Машинная лингвистика: от перевода со словарём к нелинейным динамическим системам

Л.Л. Волкова
Кафедра «Информационные технологии и автоматизированные системы» Московского института электроники и математики Научно-исследовательского университета «Высшая школа экономики»
liliyavolkova@itas.miem.edu.ru

Аннотация. В статье дан краткий обзор ключевых этапов развития машинной лингвистики в разрезе анализа и синтеза текста. Выделены проблемы работы с языком, являющиеся фундаментальными ограничениями, отделяющими существующий уровень развития отрасли от качественно нового. Рассмотрены перспективные теории, предлагающие новый подход к рассмотрению языка и открывающие возможность заглянуть за барьер машинной лингвистики.
Ключевые слова: машинная лингвистика, анализ текста, автоматическая обработка текста, динамические системы, синергетика, фракталы, тесный мир

Машинная лингвистика, беря начало в пятидесятые с первых опытов машинного перевода, бурно развивается, производя новые методики автоматической обработки текстов, в тесной связи с аппаратами дискретной математики, дискриминантного анализа, задач оптимизации. Однако отрасль вплотную подходит к барьеру, для перехода через который нужно создать качественно новые системы, позволяющие иначе взглянуть на язык. Речь о переходе на более высокий по сравнению, в частности, с синтаксическими правилами уровень, на котором строение языка будет рассматриваться с точки зрения более общих законов, с естественно-языкового ракурса. Это хорошо подмечено в [5]: «Проблема не в том, что имеющиеся теории находят неверные ответы на поставленные ими вопросы, а в том, что они неверно ставят сами вопросы. Нужен подход, который мог бы дать разумный ответ на вопрос: «Как организован язык для выражения значения?», а не на вопрос: «Как организованы синтаксические структуры, если их рассматривать изолированно?»

Предыстория машинной лингвистики состоит из трудов составителей словарей и сводов правил для каждого языка, представляющих собой входные данные для многих методов обработки языков и текстов. Этот краеугольный камень заложен в фундамент отрасли, эклектичное строение которой кропотливо возводится и перестраивается инженерами и по сей день при самом деятельном участии лингвистов. Первые отголоски надвигающейся эпохи машинного перевода раздались ещё в XVII в., когда Лейбниц и Декарт выдвинули предположение о существовании некоего кода, соединяющего между собой слова разных языков [6]. Впервые мысль собственно о возможности машинного перевода высказал Чарльз Бэббидж, разработавший в 1836-1848 гг. проект механической цифровой аналитической машины: он предложил использовать память объемом в тысячу 50-разрядных десятичных чисел (по 50 зубчатых колес в каждом регистре) для хранения словарей [7]. Начало же машинной лингвистике было положено в 50-е гг. XX века с возникновением задачи машинного перевода. Так с создания подстрочников, сначала с крайне ограниченным словарём и минимальным количеством правил, в дальнейшем дополнявшихся и обрабатывавшимися постоянно совершенствующимися методиками в США [8], Советском Союзе [9], а также Великобритании, Канаде, Франции и Германии, началась машинная лингвистика.

При изучении языка и его формализации возникло разделение языков на формальные и естественные. Данные термины можно объяснить соответственно как созданные человеком и без его участия [3]. В зависимости от полноты информации о языке различают системы с ограниченным естественным языком (ОЕЯ) и естественно-языковые (ЕЯ) системы. Системы с ОЕЯ обладают знанием об ограниченном количестве лингвистических конструкций, работают с понятиями только данной предметной области. Как следствие, в них разбирается только некоторое подмножество предложений естественного языка. ЕЯ-системы лишены подобных недостатков, однако чёткой демаркационной линии пока проведено не было. Справедливо считается, что на данный момент созданы лишь ОЕЯ-системы, вплотную приближающиеся к решению ЕЯ-задач [1].

При употреблении термина «формальный язык» подразумевается изучение и описание искусственных языков, созданных людьми для специальных целей, к примеру, языков программирования. Но, как и естественные языки, они подчиняются грамматическим правилам и имеют неоднозначности [2]. Изучение формальных языков с математической точки зрения – предмет дискретной математики и теории формальных языков. Тогда как развитие классической («непрерывной») математики было обусловлено в первую очередь решением задач естествознания, главным образом физики, «дискретная» математика развивалась в связи с изучением законов и правил человеческого мышления, что и обусловило её применение в тех областях техники, которые так или иначе связаны с моделированием мышления. Мышление реализует себя в первую очередь в языке.

