Семантический анализ и РОМИП

Георгий Дерновой

Велтон.Soft

Основная задача, которую мы ставили перед собой, было создание персональной поисковой системы с возможностью интеллектуального анализа текстов. Поиск по ключевым словам не дает возможности находить смысловое соответствия, если они описаны другими словами, поэтому для решения задачи был выбран семантический подход, суть которого сводится к следующему: по фрагменту текста строится описание, независимое от конкретных слов, но определяемое тем, что эти слова означают. В качестве основы мы использовали семантический словарь В. А. Тузова, единственное на тот момент (вероятно, сейчас тоже) описание семантики русского языка. Словарь достаточно легко адаптировался с нашим ПО, несколько измененным для решения этой задачи.

Содержание семантического словаря основывается на теории семантического анализа [1, 2, 3], разработанной проф. В. А. Тузовым на базе другой теории “Смысл - текст”, идеологом которой является И. А. Мельчук [4]. В своей работе В. А. Тузов сумел превратить, в общем, философские рассуждения Мельчука в последовательную и понятную систему, при этом, к некоторому сожалению, переняв и ее недостатки. Тем не менее, на наш взгляд, без этой работы было бы вообще невозможно понять, возможен ли семантический анализ, и в каком направлении следует идти. То, что стало возможным обозреть весь язык в рамках одной формальной(!) системы и увидеть глобальную таксономию понятий, пусть и в упрощенном виде, - колоссальное достижение.

Семантический словарь В. А. Тузова содержит более ста тысяч лексических единиц, разбитых на 3 семантических (смысловых) уровня:

1. ***Фундаментальный.*** Состоит 1500 иерархических классов, и небольшой набор – около 2 десятков – базисных функций и ~200 их вариаций.
2. ***Вариативный***. Состоит из 23000 классов, которые тесно связаны с фундаментальным уровнем и являются вариациями фундаментальных понятий. Описываются на основе понятий фундаментального уровня.

3. ***Описательный.*** Слова и понятия, имеющие смысл, выходящий за рамки фундаментального и вариативного, описываются на основе понятий уровней 1 и 2.

Основная сложность языка – описание внутренней структуры понятий, которыми он оперирует. В таком описании просматриваются два уровня – поверхностно-семантический и концептуальный уровень, зависящий от предметной направленности текста. Поверхностно-семантический уровень позволяет разбить все слова языка на смысловые кластеры, в которые попадают слова, тесно связанные с базовым, не упрощаемым понятием на уровне поверхностно-семантического описания. Слова в кластере имеют значения, описываемые через данный базовый смысл с помощью базисных функций (фундаментальных отношений) и понятий фундаментального и вариативного уровней. Концептуальный уровень содержит информацию о глубинной связи понятий (знания), которая явно в языке не присутствует, но без которого анализ никогда не сможет быть интеллектуальным.

В процессе изучения и адаптации словаря к работе были обнаружены сложности, не позволяющие подняться выше некоторого (невысокого) порога интеллектуального поиска.

Вот основные:

- часто близкие по смыслу понятия имели различные, непересекающиеся определения;
- не было ясного способа вычисления семантической близости/разности понятий, не говоря уже о целых предложениях, описанных на семантическом языке (СЯ);
- базисные функции при ближайшем рассмотрении такими не являлись и имели сложное семантическое значение; Например, базисная функция Prepar в словаре имеет различные смыслы:
  A. Делать годным;
  B. Выполнять;
  C. Собираться сделать;
  D. Делать возможным.

  Очевидно, что оно никак не базовое (интуитивно, по крайней мере).
- в семантическом языке не было механизма описания понятий, представляющих граф; *пример из словаря:*

  АППЕТИТНЫЙ
  N%~АППЕТИТ$1303(Caus_a1(НЕЧТО$1~!%1,IncepFunc(Ж

ЕЛАНИЕ$1303(Hab(!Для,ПИЩА$124/1))))) в переводе на русский:

Нечто пробуждает желание у кого-то(!Для) иметь пищу.

Здесь не использован механизм ссылок, поэтому смысл отличен от правильного: Нечто пробуждает желание кого-то съесть это НЕЧТО.

Т. е. НЕЧТО совпадает с пищей, потому как вряд ли вид, к примеру, зонт может пробудить желание иметь пищу.

Без использования механизма невозможно корректно описать это понятие, т. к. один и тот же объект упоминается дважды.

- в семантическом языке отсутствовали понятия говорящего и слушающего, а также многих фундаментальных понятий, без чего многие понятия можно описать только неправильно.

