УДК 81.11 **Филологические науки**

Ван Сюефэн, *магистр*
по направлению «Медиалингвистики»
Санкт-Петербургский государственный университет
Россия, г.Санкт-Петебург

СЕМАНТИЧЕСКИЙ И СТРУКТУРНЫЙ АНАЛИЗ ТЕКСТОВ В СЕТИ ИНТЕРНЕТ

Аннотация: Статья посвящена семантическому и структурному анализу как одному из разделов компьютерной лингвистики. Автор обращается к уровням естественного языка и прослеживает алгоритмы их анализа компьютерными системами. В статье приводятся основные термины компьютерной лингвистики, используемые в процессе СемАн, которые сопровождаются примерами. Автор делает несколько выводов, касающихся значения, функций и сложностей семантического анализа.

Ключевые слова: семантический анализ, семантическое ядро, единицы текста, уровень, коллокация, сема, текст, главное слово.

Annotation: The article is devoted to semantic and structural analysis as one of the sections of computational linguistics. The author refers to the levels of natural language and traces the algorithms of their analysis by computer systems. The article presents the main terms of computational linguistics used in the process of semantic analysis, which are accompanied by examples. The author draws several conclusions concerning the meaning, functions, and complexities of semantic analysis.

Key words: semantic analysis, semantic core, text units, level, collocation, SEMA, text, main word.

Компьютерная лингвистика (КЛ) за последние десятилетия получила широкое распространение, что напрямую связано с разработкой множества прикладных программных систем. Необходимость анализа текстовой информации, в том числе в сети Интернет, и автоматической обработки текстов на естественном языке(ЕЯ)является толчком к развитию компьютерной лингвистики как области науки.

Модели языка, которые используются в КЛ, построены на основе теорий, созданных лингвистами в процессе изучения различных текстов и на основе лингвистической интуиции, или интроспекции. Модели КЛ характеризуются следующими особенностями:

- алгоритмизируемость;

-функциональность;

-общность модели, т.е. учет множества текстов;

-экспериментальная обоснованность;

-опора на словари как обязательную составляющую модели [1, с. 35].

Компьютерный анализ естественно-языкового текста отражает уровневую организацию языка:

- фонетический уровень;

- графематический уровень;

- морфологический уровень;

- синтаксический уровень;

- семантический уровень [11, с. 243].

Остановимся на последнем уровне. Семантический (от греч. σεμαντικός — обозначающий) анализ направлен на возможность определения значения слова в определённом контексте и конкретной ситуации, понимание смысла фразы. Единицей такого анализа является сема (семантический компонент, семантический множитель, семантический маркер, дифференциальный признак и др.).

Например, старуха = [взрослый] [немолодой] [женщина]; женщина = [человек] [женского пола] и т. д. В любом случае представление значения слова

в виде набора сем не дает возможности объяснить реальную сферу его употребления. Ведь никто не вздумает назвать старухой королеву английскую, хотя она и немолодая особа женского пола.

В таких случаях приходит на помощь знание семантической декомпозиции, которая объясняет и даже предсказывает черты сочетаемости анализируемого слова.

Например, мы говорим **пить кофе**, а не **есть кофе** потому, что слова **кофе** и **пить** включают сему [жидкий], а слово **есть** — нет.

Часто смысловой элемент включает два или более слов. Если частотность их совместного появления в тексте выше, чем ожидаемая вероятность, такое сочетание называют коллокацией. Она определяет и предсказывает совместное использование слов [7, с. 92]. Например, глагол **убежать** управляет предлогом **от**, а глагол **прибежать** требует предлога **к**. Или сочетаемость разных частей речи: **мощная машина** и **крепкий кофе**. В них нельзя заменить коллокации без вреда для значения.

В коллокациях выбор ключевого компонента осуществляется по смыслу, а выбор второго компонента(коллоканта) зависит от выбора первого (например, **принимать решение** — выбор глагола принимать является традиционным и зависит от существительного **решение**).

Для коллокаций характерна частичная некомпозициональность, когда значение целого не складывается из суммы значений частей. Если словосочетание полностью некомпозиционально, то оно является идиомой [i] (**бить баклуши, собаку съесть**). К коллокациям иногда относят составные топонимы и другие слитные названия (завод «Красный Октябрь», крейсер «Аврора»). Также встречаются разрывные коллокации: «**гнетёт жизнь**» и «**гнетёт меня жизнь**».

Семантический анализ является наиболее сложным направлением в автоматическом анализе текста. Устанавливаются семантические отношения между словами в тексте, объединяются семы и т.п.

Современные компьютерные технологии позволяют автоматически извлекать коллокации из текста. Это делается при помощи различных мер ассоциативной связи, которые выясняют статистическую значимость взаимного появления лексических единиц.

В семантическом анализе предложений предусматривается использование падежных грамматик и семантических падежей, или валентностей [6, с. 51]. Семантика предложения анализируется через связи главного слова (глагола) с его семантическими актантами. Например, глагол **переслать** описывается через семантические падежи пересылающего (агенса), адресата и объекта пересылки.