Теория формальных языков (ТФЯ) опирается на аппарат, базирующийся на понятии алгебраической структуры кольца [2], близкий к аппарату линейной алгебры, и на теорию графов. Многие задачи теории языков (например, задача определения языка конечного или магазинного автомата) сводятся к задаче о путях во взвешенных (размеченных) ориентированных графах, где множество меток имеет алгебраическую структуру полукольца. ТФЯ включает в себя теорию конечных автоматов и регулярных языков, а также контекстно-свободных языков, последние являются основой многих информационных технологий. Фундаментальным здесь является понятие магазинного автомата – механизма распознавания в классе контекстно-свободных языков, служащего математической основой технологий разработки синтаксических анализаторов для языков программирования. А от синтаксического анализатора недалеко до семантики формальных языков.

Согласно [2], можно выделить 3 аспекта изучения языка.

1) Синтаксис языка – рассмотрение языка как множества слов, которые являются последовательностями букв – символов фиксированного алфавита, коими могут быть как буква алфавита формального или естественного языка (таким буквам соответствуют слова в привычном понимании), так и предложения ЕЯ (или программы на языке программирования) в случае постулирования буквы как целого слова (лексемы). При этом не каждая последовательность букв будет словом, или лексемой, языка: корректность здесь обеспечивается синтаксисом языка, т.е. системой правил. Поскольку каждое слово характеризуется определённой структурой, специфичной для данного языка, необходимы механизмы перечисления, или порождения, слов и механизмы проверки принадлежности слова языку. Эти механизмы в первую очередь изучает ТФЯ.

2) Семантика языка предполагает сопоставление словам языка некого смысла. К примеру, математическая формула несёт в себе не только корректность записи, но и

некое значение. При соблюдении корректного синтаксиса можно создать случайный, некорректный смысл. Математическое понятие теории «смысла» появились сравнительно недавно.

3) Прагматика языка связана с теми целями, которые ставит перед собой человек, при этом они могут не совпадать с пространством семантических смыслов текста (к примеру, получение за речь некой суммы денег). Это дисциплина социально-философская, затрагивающая целеполагающую деятельность личности. Теорией формальных языков она не рассматривается.

Первой задачей обработки текста является определение его структуры. Только после перевода текста на естественном языке в его формальное представление, пригодное для машинной обработки [11], текст можно подвергать дальнейшей обработке и более глубокому анализу, выявлению закономерностей и пр. Здесь возникает вопрос о том, какие есть уровни изучения текста. По В.А. Звегинцеву, можно выделить следующие уровни изучения текста: дискурс (связанный текст); предложение; словосочетание; слово; морфема; слог; фонема; дифференциальный признак [12]. В системах автоматической обработки текста используются в основном первые пять уровней. Слоги иногда используются при представлении информации в морфологических словарях.

Установление структуры текста, или анализ в структурном смысле, включает в себя начальные, базовые этапы обработки документации [1]:

1) графематический анализ;
2) морфологический анализ;
3) предсинтаксический анализ;
4) синтаксический анализ;
5) постсинтаксический анализ;
6) семантический анализ.

Каждый этап представляет собой анализ определённого уровня абстракций текста: графематический – анализ графем, т.е. синтаксических и структурных единиц, морфологический – определение морфологических атрибутов и нормальных форм (т.е. исходной формы каждого слова, от которой образована данная словоформа), синтаксический – определение ролей слов и их связей между собой (существует множество лингвистических критериев, по которым могут быть выделены связи между отдельными словами в предложении [29]), семантический – анализ смысла текста; предсинтаксический и постсинтаксический анализ призваны выполнить сопутствующие синтаксическому анализу задачи, к примеру, снять неоднозначность (в частности, омонимию, некоторые механизмы снятия синтаксической и морфологической омонимии см. в [10, 32, 30, 28]), объединить или разъединить лексические единицы, в том числе разнесённые в предложении, и таким образом повысить качество синтаксического анализа. Современные системы проводят анализ текста поэтапно таким образом, что на вход каждого следующего этапа поступают результаты обработки предыдущего этапа. Самое высокое качество анализа текста на естественном языке получается при полном анализе текста.

