Семантические сети и анализ связного текста

Михайлов Д. В.

Новгородский государственный университет
имени Ярослава Мудрого

Дополнительные разделы к учебному курсу
«Системы искусственного интеллекта»

по направлению
230100.62 — Информатика и вычислительная техника

2015 г.

Определение 1

Семантика (по Моррису) означает определённые (общие) отношения между символами и объектами, представленными этими символами.

Определение 2

Прагматика изучает выразительные (охватывающие) отношения между символами и создателями (пользователями) этих символов.

Определение 3

Информационная модель в информатике — представление объектов и отношений, ограничений, правил и операций, формализующее семантику данных для выбранного домена (проблемной области).

Определение 4

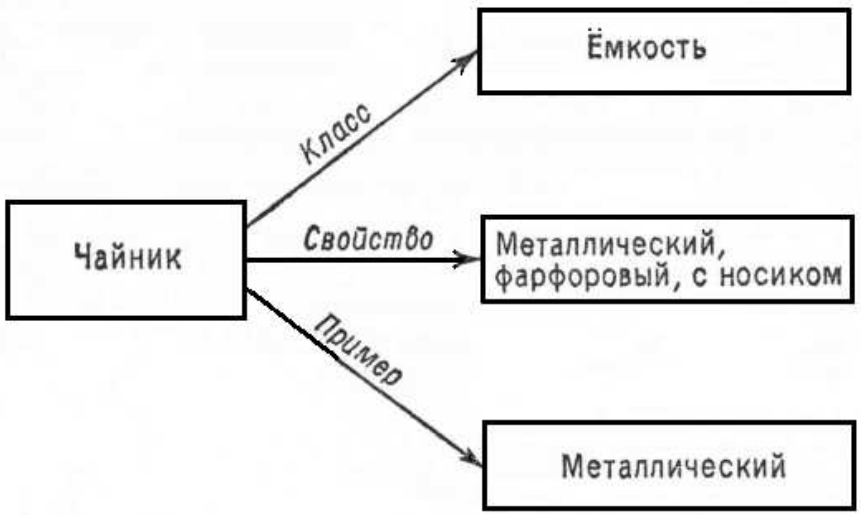
Семантическая сеть — информационная модель предметной области, заданная ориентированным графом, где вершины соответствуют объектам предметной области, а дуги — отношениям между ними.

TLC-модель — Teachable Language Comprehender.

Основная идея

Описание значений класса объекта, его прототипа и установление связей со словами, обозначающими свойства объекта.

Пример (Х. Уэно, М. Исидзука):



Примечание

Данная ассоциативная структура называется *плоскостью*, описываемые концепты объекта — *вершинами типа*, а связанные с ними ассоциативные слова — *вершинами лексем*.

Элементы, свойства и указатели в TLC-модели

TLC-модель использует представление данных *«элемент»–«свойство»*:

«элемент» $\Leftrightarrow$ *«вершина типа»;*

«свойство» $\Leftrightarrow$ *«вершина лексемы».*

Аналогия с продукционной моделью: элемент представлен заключением, свойство — структура, описывающая элемент:

$$\mathit{чайник}(X) \leftarrow \mathit{ёмкость}(X) \wedge$$
$$\wedge\big(\mathit{металлический}(X) \vee \mathit{фарфоровый}(X)\big) \wedge \mathit{имеет\_носик}(X).$$

Указатели связывают элементы и свойства. Так, в рассмотренном примере вершина типа для «чайник» имеет вершину лексемы «ёмкость», которая с помощью указателя тоже образует вершину типа (отношение «вид–род»).

При этом атрибуты свойств и их значения могут декларировать предложения вида *«Носик имеется у металлического/фарфорового»*.

В конечном итоге показанное представление данных описывает контекст: *«Чайник — это металлическая или фарфоровая ёмкость с носиком»*.

Замечание

Содержательно указателями задаются функциональные зависимости.

Средства определения функций в семантической сети (по Куиллиану):

- отношение «надмножество–подмножество»;
- частиречный индекс (наречие, прилагательное и т. п.);
- логическое «И»;
- логическое «ИЛИ»;
- исключающее «ИЛИ».

