OSTIS-2011
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

НЕКОТОРЫЕ АКТУАЛЬНЫЕ ПРОБЛЕМЫ РАЗРАБОТКИ ЛИНГВИСТИЧЕСКОГО СЕМАНТИЧЕСКОГО КОДИРОВАНИЯ

А.И. Головня (*golovnjaai@bsu.by*)
Белорусский государственный университет информатики и радиоэлектроники, г.Минск, Республика Беларусь

Создатели систем искусственного интеллекта не всегда точно могут определить свои цели при разработке той или иной проблемы, связанной с построением компьютерной интеллектуальной системы, поэтому важно найти плюсы и минусы в разработке семантического кодирования в работах известных ученых. Определить что же такое «смысл», каково его значение при разработке компьютерных интеллектуальных систем и систематизировать то, что, казалось бы, не поддается систематизации.

Ключевые слова: знание, классификация, семантическое кодирование, смысловые варианты, смысл.

Введение

Семантическое лингвистическое кодирование – это очень сложная и трудная задача, которая предполагает найти пути решения следующих задач: 1) сформулировать наиболее общие принципы создания семантического кода; 2) показать в общем виде на примерах его возможности; 3) обнаружить пути возможного сжатия многотысячного словаря словоформ. Эти задачи требуют новых решений от разработчиков интеллектуальных систем.

Попытки создания автоматизированных систем для обработки текстов уже предпринимались и описаны в работе группы сотрудников Йельского университета в работе «BORIS – экспериментальная система глубинного понимания текстов» [НЗЛ, с. 106–160]. Создание именно такой системы, которая оперирует словарем, содержащим около 400 слов и выражений, и позволяет отвечать на вопросы к тексту объемом в страницу, по мнению профессора В.А. Карпова, не представляет сложности [Карпов, 2003]. Подобная система может быть построена лингвистические за три-четыре месяца и за два месяца отлажена уже в программном виде. При этом можно использовать тот же самый текст только на русском языке, но строить систему на качественно новых принципах. При всей искусственности текста, а он практически содержит избыточную информацию, в реальных текстах такой информации не бывает, приведенная выше работа может считаться аналогом систем глубинного понимания. При одном условии: если мы решим, что такое машинное понимание? На наш взгляд понимания здесь нет: в машинной памяти просто указаны все разрешенные (правильные) связи. И именно в рамках разрешенного строятся вопросы к тексту. Данная система не в состоянии порождать вопросы и ответы на сходные ситуации, описанные, допустим, во втором рассказе на аналогичную тему, в котором будет использоваться другой синтаксис и, следовательно, предложения будут иметь другой вид, тем более, если это будет текст другого рода, не о разводе, например, а о продаже партии сыра. При построении такой системы ее разработчики шли не от теории к практике, а от текста – к способам его препарирования с целью получения ответов на предполагаемые (будущие) вопросы. Знания о

Библиотека БГУИР

мире как таковом в данной системе отсутствуют. Имеется лишь тот их фрагмент, который связан непосредственно с темой *развод*. И при этом даже эти знания ограничены.

О необходимости семантического кодирования говорится практически во всех работах, связанных с созданием систем искусственного интеллекта. Но функции, предназначение кода и принципы его построения настолько разнятся, что выбор какого-то конкретного направления для анализа и дальнейшего развития сильно затруднен.

Причины такой ситуации кроются, на наш взгляд, в двух основных моментах. Во-первых, целостная картина мира членится на отдельные части множеством предметных областей. Именно поэтому ученые часто говорят о «мозаичном» знании. Так, в монографии французского исследователя А. Моля целый раздел посвящен мозаичной культуре и средствам массовой коммуникации и информации, создающими именно мозаичную картину мира [Моль, 1973, с.119-125]. Чаще всего эти предметные области не связаны друг с другом и в силу отсутствия общей системной терминологии эта разобщенность мешает формализовать даже имеющиеся знания о мире. Это происходит из-за того, что в рамках предметной области мы вновь сталкиваемся с делением на подобласти знания, те членятся еще больше и больше далее. Наконец наступает момент, когда полученное знание становится настолько специфичным, что становится непонятным даже большинству специалистов данной области (вспомните К.Пруткова, который сказал: «Специалист подобен флюсу», т.е. односторонний). При мельчайшем делении науки на отрасли, ветви и прутики теряется сам предмет исследования, вот здесь-то и начинает действовать и проявляться субъективизм исследователя, так как от общей задачи он удален предельно. И возникает вопрос: а нужно ли такое знание? По крайней мере, существуют данные, что большая часть книг Ленинской библиотеки в Москве ни разу не была востребована [Уемов, 1978].