Анализ, или разбор, текста – это прямое направление обработки текста. С ним тесно связано обратное направление – синтез текста, чьи этапы обратны этапам анализа. Анализ текста используется более всего: текст перед практически любой обработкой требует разбора – для извлечения информации (выделение ключевых слов и маркеров, определение эмоциональной окраски, заполнение фреймов по содержанию и др.), классификации и кластеризации, выявления характерных

особенностей текстов. Синтез же (следуя за анализом) используется в таких целях, как машинный перевод (здесь особенности языка учитываются в процессе анализа и синтеза, а внутреннее представление текста строится унифицированным и языконезависимым, чтобы синтезировать текст по правилам конкретных грамматики и синтаксиса) и генерация отклика диалоговых и рекомендательных систем.

Начальные этапы обработки текстов реализуются при помощи различных подходов и методик (к примеру, деревья, а также разработанный сотрудником IBM Джоном Бэкусом метаязык для описания языков программирования, по сути, мало чем отличающийся от контекстно-свободных грамматик [34, 35], который, будучи уточнён датским учёным Питером Науром, был назван формой Бэкуса-Наура, или бэкусовской нормальной формой [37, 38]). И анализ, и синтез текстов требуют повышения качества, которое достигается зачастую разработкой методик ускорения и оптимизации работы предсинтаксического и постсинтаксического анализа (к примеру, [41, 42]). Впрочем, все этапы начальной обработки текста открыты для доработки и переработки. На этапе морфологического анализа в некоторых случаях для повышения производительности системы выделяется только неизменяемая часть токена (стемматизация) [39, 31, 33], однако для систем с полным анализом текста такой информации недостаточно, поэтому в таких системах на этапе морфологии выделяются все возможные лексические характеристики каждого токена [36].

Машинный перевод сегодня не сводится к простой замене слов их иностранными аналогами, как это было когда-то. Существуют набирающие популярность за счёт сравнительной дешевизны улучшения результата статистические системы и системы, основанные на анализе текста. Впрочем, эти системы зачастую используют на разных этапах функционал друг друга. Текст документа разбирается «на запчасти», и на основании выделения зависимостей и подчинения в нём осуществляется перевод. К примеру, отечественные ученые из института Прикладной Математики им. Келдыша разработали свою нотацию для системы автоматического перевода «Crosslator 2», которая является расширением форм Бэкуса-Наура и позволяет учитывать согласование слов в предложении [41, 43, 40]. В данной нотации слово в предложении на естественном языке представляет собой входной символ грамматики, в котором указана нормальная форма слова и набор морфологических атрибутов [45, 44]. То есть осуществляется переход от естественного языка к его формальному представлению, анализ, решение прикладной задачи и синтез результата.

Существенным является доступ к базам корпусов языков и размеченных текстов для использования в программных средствах и машинного обучения. В этом смысле время работает на лингвистов, поскольку появляются новые размеченные корпуса текстов, дающие возможность для проведения исследований, тренировки и обучения методов [21].

В настоящее время большой удельный вес имеет статистический подход. Обучившись на внушительной выборке данных, порой меньшими затратами достигается сопоставимый со структурными, проникающими в суть конструкции языка методами результат. Уже не редкость системы с высоким качеством (хотя оценки разнятся на разных корпусах, как отмечено, к примеру, в [22]) и точностью порядка 80 % и даже 90 % (в частности, это отражено в результатах РОМИП – ежегодного Российского семинара по Оценке Методов Информационного Поиска [23]), и борьба подчас идёт за доли процентов.

Однако перспективы с точки зрения развития языка такой «плоский» подход не имеет, поскольку язык – материя гибкая и развивающаяся, что отражается в статистике лишь со временем, уже де факто. Будущее за системами де юро, синтезирующими знание, в противоположность решению насущных задач здесь и сейчас, достаточному, но не удовлетворительному с точки зрения фундаментальной науки. Статистические методы, вносящие огромный вклад в машинное обучение и выявление числовых закономерностей в текстах, скорее инструмент познания, но синтезирующей функции в своей базе они не несут.

С большим трудом преодолеваются, чаще, увы, статистически, такие «граничные условия», как снятие неоднозначностей (большой процент ошибок разбора текстов приходится на омонимию), непрозрачность в связи с всё ещё недостаточным развитием семантического анализа связей в текстах, ошибки в данных (Errare humanum est, или как обрабатывать слова с четырьмя ошибками в трёх буквах), наконец. Неизвестные слова – новые, заимствованные, имена собственные, окказионализмы. Даже в статистическую систему, способную, предположим, собрать в корпусе текстов информацию об употреблении нового термина или, более того, становящегося популярным окказионализма – слова или даже модели управления (к примеру, непереходный глагол с переходным управлением: «его ушли с работы»), – нужно вложить знания о механизмах работы языка. А это уже рассмотрение по существу, где сбор статистики лишь вспомогательный инструмент (при этом не опорный для увечной разработки).

