КОГНИТИВНЫЙ ПОДХОД К АНАЛИЗУ ТЕКСТОВ В ТЕХНОЛОГИИ АВТОМАТИЧЕСКОГО СМЫСЛОВОГО АНАЛИЗА ТЕКСТОВ TEXTANALYST

А.А. Харламов - Институт высшей нервной деятельности и нейрофизиологии РАН

Abstract

Две структуры в мозге человека ответственны за хранение и обработку специфической информации: кора большого мозга и гиппокамп. В колонках коры хранится информация о событиях, в гиппокампе – о связях событий в рамках более крупных событий. Колонки коры упорядочены в иерархическую структуру таким образом, что формируют парадигматическое многоуровневое представление информации, в котором поуровневые словари более высокого уровня являются грамматиками для более низкого уровня. Слова словаря более низкого уровня составляют парадигму для слов словаря более высокого уровня. Гиппокамп, имеющий структуру, сходную с искусственной нейронной сетью Хопфилда, формирует из событий, хранящихся в коре, ассоциативную сеть, которая легко превращается в смысловую сеть путем пересчета весов событий. Анализ текста, использующий на нижних уровнях парадигматическое представление, в иерархии которого представлены морфологический, лексический и синтаксический уровни, и на верхнем уровне формирующий сеть ключевых понятий с последующей перенормировкой их весов, позволяет реализовать автоматическое реферирование, сравнение, и, таким образом, классификацию текстов. На основе этих представлений реализована технология автоматической смысловой обработки текстов TextAnalyst®, позволяющая выявить ключевые понятия текста в их взаимосвязях, реализовать их аннотирование и смысловое сравнение (классификацию). Реализованы продукты, использующие функциональность этой технологии: персональный – TextAnalyst®, и библиотека COM модулей – TextAnalyst® SDK [Kharlamov, 2012].

Введение

Однородная обработка специфической информации в головном мозге человека осуществляется, в основном, в двух структурах: в колонках коры большого мозга и в гиппокампе. В колонках хранится информация о событиях. Она упорядочена по ассоциации таким образом, что близко хранится близкая по форме информация [Глезерман, 1986]. Кроме того, она упорядочена по иерархии: чем выше, тем более общего характера информация хранится и обрабатывается [Бехтерева, 1978]. На каждом уровне иерархии создаются словари событий своего уровня. Они связаны так, что слова более высокого уровня являются грамматиками для слов более низкого уровня.

Колонки коры, помимо нейронов других типов, состоят, в основном из пирамидных нейронов третьего слоя, которые, будучи элетронекомпактными, осуществляют временную суммацию сигналов. Искусственные нейронные сети на основе нейронов с временной суммацией сигналов моделируют колонки коры. Они реализуют многоуровневую структурную обработку информации на основе ассоциативного преобразования, в результате которой формируется иерархическое представление в виде множества автоматически выявляемых словарей событий различной частоты встречаемости, причем более верхние уровни представлений являются грамматиками для более низких уровней [Харламов, 2006].

Гиппокамп, имеющий структуру, состоящую из множества независимых образований, моделируемых искусственной нейронной сетью Хопфилда, хранит в каждой такой структуре связи событий колонок коры в рамках более крупного события, ситуации [Rumelhart et al., 1986].

Колонки коры – функциональные образования, состоящие из объединений нейронов (преимущественно – пирамидных нейронов 3-го слоя), которые связаны единым входом – общим их иннервирующим аксоном из специфических областей таламуса, и единым управлением – все пирамидные нейроны 3-го слоя иннервируются единой

горизонтальной клеткой 2-го слоя, которая получает неспецифическую информацию из неспецифических зон таламуса. В соседних колонках хранится информация о близких событиях [Хьюбель, 1988]. Колонки объединены в гиперколонки теми же общим специфическим таламическим аксоном и общей горизонтальной клеткой. Основную функцию колонок выполняют пирамидные нейроны 3-го слоя коры.

Пирамидные нейроны 3-го слоя коры являются электронекомпактными нейронами, то есть нейронами, которые выполняют временную суммацию сигналов [Радченко, 1969; Roll, 1964] (учитывают временную структуру последовательности сигналов), в отличие от нейронов с пространственной суммацией, которые учитывают только пространственный код сигналов [Rosenblatt, 1962]. Пирамидные нейроны отличаются друг от друга временным кодом – адресом, и, потому, могут возбуждаться избирательно. Любой фрагмент входной информационной последовательности, поступающей на колонку, адресуется к своему нейрону. Обладая пластичностью, пирамидные нейроны могут запоминать и распознавать приходящую на них информацию.

Пирамидная клетка третьего слоя коры моделируется нейроном с временной суммацией сигналов [Харламов, 2006]. Обобщенный дендрит такого нейрона со специфическим для него распределением возбуждающих и тормозных синапсов – адресом нейрона – откликается строго на последовательность входных сигналов, соответствующую адресу, то есть возбуждается избирательно. Адрес нейрона соответствует возбуждающему его фрагменту входной последовательности длины n, и моделирует точку n-мерного пространства, координаты которой соответствуют адресу нейрона (последовательности нулей и единиц, где «1» соответствует возбуждающему синапсу, «0» - тормозному). Объединение из 2^n пирамидных нейронов, состоящее из нейронов с разными адресами, моделирует некоторую область n-мерного сигнального пространства. Любой фрагмент любой входной последовательности адресуется к одному из нейронов этого объединения, где и запоминается, а входная последовательность отображается в последовательность сработавших нейронов – траекторию в сигнальном пространстве (последовательность точек с координатами, соответствующими n-членным фрагментам этой входной последовательности).

Колонки коры, в зависимости от расстояния от рецепторных органов, формируют иерархию представлений, увеличивающихся по сложности снизу-вверх. В проекционных зонах коры различных анализаторов информация от рецепторов, после переключения в подкорковых ядрах, поступает на колонки первичной проекционной зоны [Глезер, 1985], формируя в них так называемые простые рецептивные поля, затем поступает на колонки вторичной проекционной зоны, формируя в них сложные рецептивные поля, и затем, на колонки третичной проекционной зоны, формируя на них сверхсложные проекционные поля. Простые, сложные и сверхсложные рецептивные поля соответствуют сформированным в колонках соответствующих уровней иерархии представлениям событий разного уровня сложности. Причем, информация от простых рецептивных полей входит в сложные рецептивные поля, а информация от сложных рецептивных полей входит в сверхсложные рецептивные поля.

События, отображенные в колонках коры, запоминаются и распознаются в совокупности не только по ассоциации, но и с помощью механизма ассоциативной памяти гиппокампа [Виноградова, 1975]. Гиппокамп представляет собой парный орган, расположенный в глубине головного мозга. Вдоль длинной оси гиппокампа располагаются так называемые ламели, каждая из которых имеет в своей структуре поле, соответствующее по архитектуре ассоциативной памяти Хопфилда [Rumelhart et al., 1986], которая содержит информацию о событиях, хранящихся в колонках коры, в их совокупности в более крупных событиях, ситуациях. Поле CA_3 [Hopfield, 1982], которое в архитектуре ламелей гиппокампа отвечает за хранение информации о связях, имеет структуру искусственной нейронной сети Хопфилда – ассоциативной памяти. Отдельные нейроны поля CA_3, также имеющие ассоциативную адресацию, как и пирамиды 3-го слоя

коры, и связанные по ассоциации с событиями, хранящимися в колонках коры, образуют ассоциативную сеть более крупного события, ситуации, в которой вершинами являются упомянутые события, хранящиеся в коре, в их ассоциативных взаимосвязях.

Использование парадигматического представления информации, характерного для колонок коры, для хранения текстовой информации морфологического, лексического и синтаксического уровней, и формирование на семантическом уровне ассоциативной сети ключевых понятий с последующей перенормировкой весов понятий в соответствие с их смысловой значимостью в тексте, как в гиппокампе, позволяет реализовать технологию автоматического смыслового анализа текстов, с помощью которой можно автоматически извлекать ключевые понятия текста (слова и устойчивые словосочетания), формировать семантическую сеть ключевых понятий со взвешенными понятиями и связями, автоматически реферировать текст, сравнивать тексты по смыслу (следовательно, классифицировать их), кластеризовать корпус текстов по темам.