Отношения для группировки элементов (зависят от мира):

- близость;
- следствие;
- предпосылка;
- сходство.

Определение 5

Мир — формальная теория, полностью задаваемая участвующими в ней элементарными объектами (алфавит), правилами построения выражений (формул) из них, аксиомами касательно формул и набором правил вывода.

Классификация семантических сетей

По количеству типов отношений:

- однородные — представляют только один тип отношений, например, **IS_A** («род–вид»);
- неоднородные — количество типов отношений больше одного.

По арности отношений:

- сети с бинарными отношениями (связывающими ровно два понятия);
- сети с отношениями произвольной арности.

По масштабу:

- для решения отдельных задач — наиболее распространены в системах искусственного интеллекта;
- для решения групп задач по заданной области знаний — используются в качестве базы создания конкретных систем;
- глобальная семантическая сеть (пример — семантический Web).

Могут быть произвольными, но внимания заслуживают те из них, которые актуальны для многих миров.

Наиболее важные типы связей:

- **IS_A** — показывают отношения включения и позволяют объединять в сеть иерархию понятий, в которой узлы низких уровней наследуют свойства узлов более высоких уровней;
- **PART_OF** — показывают отношения «часть–целое».

Примечание

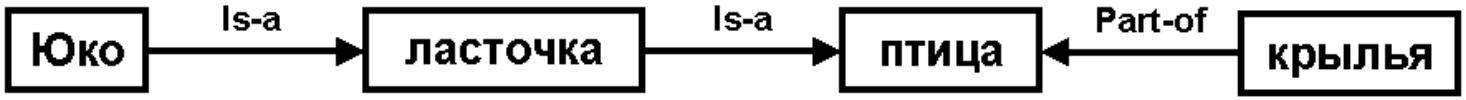
Совокупность фреймов, описывающих некоторую предметную область, образует семантическую сеть с единственным типом связей — **IS_A**.

Определение 6

Вывод множества фактов с помощью отношения **IS_A** в семантической сети называется *наследованием свойства*, а сама ветвь **IS_A** называется *ветвью наследования свойства*.

Иерархия наследования

Пример наследования свойства.

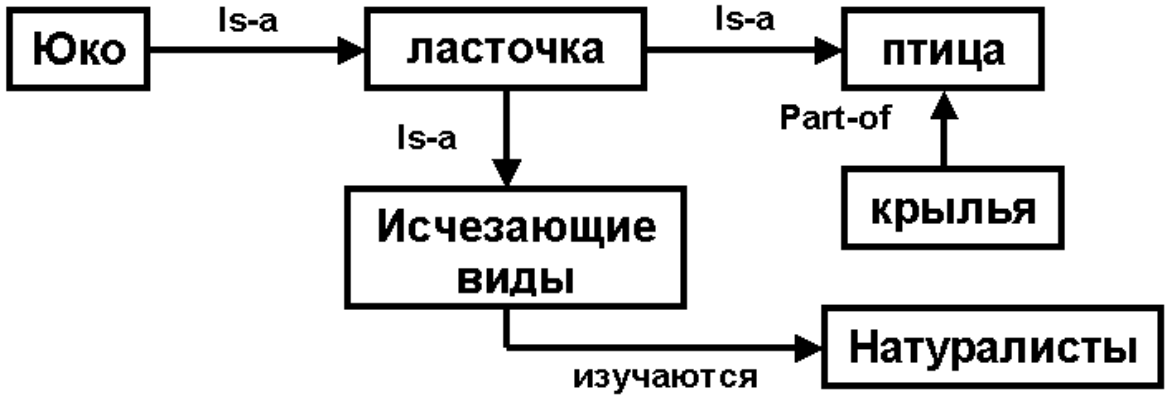


Здесь факт «Юко имеет крылья» может быть выведен из фактов:

- некоторую ласточку зовут Юко;
- все ласточки — птицы;
- все птицы имеют крылья.