Пусть в ходе конвергентной эволюции язык подчас становится похож на другой, не родственный ему вид, – развитие и внутренние законы всё же специфичны для каждого конкретного языка. И именно умение синтезировать новые правила и, сообразуясь с ними, знания – та черта, которая отличит отечественные и зарубежные системы, сиюминутно функционирующие по принципу «авось» (см. известный пример про трёх белых котят и одного политкорректно чёрного), от подлинных демиургов машинной лингвистики, чей приход мы с нетерпением ждём и по мере сил стараемся приблизить.

А чтобы разработать таких демиургов (др.-греч. Δημιουργός – мастер, ремесленник, творец), то есть хорошие системы из гильдии машинной лингвистики, усвоившие секреты ремесла, ноу-хау языка, а именно модели формирования языковых конструкций и их генерации, нужно потрудиться над умением применять, агрегировать и синтезировать системы на правилах (если не сказать «правильные» системы) функционирования языка и тем самым подняться на новую ступень на пути от систем с ограниченным естественным языком к естественно-языковым системам, лишённым многих ограничений [1] именно благодаря большей свободе синтеза. Тогда мы, повернув вспять ход этимологической истории термина «демиург», начнём новый виток спирали: обратно от концепции всемогущего творца к именованию класса ремесленников, как это и было в античном обществе.

Впрочем, предсказуемость языковых конструкций будет ограничена, какими бы оптимистичными ни были взгляды неофитов, энтузиастов и экспертов по искусственному интеллекту. Язык, возможно, является **нелинейной динамической системой**.

Как отмечено в [4], до 60-х гг. предполагалось, что есть два класса процессов. Первые описываются динамическими системами, в которых будущее однозначно определяется прошлым, вторые – системами, в которых будущее не зависит от прошлого. В первых, детерминированных, предполагалась полная предсказуемость,

во вторых – полностью случайные события. В 70-е гг. был выделен третий класс процессов: формально они описываются динамическими системами, но их поведение может быть предсказано только на небольшой промежуток времени. Ограниченность горизонта прогноза связана с чувствительностью к начальным данным (малые изменения могут повлечь лавинообразные изменения и, в конечном итоге, катастрофу) и с непериодическим движением в детерминированных системах, т.е. динамическим хаосом. Траектории движения частиц расходятся со скоростью, определяемой так называемым ляпуновским показателем. Но те же явления можно отметить и в историческом движении языка и его частиц.

Язык развивается вместе с человеком, он неразрывно связан с коллективным опытом и сознанием [24]. Под воздействием миграций, войн, научно-технического прогресса (в частности, появления радио, телевидения, компьютеров, интернета) языки видоизменяются, приобретая новые черты [25]. И некоторые факторы, к примеру, переселение народов [26], могут вызвать лавинообразные изменения в языке, будь то орфография или семантика. Даже символьные обозначения, как, например, путь от египетских иероглифов и их наследников, финикийской клинописи, к символьной записи иврита, от которой произошло арабское письмо, и греческому алфавиту, от которого, в свою очередь, произошли кириллица с одной стороны и латиница с другой. Подобные явления происходят в пространстве смыслов слов, моделей управления и, далее, построения речи. А каждое изменение входных данных влечёт всё большую разницу между выходными результатами, то есть языками и правилами в них [26].

Один из основных признаков системы состоит в её структурированности, в целесообразности связей между её элементами [3]. Структурные взаимосвязи в тексте, результате функционирования сложной системы языка, описываются, в первую очередь, синтаксисом и семантикой. На примере их сложности и подчас лавинообразного роста вариантов можно увидеть их сложность и нелинейность. Рассмотрим ряд приложений автоматической обработки текстов. В частности, рассмотрим систему, производящую по словоформе поиск нормальной формы слова. В системах, в которых поиск организован в упорядоченном по частотности дереве (по первой букве по убывающей частоте, по второй букве и т.д.), скорость нахождения допустимых словоформ выше, чем при поиске в словаре, а главное, в них высвечивается такая проблема, как увеличение допустимых конструкций на порядки при увеличении длины строки (при генерации всех возможных цепочек из алфавита рост количества допустимых цепочек, если грамматика не учитывает правил отбраковки некорректных для человеческого языка цепочек, будет экспоненциальным). А поскольку язык развивается, такой словарь словоформ никогда не будет полным. Взять хотя бы приставки и суффиксы, которые приводят к катастрофическому разрастанию и без того немалого «ядра» языка в виде стемов до бесконечного количества словоформ (будем считать это на данный момент машинной бесконечностью, поскольку система, владеющая всеми порождающими правилами для корректных цепочек языка, не существует). При этом если говорить только о рассмотрении грамматически верного текста, то на существующих базах правил ещё можно получать хорошие результаты. А теперь возьмём тексты с ошибками. Здесь требуются методы предположения об ошибке. Пусть мы допускаем 1-3 ошибки. А если их больше, речь может идти не только о неграмотности (здесь можно было бы положить, что настолько неграмотные тексты разбору не подлежат), но и о словах, не наличествующих в словаре. К ним могут относиться и новые термины, и