Второй момент. При разговоре о знаниях и попытке их формализовать не учитывается сама возможность приращения знаний за счет их моделирования. Знания приобретаются лишь за счет накопления новой информации, а не за счет логического вывода новых знаний на базе старых. Между тем, ребенок, осваивающий мир и одновременно язык моделирует и очень много, так как у него в память введена минимальная часть необходимой информации, он практически не может слышать от родных и соседей большей части слов и их форм.

Причины создания такой научной ситуации совершенно разноплановы. Одна из них предельно важна и на ней мы остановимся подробнее – это отсутствие математического аппарата в исследовании, эмпиризм, диктат практики. Даже у таких крупных ученых, как Н.И. Вавилов, можно найти строчки такого типа: «Дальнейшие исследования более точно установят закон гомологической изменчивости у растений и животных, и может появиться возможность приведения тех же рядов в математическое выражение» [Вавилов, 1987, с. 53]. Минимальная модификация треугольника Паскаля решает эту проблему не только для гомологических рядов, а вообще для N объектов одного или разных родов. Само отсутствие математического аппарата может быть и от неведения, что этот аппарат в твоих руках, но ты его не видишь. Блестящий пример именно такого классификационного аппарата-«невидимки» мы находим у аргентинского писателя Борхеса. В эссе «Аналитический язык Джона Уилкинса» Борхес в качестве примера неудачных классификаций объектов мира приводит следующую, приписывая ее со слов Франца Куна какой-то древней китайской энциклопедии.

Так, «Животные делятся на а) принадлежащих Императору, б) набальзамированных, в) прирученных, г) сосунков, д) сирен, е) сказочных, ж) отдельных собак, з) включенных в эту классификацию, и) бегающих как сумасшедшие. К) бесчисленных. Л) нарисованных тончайшей кистью из верблюжьей шерсти, м) прочих, н) разбивших цветочную вазу, о) похожих издали на мух» [Борхес, 1994, с. 87]. Весьма забавный пример классификаций, не правда ли? Но если показать небиологу, что в одно семейство у Вавилова попадают растения типа ржи, кукурузы и пырея, то большая масса будет возражать против такой классификации, хотя она и непротиворечива.

Борхес в своем творчестве показывает знакомство с работами математика Георга Кантора, Бертрана Рассела и другими. Но он все же писатель. Другое дело, когда специалист

Библиотека БГУИР

по ядерной физике академик Л.Б. Окунь в словарике на термин классификация приводит тот же пример из книги Борхеса и завершает свою мысль-отношение к классификациям подобного рода следующим пассажем: «Если классификация какого-либо раздела физики чем-то напоминает Вам эту классификацию, то, значит, Вы еще недостаточно овладели этим разделом» [Окунь, 1988, с. 177]. И в этом же словарике присутствует термин группа и его объяснение. Достижения физики последних десятилетий связаны именно с групповыми представлениями. Так вот классификация Борхеса может представлять собой группу двенадцатого порядка. В ней должно быть 4096 подсистем (два в двенадцатой степени). Нам даны из них две подсистемы: единичный элемент группы – пункт з) включенных в эту классификацию (12 плюсов), обратным элементом (12 минусов) будет пункт м) прочих (другими словами – не включенных в классификацию). Далее не сложно полагать, что может существовать и вся комбинаторика 12 остальных элементов группы по 2, по 3, по 4, …, по 11. Другими словами, могут быть животные, принадлежащие императору и одновременно бегающие как сумасшедшие; сирены могут быть сказочными и нарисованными тончайшей кистью из верблюжьей шерсти и т.п. При учете того, что отдельные собаки (в другом переводе – бродячие собаки) уже не принадлежат или не принадлежали императору– введение плюсовых и минусовых признаков контрарного плана может уменьшить порядок группы.

