Самоорганизация в семантическом пространстве глаголов естественного языка

Коршаков А. В.

ВНЦ «Курчатовский институт», Москва, narano@mail.ru

Аннотация. Объектом анализа в статье являются множества английских глаголов. Глаголы рассматриваются как объёмные тела в неком «семантическом» пространстве, на котором введена определённая метрика. В основе определения расстояний в упомянутом пространстве положено число существующих адекватных взаимообратных переводов у обрабатываемых слов-глаголов на ряд (13) других индоевропейских языков. Целью такого исследования было подтвердить или опровергнуть существование нерегулярных структур в заполнении объёмными телами (глаголами) или кластеров, а также прояснить особенности строения семантического пространства, влияющие на расположение в нём слов. Для этих целей использовалась программная эмуляция нейронной сети Хопфилда, в которой каждый нейрон соответствовал одному из рассматриваемых глаголов. Веса нейронов задавались в соответствии с «расстоянием» между понятиями в смысле выбранной метрики. Рассмотрено большое число слов, среди которых выделяется группа «многозначных» и группа с «конкретным» значением или смыслом. Причём последняя группа не образует очевидных кластеров в пространстве глаголов. Глаголы естественных языков, символизирующие набор базовых, имеющих особый смысл для повседневной жизни действий, на которые способен человек, распределены равномерно по семантическому пространству.

1. Введение

Компьютерная обработка текста и человеческой речи, в прикладном смысле, сталкивается со многими проблемами, связанными с многозначностью слов. Одним из аспектов этой проблемы является (возможно, наиболее острый её аспект) работа с глаголами, так как многие из них и, особенно группа наиболее часто встречающихся в речи, имеет множество переводов с одного языка на другой. Например, английский глагол «Take» (*«Брать»*) может иметь более чем 20 различных выражений в других языках.

Всё множество глаголов может быть упорядочено в соответствии с убыванием числа возможных переводов на иностранные языки, и только около 300-й позиции этого списка число различных переводов внутри семейства индоевропейских языков падает до 2-4. Возможно, что такая картина сохраняется и для языков других семейств и структур. Такое явление легко прослеживается при анализе многоязыковых online-словарей, существующих в Интернете. Выбор необходимого и адекватного перевода из набора, который предоставляет словарь (в случае перевода с одного языка на другой), – в некоторых случаях достаточно трудная задача для человека, а компьютер здесь сталкивается и вовсе с непреодолимыми трудностями. Человек при распознавании речи использует, например, контекст (а ещё интонации голоса, мимику и многое другое), в котором

произнесено слово, чтобы исключить двусмысленность. Под контекстом здесь следует понимать некий блок окружающего слово текста или состояние окружающей среды (например, при анализе смысла одиночного, короткого высказывания – слова). Компьютер также должен быть способен решать подобные задачи. Это требует формализации процедуры обработки нечёткой информации, содержащейся в такой категории как контекст.

Обычно слова рассматриваются только как состоящие в определённом словаре, и никаких иных свойств, кроме того, что они собственно означают, и ещё, быть может, их порядкового номера в последовательности, упорядоченной в соответствии с алфавитом, им не приписывается. Такие словари, составленные по принятым в современном мире правилам, представляют собой список состоящих в них слов и статей, относящихся к ним. Такие пары «слово-статья» более или менее независимы друг от друга. В лучшем случае их связь выражается в указании синонимов друг друга. Более тонкие взаимоотношения слов и представленных ими понятий рассматриваются как явления, проявляющие себя только в огромных корпусах текста, содержащих миллионы слов, и нередко только в статистическом смысле [1]. Устроенная в соответствии с этими принципами процедура работы с языковой информацией кажется трудоёмкой и не слишком удобной, так как каждый раз приходится обращаться к большим объёмам текстовых данных, имеющих лишь опосредованное отношение к выполняемой прикладной задаче, – например, переводу (межъязыковому) конкретного текста с конкретным смыслом. Трудоёмкость становится очевидной при рассмотрении задач, требующих решения «на лету», например, в задачах интерпретации речи. Необходим более лаконичный способ представления лингвистической информации. Есть основания полагать, что такой способ существует, так как кажется сомнительным, чтобы человеческий мозг, вполне успешно справляющийся с оговоренным кругом задач, каждый раз при работе с речью прибегал бы к статистическому анализу всей, хранящейся в памяти лингвистической информации.

Можно выделить два пути, ведущих к получению данных относительно вида представления лингвистической информации в сознании человека. Среди существующего множества методов изучения нервных процессов наиболее эффективными можно назвать техники трехмерной визуализации активности нервной ткани [2, 3]. Это первый путь. Он состоит в изучении нервной системы, например, при помощи томографов ядерного магнитного резонанса. Он даёт определённые знания о процессах, протекающих в мозге человека во время обработки лингвистической информации, но сопряжён с целым рядом трудностей различного характера. Другой путь заключается в том, чтобы исследовать непосредственно продукт этих (языковых) процессов, то есть речь, учитывая фактор представляемой в ней информации.

Предыдущие исследования [4] показывают, что слова и отражаемые ими понятия, связаны сильнее, а, главное, глубже, чем это может показаться при рассмотрении структур «словарного» типа. Нужно скорее говорить не о жёсткой связи слов и отражаемых ими понятий, а об их «нечёткой» связи, не оперируя при этом только терминами «одно и тоже», для синонимов, и, «означают существенно разное», для слов, не имеющих общих значений даже в переносных смыслах.