малоупотребимые слова, а если речь идёт не о сухих текстах с минимумом лексики, а о художественной литературе или публицистике, то и объём словаря для полного разбора требуется больший, и редких слов больше, а окказионализмы и подобные им явления в словаре всё равно не фиксируются. Следовательно, разбор требует иного, не словарного подхода, а скорее словообразовательного, далее – образующего предложения, семантические структуры и т.д.

Словообразование само по себе представляет собой нелинейную систему с изменяющимися и достаточно свободно соединяющимися структурными элементами, то же касается и синтаксиса. Опущенные члены предложения существенно усложняют разбор, а вкупе с растущим за счёт неоднозначностей (число которых дополнительно увеличивается за счёт отсутствующих узлов дерева разбора) количеством претендентов на корректный вариант разбора задача синтаксического анализа усложняется дополнительно, что показывает, насколько несовершенны ещё методы разбора, при всех существенных достижениях по улучшению качества разбора.

Многое в строении и истории развития естественных языков свидетельствует о том, что человеческие языки являются носителями черт нелинейных динамических систем. Возможно, развитие теории синергетики и прогноза [4] претворит в реальность пророчество Анри Пуанкаре о том, что в будущем можно будет предсказывать новые физические явления, исходя из общей математической структуры описывающих эти явления уравнений. Возможно, и развитие человеческих языков тоже можно будет предсказать и полностью описать математической моделью. Дело за экспериментами и фундаментальными исследованиями.

Новые теории в машинной лингвистике

Как отмечается в [3], одним из направлений перестройки в высшем образовании и систематизации науки является преодоление недостатков узкой специализации, усиление междисциплинарных связей, развитие диалектического видения мира, системного мышления.

Анализ текстов на сегодняшний день в некотором роде представляет собой работу с фреймовой моделью [27]: предложения и языковые единицы рассматриваются по некоторому ограниченному множеству шаблонов в рамках стереотипных ситуаций. В слоты фрейма (возможно, содержащие ссылки на другие фреймы) записываются извлечённые данные о структуре слов (выделяются начальные формы и заполняется список параметров в зависимости от части речи), предложений (синтаксические роли и связи), содержания (переход к семантике; сейчас решаются такие задачи, как извлечение информации из текстов и заполнения слотов фрейма: к примеру, для близкой к некому рассматриваемому шаблону статьи о государстве из текста выделяются сведения о площади, численности населения, климате и т.д.). Однако возможность развития системы правил ограничена самой шаблонностью такого подхода. «Кругозор» такой системы расширяется разовым пополнением базы правил, и это всё же система с ограниченным естественным языком.

Существуют новые для машинной лингвистики математические теории, перспективные для познания устройства языка, в частности, теория фракталов и теория тесного мира.

Фракталы

Теория фракталов, как правило, рассматривается как подход к статистическому исследованию, который позволяет получать важные характеристики информационных потоков, не вдаваясь в детальный анализ их внутренней структуры и связей. Одним из основных свойств фракталов является самоподобие (скейлинг). Как показано в работах С.А. Иванова [20], для последовательности сообщений тематических информационных потоков в соответствии со скейлинговым принципом количество сообщений, резонансов на события реального мира пропорционально некоторой степени количества источников информации (кластеров) и итерационно продолжается в течение определенного времени. Точно так же, как и в традиционных научных коммуникациях, растущее множество сообщений в Интернет по одной тематике во времени представляет собой динамическую кластерную систему, возникающую в результате итерационных процессов. Этот процесс объясняется републикациями, прямой или совместной цитируемостью, различными публикациями – отражениями одних и тех же событий реального мира, прямыми ссылками и т.д. Кроме того, для большинства тематических информационных потоков наблюдается увеличение их объемов, причем на коротких временных интервалах рост линейный, а на длительных – экспоненциальный.