Проблема

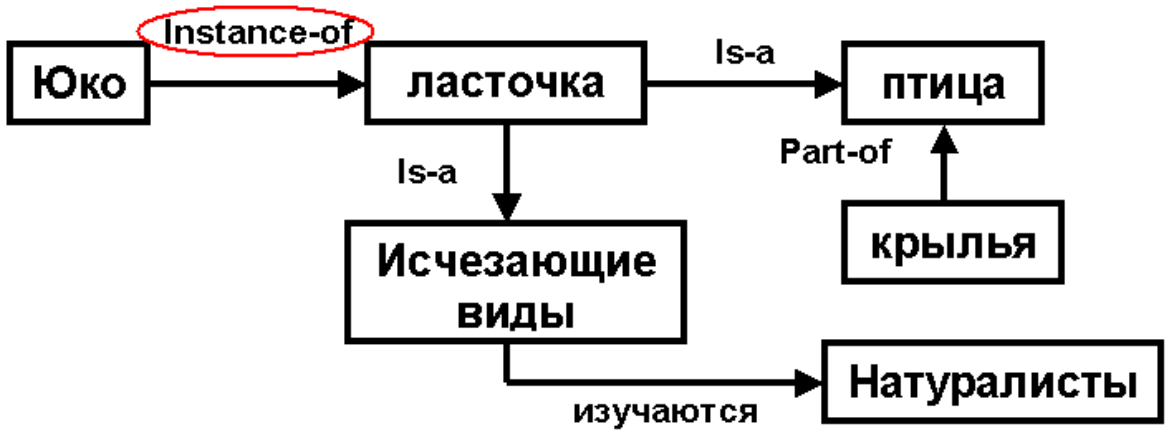
Наследование атрибутов между разными иерархическими уровнями.



Так, при показанном расширении семантической сети *не гарантирована правильность* заключения «Юко изучается натуралистами».

Управление наследованием в семантической сети

- использование связей типа **INSTANCE_OF** для разграничения в сети вершин концепта как некоторого понятия и его экземпляра (instance). Пример:



- разделение атрибутов класса на атрибуты определения и атрибуты свойства, последние при этом отображаются в качестве отношений между классами и не наследуются классом нижнего уровня;
- наследование значений атрибута с помощью связи **IS_A**, а самого атрибута — с помощью связи **INSTANCE_OF**.
- введение процедур, определяющих действия над дугами (связями) и вершинами.

Процедурные семантические сети (Hector J. Levesque и др.)

Основные действия, определяемые процедурами над дугами (связями):

- установление связи;
- аннулирование связи;
- подсчёт числа вершин, соединённых заданной дугой;
- проверка наличия–отсутствия связи между заданными вершинами.

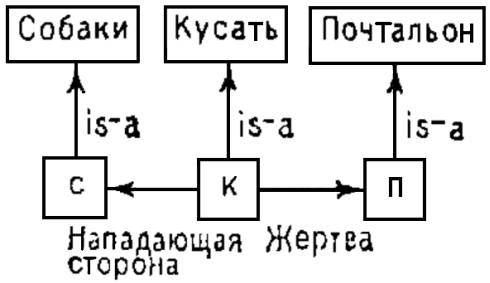
Примеры действий над вершинами:

- определение экземпляра класса;
- аннулирование экземпляра;
- подсчёт числа экземпляров, принадлежащих классу;
- проверка принадлежности экземпляра к некоторому классу.

Основная идея (Г. Хендрикс)

Вершины экземпляров — в качестве переменных, границы их действия определяются иерархически упорядоченным множеством пространств.

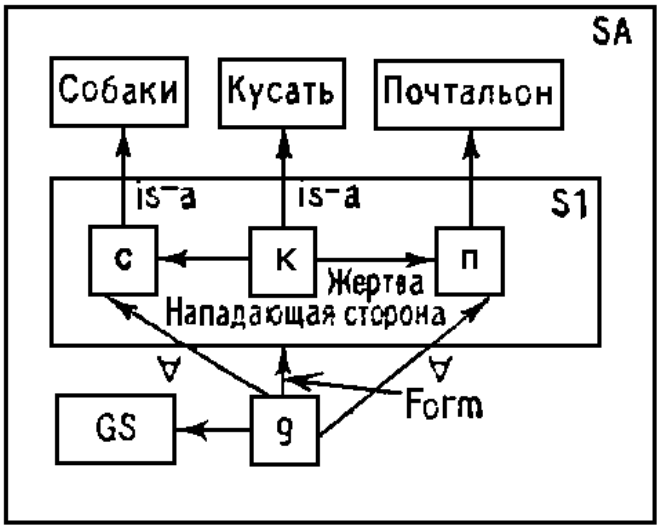
Пример исходного факта: *«Собака укусила почтальона»*.



