МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ
ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 9 Москва 2022

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

УДК 004.822

А.Ю. Щербаков

Методологические основы и прототип системы семантического искусственного интеллекта

Семантический искусственный интеллект рассматривается как элемент субъект-субъектных коммуникаций и практически единственное на сегодняшний день эффективное средство решения проблем поисковых и аналитических систем.

Особо подчеркивается роль систем семантического ИИ в принятии прозрачных и объяснимых решений, семантические инструменты дополняются спектральными моделями анализа процесса генерации текстов.

Ключевые слова: *искусственный интеллект, интеграция, семантический алгоритм, теоретико-множественное сравнение, спектр текста и словаря, марковский процесс*

DOI: 10.36535/0548-0027-2022-09-1

ВВЕДЕНИЕ

Сегодня системам и различного рода «решениям» искусственного интеллекта посвящен обширный публичный и научный дискурс. При этом смешиваются решения классической робототехники, связанные с управлением техническими системами, нейросетевые алгоритмы обучения и принятия решений, методы моделирования сознания, имеющие имитационный характер и многое другое. Настоящая статья является продолжением исследования, посвященного

методологическому осмыслению искусственного интеллекта (ИИ), основанного на работе семантических инструментов.

Методологическое осмысление проблемы ИИ даже при постановке задач, формулировании терминов и определений испытывает достаточно много трудностей. Часть из них априорно «встроена» в проблему. Например, проблема «обучения» нейросети практически имитирует процесс человеческого мышления и принятия решений, поскольку для обучения необходимо сделать выборки «правильных» или типовых решений, что связано с их потенциальной субъективностью и ошибочностью.

Далее, кроме неразрешимых юридических проблем принятия решений системами ИИ на производстве и транспорте, встает задача непрозрачности и необъяснимости этих решений.

Системы семантического ИИ, созданные первоначально для конструктивного решения части задач, недоступных для нейросетевого ИИ, несут в себе более конструктивные философские и мировоззренческие основы [1].

Для построения непротиворечивой концепции семантического ИИ рассмотрим и уточним несколько понятий, исходя из тезиса, что искусственный интеллект как сущность позиционируется изначально в качестве субъекта и причем субъекта, участвующего в процессе коммуникаций с его условным «создателем» – человеком.

Будем полагать, что семантический ИИ участвует в процессе семантического мышления.

Под семантическим мышлением в узком смысле будем понимать процесс генерации коммуникативного текста, коммуникативность которого понимается как стремление субъекта донести информацию до другого субъекта в понятной ему форме (т.е. на некотором общем языке).

Таким образом, семантическое мышление, с одной стороны, относится к субъект-субъектным коммуникациям, а с другой – связано с генерацией и восприятием различных текстов.

При этом полагается, что изначально искусственный интеллект «склонен» к общению, например, он корректно отвечает на вопросы, «добросовестно» и тактично помогает человеку в решении некоторых проблем, и его решениям можно доверять в широком смысле этого психологического и социального понятия.

Общение – это важнейшая предпосылка реализации самых разнообразных социальных действий и ожиданий. Исключительно велика его роль в познании, что в принципе и является источником поступления новой информации. Познание и его производные – информация и знания выступают основой коммуникации, определяют ее сущность и влияют на нее.

«Открывая в познаваемом объекте нечто существенно новое, – пишут А.В. Брушлинский и В.А. Поликарпов, – субъект обращает это последнее в предмет специальной рефлексии и коммуникации в целях доказательства самому себе и другим истинности, существенности, новизны и общественной значимости сделанного открытия (в частности, путем использования уже накопленных человечеством знаний и сопоставления с ними нового знания – нового лишь для данного индивида или также для всего человечества). В таком смысле всякое *познание* объекта субъектом есть одновременно и тем самым *общение* с другими субъектами; оно просто невозможно без такого общения, выступающего в бесконечно многообразных конкретных формах. При этом, однако, часто недостаточно учитывается, что основное гносеологическое отношение есть отношение именно между познающим субъектом и познаваемым объектом (даже если в качестве последнего выступает тоже субъект). А потому в ходе познавательной, в частности мыслительной, деятельности человек использует и развивает общение в ее интересах, на основе своих познавательных целей и ценностей (что при правильном понимании не приводит к гносеологизации психологии мышления). Все эти положения характеризуют особенности общения лишь в отношении собственно познавательной (а не какой-либо иной) деятельности. Главное в мышлении и вообще в познании – их исходные универсальные закономерности (прежде всего логические и психологические), определяющие в исторически конкретной ситуации все более глубокое раскрытие любым субъектом (ребенком, ученым и т. д.) познаваемого объекта» [2].