Фрактальная размерность в кластерной системе, соответствующей тематическим информационным потокам, показывает степень заполнения информационного пространства сообщений в течение определенного времени. Между количеством сообщений и кластеров проявляется свойство сохранения внутренней структуры множества при изменении масштабов его внешнего рассмотрения. По мнению С.А. Иванова, все основные законы научной коммуникации, такие как законы Парето, Лотки, Бредфорда, Зипфа, могут быть обобщены именно в рамках теории стохастических фракталов.

Исследование информационного потока в его развитии даёт возможность проследить изменения в корпусах текстов и в языке. При проведении информационно-аналитических исследований на основе обработки информационного потока, формируемого в Интернет [17], особо актуальной оказывается задача автоматического извлечения из текстов фактографической информации [18]. При этом ввиду значительных объемов и динамики информационных потоков контент-анализ осуществляется сегодня с использованием современных информационно-аналитических систем. По наблюдению Д.В. Ландэ [15], следует признать, что изначальные парадигмы поисковых систем и контент-анализа, сформированные десятилетия тому назад, уже не отвечают реальной ситуации. Один из подходов к решению задачи извлечения фактов из текстовых документов и выявления их взаимосвязей базируется на технологии контент-мониторинга, который можно рассматривать как непрерывный во времени содержательный анализ информационных потоков с целью получения необходимых качественных и количественных информационных срезов. Именно непрерывная аналитическая обработка сообщений является самой характерной чертой этого подхода, который позволяет извлекать факты из тестов, выявлять новые понятия, формировать разнообразные статистические отчеты. Названные задачи сегодня охватываются двумя основными технологиями – извлечением фактографической информации из текстов (Information Extraction [18]) и глубинным анализом текстов (Text Mining [19]).

Однако, основываясь на структуре текста, можно предположить, что фрактальные свойства присущи и внутреннему устройству текста. Самоподобие в

текстах на естественном языке не формализовано, и это требует глубинного исследования в тесной связи с правилами построения текста.

Теория тесного мира

Проведённые недавно испанскими и американскими учёными исследования показали, что свойства тесного мира (англ. small world), наблюдаемые во многих биологических и социальных системах, присущи и текстам [46]. Более того, построение тесного мира высвечивает такие свойства текста, как самоподобие и стойкие тенденции связности слов [47-49].

Слова в человеческом языке взаимодействуют в предложениях не случайным образом: они связаны. Человек способен построить огромное количество предложений из ограниченного количества структурных единиц, процесс построения крайне быстр и всё ещё недостаточно изучен, чтобы создавать достаточно близкие к человеческой способности прототипы генерирующих речь систем. Совместное употребление слов в предложениях отражает организацию языка и, таким образом, может способствовать его описанию в терминах графа взаимодействий слов. Такие графы, как показано в [46], отражают два важных свойства, открытых для различных по происхождению систем:

1) эффект малого мира (в частности, среднее расстояние между двумя словами, или среднее минимальное количество дуг (связей) на пути от одного слова к другому, оценивается в пределах от 2 до 3, хотя человеческий мозг способен хранить тысячи и тысячи таких связей);

2) новое слово (вершина) при расширении графа будет связано с существующим словом с вероятностью, пропорциональной количеству связей последнего с другими словами.

На основании исследований в [46] показано, что существуют доселе неизвестные особенности организации языка, которые могут отражать эволюцию и социальную историю языков и истоки их гибкости и комбинаторной природы. Таким образом, развитие данного направления исследований перспективно для углубления знаний о структуре языка.

Свойство самоподобия информационного пространства (см. выше) [13-15], может быть более подробно изучено по графам тесного мира. Выявление самоподобия в текстах и информационных потоках позволит повысить эффективность анализа текстов, выявления зависимостей и извлечения информации, что не только полезно для решения прикладных задач компьютерной лингвистики, но и видится существенным подспорьем в установлении общих законов языка и построения речи. Таким образом, как отмечено в [50], теория тесного мира позволит выявлять и изучать особенности построения текста, а также его семантическую структуру.

В заключение стоит сказать, что рассмотрение языка как нелинейной динамической системы (см. ранее) также может внести весомый вклад в машинную лингвистику. Основываясь на прочном фундаменте проведённых исследований языка и текстов и разумно используя новые для машинной лингвистики взгляды на естественно-языковой мир, найдя тот кардинально отличающийся подход к рассмотрению использования аппарата языка, мы, в лучших традициях Карла Сагана, совершим переход от последовательности простых чисел к объёмному восприятию информации, дающей нам ключ к удивительным путешествиям к центру Знания о Языке.