Данная сеть строится введением вершин экземпляров *«с»*, *«к»* и *«п»*, соответственно.

Слова *«собака»*, *«укус»* и *«почтальон»* здесь обозначают классы.

Квантифицируем: *«Каждая собака кусает каждого почтальона»*.



Здесь вершина g представляет утверждение

$$\forall X : собака(X)\ (\exists Y : почтальон(Y) \wedge кусать(X, Y))$$

и ограничивает с помощью дуги Form пространство S1; GS — это $\forall X : собака(X)$.

При этом:

- S1 содержится в пространстве SA;
- для поиска S1 из SA допускается использовать только дугу Form.

Применение и преимущества разделения семантической сети

- **системы понимания речи** — для условного обозначения правил, определяющих категорию объекта;
- **классификационные системы** — сжатие избыточной информации запоминанием совместно используемых данных на уровне категорий;
- **диалоговые системы** — при семантическом разборе введённого пользователем текста, а также синтезе ответа атрибуты классов объектов связывают с классами глаголов, направляя разбор/синтез.

Падежная рамка и предикат ситуации

При расширении семантической сети в ней возникают отношения, которые отличны от **IS_A**, **PART_OF** и **INSTANCE_OF** и наследуются классами нижнего уровня, связанными по **IS_A** с классами верхнего уровня.

Замечание

Указанные отношения используются при формальном описании разных аргументов предиката одной и той же ситуации.

Определение 7

Падежная рамка — вершина семантической сети, которая определяет различные аргументы предиката ситуации.