При помощи семантического анализа решаются проблемы многозначности (омонимии), которая неизбежно возникает при автоматическом анализе на любом языковом уровне [12, с. 192].

Виды омонимии:

-лексическая: совпадает звучание и/или написание слов, не имеющих общих элементов смысла, например, **лук** —растение и оружие;

-морфологическая: совпадают формы одной и той же лексемы, например, словоформа **шарф** встречается в именительном и винительном падежах данного существительного;

-лексико-морфологическая омонимия, наиболее частотная: совпадают словоформы двух разных лексем, например, **печь** — глагол в форме инфинитива и существительное **печь** в единственном числе, именительном и винительном падежах;

-синтаксическая: случаи, когда синтаксическая структура имеет несколько толкований, например: Flying planes canbe dangerous (хрестоматийный пример Хомского, где словоформа Flying интерпретируется как прилагательное или как существительное [10, с. 82]).

В компьютерной лингвистике производится автоматический синтез, отличающийся от обычного производства связного текста обратным порядком:

1) семантический синтез (переход от смысловой записи фразы к ее синтаксической структуре);

2) синтаксический (переход от синтаксической структуры фразы к представляющей фразу цепочке лексико-грамматических характеристик словоформ);

3) морфологический (по нормальной форме слова и его параметрам программа находит соответствующую словоформу);

4) графематический (объединение слов в единый текст, контроль соответствия фрагментов входного текста фрагментам выходного).

Рассмотрим подробнее теорию лингвистических моделей «Смысл⇔Текст». Она видит ЕЯ в качестве особого рода преобразователя, выполняющего переработку заданных смыслов в соответствующие им тексты и заданных текстов в соответствующие им смыслы. Смысл в данном случае - это инвариант всех синонимичных преобразований текста. Связный фрагмент речи без деления на фразы и словоформы отображается в виде специального семантического представления, которое состоит из двух компонент: семантического графа и сведений о коммуникативной организации смысла.

Метод семантических падежей - это логическое продолжение метода компонентного анализа. Чарльз Филмор в рамках компонентного анализа выделил семантически неделимые смысловые единицы (атомы смысла). Теория лингвистики принимает за центральную единицу смысла предикат, в основном это действие, выражаемое глаголом. По Филмору атрибуты, относящиеся к предикату, играют важные семантические роли, или семантические падежи [9, с. 23]:

- агента, который совершает действие
- объекта, с которым совершают действие
- адресата, для которого предназначено действие
- пациента - вещи, испытывающей эффективность действия
- результата
- контрагента - сила, против которой направлено действие

- источника - исходного состояния объекта

- инструмента - физической причины действия, или стимул

Семантический компонент (СемКомп) — главная составляющая Системы автоматического понимания текста. Его функция заключается в согласовании трёх разных «языков»:

а) языка созданных Системой лингвистических структур, которые он получает на входе;

б) языка предметной области текста, чьи термины необходимо использовать в языковом продукте на выходе;

в) языка пользователя, для которого Система АПТ должна выдать информацию.

Внутренние единицы текста, которые рассматриваются в процессе лингвистического анализа, строятся методом «снизу-вверх» и в реальных системах АПТ доходят лишь до синтактико-семантических представлений отдельных предложений. Лингвистам под каждую конкретную задачу приходится строить новую систему «перевода» между вариациями СинП или СинСемП и ещё более многочисленными единицами баз данных.

Внешние единицы рассматриваются в текстах методом «сверху-вниз», который является простым узнаванием заданных извне лексических единиц, как формируется поисковый запрос в информационно-поисковых системах. Два языка — внутренний и внешний — не способны «договориться» друг с другом, и системы АПТ заняты преодолением этого разрыва. Поэтому важно найти решение данной проблемы.

Наибольшие споры лингвистов наблюдаются в их воззрениях на семантической представление целого текста: кто-то называет так любую из структур семантического уровня (в том числе СемП целого текста), а кто-то-результат первичной семантической интерпретации, т.е.любую промежуточную структуру на пути к СемП целого текста.

Мы предлагаем в информационно-лингвистической модели понимания текста рассматривать структуры двух уровней: семантическое представление

текста и информационное представление (ИнфП) текста. Эти два термина отражают суть текста с разных сторон: СемП — это результат чисто лингвистического понимания, выраженный лексическим материалом текста; ИнфП — его внешнее представление, показывающее, как текст воспринимается внешней средой. Отличие ИнфП от СемП текста в том, что в нем отражен в сжатом виде только тот его фрагмент, который соответствует заданной извне точке зрения. ИнфП должно быть выражено в единицах воспринимающей системы («встречного текста»).

Критерий сжатия в СемП является результатом лингвистического сжатия, совершаемого в зависимости от грамматических и словарных средств.

Н. Н. Перцова в своей монографии утверждает, что модель понимания ЕЯ должна содержать наряду с поверхностно-семантическим компонентом и глубинно-семантический компонент(ГСемП), объединяющий информацию собственно языковую и энциклопедическую (= общность сведений о действительности, которые имеются у отправителя и получателя текста) [2, с. 249]. Это учитывается системами типа «вопрос-ответ».