В качестве подтверждения существования сильной связи слов, их смыслов и даже составляющих их частей, можно привести тот факт, что частота использования символов языка, то есть составных частей речи (звуки речи, фонемы, собственно слова) в различных языках (в [5] это показано на примере слов повседневного лексикона индоевропейской группы языков), подчиняется одинаковой и строгой функциональной зависимости [5-7]. Кривые частот встречаемости символов монотонно падают от максимального до почти нулевого значения, соответствующего наиболее редкому символу. Кривые, полученные, более чем для 1000 слов упомянутой группы языков, совпадают с высокой точностью, и рассогласования не превышают 5%, что укладывается в коридор статистической погрешности [8]. Для каждого из рассматриваемых языков последовательность символов специфична, что является следствием различия языков.

Употребление определённых фонем в речи является подписью каждого языка. Некоторые языки избегают употребления звуков, свойственных оставшемуся большинству.

Такие зависимости демонстрируют строгую математическую основу процесса формирования слов. Вероятнее всего, алгоритмы, применяемые мозгом для распознавания и генерации речи, независимы от конкретного языка, и могут настраиваться на определённую кодировку фонетических символов с различной степенью эффективности.

2. Многообразие глаголов в тезаурусах естественных языков

Человеческий язык или естественный язык – это открытая система, готовая принимать или «изобретать» новые слова или же отбрасывать старые слова. Множество существующих слов, имеющих специальное значение, могут быть просто неизвестны среднестатистическому пользователю лексической и семантической основ языка. То есть, язык слишком мощное средство, чтобы охватить его целиком. При исследовании такой сложной и динамичной системы, следует каким-то образом ограничить область анализа, которая, тем не менее, должна содержать, основные свойства языка. Одним из путей такого ограничения является подход, при котором ограничиваются работой с малой частью языка, соответствующей какой-то конкретной прикладной проблеме. Например, если иметь в виду задачу распознавания речи, то такой малой «областью интересов» семантики языка могут быть имена числительные и географические названия. Область применения таких систем распознавания – автоматические информационные бюро в аэропортах и голосовой набор номеров в аппаратах сотовой связи. К сожалению, на сегодняшний день, технические решения, относящиеся к этим областям применения распознавания, используют некие механистические принципы, а не внутренние закономерности языка, и, как правило, работают не идеально.

Возможно, что это явление есть следствие неверного выбора области семантики, корректно представляющей язык и отношение между его частями – словами. Для успешного решения задачи предлагается использовать критерием выбора подгрупп семантик для анализа не критерий «области практического применения», а критерий «исчерпывающего описания законов взаимодействия элементов языка внутри системы». То есть, необходимо для начала сориентировать

русло решения от прикладных проблем в сторону теоретического обоснования решения.

Кажется разумным, в этом контексте, обратиться к такому словарному множеству языка как глаголы.

Глаголы составляют от 20% до 25% всех слов (при рассмотрении выборки из 10000 наиболее часто используемых слов английского, немецкого, голландского и французского языков) [9]. В противовес именам числительным глаголы чрезвычайно полисемичны (многозначны), то есть, их употребление возможно в языке во множестве значений и контекстов. Это является одним из обстоятельств (а может и основным обстоятельством), существенно усложняющим разработку технических средств и программного обеспечения обработки текста и речи. Сюда относятся все виды устройств и программного обеспечения, работающие только с текстом, только с речью, их комбинации, программы по генерации текста по распознанной речи и генерации речи по имеющемуся текстовому файлу. С другой стороны, если посмотреть на явление полисемии (многозначности), как на характеризующее язык или мышление вообще, то можно использовать его для построения системы отношений между глаголами, а в последствии и между словами вообще, что позволит построить более прогрессивные алгоритмы для работы с естественным языком и алгоритмы понимания человеческой речи.

3. Пространство глаголов

Рассмотрим множество наиболее распространённых в речи глаголов (например, для английского языка), таких как: *TAKe, GET, SET, GO, PUT, RUN, MAKe, MOVe, HOLD, COMe, DO, TURN*.... Эта последовательность может быть переведена на русский язык как: *БРАТЬ, ПОЛУЧАТЬ, УСТАНАВЛИВАТЬ, ИДТИ*, и т. д. При попытке продолжить эту цепочку для русского языка, используя любой англо-русский словарь, можно получить представление о степени полисемии этих слов.

Переведём эти слова на родственные индоевропейские языки, используя обычную процедуру перевода и классические словари, составленные представителями лингвистической науки. Подобная процедура может быть выполнена в полуавтоматическом режиме [10]. После чего, аналогичным образом, с использованием словарей родственных языков полученные слова переведём обратно на английский язык. В результате таких действий мы получим множество английских глаголов, мощностью значительно большей, чем исходное. После нескольких шагов такой процедуры процесс расходится и уже на втором шаге мы имеем около 300 слов, при начальных 15. При сопоставлении массивов «взаимообратных переводов» становятся очевидными тонкие и чрезвычайно густые связи между глаголами внутри языка. Число переводов для каждого слова можно рассматривать как меру в некотором пространстве. Вообще говоря, для лучшего представления пространственных отношений между словами, как смысловыми понятиями в неком абстрактном «смысловом» пространстве, удобно представлять каждое понятие как объёмное тело, в общем случае произвольной формы. Тогда всё смысловое пространство можно представить заполненным этими телами, которые по некоторым признакам, как выяснится в даль-