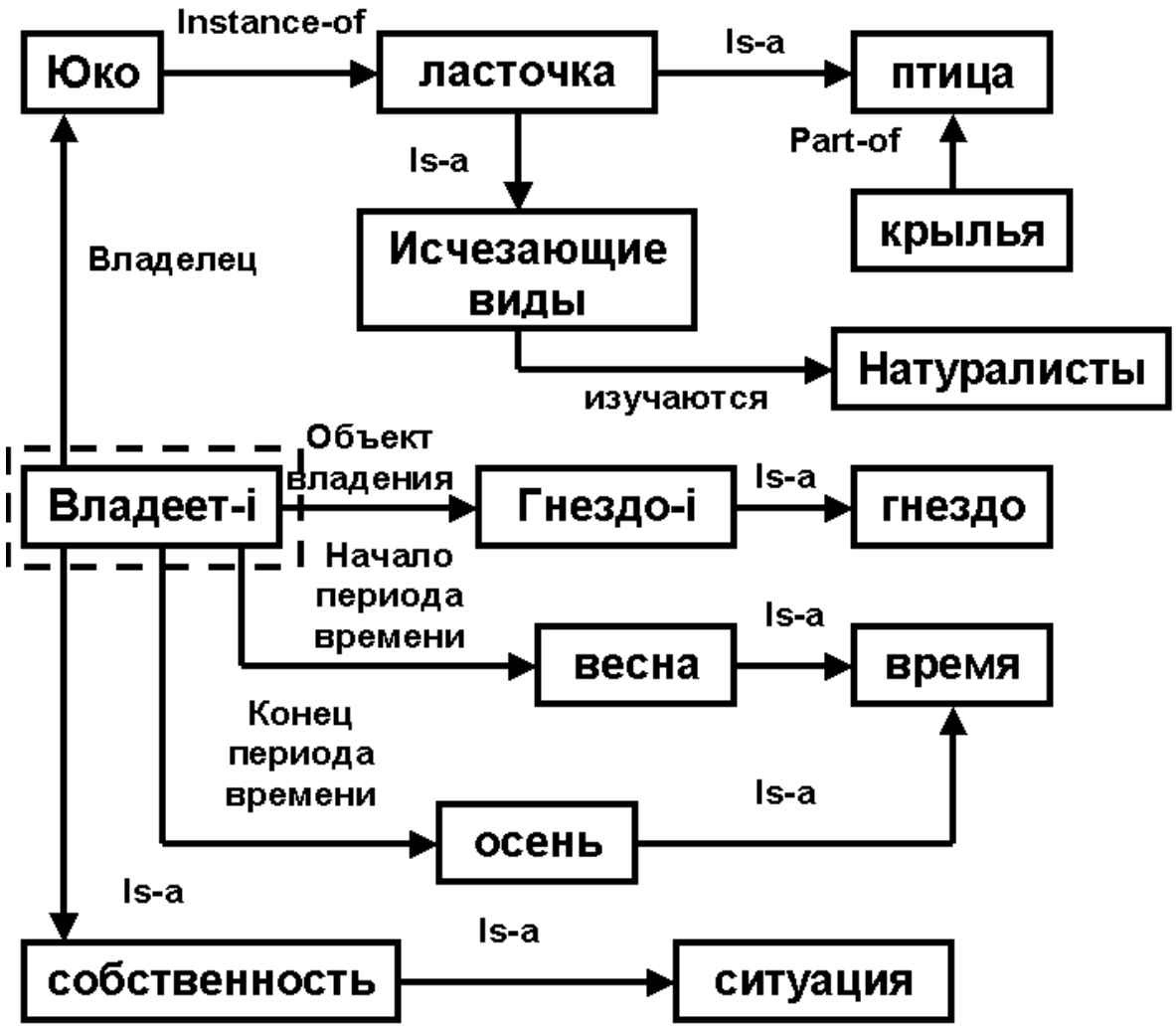
Преимущество

Возможность наследования:

- ожидаемых значений;
- значений по умолчанию

для атрибута в вершине экземпляра.

Диаграмма представления для примера семантической сети, включающей падежную рамку



Вывод в семантической сети

Определяется с помощью использующих её процедур.

Наиболее типичный способ вывода основан на сопоставлении частей сетевой структуры, которое может быть определено рекурсивно, см. описание примера реализации семантической сети на языке Лисп.

При этом вывод, как правило, происходит в три этапа:

- выделение составных частей запроса (лексем) и определение их семантической ориентации по словарю;
- построение семантической сети запроса;
- сопоставление семантических сетей запроса и области знаний.

Замечание

Поскольку семантическая сеть запроса строится на основе семантической сети области знаний, то построение сети запроса можно заменить поиском отношений между выделенными из запроса концептуальными объектами по сети области знаний.

Определение 8

Вывод в семантической сети, суть которого есть поиск узла пересечения дуг, идущих из двух различных узлов, называется *перекрёстным*.

Пример сопоставления запроса с фрагментом базы знаний

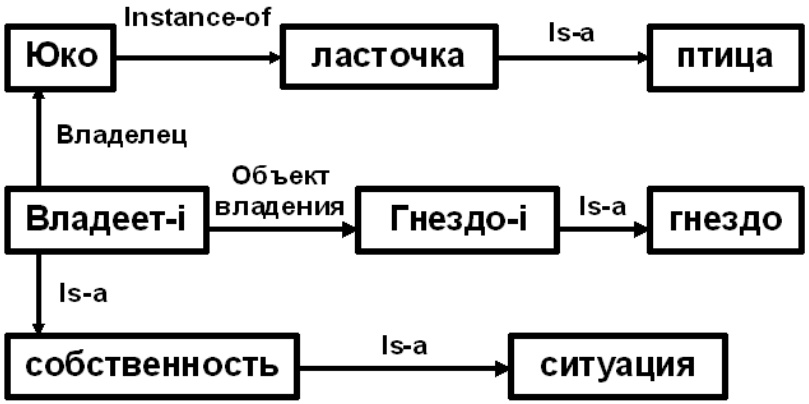


Рис. 1: Фрагмент базы знаний

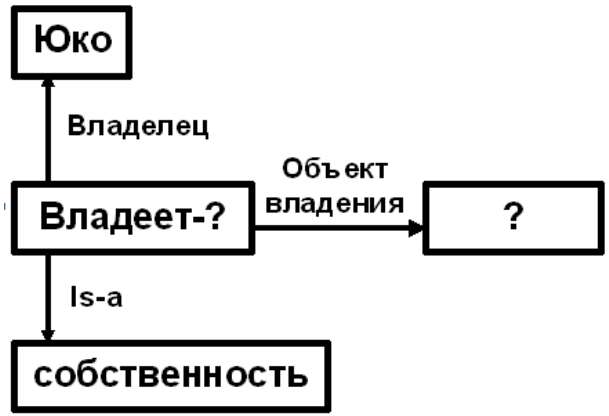


Рис. 2: Подсеть запроса «Чем владеет Юко»?