Следует заметить, что важнейшая характеристика семантического анализа текста (даже с учётом неполного знания того конкретного вида окончательного СемП текста) - определение структуры минимальных и максимальных единиц текста.

Минимальной единицей можно считать выражение, соответствующее формуле семантического языка — P (A, B). Если в формуле использовать лексические выражения, получится элементарная ситуация (ЭСит). Проводим локальный СемАн в пределах каждого предложения; из этих отрезков складывается семантическое пространство текста; из него вырастает ситуативное представление (СитП) и происходит новое деление структуры текста — на высказывания. Такой локальный анализ характеризуется обращением к отдельным предложениям, или высказываниям [14, с. 41].

Максимальной единицей считается семантический граф целого текста: происходит преобразование первичной текстовой структуры СемПрост в

СитП; а высшей единицей, которая представляет текст во внешней среде, является текстовый факт (ТФ).

В локальный СемАн входят следующие части:

1) прямое толкование единиц СинП;

2) анализ лексических валентностей;

3) толкование слабых связей;

4) интерпретация всех элементов ситуативного представления, или создание первичного СитП [5, с. 267].

Говоря о СемУзле, или синтаксической группе, необходимо привести термин **лексическое ядро(ЛЯ),** которым обозначается главное слово синтаксической группы.

Проверка «семантического ядра» выявляет лексический и синтаксический состав текста, а также частотность единиц. Самые частотные является ключевыми и задают тему текста.

Поисковые роботы анализируют ключевые слова и словосочетания и соотносят содержимое страницы и запросы пользователей — фразы, введённые в поисковую систему [3, с. 64].

Далее в СемАн включается семантический словарь, который переводит язык синтаксических структур на язык СемП, производя операции с учётом семантических категорий и смысловых валентностей единиц, которые образуют словарные входы. Этот процесс позволяет осуществить постройку более содержательных СемУзлов.

Основной движущей силой СемАн текста являются взаимоотношения валентностей единиц текста. Мы говорим о формальных, композиционных и содержательных (ситуативных либо энциклопедических) СемО.

Связи, участвующие в образовании продолжений фраз текста до нормального вида, считаются семантически сильными [6, с. 130]. Их особенности:

а) форма поверхностного выражения (в этом случае они не считаются абсолютно сильными);

б) значения связи;

в) семантические характеристики зависимой группы.

Общепринятым является разделение связей на:

а) абсолютно сильные (синтаксически + семантически);

б) только семантически сильные (валентность содержат обобщенные СемО);

в) факультативные (слабые) [13, с. 99].

Глобальный уровень СемАн текста даёт возможность создавать единицы (СемУзлы), используя материал разных предложений. В результате имеем получение единого связного графа. В процессе глобального анализа происходит сжатие лексического материала текста, поскольку результат анализа должен отличаться неизбыточностью [4, с. 22].

Глобальный анализ является завершающим этапом работы локальных механизмов.

Итак, приходим к выводам:

1) Семантический анализ является трудной математической задачей, сложность которой заключается в неспособности компьютера правильно интерпретировать образы, заложенные человеком в тексте.

2) Результаты качественного семантического анализа будут полезны для анализа спроса на товары по покупательским отзывам, в поисковиках, онлайн-переводчиках и т.д.

3) СемАн текста работает с его семантическим ядром и статистическими показатели. Качественно сформированное СемЯ быстро продвигает статью в поисковой системе.

4) СемАн можно провести и в поисковых системах, которые после определения смысла текста предлагают найденные материалы.

Библиографический список:

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011.

2. Леонтьева Н.Н. Автоматическое понимание текстов Системы модели ресурсов — Академия , 2006.

3. Боярский К. К. Введение в компьютерную лингвистику. Учебное пособие. – СПб: НИУ ИТМО, 2013.

4. Боярский К. К., Каневский Е. А. Вега — компьютерная система классификации и анализа текстов. Lambert Academic Publishing, 2011.

5. Щипицина Л .Ю. Информационные технологии в лингвистике : учеб. пособие / Л.Ю. Щипицина. — М. : ФЛИНТА : Наука, 2013.

6. Белоногов Г.Г. Компьютерная лингвистика и перспективные информационные технологии. М.: Русский мир, 2004

7. Марчук Ю.Н. Компьютерная лингвистика. учеб. пособие. М.: АСТ Восток – Запад, 2007.

8. Зубов А.В., Зубова И.И. Информационные технологии в лигвистике: учеб.пособие. М.: Академия, 2004.

9. Лахути Д.Г., Рубашкин В.Ш. Семантический (концептуальный) словарь для информационных технологий // Научно-техническая информация. 2000.

10. Падучева Е.В. Динамические модели в семантике лексики. М.: Языки славянской культуры, 2004.

11. Тузов В.А. Компьютерная семантика русского языка. СПб: Изд-во СПбГУ, 2003.

12. Рубашкин В.Ш. Представление и анализ смысла в интеллектуальных информационных системах. М.: Наука, 1989.

13. Кобозева И. М. Лингвистическая семантика. – М., 2000.

14. Мельчук И. А. Опыт теории лингвистических моделей «Смысл <=> Текст». М., 1974 М., 1999.