Список литературы

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. — 272 с.
2. Белоусов А.И., Ткачев С.Б. Дискретная математика: Учеб. для вузов / Под ред. В.С. Зарубина, А.П. Крищенко. — 4-е изд., исправл. — М.: Изд-во МГТУ им. Н.Э. Баумана, 2006. — 744 с. (Сер. Математика в техническом университете; Вып. XIX).
3. Перегудов Ф.И., Тарасенко Ф.П. Введение в системный анализ: Учеб. пособие для вузов. — М.: Высш. шк., 1989. — 367 с.
4. Малинецкий Г.Г. Синергетика и прогноз. // Будущее прикладной математики. Лекции для молодых исследователей / Под ред. Г.Г. Малинецкого. — М.: Едиториал ЦРСС, 2005. — 512 с. — С. 374-403.
5. Виноград Т. Программа, понимающая естественный язык. — М.: Мир, 1976. — 296 с.
6. Hutchins, J. Machine Translation: past, present, future. — Chichester: Ellis Horwood, 1986.
7. Апокин И.А., Майстров Л.Е., Эдлин И.С. Чарльз Бэббидж. — М.: Наука, 1981. — 128 с.
8. MacDonald N. Language translation by machine - a report of the first successful trial. // Computers and automation. 1954, v. 3, № 2, p. 6-10.
9. Кулагина О.С. Исследования по машинному переводу. — М.: Наука, 1979. — 320 с.
10. Мельчук И.А. Опыт теории лингвистических моделей «Смысл→Текст». — М.: Школа «Языки русской культуры», 1999. — I-XXII, 346 с.
11. Крысин Л.П. Лингвистический процессор для сложных информационных систем. — М.: Наука. — 1992.
12. Попов Э.В. Общение с ЭВМ на естественном языке. — М.: Наука, 1982. — 360 с.
13. Ландэ Д.В. Фрактальные свойства тематических информационных потоков из Интернет. // Регистрация, хранение и обработка данных. — Киев, 2006. — Т. 8, No 2. — С. 93 - 99.
14. Додонов А.Г., Ландэ Д. В. Самоподобие массивов сетевых публикаций по компьютерной вирусологии. // Реєстрація, зберігання і обробка даних, 2007, Т. 9, — N 2. — С. 53-60.
15. Ландэ Д. Фракталы и кластеры в информационном пространстве. // Корпоративные системы, №6'2005. — С. 35-39.
16. Ландэ Д.В. Выявление понятий и их взаимосвязей в рамках технологии контент-мониторинга. // Реєстрація, зберігання і обробка даних, 2006, - Т. 8, - N 4.
17. Ландэ Д.В. Основы интеграции информационных потоков. — К.: Інжиніринг, 2006. — 240 с.
18. Ralph Grishman. Information extraction: Techniques and Challenges. In Information Extraction (International Summer School SCIE-97) // Springer-Verlag. — 1997.
19. Гершензон Л. М., Ножов И. М., Панкратов Д. В. Система извлечения и поиска структурированной информации из больших текстовых массивов СМИ. Архитектурные и лингвистические особенности. Труды Международного семинара «Диалог'2005» (Звенигород, 1–6 июня 2005 г.). — М.: Наука, 2005.

20. Иванов С.А. Стохастические фракталы в Информатике. Научно-техническая информация. — Сер. 2. — 2002. — № 8. — С. 7–18.

21. Баранов А.Н. Введение в прикладную лингвистику. — М.: Издательство ЛКИ, 2007. — 360 с.

22. Sharoff S., Nivre J. The proper place of men and machines in language technology. Processing Russian without any linguistic knowledge. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25 - 29 мая 2011 г.). Вып. 10 (17). — М.: Изд-во РГГУ, 2011.

23. Маслов М. Ю., Пяллинг А.А. КС-классификатор и дорожка классификации веб-сайтов РОМИП'2010. // Труды РОМИП 2010. Под ред. И.С. Некрестьянова. — Казань, 2010. — 200 с. — С. 80-97.

24. Абаев В. И. О термине «естественный язык». // Вопросы языкознания. — М., 1976, № 4. — С. 77-80.

25. Практическая транскрипция личных имен в языках народов мира / [отв. ред. Э.С. Клышинский]; Ин-т прикладной математики им. М.В. Келдыша РАН. — М.: Наука, 2010. — 679 с.

26. Плунгян В.А. Почему языки такие разные. — М.: АСТ-ПРЕСС КНИГА, 2010. — 272 с.