нейшем, могут быть сгруппированы в кластеры. Отдельные тела могут смыкаться или даже взаимно проникать. Можно представить и некоторые незаполненные части пространства понятий, как области, которые не соответствуют никакой фонетической конструкции естественного языка. В первом, обсуждаемом здесь, приближении предполагается примерно сферическая форма тел-понятий. Таким образом, можно ввести в рассмотрение понятие расстояния между словами-понятиями. Расстояние, по определению, это геометрическое понятие, содержание которого зависит от объектов, для которых оно определяется. В соответствии с этим, расстояние между двумя точками есть длина соединяющих их прямой, в общем случае, многомерной [11]. Интуитивно расстояние между словами-понятиями можно рассматривать как длину отрезка, соединяющего точки центров сфер, как виртуальных образов слов-понятий в семантическом пространстве.

Однако, если существуют расстояния, необходимо ввести также метрику, посредством которой это расстояние может быть измерено. Метрика вводится следующим образом. Число взаимообратных переводов $\rho_{i,j}$ для двух разных глаголов i, j исходного списка можно рассматривать как характеристику расстояния между ними.

Пользуясь такой посылкой, как введение концепции «расстояния», можно сформировать «матрицу расстояний» для достаточно большого множества слов-понятий, которые могут попасть в поле рассмотрения.

Матрица расстояний, как основная характеристика обсуждаемого множества глаголов, и, как следствие, пространства, в котором они располагаются, строилась для каждой доступной для анализа пары глаголов для 13 родственных языков (болгарского, русского, чешского, польского, латвийского, шведского, немецкого, голландского, французского, итальянского, румынского, албанского и греческого). В качестве базового языка для рассмотрения, был выбран английский язык, но мог быть выбран и любой другой. Результат представляет собой квадратную, симметричную, положительно определённую матрицу, с длиной грани, равной числу рассматриваемых слов.

Для удобства работы с такими расстояниями, данными только в «дискретном» представлении, их желательно перевести в континуальное пространство. Это также способствует снижению чёткости определений значений слов и связей между ними, что более соответствует действительности («границы» понятия «размываются» и перестают быть резкими).

Для дальнейшего необходимо ввести ещё одну дополнительную величину – «объём» слова N_i. Объём равен общему числу переводов на все родственные языки для одного конкретного глагола. После расчёта этих величин для каждого из слов определим расстояние $D_{i,j}$ между глаголами-понятиями по следующей формуле:

$$D_{i,j} = -\ln\left(\frac{\rho_{i,j}^{\ 2}}{N_i N_j}\right), \qquad (1)$$

где $\rho_{i,j}$ – число взаимообратных переводов для глаголов i, j , N_i, N_j – объёмы

этих слов, $D_{i,j}$ – расстояние между словами-понятиями. Деление на объёмы слов производит нормировку расстояния, а функция логарифма выполняет здесь роль «сглаживателя» границ значений.

Таким образом, в нашем распоряжении имеется массив расстояний между точками в пространстве, которые это пространство и определяют. Для целей поиска возможных интересных топологий внутри этого пространства можно поставить задачу определения координат этих точек, каждая из которых соответствует понятию. Из сказанного следует, что мы имеем набор, вообще говоря, n-мерных объёмных тел, расположенных в n-мерном пространстве, о которых известны только расстояния между их «геометрическими» центрами. Необходимо найти координаты центров этих тел. Для решения введем соответствующую систему координат. Воспользуемся декартовой 4-мерной системой координат, причём предполагаем, что 4-я координата будет принимать значения, малые по сравнению с первыми 3-мя (а в отдельных «стабильных» случаях вообще принимать только нулевые значения), неся, таким образом, информацию о флуктуациях системы. Результаты расчётов, приведённые ниже, показали, что такой выбор был оправдан в том смысле, что «понятия» в большинстве случаев размещались и в более жёстких рамках пространств с меньшей размерностью.

Далее, по определению, в декартовых координатах расстояние между i-й и j-й точками равно:

$$D_{i,j} = \sqrt{\left(x0_i - x0_j\right)^2 + \left(x1_i - x1_j\right)^2 + \left(x2_i - x2_j\right)^2 + \left(x3_i - x3_j\right)^2}\,, \qquad (2)$$

где $x\{k_i\}$ – k-я координата i-й точки, $D_{i,j}$ – расстояние между i-й и j-й точкой (словом), определяемое заданным элементом «логарифмированной матрицы расстояний» по формуле (1).

Выражение (2) даёт систему нелинейных уравнений, которая с увеличением числа глаголов, начиная с некоторого момента, становится переопределённой. Её решением и будут искомые координаты.

Система решалась методом простой итерации:

$$\mathbf{x}^{(k+1)} = \mathrm{F}(\mathbf{x}^{(k)})\,. \qquad (3)$$

Для поиска решения переопределённой системы использовались дополнительные соображения, связанные со способом выбора уравнений для решения в каждом конкретном шаге процесса, связанным с добавлением в рассмотрение нового слова. В решаемой на каждом этапе процесса системе, всегда присутствовало уравнение, отвечающее расстоянию от слова (из множества базовых слов-понятий), называемого «главным» («центром кластеризации», «кластерообразователем»), это слово выбиралось экспериментатором, и от него начинались все «построения». Таким образом, при решении системы первым рассматривалось уравнение относительно расстояния от центра кластеризации. «Вторым» типом уравнений системы были соотношения для расстояний между добавляемым понятием, и словами, уже принадлежащими кластеру, безотносительно к центру.