Созданная технология легла в основу двух продуктов, разработанных московской компанией Microsystems, Ltd.: TextAnalyst, предназначенного для персонального использования, и библиотеку COM модулей TextAnalyst SDK для встраивания в персональные приложения.

В главе, посвященной автоматической технологии обработки текстовой информации представлены следующие разделы. Раздел 1 посвящен когнитивным аспектам обработки лингвистической информации в мозге человека, в том числе рассматриваются искусственные нейронные сети, моделирующие обработку информации в колонках коры и гиппокампе. Представлен формализм искусственной нейронной сети из нейроподобных элементов с временной суммацией сигналов, реализующий структурный ассоциативный подход к обработке информации. Показаны этапы формирования ассоциативной (однородной семантической) сети. В разделе 2 описывается технология TextAnalyst автоматической смысловой обработки текстовой информации. В разделе 3 описан персональный продукт для анализа текстов TextAnalyst. В разделе 4 – намечено направление дальнейших исследований.

1. Когнитивные аспекты обработки лингвистической информации

Если рассмотреть известную иерархию уровней обработки лингвистической информации, то можно усмотреть в ней некоторые аналогии представлению и обработке информации в мозге человека. Эта иерархия для обработки речевой (текстовой), то есть лингвистической, информации имеет следующий состав (снизу-вверх):

1. Акустико-фонетический (графематический) уровень.
2. Морфологический уровень.
3. Лексический уровень.
4. Синтаксический уровень.

Далее идут еще два, уже надлингвистических уровня представления информации:

5. Семантический уровень.
6. Прагматический уровень.

Первые четыре уровня характерны тем, что информация в них представленная как бы дополняет друг друга в разных комбинациях. Это так называемое парадигматическое представление. В таком представлении элементы нижнего уровня могут быть равнозаменяемы в элементах более высокого уровня (которые, как мы помним, являются грамматикой для слов словаря нижнего уровня). Элементы морфологического уровня (флективные морфемы - окончания) вместе с элементами лексического уровня (корневыми основами) составляют целые слова, флективные структуры предложений, являющиеся представлениями синтаксического уровня, вместе с элементами лексического уровня (корневыми основами) составляют целые предложения. На первом уровне мы имеем такую же картину, только для этого необходимо более подробно рассмотреть звуковую природу речи. Акустико-фонетический уровень представления

речевой информации разбивается на два подуровня: на первом подуровне выявляется фонемная структура речевой волны (формируется словарь фонем), на другом – транземная (от слова «транзема», что обозначает переходные отрезки речевой волны между фонемами) – формируется словарь транзем. Такая обработка характерна для представления информации в иерархии колонок коры.

Два надлингвистических уровня по обработке отличаются от первых четырех лингвистических уровней. Мы рассмотрим только обработку и представление информации на пятом – семантическом – уровне. На семантическом уровне лингвистическая информация характеризуется попарной сочетаемостью корневых основ. Именно пары слов являются характерными элементами представления этого уровня. Но множество пар легко преобразуется в сеть. Второй элемент первой пары, связанный с первым элементом второй пары, затем через свой второй элемент связывается с первым элементом какой-то из других имеющихся пар, и наконец, где-то цепь замыкается и второй элемент последней пары оказывается связанным с первым элементом одной из уже упоминавшихся пар.

Такое представление не случайно. В нем каждое событие связано с некоторым множеством ближайших ассоциантов, которые являются его семантическими признаками. Не случайно, именно сети являются наиболее удобным представлением семантической информации [Sowa, 1992].

Обработка внутренне структурированной информации различных модальностей (в том числе текстовой), имеющей многоуровневую структуру, с помощью нейронных сетей на основе нейроподобных элементов с временной суммацией сигналов [Харламов, 2006] позволяет автоматически сформировать словари событий разной частоты встречаемости (словари разных уровней). Такая обработка сводится к отображению F информационной последовательности A, в многомерное пространство R^n, моделируемое нейронной сетью - множеством нейроподобных элементов с соответствующими адресами, в результате которого информационная последовательность преобразуется в последовательность сработавших нейронов - в связанную последовательность точек многомерного пространства – траекторию $\hat{A}$.

1.1. Искусственная нейронная сеть из нейроподобных элементов с временной суммацией сигналов – модель колонки коры

Нейроподобный элемент с временной суммацией сигналов (см. рис. 1) [Харламов, 2006] является последовательным развитием нейроподобного элемента А.Н. Радченко [Радченко, 1969], возникшего на основе модели W. Rall [Rall, 1964], которая, в свою очередь, возникла на основе представлений D.A. Sholl [Sholl, 1953].

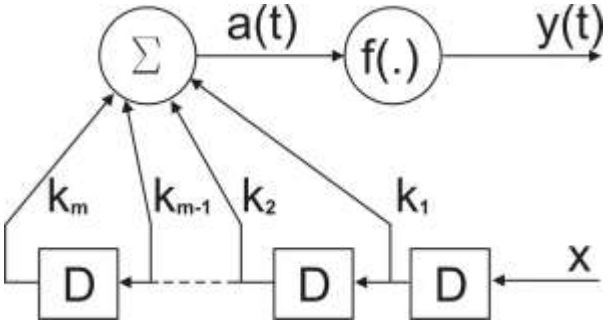

Рис. 1. Нейроподобный элемент с временной суммацией сигналов. Здесь D – элементы задержки, w_m – коэффициенты связей, $f(*)$ – нелинейный элемент, например, пороговое преобразование.

Такой нейрон выполняет свертку фрагмента последовательности длины *n* символов - $(a_{t-n+1}, a_{t-n+2}, ..., a_t)$, $a_t \in \{0,1\}$, с последовательностью весовых коэффициентов $(w_{t-n+1}, w_{t-n+2}, ..., w_t)$, $w_t \in \{0,1\}$:

$$S = \sum_{i=1}^{n} a_{t-n+i} w_i \,, \tag{1}$$

Свертка будет иметь наибольшее значение, если *n*-членный фрагмент входной последовательности соответствует последовательности весовых коэффициентов нейрона, то есть если $w_t = -1$, то $a_t = 0$, а если $w_t = +1$, $a_t = 1$. Такой фрагмент последовательности называется адресом нейрона. Наибольшее значение свертки равно числу единиц в адресе - $\sum_{single}$.

В качестве нелинейной функции используется пороговое преобразование $f(*) = H_{adr}$ с порогом h_{adr}. Если порог равен числу единиц в адресе $h_{adr} = \sum_{single}$, то нейрон будет откликаться строго на свой адрес. То есть он моделирует одну из точек *n*-мерного сигнального пространства R^n. В случае бинарной входной последовательности - это вершина *n*-мерного единичного гиперкуба G_s.

Для понимания работы нейронной сети на основе такого нейрона [Kharlamov at al, 2004] представим его в упрощенном виде (см. рис. 2). Здесь вместо бинарного регистра сдвига используется многоразрядный регистр сдвига [Харламов, 2008], который можно назвать обобщенным дендритом [Радченко, 1969].

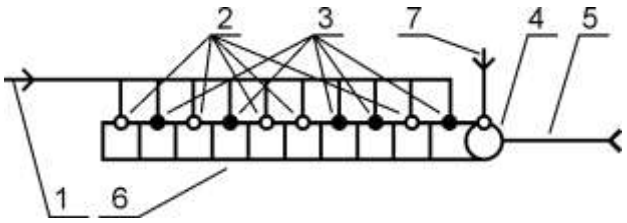

Рис. 2. Упрощенное представление нейрона с временной суммацией. Здесь 2 и 3 – возбуждающие и тормозные синапсы, соответственно, 4 – пороговое преобразование, 6 – многоразрядный регистр сдвига.

Объединение таких нейронов с разными адресами (см. рис. 3) моделирует n-мерный единичный гиперкуб в сигнальном пространстве. В сети такого вида информация запоминается [Hebb, 1949] за счет изменения состояний или связей соответствующих нейроподобных элементов.