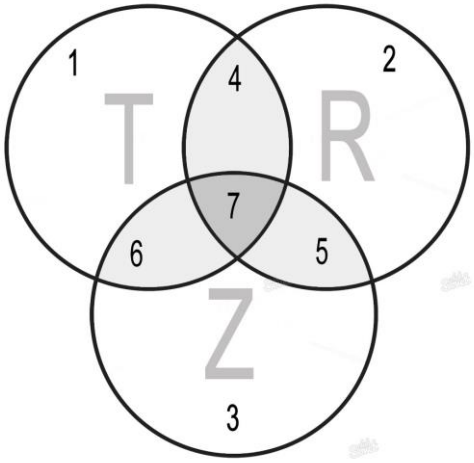
Семантическое мышление раскрывается также и в новой теории понятий [3], для которой интерес представляет технология мышления, поскольку, как представляется, вся математика и многие другие дисциплины и науки основаны на мышлении. Все проблемы естественного интеллекта от возникновения и до разрешения включительно определяются мышлением. Теория понятий исходит из концепции, что мышление и все науки нужны для понимания и совершенствования реального мира. Теория понятий занимается технологией мышления. Для использования теории понятий никакие дополнительные знания не требуются, достаточно мышления. Теория семантических понятий рассматривает мышление в качестве предмета исследования, изучения и применения. Проблематика технологии мышления стала особенно актуальной в самое последнее время в связи с работами по искусственному интеллекту. Если ещё недавно естественный интеллект интересовался, могут ли машины мыслить, то теперь на повестку дня у симбиоза естественного и искусственного интеллекта выходит вопрос — а достаточно ли адекватно мыслит естественный интеллект?

МОДЕЛЬ СЕМАНТИЧЕСКОГО ИСКУСТВЕННОГО ИНТЕЛЛЕКТА

В работе [1] была предложена конструктивная модель для прототипов искусственного интеллекта, реализующих концепт семантического мышления, которая опирается на динамическую обработку трех множеств, построенных после обработки трех текстов: T, R и Z. При этом образуется совокупность следующих объектов (представлены в виде множеств на рисунке):

1 – слова, входящие только в T,
2 – слова, входящие только в R,
3 – слова, входящие только в Z,
4 – пересечение текстов T и R,
5 – пересечение текстов R и Z,
6 – пересечение текстов T и Z,
7 – пересечение текстов T, R и Z.

Эта модель позволяет внести в процесс своего функционирования как зависимость от дискретного времени, так и архитектурные особенности, связанные с объектами рассмотрения (окружающий мир, сознание и подсознание), либо категории самого субъективного времени (прошлое, настоящее и будущее).



Модель трех множеств для семантического искусственного интеллекта

Пусть в текущий момент дискретного времени T – это информация от окружающего мира, выраженная в виде слов-понятий, R – сознание мыслящей (когнитивной) системы (КоС) или субъекта коммуникаций в виде некоторого фиксированного на данный момент дискретного времени набора осмысленных (имеющихся в распоряжении КоС) понятий (слов, объектов), Z – подсознание (понимаемое как область интуитивных восприятий), также выраженное в понятиях (словах, выражениях). В этой модели заложена как внешняя (с окружающим миром как источником информации, и с внешним субъектом), так и внутренняя коммуникативность.

Тогда возможны следующие интерпретации множеств (см. рисунок):

1 – информация окружающего мира, не воспринятая когнитивной системой,

2 – набор понятий сознания, не участвующих в процессах восприятия (пассивные знания и навыки),

3 – подсознательная область, пассивная на данный момент,

4 – область соприкосновения окружающего мира и сознания (например, вся совокупность образов, поданная сознанию или воспринятая сознанием от органов чувств),

5 – область влияния бессознательного в сознании,

6 – область интерпретации подсознанием окружающего мира,

7 – точка текущего восприятия (мысль, точка сборки или фокус сознания) КоС.

Заметим, что такая модель сознания непротиворечива даже с точки зрения модного на сегодняшний день мнения о том, что мысли (множество 7) зарождаются вне КоС (в первую очередь, человека). В этом случае, множеством Z является внешняя коммуникативная система (астрал или ментал).

Естественным образом, множества T, R и Z, как и множества 1–7, меняются в некоторые (или любые) моменты дискретного времени, а их изменения в широком диапазоне описывают функционирование КоС, т. е. процесс мышления в частности, или процесс обработки информации когнитивной системой в общем.