Уравнения, соответствующие словам, «бесконечно» удалённым от центра данного смыслового кластера (или центра тела главного слова) давали в ходе расчётов недопустимые значения координат или были значительно удалены от центра кластера. «Бесконечно» удалённые слова – слова, расстояния между которыми, согласно матрице расстояний, равны величине, принятой в расчётах за бесконечность. Фактическое значение «условной бесконечности» выбиралось как много большее, чем максимальное расстояние, присутствующее в матрице расстояний.

Таким образом, слова, не связанные с данным «кластеризующим» центром, «не подходили кластеру» и исключались из рассмотрения. Однако исключались не абсолютно, и при добавлении в кластер нового слова, из еще не обсчитанного списка, предъявлялись системе вновь, в надежде, что только что добавленное слово стабилизирует решение. И иногда действительно встречались слова, которые могли присутствовать в кластере (то есть иметь значения координат близкие к таковым для центра кластеризации) только в случае присутствия своего «соседа» – например слова *«Drag»* и *«Haul»*. (*«Тащить»* и *«Тянуть»*) Уравнения, соответствующие словам, которые отвергались всё время, вплоть до окончания процесса всегда оказывались чуждыми по смыслу данной выборке или кластеру, – например, из выборки *«Destroy»* (*«Уничтожать»*) на бесконечность всегда уходило слово *«Connect»* (*«Соединять»*).

После выбора уравнений для решения запускался итерационный процесс, который продолжался при постоянном контроле сходимости к стабильному решению.

После того, как очередное решение системы уравнений было получено, то есть, было получено очередное приближение кластера с выбранным «ключевым словом», к нему применялся ряд критериев, по которым определялось, соответствует ли полученному приближению определение понятия кластера как совокупности точек, отвечающих словам, группируемым по принципу похожего смысла. Приведём некоторые из них.

Решаемая система уравнений переопределена, и, следовательно, получить точное решение принципиально невозможно. Поэтому, прежде всего, проверялось, не превышает ли погрешность полученного решения некоторого интуитивного порога, выставляемого оператором. Таким порогом чаще всего выбиралась погрешность в 100% в расчёте на одно кластеризуемое слово-понятие, содержащееся в кластере. Это означает, что точка, соответствующая слову, проверяемому на совместимость с кластером, может отклониться от своего среднего значения решения на его удвоенную величину и всё равно остаться членом кластера. Такой выбор «порога принадлежности» является эмпирическим и позволяет получать смысловые агрегации, состав которых не противоречит истинному значению глаголов.

К словам в кластере также применялся критерий отсутствия доминирующей роли осцилляций в кластере. Последнее означает, что осциллирует не более чем определённый процент от всех слов кластера. Под осцилляциями понимается поочерёдное принятие словом (центром объёмного тела) 2-х наборов координат от итерации к итерации. Этот процесс купировался добавлением нового слова в кластер, обладающего для осциллирующего, стабилизирующими свойствами. Такое, однако, происходило далеко не всегда.

При не выполнении хотя бы одного из этих критериев вновь добавляемое к кластеру слово считалось непригодным и отбрасывалось, в соответствии с методом, описанным выше.

Если такую вычислительную процедуру применять «локально», то есть, не выходя за определённые границы по количеству одновременно рассматриваемых уравнений, то она сходится к разумному решению в пределах допустимых величин погрешности. Разрешение такой ограниченной системы уравнений даёт величины координат точек для не слишком многочисленной группы глаголов. Стоит отметить, что при попытке выйти за границу, обусловленную, возможно, некоторыми внутренними свойствами рассматриваемой группы, наблюдается катастрофическая потеря точности вычислений и итеративный процесс расходится. В ходе экспериментов было рассмотрено большое количество таких образований или групп глаголов (кластеров). Оказалось, что все они обладают двумерной геометрией, или, (иногда, в зависимости от выбора центра кластеризации) различаются взаимным расположением и координатами точек внутри каждой группы и достаточно точно вписываются в плоскость. Из-за явления расходимости вычислительной процедуры при рассмотрении больших групп понятий расчёты носили локальный характер, и, следовательно, об ориентации полученных плоскостей в пространстве, без дополнительных допущений относительно привязки к плоскостям системы координат, нельзя говорить даже в относительном смысле.

В качестве примеров описанных плоскостных топологий в пространстве понятий можно привести, например, кластер «Уничтожать-Ломать», где наблюдалось, как в ходе итерационного процесса точка «*BREAK*» (*«Ломать»*) притягивала в свою группу точки пространства, отвечающие следующим понятиям: *«CRUSH»*, *«SHATter»*, *«WRECK»*, *«deSTROY»*, *«SMASH»*[1], которые потом и вошли в результирующий «кластер». Также характерна и стабильность кластера «Тянуть-Тащить» (*«PULL»*), с членами своего семейства: *«TUG», «DRAW», «DRAG», «HAUL», «TRAIL»*[2]. Хорошо видно, что приведённые группы слов, внутри каждой группы, несут вполне определённо общее значение, и не являются при этом синонимами, в словарном смысле этого слова, по крайней мере в некоторых языках из группы рассмотренных. Это кажется особенно замечательным, если учесть, что такие образования были получены из стандартных словарей, путём применения к ним чисто математических процедур. Структуры кластеров «Уничтожать-Ломать» и «Тянуть-Тащить» представлены на рис. 1,2.