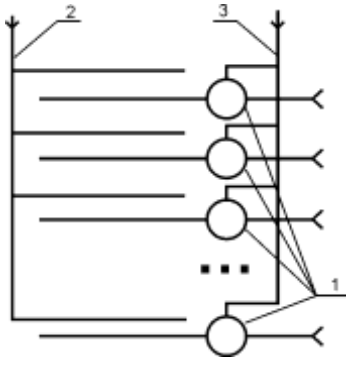

Рис. 3. Нейронный пучок. Здесь 1 - нейроны пучка, имеющие обобщенные дендриты с разными адресами от (000...0) до (111...1), 2 - общее афферентное волокно.

Нейроны с различными распределениями синапсов на дендритах (с разными адресами) можно избирательно возбуждать, подавая на них последовательности с разным распределением в них импульсов и межимпульсных интервалов. При этом пороги возбуждения нейронов, которым эти дендриты принадлежат, должны быть равны числу возбуждающих синапсов на своих обобщенных дендритах (такова может быть максимальная накопленная в каждом дендрите свертка – сумма входных сигналов, если весовые коэффициенты всех синапсов равны «+1» или «-1»).

Другими словами, такая нейронная сеть отображает последовательность импульсов и межимпульсных интервалов в последовательность сработавших нейронов. И, поскольку, в бинарном случае имеется ровно 2^n различных двоичных адресов, последовательность сработавших нейронов представляет собой последовательность вершин n-мерного единичного гиперкуба G_s, то есть траекторию в многомерном пространстве R^n.

1.2. Ассоциативное преобразование. Парадигматическое представление информации

Это отображение обладает свойством ассоциативности обращения к траектории: как только в информационной последовательности появляется повторяющийся фрагмент, траектория возвращается к ранее пройденному участку. Запоминание числа прохождений траекторией точек многомерного пространства, с последующим применением порогового преобразования, позволяет выявлять фрагменты траектории заданной частоты появления, которые составляют словари событий входной информации заданной частоты встречаемости $\{\hat{B}_i\}$. Для лингвистической информации это, например, словари флективных морфем, корневых основ, синтаксических групп. Выявленные таким образом лингвистические единицы в дальнейшем можно использовать для обработки текстовой информации. Словарь флективных морфем можно использовать для морфологического анализа, словарь корневых основ – для выявления ключевых понятий в тексте и формирования однородной (ассоциативной) семантической сети [Харламов и др., 2008], словарь синтаксических групп – для формирования неоднородной семантической сети.

Рассмотрим это отображение подробнее на примере обработки двоичной информации.

Преобразование, реализующее свойство ассоциативности обращения к информации

Пусть мы имеем n-мерное сигнальное пространство R^n и в нем - единичный гиперкуб $G_e \in R^n$.

Рассмотрим преобразование F двоичной последовательности A в n-мерное пространство R^n таким образом, что каждому n-членному фрагменту последовательности соответствует точка в R^n - $\hat{a}(t)$, с соответствующими n-членному фрагменту координатами, а всей последовательности A соответствует последовательность точек: $\hat{A} = \{...,(a(-n-1),a(-n),...,a(-2)),(a(-n),a(-n+1),...,a(-1)),$ $(a(-n+1),a(-n+2),\ldots,a(0)),(a(-n+2),a(-n+3),\ldots,a(1)),(a(-n+3),$ $a(-n+4),...,a(2),...,(a(-n+t),a(-n+1+t),...,a(t)),...\} = = (...,\hat{a}(-2),\hat{a}(-1),\hat{a}(0),\hat{a}(1),\hat{a}(2),...,\hat{a}(t),...)$ - траектория:

$$\hat{A} = F(A), \qquad (2)$$

здесь F - обозначает отображение в сигнальное пространство. Отображение F является основой для осуществления структурной обработки информации.

Пример. Пусть последовательность $A = (101100101011)$ отображается преобразованием F в трехмерное пространство. Тогда последовательности A соответствует

последовательность вершин трехмерного единичного куба: $\hat{A} = F(A) = (\hat{a}(1) = 001,$
$\hat{a}(2) = 010, \hat{a}(3) = 101, \hat{a}(4) = 011, \hat{a}(5) = 110, \hat{a}(6) = 100, \hat{a}(7) = 001,$
$\hat{a}(8) = 010, \hat{a}(9) = 101, \hat{a}(10) = 010, \hat{a}(11) = 101, \hat{a}(12) = 011, \hat{a}(13) = 110, \hat{a}(14) = 100, \hat{a}(15) = 000)$,
а траектория будет иметь вид, представленный на рис. 2.

Преобразование F обладает свойством ассоциативности обращения к точкам траектории $\hat{A}$ ассоциацией по n-членному фрагменту последовательности A (то есть - по его содержанию): любые n символов сразу же адресуют нас к соответствующей точке траектории.

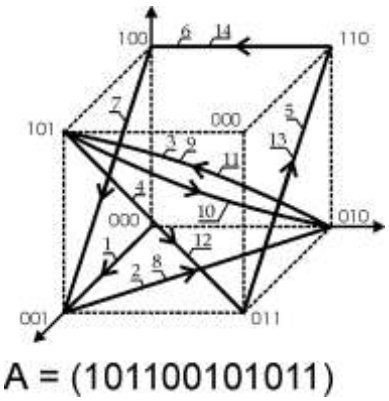

Рис. 4. n-мерный единичный гиперкуб в случае n=3 с траекторией, соответствующей последовательности A.

Можно выбрать размерность пространства R^n такой, что некоторая последовательность конечной длины может быть отображена в это пространство без пересечений, то есть каждому n-членному фрагменту последовательности символов в R^n будет соответствовать точно одна точка траектории. В этом случае существует обратное F преобразование F^{-1} траектории $\hat{A}$ в исходную последовательность A:

$$A = F^{-1}(\hat{A}), \tag{3}$$

В общем случае среди n-членных фрагментов информационной последовательности может встретиться уже ранее встречавшийся n-членный фрагмент, и траектория в этом случае пройдет через вершину, уже принадлежащую ей, то есть пересечется с самой собой. В этой точке возможно более одного продолжения траектории. Для двоичной последовательности продолжений может быть не более двух.

Ассоциативность преобразования F позволяет сохранить топологию структуры преобразуемой информации. Действительно, одинаковые фрагменты последовательности преобразуются в одну и ту же траекторию, разные - в разные траектории.

Запоминание информации. Воспроизведение. Авто- и гетероассоциативность

Пусть есть две синхронно отображаемые в многомерное пространство R^n последовательности A и J. Траектория последовательности A (назовем ее несущей) в сигнальном пространстве может быть использована для запоминания в точках соответствующей ей траектории $\hat{A}$ символов синхронизированной с ней информационной последовательности J. Введем в вершинах гиперкуба G_e функцию памяти M, ставящую в соответствие каждой вершине $\hat{a}(t) \in \hat{A}$, соответствующей t-му символу последовательности A, двоичную переменную $j(t+1)$, являющуюся $(t+1)$-м символом двоичной последовательности J.

$$M\{\hat{a}(t), j(t+1)\} = [\hat{a}(t)]_{j(t+1)}. \tag{4}$$

Мы имеем, таким образом, траекторию $\hat{A}$, обусловленную последовательностью J. $[*]$ - обозначает обусловленность.

$$[\hat{A}]_J = M\{F(A), J\}. \tag{5}$$

Другими словами, последовательность J записывается в точках траектории $\hat{A}$ (в ассоциации с траекторией A).

Можно осуществить восстановление информационной последовательности J по обусловленной ею траектории $[\hat{A}]_J$ и несущей последовательности A:

$$J = M^{-1}\{[\hat{A}]_J, F(A)\}, \tag{6}$$

где в каждой точке $\hat{a}(t) \in \hat{A}$: $M^{-1}([\hat{a}(t)]_{j(t+1)}, a(t)) = j(t+1)$ (здесь M^{-1} - воспроизведение). При этом развертывание в траекторию несущей последовательности позволяет обратиться к информации, записанной в точках траектории, то есть к информационной последовательности. Такой способ записи назоваем гетерассоциативной записью, а воспроизведение - гетероассоциативным воспроизведением.

