / Серия «Компьютерная лингвистика». Вступ. статья А.И. Новикова. М.: МГУП, 2002.
8. *Харламов А.А., Ермоленко Т.В.* Автоматическое формирование неоднородной семантической сети на основе выявления ключевых предикатных структур предложений текста // Труды Международной научно-технической конференции «Открытые семантические технологии проектирования интеллектуальных систем» (OSTIS'2012), — Минск: 2012. С. 385 — 390.
9. *Городецкий Б.Ю.* Актуальные проблемы прикладной лингвистики // Новое в зарубежной лингвистике. Вып. XII. М., 1983. [Gorodetsky B.Yu. Actual'nye problemy pricladnoy lingvistiki // Novoye v zarubeshnoy lingvistike. Vyp. XII. M., 1983].
10. *Alexander A. Kharlamov, Tatyana V. Yermolenko, and Andrey A. Zhonin.* Modeling of Process Dynamics by Sequence of Homogenous Semantic Networks on the Base of Text Corpus Sequence Analysis// SPECOM 2014, Novy Sad, Serbija, 2014. LNAI 8773 Springer — Pp 300 — 307.

TEXT SEMANTIC NETWORK AS A SITUATION MODEL

Kharlamov A. A.,
Doctor of Technical Sciences, Senior Researcher, Institute of Higher Nervous Activity RAS, Moscow, Professor, Department of Applied and Experimental Linguistics, MSLU, Moscow, Professor, School of Software Engineering HSE, Moscow

Abstract

As a rule an individual text describes an individual situation. Then it is possible to correlate a situation with its description with a high degree of certainty. Any text can be represented by several ways. One of them is to form a semantic network from the text [1] which, in its turn, will represent the situation described by the text. To make the representation more efficient the semantic network should be formed automatically. This paper presents a way of describing a text with a homogenous semantic network, formed from the text. First of all, we will show that any text may be represented as a non-homogeneous semantic network. It is clear that separate sentences of the text may be represented as a graph by extended predicate structures. If we combine all of the extended predicate structure graphs of all sentences of the text and match the corresponding graphs nodes, we will get a graph representing the text semantics. If we remove arc annotation in a non-homogeneous network we will get a homogeneous (associative) semantic network. A homogeneous semantic network can be obtained automatically by text analysis. Actually, a semantic network is inherently a static representation of the situation, but if we go beyond the semantic network we can easily model the situation dynamics. For this we construct more than one semantic network using several text corpora which describe the situation the day before yesterday, yesterday, today. If we connect their similarly-named nodes we will get the situation dynamic representation. A homogeneous semantic network, formed automatically, may represent the structure (the semantics) of a text. That is why a tool for automatic analysis of the text semantics is an instrument of the analysis of situations.

Keywords: semantic text network as a situation model, text as a situation model, extended predicate structure of sentences, semantic structure of text, homogenous (associative) semantic network, text thematic structure, situation dynamics analysis, situation automatic analysis tool.