Можно думать, что подобные группы слов просто «покрывают» некоторую область пространства понятий, которая отвечает определённой смысловой нагрузке, передать каковую посредством языка от одного человека к другому, можно используя одно из слов, входящих в группу. Соответствующие слова другого языка «покрывают» ту же самую область смыслового пространства.

---

[1] Соответственно: *«Давить», «Раздробить», «Крушить», «Уничтожать», «Разбивать»*.
[2] Соответственно: *«Дёргать», «Волочить», «Тащить», «Тянуть», «Тащиться (идти сзади)»*.

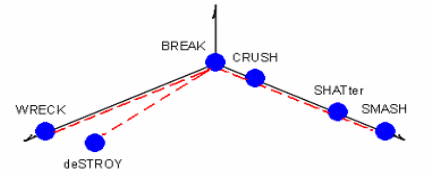

Рис. 1. Кластер «Уничтожать-Ломать»

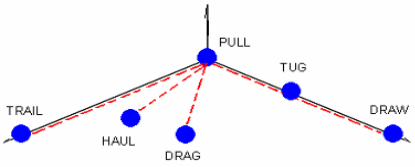

Рис. 2. Кластер «Тянуть-Тащить»

Согласно исследованиям методами магнитно-резонансной томографии (МРТ) высшей нервной деятельности людей, разговаривающих на двух разных языках с раннего детства, корковые зоны, активные при речи на первом или втором языке в значительной степени совпадают, что может говорить об обращении к одному «полю» памяти при формировании высказываний. И хотя этот факт всё ещё является предметом оживлённых дискуссий, любопытно обратить внимание на следующее. Так в [12], при помощи контрастной функциональной МРТ по определению уровня кислорода в крови исследовалась нервная деятельность людей бегло говорящих на двух сильно различных языках – английском и мандарине (диалект китайского). Картина активности коры головного мозга регистрировалась в то время, как субъекты называли или договаривали слова на обоих языках, по приведённому визуальному стимулу или высказанному началу слова. При работе на каждом из двух языков наблюдалась активность в префронтальной, височной, теменной и добавочной моторной областях. Во всех случаях местоположение пиковой активности совпадало. По крайней мере на уровне работы с единичными изолированными словами общая степень регистрируемой активности, или количество активных вокселей[1], в опытах была характерна для вызванной активности. Разница в степени активности для разных языков имелась, однако, существенной разницы в местоположении очага активности между двумя языками выявлено не было (рис. 3). Особенно замечательно то, что разницы не было выявлено, в том числе и в префронтальных «лингвистических» областях, которые традиционно считаются ответственными за генерацию смысловой составляющей речи. Такая картина активности не зависит от возраста и времени начала полноценного использования второго языка.

То, что в каждой смысловой группе или кластере содержится несколько слов, возможно, является причиной того, что часто бывает трудно подобрать подходящий перевод для конкретного слова. Слова, не имеющие «общего» смысла со словами рассматриваемого кластера, расположены «бесконечно» далеко от кластеризующего центра (расстояние между такими словами равно «условной бесконечности») и в ходе итерационного процесса исторгались из текущей смысловой структуры. С другой стороны, доступность множества слов, состоящих в «соседстве», делают возможными несколько версий перевода определённого предложения.

[1] Воксель – объёмный эквивалент пикселя (элементарного представимого в ЭВМ элемента изображения)

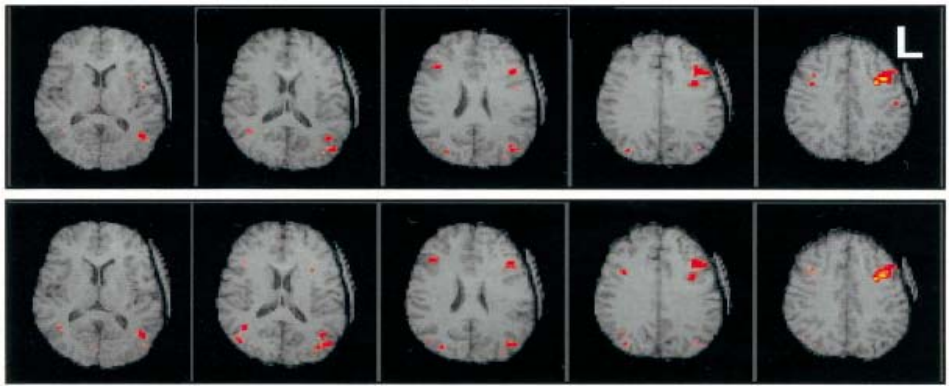

Рис. 3. Типичные картины активности при использовании английского и мандарина (Соответственно верхняя и нижняя линейки). Левое полушарие с правой стороны [12].

Как бы то ни было, при возможности локального картирования пространства понятий, используя метод решения системы нелинейных алгебраических уравнений, полное картирование или, что, то же самое, нахождение связей между различными кластерами, возможно либо путём сопоставления локальных данных, либо с применением некоторой другой вычислительной процедуры. Причины локальности геометрической модели, возможно, связаны с тем, что пространство понятий, само по себе подчиняется иной, чем евклидова, геометрии. Быть может, более уместным тут было бы рассмотрение точек, соответствующих словам-понятиям, на сферической поверхности.