Если в качестве обуславливающей последовательности используется та же последовательность, что и несущая, то есть в точках траектории в сигнальном пространстве записываются символы этой же последовательности, - имеем случай самообусловливания: то есть, если $J \equiv A$, $M\{\hat{a}(t), a(t+1)\} = [\hat{a}(t)]_{a(t+1)}$:

$$[\hat{A}]_A \equiv [\hat{A}] = M\{F(A), A\}. \tag{7}$$

Аналогично (6):

$$A = M^{-1}\{[\hat{A}]_J, F(A)\}. \tag{8}$$

В этом случае можно восстановить исходную последовательность, начиная с одной из точек траектории:

$$A = M^{-1}\{[\hat{A}], \hat{a}(t) \in F(A)\}. \tag{9}$$

Действительно, имея n-членный фрагмент последовательности $\hat{a}(t) = (a(t-n+1), a(t-n+2), ..., a(t))$, мы обращаемся к одной из точек $\hat{a}(t)$ траектории $\hat{A}$. В этой точке записана информация $M^{-1}\{[\hat{a}(t)], t\} = a(t+1)$, соответствующая следующему символу последовательности A, породившей траекторию $\hat{A}$. Добавляя к $(n-1)$-му символу предыдущего n-членного фрагмента новый символ $a(t+1)$, мы получаем новый

n-членный фрагмент $(a(t-n+2), a(t-n+3), ..., a(t+1))$, по которому осуществляется обращение к следующей вершине траектории: $\hat{a}(t+2)$. В ней считывается следующий символ последовательности $M^{-1}\{[\hat{a}(t+1)], t+1\} = a(t+2)$. И так далее до конца последовательности или до ближайшего ветвления траектории. Такая запись называется автоассоциативной записью, а воспроизведение - автоассоциативным воспроизведением.

Таким образом, использование функции M совместно с преобразованием F, обладающим свойством ассоциативного обращения к информации, позволяет реализовать ассоциативную память с возможностью авто- и гетероассоциативной записи/воспроизведения информации.

Формирование статистической модели. Забывание

Несколько усложнив функцию памяти M, мы можем реализовать, наряду с функцией ассоциативной записи/воспроизведения, механизм статистической обработки информации. Для этого заменим триггер регистрации следующего символа $a(t+1)$ последовательности A двумя счетчиками, фиксирующими число прохождений траекторией заданной точки в заданном направлении: C_0 - для переходов в "0" и C_1 - для переходов в "1". Введем также пороговое преобразование H, позволяющее восстановить по значению функции H в точке многомерного сигнального пространства, определенной ее координатами $\hat{a}(t)$ - значение наиболее вероятного перехода в следующую точку – в "0" или в "1": $a(t+1)$. Такой механизм памяти чувствителен к числу прохождений заданной точки в заданном направлении. Он позволяет характеризовать каждую точку траектории с точки зрения частоты появления во входной информации сочетания $(\hat{a}(t), a(t+1))$. При этом применение порогового преобразования H позволяет воспроизводить информацию заданной степени достоверности.

Этот механизм работает следующим образом. В случае прохождения более одного раза одного и того же фрагмента траектории, счетчики числа переходов каждой из точек этого фрагмента траектории запоминают число прохождений. При запоминании счетчики изменяют свои состояния в зависимости от направления перехода:

$$M\{\hat{a}(t), a(t+1)\} = [\hat{a}(t)] = C_{\hat{a}(t)} = \tag{10}$$
$$= \begin{cases} C_0(t) = C_0(t-1)+1, C_1(t) = C_1(t-1) \mid a(t+1) = 0; \\ C_0(t) = C_0(t-1), C_1(t) = C_1(t-1)+1 \mid a(t+1) = 1. \end{cases}.$$

При воспроизведении анализируются состояния этих счетчиков, и текущий символ формируется в зависимости от выполнения порогового условия:

$$a(t+1) = HM^{-1}\{[\hat{a}(t)], t\} = HM^{-1}\{C_{\hat{a}(t)}(t)\} = \tag{11}$$
$$= \begin{cases} 0 \mid C_1 - C_0 < 0; \\ 1 \mid C_1 - C_0 \geq 0. \end{cases}.$$

Наряду с запоминанием - неуменьшением значений счетчиков C_0 и C_1 (2.9), возможно забывание - равномерное уменьшение значений счетчиков во времени со скоростью изменения их содержимого значительно меньшей, чем при запоминании:

$$M\{\hat{a}(t), a(t+1)\} = [\hat{a}(t)] = C_{\hat{a}(t)} = \tag{11}$$

$$= \begin{cases} C_0(t) = C_0(t-1) + d_1, C_1(t) = C_1(t-1) - d_2 \mid a(t+1) = 0; \\ C_0(t) = C_0(t-1) - d_2, C_1(t) = C_1(t-1) + d_1 \mid a(t+1) = 1. \end{cases}.$$

где $d_1 >> d_2$. Введение забывания позволяет устранить случайные точки на траектории, не подтверждающиеся в процессе дальнейшего обучения.

Формирование словаря

Механизм памяти, чувствительный к числу прохождений заданной вершины в заданном направлении (механизм статистической обработки), является инструментом для анализа входной последовательности с точки зрения повторяющихся ее частей. Как было показано выше, одинаковые фрагменты последовательности отображаются преобразованием F в одну и ту же часть траектории.

Если мы имеем класс последовательностей $\{A\}$, в которых в разных комбинациях встречаются подпоследовательности $\{B_i\}$, то, отображая последовательности класса $\{A\}$ в n-мерное пространство и применяя к ним пороговое преобразование, мы сформируем множество траекторий $\{\hat{B}_i\}$, соответствующее множеству последовательностей $\{B_i\}$ - словарь.

Можно сказать, что преобразование $HM^{-1}MF$ при взаимодействии с входным классом $\{A\}$ формирует словарь, характеризующий траектории, соответствующие подпоследовательностям входного класса в пространстве данной мерности:

$$\{\hat{L}\} = HM^{-1}MF(\{A\}). \qquad (12)$$

В зависимости от величины порога h преобразования H слова словаря могут быть либо цепями, либо графами. Чтобы отразить этот факт, мы будем обозначать слова словаря символом $\hat{L}$ в отличие от траекторий $\hat{B}$.

Формирование синтаксической последовательности. Многоуровневая структура

Сформированный словарь часто встречающихся событий может быть использован для детектирования старой информации в потоке новой. Для этого необходимо поглощение фрагментов входной последовательности A, соответствующих словам словаря, и пропускание новой, относительно словаря, информации. В результате появляется возможность реализовать структурный подход к обработке информации.

Для решения задачи детектирования преобразование F^{-1} модифицируется для придания ему детектирующих свойств. Преобразование F_c^{-1} взаимодействует с входной последовательностью $\tilde{A}$, которая содержит, наряду со старой, некоторую новую информацию. Если на основании множества входных последовательностей A ранее был сформирован словарь $\{\hat{B}\} = HM^{-1}MF(\{A\})$, то использование преобразования F_c^{-1} позволяет сформировать так называемую синтаксическую последовательность или последовательность аббревиатур - C, характеризующую связи слов B словаря $\{B\}$ в последовательности A. Здесь $\{B\}$ есть множество подпоследовательностей, соответствующих всем цепям слов $\hat{B}$ словаря $\{B\} = F^{-1}(\{\hat{B}\})$.

В результате такого взаимодействия происходит формирование последовательности C, в которой заменяются нулями те части последовательности $\tilde{A}$,

соответствующие которым части траектории $\hat{\tilde{A}} = F(\tilde{A})$, совпадают с частями траектории $\hat{A}$. Другими словами, во входной последовательности $\tilde{A}$ заменяются нулями символы, соответствующие которым точки траектории $\hat{\tilde{A}}$ совпадают с точками сформированной ранее траектории $\hat{A} = F(A)$: $C = (..., c(-1), c(0), c(1), ..., c(t), ...)$, где:

$$c(t) = \begin{cases} \tilde{a}(t) \mid \hat{\tilde{a}}(t) \neq \hat{a}(t); \\ 0 \mid \hat{\tilde{a}}(t) = \hat{a}(t). \end{cases} \quad (13)$$

Здесь $\hat{a}(t) \in \hat{A}$, а $\hat{\tilde{a}}(t) \in \hat{\tilde{A}}$, или в другой записи:

$$C = F_c^{-1}(\tilde{A}), HM^{-1}(\{A\})). \quad (14)$$

Таким образом, отображение F_c^{-1} позволяет устранить из входной последовательности $\tilde{A}$ некоторую информацию, содержащуюся в словаре $\{\hat{B}\}$. Тем самым создается предпосылка построения многоуровневой структуры для лингвистической (структурной) обработки входной информации. Синтаксическая последовательность C, содержащая только новую, по отношению к словарю данного уровня, информацию, становится входной для следующего уровня. На следующем уровне, подобно описанному выше, из множества синтаксических последовательностей $\{C\}$ формируется словарь $\{\hat{D}\}$ и множество синтаксических последовательностей следующего уровня $\{E\}$ (см. рис. 5). Мы имеем стандартный элемент многоуровневой иерархической структуры: такая обработка с выделением поуровневых словарей может происходить на всех уровнях. Словарь следующего уровня является, в этом случае, грамматикой для предыдущего уровня, так как его элементами, при соответствующем выборе размерностей пространств этих уровней, являются элементы связей слов предыдущего уровня. Нечто подобное наблюдается на разных уровнях переключений в слуховой коре при восприятии речевой последовательности [Бехтерева, 1978].