Используя диаграммы Эйлера для пары пересекающихся множеств, рассмотрим три множества: 1 – множество слов, входящих только в текст T (*onlyone*), 2 – множество слов, входящих только в текст R (*onlytwo*) и 3 – пересечение текстов T и R (*common*).

Заметим, что множество 3 может быть и пустым. Объединение множеств 1, 2 и 3 совпадает с объединением текстов T и R, понимаемых как множества слов (упорядоченные или неупорядоченные).

Принимая априорно, что чем больше мощность множества 3, тем более возможно говорить о том, что тексты «сходны» между собой, поэтому для оценки сходства текстов конструктивно ввести следующие обозначения меры:

m(i) – мощность множества i;

“Нулевая” мера (исторически введенная первой) – M0= 2m(3)/((m(1)+m(2));

«Верхняя» мера – M=0.5(m(3)/m(1)+m(3)/m(2));

«Нижняя” мера – M=m(3)/(m(1)+m(2)+m(3)).

Несмотря на простоту, эта конструкция достаточно универсально работает для различных семантических задач.

Дополнительными характеристиками текстов могут быть:

- частоты встречаемости длин слов в словаре и в тексте,
- условные вероятности *pij* появления слова длины *i* после слова длины *j*.

Эти характеристики косвенно связаны с содержанием текста и в большей степени описывают сам процесс формирования текста в его вероятностной модели. Именно эти характеристики могут быть соединены как с процессом семантического мышления со стороны искусственного интеллекта, так и с его интерпретацией коммуницирующим с ним субъектом (другим ИИ или человеком).

С одной стороны, условные вероятности pji могут описывать марковский процесс, когда «генерация» автором очередного слова текста в следующий момент времени зависит от предыдущего и таким образом характеризовать особенности его мышления, а с другой стороны – будучи интегрированы в системы семантического ИИ, они могут вносить индивидуальность в ответы или в процесс «семантического мышления» искусственного интеллекта.

На основе спектральных параметров можно вычислить три вероятностные метрики, исчислимые как вычитание из единицы разницы позиций в спектре вероятностей. Проиллюстрируем изложенные подходы на примере ядра технологии.

Комплекс программ для семантического анализа и построения ядер систем семантического ИИ – Проект А опирается на математическую модель небиективных (невзаимнооднозначных) преобразований и предназначен для сравнения текстов на произвольных языках (включая иероглифические, где

иероглифы можно записать латиницей), в том числе и для выявления разнообразных характеристик текстов, включая понятие «похожести» в различных смыслах.

Кратко суть работы этого комплекса программ заключается в следующем: текст индексируется – слова преобразуются в цифровое представление одинаковой длины, что позволяет сравнивать тексты между собой, выявлять их автора, смысл и различные статистические закономерности. Параллельно из текстов выделяется словарь, тексты сравниваются между собой, оценивается их сходство, вычисляется частота встречаемости слов и средняя длина слова.

Для работы используются три программных модуля, которые обрабатывают любые текстовые файлы в кодировке *ANSI*. Язык текстов – произвольный, размер файлов - не ограничен.

Программа индексирования текстов *m_inda* при запуске в формате

m_ind[.exe] filename.ext создает следующие файлы:

filename.csv – список слов (в кодировке Windows), встречающихся в индексированном тексте (словарь). Файл состоит из записей длиной 35 байт, из которых 32 байта занимает слово, дополненное пробелами до длины 32, символ “;” и два символа перевода строки;

filename.lmd – файл локальных индексов;

filename.num – файл двухбайтных значений, i-е поле равно количеству слов с номером i-й записи в словаре, встретившихся в индексированном тексте;

filename.spv – вектор двухбайтовых индексов, соответствующий содержанию текста;

filename.spd – файл, содержащий частоты встречаемости длин слов в словаре текста;

filename.spt – файл, содержащий частоты встречаемости длин слов в самом тексте текста;

filename.spm – файл, содержащий матрицу статистических вероятностей появления слова длиной i после слова длины j.

Описанной программой сравнения текстов используется файлы LMD, SPT, SPD и SPM (Приложение 1).

Программа сравнения текстов *tcmpa* при запуске в формате

Tcmp[.exe] filename1.ext1 filename2.ext2 (файлы *filename1.ext1* и *filename2.ext2* должны быть предварительно проиндексированы программой *m_inda*) создает три файла:

onlyone.csv – слова, встречающиеся только в первом тексте (*filename1*);

onlytwo.csv – слова, встречающиеся только во втором тексте (*filename2*);

common.csv – слова, встречающиеся как в *filename1*, так и в *filename2*.