Для картирования всего пространства понятий был выбран путь, связанный с использованием нейронной сети Хопфилда.

4. Сетевая система глаголов

Метод подразумевал следующие действия с имеющейся в распоряжении нормированной матрицей расстояний. Для рассмотрения пространства понятий, в котором состоят глаголы и предоставленного для анализа списка, строилась полносвязная нейронная сеть Хопфилда, каждый нейрон в которой соответствовал конкретному глаголу списка, то есть число нейронов в сети (или размерность сети) было равно числу рассматриваемых слов. Активное или пассивное состояние нейронов говорило о присутствии глагола в конкретной группировке (кластере) или исключении его из неё. Величины или веса связей между нейронами выставлялись в соответствии с обучением, которое, в свою очередь проводилось на основе нормированной логарифмической матрицы расстояний между глаголами. Процедура обучения будет описана ниже. Из рассмотрения были исключены первые несколько английских глаголов рассматриваемого списка. Взаимосвязи с прочими словами, для этой группы глаголов слишком сложны для анализа в этом приближении, что было причиной чрезмерной «загруженности» сети значащими образами, и, вело к преобладанию «эпилептической активности» (все нейроны сети находились в активном состоянии) при функцио-

нировании сети среди всех возможных картин активности. Значения в речи глаголов, входящих в исключённую группу, часто модулируются приставками и предлогами, анализ которых для прочих слов не рассматривался в этом исследовании. Кроме того, из рассмотрения преимущественно исключались глаголы, участвующие в кластерах, полученных при исследовании пространства понятий при помощи системы нелинейных алгебраических уравнений.

Таким образом, все рассмотренное поле английских глаголов можно метафорически уподобить сетчатке глаза. В ней, как известно, существует область «острого зрения» и область «нормального зрения». При нейросетевом анализе основной упор делался на исследование свойств «нормальной области», то есть таких частей пространства, которые заполнены словами, не вошедшими ни в один из смысловых кластеров. Топологические структуры, полученные методом решения нелинейных алгебраических уравнений, можно уподобить области «острого зрения», где взаимосвязи между отдельными частями «изображения» хорошо видны.

Начальная выборка для английского языка составляла 256 глаголов, после исключения, оговоренных выше подмножеств, остаток составил около 60 слов, которые использовались как центры кластеров при попытках их конструирования нейросетевым методом. Конкретные слова выбирались случайным образом в каждом из экспериментов, причём предпочтение отдавалось словам, не участвовавшим в кластерах, полученных вышеописанным способом.

С сетью проводилось две серии опытов для каждой выборки слов. В первой серии опытов количество обучающих паттернов ограничивалось эмпирической формулой для максимальной ёмкости памяти сети Хопфилда:

$$L = \frac{N}{4 \cdot \ln(N)}, \qquad (4)$$

где L – ёмкость памяти, N – число нейронов сети [13], $N = 60$.

Во второй серии сеть обучалась N обучающим паттернам для N нейронов. Каждый обучающий паттерн представлял собой «гипотетическое» множество слов с родственным смыслом. Тренировочные паттерны синтезировались в изобилии для каждого слова как «центра кластеризации». Не входящими в данную группу словами считались слова, расположенные бесконечно удалённо от центра кластеризации согласно нормированной логарифмической матрице расстояний (такие слова не имели общих переводов на другие рассматриваемые языки). Технически процедура формирования обучающих паттернов для нейронной сети выглядит следующим образом. Сначала выбиралось слово – центр кластеризации. В этой роли, как уже говорилось, по очереди выступают все слова рассматриваемой выборки. После чего формировалась «гипотеза» кластера, первоначально состоящая из всех остальных слов рассматриваемой выборки, кроме текущего центра кластеризации. После этого из гипотезы кластера изымались те слова, расстояния между которыми и словом-центром кластеризации, согласно нормированной логарифмической матрице расстояний было равно бесконечности.

Процесс обучения формирует зоны притяжения (аттракции) некоторых то-

чек равновесия, соответствующих обучающим данным. При использовании ассоциативной памяти мы имеем дело с множеством обучающих векторов $\overline{X}$, которые, в данной работе, формировались вышеописанным способом, и, которые в результате проводимого обучения определяют расположение конкретных аттракторов, то есть конкретных картин активности с характерным набором активных нейронов, выражающих принадлежность соответствующих им слов рассматриваемому кластеру, символическим выражением которого является наблюдаемый аттрактор. Каждый нейрон имел биполярную пороговую функцию активации.

Для обучения сети использовалась классическая процедура обучения нейронных сетей Хопфилда – обобщённое правило Хэбба, в соответствии с которым [14]:

$$w_{i,j} = \frac{1}{N}\sum_{k=1}^{p} x_i^{(k)} \cdot x_j^{(k)}, \qquad (5а)$$

$$T_i = -\sum_{k=1}^{p} x_i^{(k)}, \qquad (5б)$$

где p – число обучающих векторов в матрице $\overline{X}$, $w_{i,j}$ – веса синапсов нейронов, T_i – пороги нейронов.

Фаза обучения сети Хопфилда ориентирована на формирование таких значений весов, при которых в режиме функционирования задание начального состояния нейронов, близких к одному из обучающих векторов, приводит в результате функционирования сети к стабильному состоянию, в котором реакция нейронов остаётся неизменной в любой момент времени.