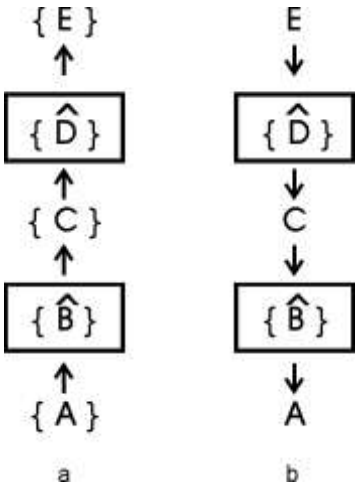

Рис. 5. Стандартный элемент многоуровневой иерархической структуры. На его вход поступает множество последовательностей $\{A\}$, формирующих в нижнем уровне словарь $\{\hat{B}\}$ и на его выходе - множество синтаксических последовательностей $\{C\}$, являющихся входными для верхнего уровня. В верхнем уровне на основе множества синтаксических последовательностей $\{C\}$ формируется словарь $\{\hat{D}\}$, а на его выходе - множество синтаксических последовательностей $\{E\}$.

Распознавание

Под распознаванием понимается процесс принятия решения о степени совпадения входной информации с ранее запомненной. Распознавание предполагает

предшествовавший ему процесс обучения. В основе механизма распознавания лежит сравнение входной последовательности $\tilde{A}$ и наиболее близкой ей, из запомненных, последовательности A, которая начинает воспроизводиться с помощью преобразования $HM^{-1}MF$ в ответ на входную последовательность $\tilde{A}$, с вычислением меры близости по Хеммингу:

$$D_x = \| \hat{\tilde{A}} - \hat{A} \| . \tag{15}$$

Вычисление D_x осуществляется суммированием расстояния по Хеммингу между соответствующими n-членными фрагментами входной и воспроизводимой последовательностей, полученных на каждом шаге:

$$D_x = \sum_T d(t), \; d(t) = \| \hat{\tilde{a}}(t) - \hat{a}(t) \| . \tag{16}$$

где T - длина траектории. Решение о совпадении с заданной степенью точности принимается сравнением с порогом по распознаванию.

1.3. Формирование семантической сети

Формирование многоуровневого представления

Рассмотрим формирование многоуровневого представления при обработке текстовой информации. Предметная область, представленная в виде множества текстов, с помощью описанного выше формализма подвергается статистическому анализу, в результате которого выявляются его словарные компоненты разных уровней.
При обработке текстов обычно рассматриваются следующие уровни обработки информации: морфологический, лексический, синтаксический, а также - семантический уровень. На каждом из уровней возможно формирование несколько словарей. Мы рассмотрим только некоторые из них. На морфологическом уровне - словарь флективных морфем $\{\hat{B}\}_1$. На лексическом уровне - словарь корневых основ $\{\hat{B}\}_2$. На синтаксическом уровне – словарь синтаксем $\{\hat{B}\}_3$, представляющих собой флективную структуру синтаксических узлов с выколотыми корневыми основами. На семантическом уровне – словарь попарной сочетаемости корневых основ $\{\hat{B}\}_4$.

Рассмотрим последовательные этапы формирования вышеперечисленных словарей. Сначала формируется словарь флективных морфем $\{\hat{B}\}_1$, так как они являются наиболее часто встречающимися языковыми единицами.

На следующем этапе формируется словарь корневых основ слов $\{\hat{B}\}_2$. После того, как сформировался словарь флективных морфем, фильтрация словарем флективных морфем множества текстов приводит к формированию словаря корневых основ, так как в результате взаимодействия множества текстов $\{A\}_1$ со словарем флективных морфем возникает множество синтаксических последовательностей первого уровня $\{C\}_1$ с купюрами вместо флексий, - множество последовательностей корневых основ $\{A\}_2 \equiv \{C\}_1$.

Далее формируется словарь синтаксического уровня $\{\hat{B}\}_3$. Этот словарь формируется фильтрацией через словарь корневых основ слов $\{\hat{B}\}_2$ фрагментов текста, соответствующих по длине предложению. Полученные при этом цепочки флективных

морфем – кластеризуются на подклассы по графемной структуре. Эти подклассы являются частями синтаксических классов, соответствующих основным синтаксическим узлам.

После того, как сформирован словарь синтаксем, формируется словарь попарной (смысловой) сочетаемости слов $\{\hat{B}\}_4$ [Рахилина, 2000], которая определяет семантику текста. При этом строится частотный портрет текста, то есть выявляются частоты p_i встречаемости корневых основ понятий и их устойчивых сочетаний, и частоты p_{ij} их попарной встречаемости в предложениях текста.

Перенормировка семантических весов

При формировании сети на основе большого корпуса текстов получаются корректные весовые характеристики понятий: частота их встречаемости приближается к их смысловому весу. При анализе малых по объему текстов частота встречаемости уже не характеризует важности понятия. В этом случае весовые характеристики понятий ассоциативной сети необходимо перенормировать. При этом на каждой итерации перенормировки понятия, связанные с понятиями, имеющими большой вес, свой вес увеличивают. Другие их равномерно теряют.

Сформированное первоначально статистическое представление текста – сеть слов с их связями перенормируется с помощью итеративной процедуры, аналогичной алгоритму сети Хопфилда [Hopfield, 1982], что позволяет перейти от частотного портрета текста к ассоциативной сети ключевых понятий текста:

$$w_i(t+1) = (\sum_{\substack{i \\ i \neq j}} w_i(t) w_{ij}) \sigma(\bar{E}), \qquad (17)$$

здесь $w_i(0) = \ln p_i$; $w_{ij} = \ln p_{ij} / \ln p_j$ и $\sigma(\bar{E}) = 1/(1+e^{-k\bar{E}})$ функция, нормирующая на среднее значение энергии всех вершин сети $\bar{E}$. Полученная числовая характеристика слов – их смысловой вес – характеризует степень их важности в тексте.

Сравнение семантических сетей. Классификация текстов

В результате получается так называемая ассоциативная (однородная) семантическая сеть N как совокупность несимметричных пар понятий $< c_i c_j >$, где c_i и c_j – понятия, связанные между собой отношением ассоциативности (совместной встречаемости в некотором фрагменте текста) [Харламов и др., 2008]:

$$N \cong \{< c_i c_j >\}, \qquad (18)$$

В данном случае отношение ассоциативности несимметрично: $< c_i c_j > \neq < c_j c_i >$.

Семантическая сеть, описанная таким образом, может быть переописана как множество так называемых звездочек $z_i = < c_i < c_j >>$:

$$N \cong \{z_i\} = \{< c_i < c_j >>\}, \qquad (19)$$

Под звездочкой $z_i = < c_i < c_j >>$ понимается конструкция, включающая главное понятие c_i, связанное с множеством понятий-ассоциантов $< c_j >$, которые являются семантическими признаками главного понятия, отстоящими от главного понятия в

ассоциативной сети на одну связь. Ассоциативные связи направлены от главного понятия к понятиям-ассоциантам.

Представленные таким образом семантические портреты текстов можно сравнивать между собой. Для этого необходимо вычислить степень пересечения их семантических сетей.

Если предварительно подготовить модели предметных областей в виде множеств описывающих их текстов (рубрики), можно автоматически классифицировать текст отнесением его к одной или нескольким рубрикам путем сравнения семантической сети текста с сетями этих рубрик.