Создатели систем искусственного интеллекта не всегда точно могут определить свои цели при разработке той или иной проблемы, связанной с построением компьютерной интеллектуальной системы. В отношении семантического кодирования отмечается та же самая ситуация: осознается, что семантический код нужен, но для чего? – Чтобы закодировать семантику каждого отдельного слова. Возникает вопрос, а что это даст? И тогда-то приходится хвататься за ряд терминов, которые по мнению разработчиков и теоретиков помогут им в решении проблемы. Самые употребительные среди них – *смысл, текст, информация, знание, понимание,* которые нуждаются хотя бы в кратком анализе.

С использованием термина *смысл* возникает больше всего хлопот. Он четко не определен и поэтому существуют целые монографии типа «Опыт теории лингвистических моделей «СМЫСЛ – ТЕКСТ» [Мельчук, 1974], в которых с помощью графов представлены смыслы предложений типа «*Иван вчера твердо обещал Петру, что даст ему нужную книгу*» и вариантов – «*Вчера Иван сообщил Петру, что обязпательно снабдит его книгой, в которой он нуждается*» [Мельчук, 1974, с. 184–186]. Особенно интересны графы, представляющие семантические представления в Приложении: «*Ваня твердо обещал Пете вечером принять Машу самым теплым образом*», «*Ваня дал Пете обещание, что вечером он непременно окажет Маше самый сердечный прием*», «*Ваней было твердо обещано Пете, что вечером Машу ждет у него самый теплый прием*» и т.п. [Мельчук, 1974, с. 302 – 310]. За усложненной авторской терминологией цель работы и смысл отыскивается с трудом. Базирующиеся на данной работе последующие работы еще более затемняют проблему смысла.

Проанализируем для примера описательную работу Варшавской А.И. типа «Смысловые отношения в структуре языка», сделанную на английском материале [Варшавская, 1984]. В ней содержатся такие, например, отношения: локативные, часть целого, принадлежности, предшествования-следования и одновременности, причинно-следственные, цели, условия, паратактические (уступительные, противительные, соединительные, разделительные). Во-первых, указанные отношения существуют не в языке, а в Универсуме между объектами, входящими в него. Язык только отражает эти отношения. Во-вторых, эти отношения общеизвестны: ясно, что вещи где-то локализуются, кому-то принадлежат, что в силу их сложности они членимы на искусственные их составляющие или элементы состава, что события могут рассматриваться по мере их наступления, что человеческая деятельность характеризуется целью, что наступление того или иного события имеет более явную или менее явную причину, что событие происходит при определенных условиях. Паратактика не представляет собой учение о союзах: и в седьмом классе уже знают соединительный союз *И,* разделительный союз *ИЛИ* и т.п.

Библиотека БГУИР

Далее Варшавская А.И. локативные отношения делит на две подгруппы:

А – конкретное вещественное значение:
1. космос, пространство, воздух, атмосфера и т.д.;
2. Земля, континент, моря, океаны, горы, холмы, реки, леса, поля, луга и т.п.;
3. север, юг, Бостон, Альпы и т.п.;
4. графство, Америка, штат, район, территория, провинция и т.п.;
5. город, деревня, пригород, улица; резиденция, дом, площадь, бар; комната, кухня, угол, клозет и т.п.

Б – существительные с обобщенным значением места:
1. место, позиция, местонахождение и т.п.;
2. со значением расстояния: дистанция, расстояние, направление и т.п.

Автор претендует на классификационное разделение семантик, а в результате его приходит к смешению. Причина – давление, диктат семантики (локативное отношение) и неучет обычной грамматики – разные предлоги будут выражать разные семантические локативные значения. Аналогичны и классификации части-целого, принадлежности, предшествования-следования и одновременности, причинно-следственные, цели, условия, паратактические (уступительные, противительные, соединительные, разделительные) [Варшавская, 1984, с. 33 – 58].