Ход сопоставления:

1. Поиск вершины для «владеть».
2. Найдена падежная рамка «Владеет-i», ветвь аргумента «Владелец» направлена к вершине «Юко».
3. Соединение с узлом, входящим для вершины «Владеет-i» по ветви аргумента «Объект владения».
4. Возвращается значение экземпляра для понятия «гнездо», т. е. конкретное «Гнездо-i».

- имеет четыре сети для основных знаменательных частей речи: существительных, глаголов, прилагательных и наречий;
- базовой единицей является *синонимический ряд* (т. н. *синсет*, англ. *synset*), объединяющий слова со схожим значением;
- каждый синсет дополнен дефиницией (толкованием) и примерами употребления слов в контексте;
- слово или словосочетание может появляться более чем в одном синсете и иметь более одной частиречной категории;
- каждый синсет содержит список слов-синонимов или синонимичных словосочетаний и указатели для отношений между ним и другими синсетами;
- слова с несколькими значениями включаются в несколько синсетов и могут быть отнесены к различным синтаксическим и лексическим классам.

Официальная страница WordNet

- **гипероним:** $breakfast \to meal$ ($завтрак \to приём\ пищи$);
- **гипоним:** $meal \to lunch$ ($приём\ пищи \to обед$);
- **has-member:** $faculty \to professor$ ($факультет \to профессор$);
- **member-of:** $pilot \to crew$ ($пилот \to экипаж$);
- **мероним:** $table \to leg$ ($стол \to ножка$);
- **антоним:** $leader \to follower$ ($лидер \to последователь$).

Замечание 1

Приписанные синсету отношения могут выполняться одновременно (конъюнкция отношений) или выборочно (отношения дизъюнктивны).

Замечание 2

Обычно неявно предполагают конъюнктивность меронимов и гиперонимов и дизъюнктивность гипонимов.

Замечание 3

Дизъюнктивность/конъюнктивность для совокупности отношений иногда желательно указать явно, ср. *пропеллер–самолёт–реактивный двигатель*.

- информационно-поисковые тезаурусы описывают определённую предметную область, WordNet содержит информацию о значениях общей лексики языка, но в то же время возможно создание тезаурусов типа WordNet и для конкретных предметных областей;
- в информационно-поисковых тезаурусах практически не представлена многозначность языковых единиц, в WordNet слово включается во всех определённых для него значениях;
- в WordNet отсутствует ограничение глубины понятийной иерархий;
- для информационно-поисковых тезаурусов характерны ограничения на включение словосочетаний, задаваемые перечнями правил.

- среди синонимов синсета выделяется доминантный синоним, представляющий собой наиболее нейтральный и частотный способ выражения соответствующего лексического значения;
- основным инструментом при разграничении значений слова является контекстный анализ;
- при принятии решений о количестве значений многозначного слова выделяются статистически значимые маркеры:
  - определённая грамматическая форма;
  - принадлежность к некоторому дереву родовидовой иерархии RussNet;
  - оба вышеуказанных показателя вместе.

  Эти признаки должны проявляться устойчиво: более чем в 33% контекстов для рассматриваемого значения в корпусе;
- значения слова, частотность появления которых в корпусе составляет менее 1% контекстов этого слова, считаются неустойчивыми и не включаются в тезаурусное описание.

Примечание

Словарь RussNet не переводился с Принстонского WordNet, а создавался как отдельный ресурс.

Особенности подготовки исходных данных:

- для задания частотного упорядочения значений многозначного слова используется разметка выборочной совокупности контекстов корпуса, выполняемая вручную;
- устойчивые словосочетания выделяются на основе меняющегося контекстного диапазона.