2. Технология автоматической смысловой обработки текста

Ранее одним из авторов была реализована технология обработки текстовой информации TextAnalyst [Харламов, 1998], позволяющая автоматически выявлять ключевые понятия в тексте на основе только информации о структуре самого текста (независимо от предметной области и для нескольких европейских языков). Для этого формируетсяся частотный портрет текста, содержащий информацию о частоте встречаемости понятий текста, представленных как корневые основы соответствующих слов, или их устойчивых словосочетаний, встречающихся в тексте, а также об их совместной (попарной) встречаемости в смысловых фрагментах текста (например, в предложениях). Частотный портрет, таким образом, содержит информацию о частоте встречаемости понятий и их попарной (в терминах их ассоциативной связи) встречаемости в тексте. Использование хопфилдоподобного алгоритма [Хорфилд,] позволяет перейти от частоты встречаемости к смысловому весу (вес связей при этом оставаетя неизменным).

Эта обработка включает несколько этапов. На первом этапе осуществлялется первичная обработка: из текста удаляется нетекстовая информация, текст сегментируется на слова и предложения, из текста удаляются стоп-слова, рабочие и общеупотребимые слова, а оставшиеся слова подвергаются морфологической обработке. Для простоты анализа морфологическая обработка производится с использованием традиционного морфологического словаря – словаря первого уровня - $\{\hat{B}\}_1$. Далее с помощью программной модели искусственной нейронной сети из нейроподобных элементов с временной суммацией сигналов формируется словарь второго уровня – $\{\hat{B}\}_2$ – словарь корневых основ (и устойчивых словосочетаний). На следующем этапе строился частотный портрет текста, то есть выявляются частоты p_i встречаемости корневых основ понятий (полученных в результате морфологического анализа) и их устойчивых сочетаний, и частоты p_{ij} их попарной встречаемости в предложениях текста (то есть формируется словарь третьего уровня $\{\hat{B}\}_3$). И, наконец, на третьем этапе, частоты встречаемости перенормировываются в смысловые веса с использованием итеративной процедуры, похожей на алгоритм искусственной нейронной сети, предложенной Хопфилдом.

В результате итеративной процедуры перенормировки наибольшие веса получают понятия, связанные с наибольшим числом других понятий с большим весом, то есть те понятия, которые стягивают на себя смысловую структуру текста. Полученные таким образом смысловые веса ключевых понятий показывают значимость этих понятий в тексте. В дальнейшем эта информация используется для выявления предложений текста, содержащих наиболее важную информацию в тексте.

В результате получается так называемая ассоциативная (однородная) семантическая сеть N как совокупность несимметричных пар понятий $< c_i c_j >$, где c_i и c_j – понятия, связанные между собой отношением ассоциативности (совместной встречаемости в некотором фрагменте текста).

Наличие весовых коэффициентов ключевых понятий и их связей позволяет вычислить вес предложений текста. Последующее сравнение этого веса с пороговым значением позволяет удалить предложения с весом меньше порога. Последовательность оставшихся предложений с большим весом представляет собой основной смысл текста и является полученным автоматически рефератом этого текста.

Как говорилось выше, семантические портреты текстов можно сравнивать между собой. Для этого необходимо вычислить степень пересечения их семантических сетей. Если предварительно подготовить модели предметных областей в виде множеств описывающих их текстов (рубрики), можно автоматически классифицировать текст отнесением его к одной или нескольким рубрикам путем сравнения семантической сети текста с сетями этих рубрик.

3. TextAnalyst – персональный продукт для смыслового анализа текста

На основе технологии автоматической смысловой обработки текстов Научно-производственным инновационным центром «Микросистемы», г. Москва было разработано семейство программных продуктов для автоматического смыслового анализа текстовой информации TextAnalyst [Kharlamov, 2012; Sullivan, 2001]. Необходимо заметить, что система TextAnalyst возникла как развитие синтактико-семантического модуля системы распознавания речи.

Разработанная система обработки текстовой информации основана на использовании структурных свойств языка и текста, которые могут быть выявлены с помощью статистического анализа, реализованного в технологии TextAnalyst. На основе этой технологии реализовано автоматическое формирование описания семантики предметной области текста, и реализуются функции организации текстовой базы в гипертекстовую структуру, автоматического реферирования, кластеризации и классификации текстов, а также функция смыслового поиска.

Использование указанной технологии позволяет автоматически, на основе анализа статистики слов и их связей в тексте, реконструировать внутреннюю структуру текста. Статистический анализ выявляет наиболее часто встречающиеся элементы текста - слова или устойчивые словосочетания. Важной особенностью используемого подхода, является возможность автоматически устанавливать взаимосвязи между выявленными элементами текста. При выявлении связей учитывается статистика попарного появления слов во фрагментах исследуемого материала. Далее статистические показатели пересчитываются в семантические с помощью итеративной процедуры, идея которой заключается в том, что при расчете весовой характеристики элемента сети учитываются весовые характеристики элементов с ним связанных, а также учитываются численные показатели связей. После пересчета статистических характеристик в семантические, понятия, которые мало соответствуют анализируемой предметной области, получают малый вес, а наиболее представительные наделяются высокими показателями. Полученная семантическая сеть отражает внутреннюю структуру текста, значимость выделенных понятий, а также, показывает степень связанности понятий в тексте. Такое представление текста получается полностью автоматически.

Семантические веса элементов сети используются при расчете смысловой близости (релевантности) текстов. На их основе возможно выделение наиболее информативных участков текста. Использование ассоциативных связей элементов сети позволяет расширять поле поиска информации. Ответ на запрос пользователя, в этом случае, может содержать информацию явно не указанную в запросе, но связанную с ней по смыслу.

Программная реализация технологии

На основе алгоритмов обработки текстовой информации, описанных в разделе 1, создана система для анализа текстовой информации. Система реализована как инструмент для автоматического формирования баз знаний на основе множества естественно-языковых

текстов. Ядро системы выполнено как программный компонент (inproc server), соответствующий спецификации Component Object Model (COM) фирмы Microsoft.

Ядро системы реализует следующие функции. Нормализацию грамматических форм слов и вариаций словосочетаний. Автоматическое выделение базовых понятий текста (слов и словосочетаний) и их взаимосвязей с вычислением их относительной значимости. Формирование представления семантики текста (множества текстов) в форме семантической сети.

В состав ядра системы, помимо блока первичной обработки, входят следующие блоки (см. рис. 6): лингвистический процессор, блок выделения понятий текста, блок формирования семантической сети, блок хранения семантической сети.

Блок первичной обработки. Задачами этого блока являются извлечение текста из файла (входного потока данных) и подготовка его к обработке в лингвистическом процессоре. Подготовка текста заключается в очистке его от символов, неизвестных лингвистическому процессору, а также в корректной обработке таких единиц текста как аббревиатуры, инициалы, заголовки, адреса, номера, даты, указатели времени.

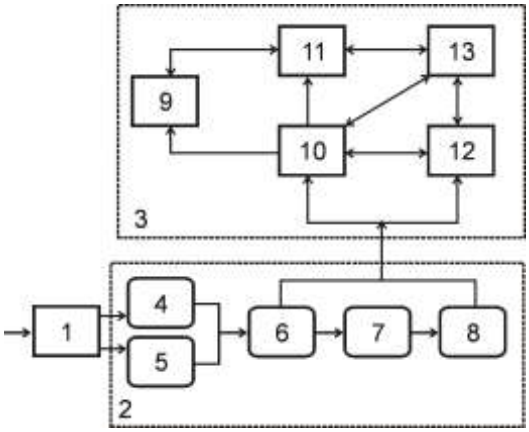

Рис. 6. Система анализа текстов содержит блок первичной обработки (1), лингвистический, и семантический процессоры. Лингвистический процессор (2) состоит из словарей: (4) слов разделителей, (5) служебных слов, (6) общеупотребимых слов, а также (7) флективных и (8) корневых морфем. Семантически процессор (3), в свою очередь, содержит: (9) блок отсылок в текст, (10) блок формирования семантической сети, (11) блок хранения семантической сети, (12) блок выделения понятий, и (13) блок управления.