В экспериментах исследовались сети с параллельной динамикой, то есть характеризующиеся синхронным функционированием нейронных элементов сети. При этом за один такт работы сети все нейроны одновременно изменяют своё состояние согласно формуле:

$$y_i\left(t+1\right) = F\left(\sum_{\substack{j=1 \\ j\neq i}}^{N} w_{j,i} \cdot y_j\left(t\right) - T\right), \qquad (6)$$

где **y** – наблюдаемая в момент времени t активность сети, F – биполярная пороговая функция активации, T – текущее значение порога нейронов, а i пробегает все значения от 1 до N.

При функционировании сети можно наблюдать только устойчивые стационарные точки и циклы длиной два. Это свойство естественным образом использовалось при отождествлении картин активности сети с искомыми группами слов – кластерами.

В первых сериях опытов упор делался на ассоциативные функции нейронной сети Хопфилда. Так как ассоциативная память демонстрирует способность к коррекции, то при представлении тестовой выборки, например, отличающейся некоторым количеством битов на отдельных позициях вектора, нейронная сеть может скорректировать эти биты и завершить процесс классификации на нужном аттракторе. Следовательно, согласно этому режиму функционирования предполагалось, что, сеть по предоставленным примерам будет способна сформировать ассоциативные связи между словами согласно их смыслу, отражённому в матрице расстояний, по которой и формировались обучающие паттерны.

Вторая серия опытов, кроме ассоциативной функции, предполагала использование некой «креативности» сети. При использовании сети Хопфилда как ассоциативной памяти, в ней неизбежно формируется «паразитная память» - набор паттернов, которым сеть изначально не обучалась. Такое явление особенно характерно для сети, в которую записывают объём информации, превышающий номинальную ёмкость её памяти L. В этой серии опытов, наборы картин активности, относящихся к паразитной памяти, рассматривались как полноценные смысловые кластеры.

В обоих случаях после обучения анализировались полученные для каждого значения порога сети, одинакового для всех нейронов, картины активности. Для этого применялась специальная автоматическая процедура просмотра всех значений порогов сети и записи всех зарегистрированных паттернов активности сети.

Процедура тотального картирования памяти нейронной сети состояла в том, что для данного эксперимента при обучении, кроме значений массива весов, рассчитывался массив порогов нейронных элементов по формуле (5б). При минимальной величине порога T_{min}, содержащейся в этом массиве, сеть гарантировано переходит в «эпилептический» режим, когда активны все нейроны. При максимальной величине порога T_{max}, сеть находится в перманентной «коме», т.е. ни один нейрон не получает достаточно возбуждающих воздействий, чтобы превзойти порог и перейти в активное состояние. Далее, на каждом шаге процедуры картирования последовательно выбиралось значение порога из диапазона $[T_{min}, T_{max}]$, и присваивалось значению текущего порога каждого нейрона сети.

При использовании такой динамики сети результат работы обсуждаемых алгоритмов оказался одинаковым, независимо от серии опытов. Тотальный просмотр всей возможной активности сети, для всех серий опытов, не выявил смысловых агрегаций слов, содержащих более чем 2 слова. Следует напомнить, что здесь рассматривались слова, не вошедшие в смысловые кластеры. Типичные картины активности сети представлены на рис. 4. Показаны только информативные области паттернов. Красным отмечены активные нейроны, символизирующие присутствие соответствующего слова в кластере. Разумеется, картины активности сети, с количеством активных нейронов, больше 2-х, существовали, но они включали в себя «слова», никак не связанные между собой по смыслу, и содержали очень большое число составляющих, охватывающих практически всю выборку слов. Данное явление можно интерпретировать как следствие гомогенности распределения глаголов в смысловом пространстве.

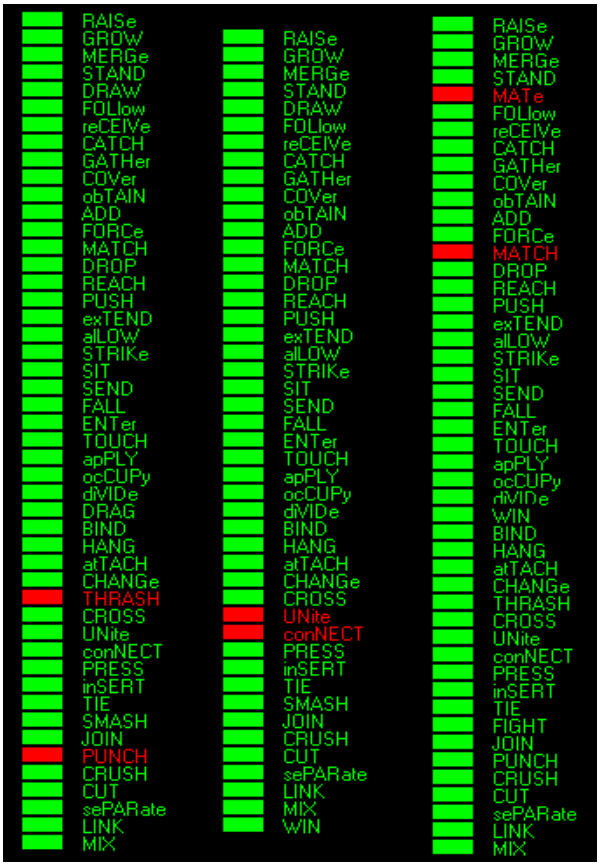

Рис. 4. Паттерны активности кластеризующей сети.