Критерии для выделения устойчивых словосочетаний:

- абсолютная частота сочетания слов;
- t-критерий Стьюдента;
- коэффициент взаимной информации.

1. словарь опирается на корпус современных текстов 1985–2004 гг. общим объёмом около 21 млн. словоупотреблений.
   Корпус включает статьи из газет и журналов на темы: *повседневная жизнь*, *экономика*, *политика*, *наука*, *культура*, *спорт*.
2. ядерная структура тезауруса задаётся примерно двумя тысячами наиболее частотных слов (существительных, глаголов, прилагательных, наречий), которые встречаются более 100 раз на миллион словоупотреблений в рассматриваемом корпусе;
3. разные значения некоторого слова упорядочиваются в соответствии со значением *частотности* — отношения числа употребления значения слова к общему числу словоупотреблений в корпусе;
4. в RussNet изначально представлена *общая лексика*, не относящаяся к терминам;
5. синсеты RussNet соотносятся с Межъязыковым лингвистическим индексом (ILI), предложенным в рамках проекта EuroWordNet.

Основное предположение

Лексическая связность может не только охватывать пары слов, но и соединять между собой группы слов текстового фрагмента, посвящённого одной и той же теме.

Определение 9

Лексическая цепочка — последовательность слов, в которой каждое следующее слово связано некоторым отношением с предшествующими словами той же цепочки.

Проблема

Выбор варианта цепочки в случае многозначного слова.

Определение 10 [Barzilay and Elhadad, 1999]

Для снижения числа вариантов цепочки Ch вводится *показатель* её *силы*

$$\mathrm{sc}\,(\mathrm{Ch}) = \mathrm{hid}\,(\mathrm{Ch}) \cdot \mathrm{len}\,(\mathrm{Ch})\,, \tag{1}$$

где $\mathrm{hid}\,(\mathrm{Ch}) = 1 - \frac{\mathrm{ndif(Ch)}}{\mathrm{len(Ch)}}$; $\mathrm{ndif}\,(\mathrm{Ch})$ — число разных слов в Ch.

Пусть $\mathrm{Avg}_{\mathrm{Ch}}$ — среднее значение величины (1) по анализируемому тексту,
σ_{Ch} — среднеквадратическое отклонение указанной величины.

Тогда для использования Ch в дальнейшем анализе необходимо, чтобы

$$\mathrm{sc}\,(\mathrm{Ch}) > \mathrm{Avg}_{\mathrm{Ch}} + 2\sigma_{\mathrm{Ch}}. \tag{2}$$

Замечание

Предполагается, что если лексические цепочки являются промежуточным представлением содержания документа и отвечают условию (2), то они же будут хорошо представлены и в аннотациях, подготовленных вручную.

Основные принципы составления аннотаций

Определение 11

Тематический узел — совокупность близких по смыслу понятий, упоминаемых в тексте. При этом т. н. основные тематические узлы моделируют главных участников ситуации, а суть текста состоит в описании взаимодействия между главными участниками.

1. Информативными и потенциально включёнными в аннотацию считаются те предложения, которые содержат минимум два понятия, входящих в состав разных основных тематических узлов текста.
2. Для каждой пары выявленных основных тематических узлов в аннотацию выбираются предложения, содержащие первое вхождение этой пары (следуя по порядку текста).

Порядок построения аннотации:

1. Формируется множество «аннотационных» фрагментов, которые не являются вопросительными или восклицательными предложениями.
2. Создаётся таблица всех возможных пар основных тематических узлов.
3. Начиная с начала текста, отбираются такие предложения, которые содержат ещё не упоминавшуюся пару разных тематических узлов.

Основные принципы составления аннотаций

Если предложение подходит для аннотации, но содержит местоимение, то:

- если предыдущее предложение входит в состав аннотации, то и текущее включается в аннотацию;
- если предыдущее предложение не входит в состав аннотации, то его проверяют на возможность включения в формируемую аннотацию. Для этого предложение либо должно не содержать местоимений, либо следовать за другим предложением, уже включённым в аннотацию;
- в остальных случаях предложение с местоимением не включается в аннотацию.