Лингвистический процессор. Лингвистический процессор осуществляет предобработку входного текста (последовательности символов в определенной кодировке) на основе априорных лингвистических знаний, общих для выбранного языка (в настоящий момент поддерживаются несколько европейских языков, помимо русского и английского), и выполняет следующие функции. Сегментацию предложений текста на основе знаков пунктуации и специальных грамматических слов, и их фильтрацию. Нормализацию слов и словосочетаний - фильтрацию флексий (окончаний) с сохранением только корневых основ. А также - фильтрацию в тексте семантически несущественных, вспомогательных слов: удаляются предлоги, числительные и самые общеупотребимые слова с широким значением. И, наконец, маркировку общеупотребимых слов.

Сегментация предложений позволяет разбить текст на участки, которые могут содержать терминологические словосочетания предметной области и избежать выделения неадекватных словосочетаний на стыках таких участков.

В результате предобработки семантически близкие словосочетания приводятся к одинаковой форме (нормализуются). Маркировка общеупотребимых слов необходима с целью исключения их выделения как самостоятельных терминов при дальнейшем анализе.

База общих языковых знаний лингвистического процессора содержит словари, по одному для реализации каждой из четырех функции: словарь слов-разделителей предложения, словарь вспомогательных слов, словарь флексий и словарь общеупотребимых слов.

Блок выделения понятий. Блок выделения ключевых понятий предметной области (слов и словосочетаний) создан на базе программной модели иерархических структур из искусственных нейронных сетей (ИНС) на основе нейроподобных элементов с временной суммацией сигналов, и реализует алгоритмы автоматического формирования частотного словаря текста.

Число уровней ИНС в иерархической структуре определяет априорно заданную максимально допустимую длину понятия предметной области и равняется двадцати.

На первом уровне иерархической структуры представлен словарь двухбуквенных специальных слов предметной области - слов, пропущенных через все фильтры лингвистического процессора, и не отнесенных к общеупотребимым, а также двухбуквенных сочетаний из слов этого словаря. Там же хранятся двухбуквенные слова общеупотребимой лексики, входящие в устойчивые словосочетания, и их начальные двухбуквенные фрагменты. Второй уровень иерархической структуры представляет ИНС, хранящие словари трехбуквенных слов и сочетаний букв из словарей специальных и общеупотребимых слов, встреченных в тексте, в виде индексов элементов соответствующих словарей первого уровня, дополненных еще одной буквой. На последующих уровнях представление информации полностью однородно - в ИНС хранятся индексы элементов хранения более низкого уровня ИНС, дополненные одной буквой.

В процессе формирования представления информации в иерархической структуре из ИНС подсчитывается частота встречаемости каждого сочетания букв в соответствующих элементах ИНС. Частота слов (сочетаний букв, не имеющих продолжения на следующем уровне) используется для последующего анализа.

Сформированное таким образом представление лексики текста подвергается затем пороговому преобразованию по частоте встречаемости. Порог отражает степень детальности описания текста. В процессе статистического анализа в иерархической структуре ИНС выделяются устойчивые термины и терминологические словосочетания, которые служат далее в качестве элементов для построения семантической сети. При этом общеупотребимые слова, а также словосочетания, содержащие только общеупотребимые слова, опускаются.

Блок формирования семантической сети. Блок формирования семантической сети реализован как база данных, в которой представляются семантические связи понятий предметной области. Поскольку типы семантических связей [Осипов, 1997] в системе не определяются, такие связи представляют собой просто ассоциативные связи.

В качестве критерия для определения наличия семантической связи между парой понятий используется частота их совместной встречаемости в одном предложении. Превышение частотой некоторого порога позволяет говорить о наличии между понятиями ассоциативной (семантической) связи, а совместные вхождения понятий в предложения с частотой меньше порога считаются просто случайными.

Элементы семантической (ассоциативной) сети и их связи имеют числовые характеристики, отражающие их относительный вес в данной предметной области - семантический вес. При достаточно представительном множестве текстов, описывающих предметную область, значения частот встречаемости понятий отражают соответствующие семантические (субъективно оцениваемые) веса. Однако, для небольших обучающих выборок, в частности, при анализе отдельного текста, не все частотные характеристики соответствуют действительным семантическим весам - важности понятий в тексте. Для более точной оценки семантических весов понятий используются веса всех связанных с ними понятий, т.е. веса целого “семантического сгущения”. В результате такого анализа наибольший вес приобретают понятия, обладающие мощными связями и находящиеся как бы в центре “семантических сгущений”.

Основные функции системы TextAnalyst

На основе результатов работы модуля индексации реализованы следующие функции обработки текстовой информации: (1) функция формирования гипертекстовой структуры, (2) навигации по базе знаний, (3) формирования тематического дерева, (4) реферирования текстов, (5) автоматической кластеризации множества текстов, (6) сравнения текстов (автоматической классификации текстов), и, наконец, (7) функция формирования ответа на смысловой зарос пользователя – формирования тематического реферата.

После формирования семантической сети исходный текст, объединенный гиперссылками с семантической сетью, становится гипертекстовой структурой. Семантическая сеть в этом случае оказывается удобным средством навигации по тексту. Она позволяет исследовать основную структуру текста, переходя от понятия к понятию по ассоциативным связям. Пользуясь гиперссылками пользователь может быстро найти множество предложений текста, содержащих эти понятия. С помощью тех же гиперссылок он может перейти от любого предложения непосредственно к его контексту в тексте. С этой же целью пользователь может пользоваться минимальным древовидным подграфом семантической сети – тематическим деревом. В нем оказываются иерархически представленными основные и соподчиненные понятия сети, причем понятия нижнего уровня объясняют содержание понятий более высокого уровня. Тематическим деревом также можно пользоваться для навигации по базе знаний - оно напоминает оглавление текста.

Семантическая сеть с числовыми характеристиками ее компонент – понятий и их связей – позволяет вычислить вес каждого предложения в тексте. Множество предложений текста, выбранных в порядке их появления в тексте, вес которых превысил некоторый пороговый уровень, можно считать рефератом текста.

Семантическая сеть исследуемого текста (или группы текстов) может быть разбита на подсети удалением из нее слабых связей. Каждая такая подсеть группируется вокруг некоторого понятия с максимальным весом в данной подсети. Это понятие обозначает тему части текста или отдельных текстов, которые оказываются сгруппированными в данной подсети. Такая автоматическая кластеризация позволяет разбить множество текстов на рубрики, а также визуализировать динамику развития этих рубрик во времени.

Используя числовые характеристики семантической сети, можно сравнивать сети двух текстов с точки зрения вычисления их пересечения (общей части). То есть можно сравнивать степень совпадения текстов по смыслу. Если в качестве одного из текстов берется целая рубрика, то имеется возможность оценить степень принадлежности исходного текста к данной рубрике, то есть автоматически классифицировать тексты.

Система для смыслового анализа текстов позволяет реализовать также смысловой поиск (сформировать тематический реферат). Функция смыслового поиска, основываясь на ассоциативном иерархическом представлении содержания информации в базе, функциях кластеризации и классификации, осуществляет выборку информации, соответствующей запросу пользователя, и структурирует ее в соответствии с близостью к запросу.

Смысловой поиск, используя ассоциации, способен выдавать пользователю информацию явно не указанную в тексте запроса, но связанную с ней по смыслу. Использование такого подхода ведет не к увеличению выдаваемой пользователю информации, а к ее тщательному отбору и анализу по главному критерию - смысловой близости к запросу.

Электронная книга

Понятие «электронная книга» (е-книга) в настоящий момент еще не вполне устоялось. Под электронной книгой сейчас понимают и просто текст книги в электронном виде, и хорошо структурированную базу данных – электронный учебник. Совершенно ясно, что чтение плоского текста с экрана – дело безнадежное, если этот текст по объему превышает

две страницы. Формирование гипертекстовой страницы вручную – дело столь же неблагодарное, да еще и не дешевое (сколько стоит дизайн простой странички в Internet?). Другими словами, существует проблема подыскания подходящего инструмента для создания е-книги.