Таким образом, за исключением области «острого» восприятия смысла, содержащего ограниченное количество, очевидно, наиболее применимых в повседневной речи, либо более значимых по какой-то другой причине, слов-глаголов, и стягивающих к себе слова с близким значением, остальное пространство понятий не содержит слов, несущих близкие значения и способных к группировке, по крайней мере, в рамках рассматриваемых моделей. Смысловые связи в этой области семантического пространства существуют только с «ближайшими» соседями, причём такие связи слабее, чем связи внутри смыслового кластера.

5. Выводы

В работе приведены два взаимодополняющих метода, работающих с трудно формализуемой лингвистической информацией. Цели функционирования этих алгоритмов несколько различны. Формирование смысловых кластеров через

анализ значений их координат в семантическом пространстве, полученных путём решения систем нелинейных уравнений, может продемонстрировать сильное сродство значений некоторых слов друг к другу. В то время как «нейросетевой» метод даёт представление о заселённости областей смыслового пространства, не занятого кластерами, словами, в эти кластеры не вошедшими.

Использование обоих методов может показать, что содержащиеся в мультиязыковых словарях лингвистические данные могут быть систематизированы с применением к ним категории «смысла», который кодирует слово внутри человеческого языка, а не простого статистического упорядочения.

Слова-понятия могут быть размещены в некотором пространстве, к которому применим достаточно эффективный математический аппарат. При локальном рассмотрении некоторые области семантического пространства выглядят как плоскости в евклидовой геометрии.

Подобная процедура смыслового упорядочения кажется полезной для её применения при организации словарей электронных систем перевода и интерпретации речи. Если заметить, что человек, занимающийся переводом, переводит смысл текста, а не слова текста, то такое усовершенствование может улучшить качество предоставляемых электронными переводчиками услуг.

В отличие от, например, [15] в данной работе фонетическая форма слова вообще не рассматривалась. Тем не менее, в упомянутой статье показано, что слова могут быть эффективно упорядочены и по признаку «фонетической формы» корня слова, причём эффективнее, чем простая сортировка по алфавиту. Это поднимает вопрос о взаимосвязи между этими двумя пространствами – фонетическим (пространством комбинаций фонетических символов) и смысловым, или об отображении одного пространства на другое.

Литература

1. McEnery, T. and Wilson, A.: *Corpus Linguistics*, Edinburgh: Edinburgh University Press. (1996)

2. Price, C., Indefrey, P., van Turennout M.: The Neural Architecture, Underlying the Processing of Written and Spoken Word Forms. In: Brown, C.M., Hagoort P. (eds.): *The Neurocognition of Language*. Oxford Univ. Press, NY (1999) 211–240

3. Simos, P.G., Breier, J.I., Fletcher, J.M, Bergman, E., Papanicolaou, A.C.: Cerebral Mechanisms Involved in Word Reading in Dyslexic Children: a Magnetic Source Imaging Approach. Cerebral Cortex 10(8) (2000) 809-816

4. Коршаков А. В.: Процедура построения пространства понятий как часть системы машинного перевода. *Нейроинформатика 2006, VIII Всероссийская Научно-Техническая Конференция,* Сборник научных трудов, часть 3, стр. 155-162.

5. Vvedensky V.L., Korshakov A.V., (2004). Visualization of the Basic Language Thesaurus. *Proc. VII International Conf. «Cognitive Modelling in Linguistics»*, Varna, pp. 308-313.

6. Введенский В. Л., Коршаков А. В.: Естественно упорядоченный алфавит индоевропейских языков. *Нейроинформатика 2004, VI Всероссийская Научно-Техническая Конференция,* Сборник научных трудов, часть 2, стр. 18-24.

7. Shannon C. E. A Mathematical Theory of Communication. *The Bell System Technical Journal,* Vol. 27, pp. 379–423, 623–656, July, October, 1948.

8. Введенский В. Л.: Математические закономерности словообразования в евро-

пейских языках. *Нейроинформатика 2005, VII Всероссийская Научно-Техническая Конференция,* Сборник научных трудов, часть 2, стр. 263-270.

9. Der Dokumenten-Server der Universitat Leipzig, http://wortschatz.uni-leipzig.de/index.html

10. Foreignword.com, The language site http://www.foreignword.com

11. *Математический энциклопедический словарь*. Москва «Советская энциклопедия» 1988.

12. Michael W. L. Chee, Edsel W. L. Tan, and Thorsten Thiel. Mandarin and English Single Word Processing Studied with Functional Magnetic Resonance Imaging. The Journal of Neuroscience, April 15, 1999, *19*(8):3050–3056

13. Головко В. А. *Нейронные сети: обучение, организация и применение*. Издательское предприятие журнала «Радиотехника». Москва 2001.

14. Осовский С. *Нейронные сети для обработки информации*. Москва «Финансы и статистика» 2002.

15. Введенский В. Л.: Сеть корней глаголов русского языка. *Нейроинформатика 2006, VIII Всероссийская Научно-Техническая Конференция,* Сборник научных трудов, часть 3, стр. 236-243.

Статья поступила 3 июля 2006 г.

После доработки 11октября 2006 г.