27. Минский М. Фреймы для представления знаний. — М.: Энергия, 1979. — 152 с.

28. Ножов И.М. Реализация автоматической синтаксической сегментации русского предложения. Диссертация на соискание ученой степени кандидата технических наук, М.: РГГУ, 2003.

29. Тестелец Я.Г. Введение в общий синтаксис. М.: РГГУ, 2001 г. 798 с.

30. Neumann G., Brauny C., Piskorski J. A Divide-and-Conquer Strategy for Shallow Parsing of German Free Texts // ANLC '00 Proceedings of the sixth conference on Applied natural language processing, 2000, pp. 239-246.

31. Porter M.F. An algorithm for suffix stripping // Program, 1980, Vol. 14, №3, pp. 130-137.

32. Przepiorkowski A., Buczynsk A. Shallow Parsing and Disambiguation Engine. [Proceedings of 3rd Language & Technology Conference, 2007]

33. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // [Proceedings MLMTA, 2003]

34. Рассел С., Норвиг П. Искусственный интеллект. Современный подход. Москва — С-т Петербург — Киев: Вильямс, 2006. — 1093 с.

35. Формальные модели анализа и распознавания языковых структур. [Сайт Международной конференций по компьютерной лингвистике «Диалог»].

36. Черненьков Д.М., Клышинский Э.С. Формальный метод пополнения словарей морфологического анализа с использованием несловарной лексики // Вестник компьютерных и информационных технологий, №3, 2011, сс. 22-28

37. Backus J.W. The syntax and semantics of the proposed international algebraic language of the Zurich ACM-GAMM Conference // Proceedings of the International Conference on Information Processing, UNESCO, 1959, pp. 125–132.

38. Knuth D.E. Backus Normal Form vs. Backus Naur Form // Communications of the ACM, 1964, Vol. 7, Issue 12, pp. 735–736.

39. Lovins J.B. Development of a stemming algorithm // Mechanical Translation and Computational Linguistics 11, 1968, pp. 22–31.

40. Востриков А.В., Клышинский Э.С., Морозов С.Н., Манушкин Е.С., Максимов В.Ю. Исследование метода автоматической генерации правил фрагментарного анализа // Сб. тезисов международной конференции MegaLing'08, Партенит, 2008

41. Галактионов В.А., Мусатов А.М., Мансурова О.Ю., Ёлкин С.В., Клышинский Э.С., Максимов В.Ю., Аминева С.Н., Жирнов Р.В., Игашов С.Ю., Мусаева Т.Н. Система машинного перевода «Кросслятор 2.0» и анализ ее функциональности для задачи трансляции знаний // Препринт ИПМ им. М.В.Келдыша РАН. Москва, 2007.

42. Гладкий А.В. Синтаксические структуры естественного языка, Изд. 2 — М.: ЛКИ, 2007. С. 12-15.

43. Жирнов Р.В., Клышинский Э.С., Максимов В.Ю. Модуль фрагментарного анализа в составе системы машинного перевода. Crosslator 2.0 // Вестник ВИНИТИ, 2005 г. НТИ. Серия 2. №8 С. 31-33

44. Клышинский Э.С., Манушкин Е.С. Математическая модель порождения правил синтаксической сегментации // Сб. трудов второй Всероссийской конференции «Знания – Онтологии – Теории», Новосибирск, 2009, Том 2., С. 182-186.

45. Манушкин Е.С., Клышинский Э.С. Метод автоматического порождения правил синтаксической сегментации для задач анализа текстов на естественном языке // Информационные технологии и вычислительные системы, 4, 2009 г., С. 57-66.

46. The small world of human language. Ramon Ferrer i Cancho and Ricard V. Solé. Proceedings of the Royal Society, Series B, Vol. 268, No. 1482, pp. 2261-2266, 2001.

47. Amaral, L. A. N., Scala, A., Barthélémy, M. & Stanley, H. E. Classes of behaviour of small-world networks. Proc. Natl Acad. Sci. USA 97, p. 11149-11152, 2000.

48. Watts, D. J. & Strogatz, S. H. Collective dynamics of “small-world” networks. Nature 393, p. 440-442, 1998.

49. Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. Nature, 406, p. 378-381, 2000.

50. Волкова Л.Л. Приложения теории тесного мира в компьютерной лингвистике. // Сборник трудов третьей Международной научно-практической конференции «Модель подготовки специалистов новой формации, адаптированных к инновационному развитию отраслей». — Душанбе, 2012.