Связная и понятная аннотация может быть построена не всегда для:

- нормативных документов сложной структуры (законы, президентские и правительственные документы, международные договоры);
- газетных интервью;
- текстов больших размеров — при наличии ограничения на длину аннотации.

Определение 12

Структурная тематическая аннотация:

- представляет содержание текста посредством описания участников его основной темы, каждый из которых моделируется совокупностью понятий, относящихся к соответствующей теме;
- содержит наиболее информативные фрагменты тематического представления текста, которое включает все понятия текста, разбитые на тематические узлы.

Структурная тематическая аннотация включает в себя:

- понятия для основных тематических узлов, упорядоченных по убыванию частотности и расположенных горизонтально;
- отметки об относительно суммированной частотности основных тематических узлов, обозначаемые, например, различным количеством символов «*»;
- отметки об относительной силе взаимоотношений между различными тематическими узлами, например: «X» — очень сильное; «Z» — сильное отношение; «.» — отношение (без указания силы).

* * * *				*ИНФОРМАЦИЯ; ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ; ИНФОРМАТИКА;*
				ДОСТОВЕРНОСТЬ ИНФОРМАЦИИ; СЛОВАРЬ
* * * *		X		*ИНФОРМАЦИОННАЯ СИСТЕМА; СОБСТВЕННОСТЬ; ПРАВО СОБСТВЕННОСТИ;*
				НАУКА И ТЕХНИКА; ЭЛЕКТРОННАЯ ТЕХНИКА
* * * *		X	z	*ФЕДЕРАЛЬНОЕ ЗАКОНОДАТЕЛЬСТВО; ЗАКОН; ЗАКОННОСТЬ;*
				НОРМАТИВНЫЙ АКТ; ОСНОВНЫЕ ГРАЖДАНСКИЕ ПРАВА
* * * *		X	z	. *ГОСУДАРСТВЕННАЯ ДУМА;;*
				СЕРТИФИКАЦИЯ; ПРОМЫШЛЕННАЯ ПОЛИТИКА;
				ОРГАН ГОСУДАРСТВЕННОЙ ВЛАСТИ;

Структурная аннотация ФЗ об информации, информатизации и защите информации РФ

****	X	.	z	.	*ГРАЖДАНИН; ЧЕЛОВЕК; НАСЕЛЕНИЕ; ТАЙНА;*
					ДЕМОГРАФИЧЕСКАЯ СИТУАЦИЯ; СЕМЕЙНАЯ ТАЙНА;
					ФИЗИЧЕСКОЕ ЛИЦО; ЧАСТНАЯ ЖИЗНЬ
****	z	X	.	.	*ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ; ТЕХНОЛОГИЯ;*
					ЭЛЕКТРОННАЯ ТЕХНИКА ;
					КОМПЬЮТЕР
***	z	.	.	.	*ПРАВА ЧЕЛОВЕКА; ПРАВА ГРАЖДАН;*
					ОСНОВНЫЕ ГРАЖДАНСКИЕ ПРАВА
					МОРАЛЬНЫЙ УЩЕРБ; РАВНОПРАВИЕ;

Заключение

Преимущества представления знаний семантическими сетями:

- максимальная близость к естественному языку описания понятий и ситуаций;
- отношения между понятиями и ситуациями образуют достаточно небольшое множество и хорошо формализуются;
- сетевая модель представления знаний позволяет определить простые и эффективные принципы аннотирования текстовых документов рассмотрением ситуаций и значимых в них понятий.

Слабые стороны:

- рост времени поиска с увеличением размеров сети («изоморфизм подграфу»);
- требуется механизм отслеживания противоречий, в частности, использованием метазнаний.

Использованные источники

1. Представление и использование знаний / под ред. Х. Уэно, М. Исидзука. Режим доступа: прямая ссылка.
2. *Лукашевич Н. В.* Тезаурусы в задачах информационного поиска[1] / Н. В. Лукашевич; Изд-во Моск. ун-та. М., 2011. 512 с.
3. *Хабаров С. П.* Представление знаний в информационных системах / С. П. Хабаров. Режим доступа: прямая ссылка.
4. Википедия — свободная энциклопедия [Электронный ресурс]. Режим доступа: https://ru.wikipedia.org/.

[1] При подготовки материала использовалась электронная версия на www.twirpx.com