В другой, уже специализированной работе, «Распознавание образов и машинное понимание естественного языка» [Файн, 1987, с. 33 – 58] речь идет, разумеется, о понимании текстов на русском языке, а не о понимании естественного языка. В.С. Файн, рассуждая о декларативном определении «смысла» [Файн, 1987, с.12 – 13], приходит к такому неутешительному выводу: «Отсутствие удовлетворительного определения для “смысла” приводит к необычному и драматическому положению: в разработках, традиционно нацеленных на выяснение смысла текста, делается попытка решить проблему, которая не может быть даже поставлена. Фактически создается совершенно сказочная ситуация, когда компетентно и масштабно строятся компьютерные системы, долженствующие “искать то – не знаю что”» [Файн, 1987, с. 17]. Приводимые далее в книге результаты применения имитационного принципа, предлагаемого автором, мягко говоря, не убеждают.

В.А. Карповым с сотрудниками НИЛТиПЛ также была предпринята попытка разобраться в сути этой проблемы. Само слово «смысл» в качестве нечетко определенного термина нами использоваться не будет. Антиподом смысла будет бессмыслица как отсутствие смысла. В качестве примера осмысленной бессмыслицы нам предлагается американский вариант в виде «*Зеленые идеи яростно спят*», где можно через метафорическое представление получить какой-то смысл. Или родное русское предложение «*Глокая куздра штеко будланула бокра и курдячит бокренка*», где также стопроцентная правильность грамматики при возможных смысловых вариантах.

Мы же с детства знаем о чепухе и бессмыслице больше, чем после университетского курса. Пример 1. «*На воротах чепуха жарила варенье, куры съели петуха в это воскресенье*». В этом предложении с грамматикой все в порядке. А вот с правильностью лексических связей – путаница, бессмыслица: *жарить на воротах нельзя, варенье варят, а не жарят, куры не могли съесть петуха*. Все это примеры лишь ситуационных норм сочетаемости. При других переменных в предложении, например, *на газу* (вместо *воротах* – обстоятельство) *соседка* (субъект вместо *чепуха*), *варила варенье* или *жарила грибы* (вместо *жарила варенье* – предикат и объект) получаем истинную или возможно истинную ситуацию, отражаемую предложением «*На газу соседка варила варенье*». Мы можем врать, т.е. не быть истинными, наша соседка в данный момент не варила варенье и вообще была в Крыму, но такое предложение нельзя отнести к разряду невозможных. Заменой всего одного предиката – *заклевали* (вместо *съели*) мы достигаем такого же возможного варианта и во второй части предложения: *куры заклевали петуха в это воскресенье*. Заменой субъекта куры на кошка получаем вариант: *Кошка съела петуха в это воскресенье*. И примеров такого рода в детстве мы наслушались от бабушек, дедушек и приятелей с целью приобретения представлений об

Библиотека БГУИР

истинности-ложности нашей информации еще задолго до знания логики. Это логика здравого смысла. Но тогда правильность неопределяемого нами смысла представляет оценку текста с точки зрения истинности-ложности, правдивости-неправдивости содержания отражаемых ситуаций.

Приняв приведенные примеры за базу дальнейших рассуждений следует развивать временно-ситуативную возможностную логику, или логику возможных ситуационных миров, отражаемых отдельным предложением или несколькими предложениями. Обозначим условие 1, при наличии которого может совершаться некоторое действие или группа действий как У1, условие 2 (другое условие, при котором совершаются качественно иные действия как У2.

Допустим, что на начальном этапе рассуждения имеются лишь два условия. Содержание самих условий на этом этапе совершенно неважно, мы строим абстрактную систему. Плюс будет означать возможность некоторого события при некотором условии, минус – невозможность события. Суть события также важна. Тогда целостная система будет состоять из четырех подсистем.