Удобство гипертекстовой структуры для представления текста на экране компьютера не вызывает сомнений, по крайней мере – по сравнению с плоским текстом. Желательно расширить его автоматическим группированием материала по темам. А также – автоматическим же выявлением тематической структуры текста. В дополнение к гипертекстовому представлению текста современные вычислительные средства предоставляют возможность его сопровождения другими мультимедийными модальностями: аудио и видео.

Кроме того, можно представить себе дополнительные возможности при создании е–книги. Такими дополнительными возможностями являются предоставление пользователю не только текста книги, но и традиционного оформления книги, включая шрифты и иллюстрации.

Нейросетевая технология для анализа неструктурированных текстов TextAnalyst, удовлетворяет большинству перечисленных принципов. Функциональность технологии позволяет автоматически сформировать индекс текста в виде перечня основных понятий и связей между ними. Формирует гипертекстовую структуру, в которой индекс является средством навигации по тексту, автоматически реферирует текст (можно также формировать реферат на заданную тему). Поэтому программа TextAnalyst может использоваться для формирования базы знаний е-книги. При наличии сформированной базы, тот же TextAnalyst может эффективно визуализировать информацию из этой базы.

Первый этап в создании е-книги в оболочке TextAnalyst – это формирование базы знаний, содержимое которой в дальнейшем будет предоставляться пользователю. Хотя обработка текста книги в оболочке TextAnalyst осуществляется автоматически – автоматически строится гипертекстовая структура текста и средство навигации по ней – тематическое дерево, требуются определенные усилия для приведения тематического дерева к наиболее удобному виду.

Так как исходный текст книги уже разбит на главы, гипертекстовая структура и тематическое дерево формируются для каждой главы в отдельности. Затем, автоматически сформированное тематическое дерево корректируется вручную: из него удаляются случайные темы, а грамматические формы слов приводятся к нужному виду. Если не пытаться добавить в структуру е-книги мультимедийной информации, дополнительного дизайна и дополнительного сервиса, можно считать, что после этого е-книга готова к использованию.

Подготовленная база готова для просмотра в оболочке TextAnalyst. Сначала пользователь выбирает и открывает одну из глав книги. Затем он может работать с тематическим деревом. Главная тема главы раскрывается содержащимися в ней подтемами. Каждая подтема также раскрывается вниз.

Каждой теме тематического дерева ставится в соответствие множество предложений, содержащих данное понятие. Далее, из любого из этих предложений можно перейти непосредственно в текст книги.

Такая ассоциативная навигация позволяет быстро познакомиться с содержанием книги на заданную глубину. Пороговые настройки позволяют изменять количество визуализируемого материала. При желании пользователь может получить в правом верхнем окне реферат выбранной главы, а, воспользовавшись функцией смыслового поиска, получить реферат на заданную тему.

При желании, отдельные понятия тематического дерева можно снабдить ссылками на мультимедийные приложения, а также каждую ссылку в текст сопроводить параллельной страницей этой книги, например, в PDF формате.

4. Дальнейшее развитие технологии. Объединение статистического и лингвистического подходов

Описанный выше анализ семантики целого текста является достаточно грубым инструментом, который не использует точной семантической информации, содержащейся в отдельных предложениях текста, но работает быстро и устойчиво. Возможно объединение описанного статистического подхода с лингвистическим, учитывающим точную семантику предложений текста. В этом случае используется та же методика формирования семантической сети, только в качестве исходного материала к этому анализу вместо обычного текста используется текст с синтаксически размеченными предложениями. Разметка осуществляется автоматически на основе правил согласования слов в синтаксических группах. В процессе такого анализа также формируется семантическая сеть, ключевые понятия которой ранжированы по их смысловой значимости в тексте, но наряду с ассоциативными отношениями между понятиями сети используются некоторые общепринятые семантические отношения, в результате чего вместо однородной (ассоциативной) сети формируется неоднородная семантическая сеть.

Заключение

Представленная технология автоматического смыслового анализа текстов TextAnalyst®, реализованная на основе нейросетевого подхода, будучи когнитивной, является статистической по своей природе, и позволяет автоматически выявлять ключевые понятия текста в их взаимосвязях, взвешенные их смысловыми весами (формировать ассоциативную сеть текста). Такое сетевое представление, в свою очередь, позволяет реализовать автоматическое реферирование текста и автоматическое сравнение (классификацию) текстов. Реализованный на основе этой технологии персональный продукт TextAnalyst® является удобным инструментом аналитика, беря на себя функцию предварительной обработки больших массивов текстовой информации. Причем, обработка текстов осуществляется по принципам, характерным для обработки текстовой информации в мозге человека. Полученное в результате такой обработки гипертекстовое представление текста с ассоциативной сетью ключевых понятий в качестве инструмента навигации по тексту, является уникальным способом нелинейного представления текста, характерным для человека, эффективно визуализирующим эту информацию.

Литература

[Kharlamov, 2012] URL: http://ww.analyst.ru.

[Глезерман, 1986] Глезерман Т.Б. Психофизиологические основы нарушений мышления при афазии // -М.: Наука, 1986.

[Бехтерева, 1978] Бехтерева Н.П. Мозговые коды психической деятельности // –Л.: «Наука», 1978.

[Харламов, 2006] Харламов А.А. Нейросетевая технология представления и обработки информации (естественное представление знаний) // - М.: «Радиотехника», 2006. - 89 с.

[Rumelhart et al., 1986] Rumelhart, D.E., Zipser D. Feature discovery by competitive learning // In: Parallel Distributed Processing // Rumelhart D.E. and McClelland J.L. and PDP Group eds. // - Cambridge, Mass.: MIT Press, 1986. Pp. 151 – 193.

[Хьюбель, 1988] Hubel D.H. Eye, brain and vision. // - New York: Scientific American Library, A Division of HPHLP, 1988.

[Радченко, 1969] Радченко А.Н. Моделирование основных механизмов мозга // - Л.: Наука, 1969.

[Roll, 1964] Rall W. Theoretical significance of dendritic trees for neuronal input-output relations // In: Neural Theory and Modelling. (Proc. of the 1962 Ojai Symp.) // Reiss R.F., ed., Stanford, Calif., Stanford University Press, 1964. Pp. 73 – 97.

[Rosenblatt, 1962] Rosenblatt F. Principles of Neurodynamics // - New York, 1962.

[Глезер, 1985] Глезер В.Д. Зрение и мышление // -Л.:, "Наука", 1985.

[Виноградова, 1975] Виноградова О. С. Гиппокамп и память // - М.: «Наука», 1975.
[Hopfield, 1982] Hopfield, J.J. Neural networks and physical systems with emergent collective computational abilities // In: Proc. Natl. Acad. Sci. 79, 1982. Pp. 2554 – 2558.
[Sowa, 1992] Sowa, J.F. Semantic networks, Encyclopedia of Artificial Intelligence / Shapiro S.C., ed. // - New York: Wiley, 1987 // revised and extended for the second edition, 1992.
[Sholl, 1953] Sholl D.A. Dendritic organization in the neurons of the visual and motor cortices of the cat // J. Anat., 87, 1953. Pp. 387 – 406.
[Kharlamov at al, 2004] Kharlamov A.A., Raevsky V.V. Networks constructed of neuroid elements capable of temporal summation of signals // In: Neural Information Processing: Research and Development // Jagath C. Rajapakse and Lipo Wang, eds. – New York: Springer-Verlag, 2004. 478 pp.
[Харламов и др., 2008] Харламов А.А., Раевский В.В. Перестройка модели мира, формируемой на материале анализа текстовой информации с использованием искусственных нейронных сетей, в условиях динамики внешней среды // Речевые технологии, N 3, 2008. Стр. 27-35.
[Hebb, 1949] Hebb D.O. The Organisation of Behavior // - New York: Wiley, 1949.
[Рахилина, 2000] Рахилина Е.В. Когнитивный анализ предметных имен: семантика и сочетаемость // - М.: Русские словари, 2000.
[Харламов, 1998] Харламов А.А., Ермаков А.Е., Кузнецов Д.М. TextAnalyst - комплексный нейросетевой анализатор текстовой информации // Вестник МГТУ им. Н.Э. Баумана. N 1, 1998г. -С. 32-36.
[Sullivan, 2001] Sullivan Dan Document Warehousing and Textmining // - New Yorl: Wiley publishing house, 2001.
[Осипов, 1997] Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии // - М.: Наука. Физматлит, 1997.