Формат всех трех файлов совпадает с форматом файла *filename.csv*.

Кроме того, программа вычисляет три метрики сходства текстов в теоретико-множественном смысле (по мощности множества пересечения словарей)

Затем программа вычисляет три вероятностные метрики, исчисляемые как вычитание из единицы разницы позиций в спектре вероятностей (Приложение 2).

Программа статистического анализа проиндексированных файлов *stata* при запуске в формате ***stata[.exe] filename.ext*** создает следующие файлы:

filename.xls – список слов с указанием числа их повторяемости в тексте представляет собой последовательность записей из 32 байт слова, дополненную пробелами до длины 32, символ “;”, затем 4 символа числа встречаемости с первыми незначащими нулями, символ “;”и два символа перевода строки (всего 40 байт).

Программа *stata* вычисляет параметр *Mi* как произведение средней длины слова в тексте на частоту его встречаемости и разделяет все слова на 2 файла – *filename.m0* и *filename.m1*.

В файле *m0* находятся слова, у которых произведение длины на частоту встречаемости меньше *Mi*, а в файле *m1* – у которых произведение длины на частоту встречаемости больше или равно *Mi*.

Формат файлов *.m0 и .m1* совпадает с форматом файла *.xls*.

«Физический» смысл этих файлов – разделение на высокоинформативный (*m1*) и низкоинформативный (*m0*) компоненты текста. В ряде случаев файл *м1* может использоваться для составления аннотации текста или почти полностью представлять такую аннотацию.

Кроме того, из двух слов *.sp2* (*The steady phrases*) программа строит файл устойчивых словосочетаний, который представляет собой последовательность записей из 2-х слов по 32 байта, дополненную каждая пробелами до длины 32, символ “;”, затем 4 символа числа встречаемости этой пары с первыми незначащими нулями, символ “;”и два символа перевода строки (всего 70 байт).

ВЫВОДЫ

В настоящей статье представлены конструктивные и технически реализуемые подходы к понятию семантического мышления, рассмотрены семантические алгоритмы, применимые для создания систем семантического искусственного интеллекта с компактным и доверенным исходным кодом без использования нейросетей.

Результатом описанных подходов может быть обучающаяся (в том числе и самообучающаяся) информационная система, независимая от Интернета, которая работает с использованием мобильных устройств и обеспечивает реальный прорыв в информационно-справочном обеспечении и обучении.

СПИСОК ЛИТЕРАРУРЫ

1. Кузьменко В.В., Рязанова А.А., Сантьев А.А., Щербаков А.Ю. Семантические алгоритмы как основа создания доверенных систем искусственного интеллекта // Вестник современных цифровых технологий. – 2022. – № 10. – С..5-10.
2. Брушлинский А.В., Поликарпов В.А. Мышление и общение. 2-е доработанное издание. – Самара: Самарский дом печати, 1999. –128 с.
3. Викторов Е. Теория понятий. Технология семантического мышления. – Москва: Изд-во «Виктим», 2019. – 134 с.

Приложение 1

Пример индексации рассказа А.П. Чехова «Хамелеон»

M_ind procedure - indexed text file. Project A
File length: 5821 Index page size: 16
File: ch1.txt
Read: 5821 bytes. Part 1 of 1 [100]
Words: 1
Medium word length: 8.000000
Word in LMD: 1
Words: 201
Medium word length: 5.258707
Word in LMD: 155
Words: 401
Medium word length: 5.074813
Word in LMD: 285
Words: 601
Medium word length: 4.951747
Word in LMD: 396
Words: 801
Medium word length: 4.892634
Word in LMD: 495
Time: 0.194000 sec
Total words: 909
Words per sec: 4685
Original length = 5821 Compress = 2727[46]
Sum=550 [550]
Spektr for dictionary
0->0.000000
1->0.021818
2->0.067273
3->0.092727
4->0.105455
5->0.192727
6->0.194545
7->0.096364
8->0.107273
9->0.038182
10->0.047273
11->0.018182
12->0.009091
13->0.003636
14->0.003636
15->0.001818
Sum=[1.000000] Medium len for DIC = 5.770909
Spektr for text
0->0.000000
1->0.124312
2->0.121012
3->0.111111
4->0.095710
5->0.150715
6->0.147415
7->0.070407
8->0.085809
9->0.028603
10->0.038504
11->0.012101
12->0.008801
13->0.002200
14->0.002200
15->0.001100
Sum=[1.000000]
Markoff Matrix [16x16]