Например, определение словаря

АНОНИМ NeHab_o1(ЧЕЛОВЕК$1241,ИМЯ$12/0172)

Это формула говорит о том, что некоторый человек не имеет имени.

Правильней будет

NeHab(Orator,ИНФОРМАЦИЯ$13154(ИМЯ$12/0172(**ЧЕЛОВЕК$1241)))

Т.е. говорящий не имеет информации об имени обсуждаемого им человека.

Правда, здесь для полноты картины нужно было бы добавить, что с большой вероятностью упоминаемый человек умышленно скрыл свое имя. В противном случае любой прохожий – аноним.

- в классификаторе отсутствует информация о семантической разнице уровней (т.е. определение, что изменяется в понятии на верхней ступени), что вносит определенную путаницу, особенно когда понятие высокого уровня включает понятие нижнего уровня не как основу, а как некоторое отношение с ним.

Например, Источник описан как производное понятие от Возникновения, Термообработка как производное от Температуры и т.д.

Источник, несомненно, связан с Возникновением, но это не обычная связь родитель – потомок, а либо Характеристика

(Возникновения), или, в худшем случае, 2-й аргумент отношения Возникновение(Что, Где, Когда, Инициатор.).

Эти и другие сложности, выявленные в процессе разработки ПО и адаптации словаря к нему, позволили вполне определенно увидеть, что нужно для семантического анализа текста на хорошем уровне. К сожалению, принципы построения словаря должны быть несколько отличными от тех, которые использовал В. А. Тузов. Главное отличие – это допустимое и рекомендуемое использование понятий высокого уровня для определения производных. На данный момент семантический язык (точнее будет сказать – семантический анализатор профессора) допускает только отношения первого порядка, т.е. в рамках этой системы нельзя определить понятие и использовать его потом для определения другого понятия. В описании понятия СЯ В. А. Тузова допустимо использовать только базовые отношения и понятия фундаментального и вариативного уровней словаря. Второе – построение непротиворечивого семантического фундамента, который наряду с лексическими функциями будет включать математические, поведенческие, и другие фундаментальные понятия. Третий, более спорный, но важный момент, - отделение семантики понятия от его лексических свойств, т. е. отделения понятия от конкретного языка. В этом случае, возможно, сбудется мечта Лейбница, когда в случае спора истина будет поддаваться строгому вычислению. По меньшей мере, для этого будет заложен фундамент.

Фазы анализа текста в Алхимике

Результатом нашего исследования семантического словаря В. А. Тузова стала поисковая система Алхимик, доступная на сайте нашей компании[3].

Процесс анализа текста состоит из трех составных частей: морфологический , поверхностно-синтаксический и собственно семантический анализ.

Морфологический анализатор осуществляет обработку текста, вычисляя необходимые для дальнейшего анализа морфологические (грамматические) характеристики каждой словоформы.

Поверхностно-синтаксический анализ производит сканирование предложения, выбирая пары слов, и в зависимости от контекста (уже имеющегося результата) строит синтаксическое дерево. При этом происходит отсев большей части неподходящих морфологических вариантов.

Семантический анализ сводится к нахождению оптимального множества семантических пар, которые хитрым образом синтезируются по синтаксическому дереву, используя формулы слов семантического словаря. Хитрым он называется потому, что подбирался экспериментальным способом по критерию ~ информативность / размер множества.

После анализа, в зависимости от решаемой задачи, следует сохранение множества (при добавлении документов в базу данных документов), либо (при запросах) отыскание множества, включающего множество запроса, либо допускающего отличие на заданную величину.

Проблемы использованного подхода.

Программа вела себя по-разному, иногда показывая удивительные результаты (можно было подумать, что это результат анализа мощного ИИ), в других случаях (не частых, но тем не менее) показывала низкий результат на простых запросах. В среднем ее результаты были сравнимы с поисковыми системами, использующими поиск по ключевым словам. Большей частью это было вызвано теми причинами, о которых говорилось выше, часть была и на нас, как разработчиков ПО (наша схема анализа отличалась от предложенной В. А. Тузовым). Анализ ситуации позволил сделать вывод, без радикального изменения словаря или ПО значительно увеличить адекватность поиска не представляется возможным. Но стало понятно и другое (хотелось бы надеяться) – что именно нужно.