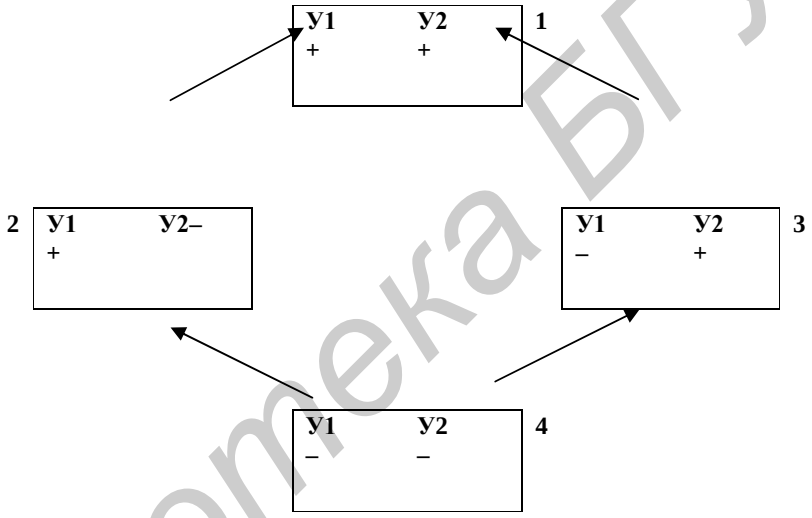

Рисунок 1 - Отражение возможных и невозможных ситуаций при определенных условиях

Интерпретация построенной системы такова: код (++) или событие, возможное при условиях У1 и У2, приравнивается к подсистеме-событию, возможному всегда)так как у нас всего два условия и оба присутствуют), коды + – и – +) будут означать подсистемы-события иногда, подсистема с двумя минусами в коде будет означать события никогда, ни прикаких условиях (из У1 и У2). Здесь важно отметить, что в одном и том же минимальном локусе в одно и то же время не могут происходить одновременнол два события, объект не может одновременно испытывать два разных состояния, иметь два взаимоисключающих признака – должно измениться время, за которое одно событие должно перейти в другое. Например невозможна ситуация: *спать-не спать*, есть ее градации *спать, дремать*; невозможна ситуация *идти – сидеть* и масса других. Нам можно возразить, что одновременно можно *сидеть* и *передвигаться*, например, *сидеть* и *ехать* одновременно. В системе, построенной на на двух признаках (статика – *сидеть*, динамика – *ехать*) эта статико-динамика будет находиться в подсистеме с двумя плюсами (там же будут и ее варианты: сидеть и плыть на лодке, сидеть и лететь в самолете и т.п.). Сидеть и ехать по отдельности будут представлять иногда-подсистемы, подсистема с двумя минусами будет представлять иные действия, не характеризующиеся статикой и динамикой. Несложно получить более подробную систему вышеуказанного типа для трех условий. Она будет состоять из 8-ми подсистем. При этом

Библиотека БГУИР

интересно, что увеличится число подсистем иногда, а *всегда-подсистема* и *никогда-подсистема* постоянно будут являться макросистемами и одновременно микросистемами.

Библиографический список

[Борхес, 2004] Борхес Хорхе Луис Сочинения в трех томах. т.2 / Х.Л. Борхес. М., 1994.

[Вавилов, 1987] Вавилов Н.И, Закон гомологических рядов в наследственной изменчивости / Н.И. Вавилов. Л., 1987.

[Варшавская, 1984] Варшавская А.И. Смысловые отношения в структуре языка. Ленинград, 1984.

[Карпов, 2003] Карпов В.А. Язык как система / В.А. Карпов. М., 2003.

[Мельчук, 1974] Мельчук И.А. Опыт теории лингвистических моделей «СМЫСЛ – ТЕКСТ» / И.А. Мельчук. М., 1974.

[Моль, 1973] Моль Абраам Социодинамика культуры / А. Моль. М., 1973.

[НЗЛ, 1989] Новое в зарубежной лингвистике, т. 24. Москва, 1989.

[Окунь, 1988] Окунь Л.Б. Физика элементарных частиц / Л.Б. Окунь. М., 1988.

[Уемов, 1978] Уемов А.И. Системный подход и общая теория систем / А.И. Уемов / в сб. Системный анализ и научное знание. М., 1978.

[Файн, 1987] Файн В.С. Распознавание образов и машинное понимание естественного языка / В.С. Файн. М., 1987.

Библиотека БГУИР