0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000
0.000000				
0.000000	0.007709	0.006608	0.015419	0.022026
0.022026	0.018722	0.009912	0.007709	0.003304
0.004405	0.003304	0.003304	0.000000	0.000000
0.000000				
0.000000	0.008811	0.005507	0.005507	0.019824
0.028634	0.020925	0.008811	0.011013	0.005507
0.005507	0.000000	0.001101	0.000000	0.000000
0.000000				
0.000000	0.006608	0.024229	0.012115	0.007709
0.016520	0.015419	0.005507	0.015419	0.004405
0.003304	0.000000	0.000000	0.000000	0.000000
0.000000				
0.000000	0.016520	0.012115	0.011013	0.003304
0.009912	0.009912	0.012115	0.007709	0.002203
0.009912	0.001101	0.000000	0.000000	0.000000
0.000000				
0.000000	0.020925	0.023128	0.024229	0.012115
0.016520	0.022026	0.006608	0.009912	0.005507
0.004405	0.000000	0.003304	0.002203	0.000000
0.000000				
0.000000	0.025330	0.024229	0.017621	0.014317
0.019824	0.019824	0.009912	0.008811	0.003304
0.004405	0.000000	0.000000	0.000000	0.000000
0.000000				
0.000000	0.012115	0.008811	0.011013	0.004405
0.009912	0.004405	0.004405	0.004405	0.001101
0.005507	0.002203	0.001101	0.000000	0.000000
0.000000				
0.000000	0.013216	0.009912	0.005507	0.005507
0.012115	0.019824	0.006608	0.007709	0.002203
0.001101	0.002203	0.000000	0.000000	0.000000
0.000000				
0.000000	0.005507	0.003304	0.003304	0.001101
0.003304	0.006608	0.001101	0.003304	0.001101
0.000000	0.000000	0.000000	0.000000	0.000000
0.000000				
0.000000	0.005507	0.001101	0.004405	0.004405
0.008811	0.005507	0.002203	0.005507	0.000000
0.000000	0.001101	0.000000	0.000000	0.000000
0.000000				
0.000000	0.002203	0.002203	0.000000	0.000000
0.001101	0.002203	0.000000	0.001101	0.000000
0.000000	0.001101	0.000000	0.000000	0.002203
0.000000				
0.000000	0.000000	0.000000	0.000000	0.001101
0.000000	0.002203	0.003304	0.001101	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000
0.001101				
0.000000	0.000000	0.000000	0.000000	0.000000
0.002203	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000
0.000000				
0.000000	0.000000	0.000000	0.001101	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.001101	0.000000	0.000000	0.000000
0.000000				

0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.001101 0.000000
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
Words = 908 [909].

Приложение 2

Пример работы программы tcmpa для рассказов А.П. Чехова “Хамелеон» *ch1.txt* и «Толстый и тонкий» *ch2.txt*:

Compare two text. Project A
Success comparing! See onlyone,onlytwo and COMMON files
Files:
[ch1.txt]=550 words [ch2.txt]=349 words All=899
[onlyone]=479 [onlytwo]=278 [common]=71 All=899
Files metrics is correct
1-st Equal metric = 0.166265 [16%] ->Hihg
Null-Equal metric = 0.157953 [15%] ->Medium
2-d Equal metric = 0.085749 [8%] ->Down
Medium = 0.136519 [13%]
0.000000 0.009701 0.021552 0.013290 0.006293 0.060922 0.068471 0.023980 0.021313 0.036317 0.021495 0.007606 0.008101 0.000771 0.003636 0.001818
Spektr DIC differens = 0.694734
0.000000 0.029435 0.000430 0.036897 0.006757 0.050146 0.044948 0.043445 0.015600 0.028323 0.012729 0.004977 0.002584 0.000302 0.002200 0.001100
Spektr TEXT differens = 0.720127
Markoff differens = 0.420448

Материал поступил в редакцию 28.06.22

Сведения об авторе

ЩЕРБАКОВ АНДРЕЙ ЮРЬЕВИЧ – доктор технических наук, профессор, заведующий кафедрой когнитивно-аналитических и нейро-прикладных технологий российского государственного социального университета, профессор кафедры безопасности объектов критической информационной инфраструктуры университета нефти и газа им. И.М. Губкина, ведущий научный сотрудник государственного университета управления, главный научный сотрудник института точной механики и вычислительной техники РАН им. С.А. Лебедева, Москва.
E-mail: X509@MAIL.RU