На момент предложения об участии у нас было 2 непохожие друг на друга системы – Алхимик и новая разработка - Ментал. В результате испытаний Алхимика был сделан вывод, что универсальную интеллектуальную систему (а именно в этом состояла основная цель), со словарем или без, построить в разумные сроки не удастся. Помимо семантического описания необходимо слишком много правил, очевидных для любого человека, но неочевидных для семантического анализатора. При запросе “веселый дворник” вполне очевидно, что “веселый человек” подходит в качестве ответа, но дворник описан в словаре как род занятия, и программа проходит мимо, казалось бы, очевидного решения. Можно добавить правило, что название профессии может выступать в роли человека, но предварительное исследование показало, что таких “очевидных” правил нужно многие тысячи.

Такое неприятное открытие натолкнуло на мысль отказаться от “универсального решения” и сосредоточиться на разработке систем семантического анализа, специализированных в некоторой предметной области. Для обеспечения расширяемости разработать технологию, позволяющую при желании легко объединять такие системы в более универсальные. Специализация заключается в детальной разработке иерархии понятий и взаимосвязей конкретной ПО, а также правил “действительности”, образующих концептуальный уровень. Этот уровень лежит за пределами семантики понятия, и без него невозможно достоверно решить, удовлетворяет ли данное утверждение заданному запросу.

Для участия в РОМИП мы решили использовать Алхимик, во-первых, из-за его тематической универсальности, во-вторых, наша работа над Менталом на данный момент далека от завершения. Главная сложность в этой работе, - подготовка данных для системы, из-за чего написанный код до сих пор не прошел всесторонней проверки.

Участие в РОМИП

Нас несколько разочаровал способ оценки результатов РОМИП. Так как наша система была ориентирована на интеллектуальный, а не пословный поиск, расчет оценки “автоматизированным” способом не мог, и, наверное, не показал, как в действительности обстоят дела, хотя, конечно, можно и ошибаться. На наш взгляд было бы корректней произвести оценку случайно выбранных фрагментов “вручную”, что, во-первых, застраховало бы от ошибок программы-оценщика, во-вторых, исключило бы примитивность конкретного программного метода оценки. Оценка таких тонких материй, как адекватность поиска может находиться только в руках человека. Когда при тестировании результатов Алхимика и Ментала между разработчиками возникает спор об адекватности конкретного результата, это нормально, потому как это связано с разным уровнем понимания проблемы и ожиданий. При использовании Алхимика один бухгалтер сильно удивлялся, считая дефектом программы, что на запрос, связанный с “Основными средствами” программа отказывалась показать ему документы, в которых упоминались подстатьи этого раздела. И коль скоро даже между людьми возникают споры о релевантности, то привлекать сюда автоматизированные средства для оценки кажется мне не совсем серьезным.

Для участия в дорожке по классификации сайтов потребовалось написать модуль сбора статистики, который делал комплексную оценку множества документов на предмет принадлежности к конкретной теме. Нам не пришлось использовать обучающее множество, т. к. алгоритм оценки в нем попросту не нуждался. Применение такого множества, на наш взгляд, оправдано может быть там, где нет понимания того, что представляют собой данные на структурном уровне, т. е. с “черным ящиком”.

Для определения тематической направленности сайта программа определяла тематическую связанность каждого документа с каждой темой. Тематический вес документа приближенно равен взвешенному по точности кол-ву предложений в документе, в которых присутствует заданная тема, с учетом размера документа. После определялись комплексные оценки сайта по каждой теме и, на основании того, существенны они или нет, и насколько разнятся эти оценки в относительных значениях, сайт относился к определенной теме (темам) или не относился ни к одной.

Работа по созданию системы интеллектуального анализа оказалась очень трудоемкой. Хотя мы имеем обнадеживающие результаты, указывающие, что есть свет в конце тоннеля, до него все еще далеко. Выход из ситуации видится в объединении усилий, направленных на разработку семантических и концептуальных словарей. Для этого нужно разработать стандарт семантического языка, на котором будут единообразно описываться понятия и знания, а также стандартизировать ПО для работы с СЯ: интерпретаторы СЯ, средства разработки, интеграции в информационные системы и др. Мы открыты для любого конструктивного сотрудничества, которое позволит сделать интеллектуальный (семантический) анализ текстов реальностью.

Литература

[1] Тузов В. А. Компьютерная семантика русского языка. Материалы Диалог-2001: http://www.dialog-21.ru/Archive/2001/volume2/2_53.htm

[2] Тузов В. А. Компьютерная лингвистика. СПБУ 1998.

[3] Тузов В. А. Прагматический анализ текстов. СПБУ 2001.

[4] Мельчук И.А. Опыт теории лингвистических моделей "Смысл-Текст".

[5] Информация о компании Велтон.Soft. http://www.soft.velton.net.ua/langs.html