2Ш

УДК 519.7

Д.Е. Шуклин
Харьковский национальный университет радиоэлектроники, Украина

Разработка системы, обрабатывающей текст естественного языка на основе семантической нейронной сети

Описана структура связей семантической нейронной сети в виде синхронизированного линейного дерева и линии времени. Предложена многоуровневая структура системы, обрабатывающей текст естественного языка. Проведено описание семиотической обратной связи, служащей для коррекции результата обработки входного текста. Обнаружено, что наличие циркулирующих по контуру семиотической обратной связи символов обрабатываемого текста и управляющих спецсимволов напоминает процесс мышления. Сделано предположение, что семиотическая обратная связь может привести к возникновению в системе внутреннего монолога и самосознания.

Введение

Прогресс информационных технологий в немалой степени зависит от решения проблемы обработки текстов естественного языка. Построение вычислительной системы, обрабатывающей текст на естественном языке, требует наличия формальной модели понимания текста естественного языка. Разработка формальной модели невозможна без решения подзадач, на которые разбивается задача понимания. Реализацию подзадачи разбора следует выполнить, разделив ее по уровням абстракции на несколько последовательных уровней разбора: морфологический разбор, синтаксический разбор и семантический анализ.

Структура линейного дерева

Для решения задач морфологического и синтаксического анализа текста, а также задач анализа словоизменения применим семантическую нейронную сеть[1], близкую по свойствам формальной нейронной сети Неймана – Маккаллока – Питтса[2]. В подсети извлечения смысла из текста отдельный нейрон обозначает элементарное понятие, соответствующее этапу обработки, к которому относится данный подслой нейронной сети. Элементарными понятиями являются любые понятия естественного языка с законченным смыслом, такие, как символ, слог, слово, словосочетание, предложение, абзац, весь текст. Различным этапам обработки соответствуют различные уровни агрегации элементарных понятий, например: символ, слог, слово, словосочетание.

В качестве структуры семантической нейронной сети, выполняющей морфологический и синтаксический разбор, применим синхронизированное линейное дерево[3], [4], [5]. Синхронизированное линейное дерево состоит из слоев нейронов.

Каждому слою соответствует фронт волны обработки. Нейроны первого слоя соответствуют первой букве слова, второго – второй и так далее. Общее количество подслоев равно максимальному количеству букв в одном слове. Первый слой состоит из нейронов, распознающих первую букву, второй слой состоит из нейронов, распознающих первые две буквы, третий – первые три буквы [3, рис. 1].

Рис. 1. Структура дизконъюнктора

Слой состоит из классифицирующего подслоя несинхронизированных дизъюнкторов и распознающего подслоя синхронизированных конъюнкторов [3]. Функции классификации реализуются с помощью агрегирующих подслоев, состоящих из несинхронизированных нейронов-дизъюнкторов. Агрегирующие подслои не синхронизированных нейронов, выполняющих функции дизъюнкции, размещаются между подслоями синхронизированных нейронов выполняющих функции конъюнкции. В результате получается многослойная структура, в которой после каждого подслоя фронта волны находится подслой агрегирования [3, рис. 2, рис. 3]. Каждый нейрон-конъюнктор имеет одну входную связь с нейроном-дизъюнктором из текущего слоя и одну входную связь с нейроном из слоя рецепторов, соответствующим текущей букве. Дизъюнктор связан с несколькими нейронами-конъюнкторами из предыдущего слоя, соответствующими предыдущим буквам слова. Каждый нейрон-конъюнктор может иметь выходную связь с неограниченным количеством нейронов-дизъюнкторов из следующего слоя обработки. Таким образом, нейроны-дизъюнкторы текущего слоя соответствуют буквам, распознанным в предыдущем кванте времени.

Целесообразно упростить структуру связей между нейронами и уменьшить количество нейронов в системе. Это можно сделать, объединив в одном логическом элементе несинхронизированные дизъюнкторы из подслоя агрегирования и связанные с их аксонами синхронизированные конъюнкторы фронта волны [4, рис.1]. У каждого такого элемента будет по два дендритных дерева, одно – выполняющее функцию дизъюнкции входных градиентных значений, второе – выполняющее функцию конъюнкции входных градиентных значений и результата функции дизъюнкции. Можно сказать, что этот логический элемент представляет собой ансамбль из двух нейронов, назовем его дизконъюнктором, (рис. 1). Дизконъюнктор обозначим кругом, его дендрит дизъюнктора обозначим слева от этого круга (рис. 1, п. 1), дендрит конъюнктора сверху или снизу круга (рис. 1, п. 2), аксон – справа круга (рис. 1, п. 3). Дополнительные семантические связи дизконъюнктора будем обозначать линиями с точкой на этом круге (рис. 1, п. 4). Для

удобства внутри круга будем записывать символы, соответствующие дендриту конъюнктора, или точку, обозначающую возбуждение дизконъюнктора.

Один подслой дизконъюнкторов соответствует смежным подслоям несинхронизированных агрегирующих дизъюнкторов и синхронизированных конъюнкторов. Можно видеть, что структура соединений дизконъюнкторов аналогична структуре соединений конъюнкторов в упрощенном синхронизированном дереве [3, рис. 1]. Слой извлечения смысла в виде синхронизированного линейного дерева можно рассматривать как конечный автомат, в котором отдельным словоформам соответствуют субавтоматы, принимающие одновременно некоторое множество субсостояний, определяемое синонимичностью или омонимичностью входной символьной последовательности [4], [5]. Благодаря параллелизму вычислений, омонимия представляется в нейронной сети как множество одновременно возбужденных нейронов, соответствующих концептам, присущим данному слову. Синонимия представляется как возбуждение одного и того же нейрона-концепта разными словами-синонимами. Совокупность возбужденных нейронов сети в каждый момент времени является результатом морфологического и синтаксического разбора.

В процессе реализации синхронизированного линейного дерева необходимо обеспечить два режима работы дерева: программирование (обучение) и обработку (анализ) текста. Этап программирования обеспечивает обучение системы конкретным правилам морфологии и синтаксиса конкретного языка. В режиме обработки текста введенные в систему правила не изменяются, а только применяются к входным текстам. Как описано в [4], в процессе программирования синхронизированного линейного дерева в нем сохраняется только та информация, которая является для этого дерева новой, а уже имеющаяся в нем информация повторно не запоминается. На момент появления на входе линейного дерева некоторой обучающей последовательности в нейронной сети уже сформирована некоторая структура. В процессе программирования линейного дерева волна обработки распространяется по нейронным связям этой структуры так же, как и в режиме обработки текста. Однако в процессе программирования активирована подсистема синтеза нейронной структуры. В отличие от режима обработки текста, при затухании волны в режиме программирования синхронизированное линейное дерево достраивается новыми дизконъюнкторами. Эти дизконъюнкторы связываются своими дендритами-дизъюнкторами с возбужденными нейронами предыдущего фронта волны, а дендритами-конъюнкторами – с возбужденными нейронами-рецепторами. Очевидно, что новые нейроны возбуждаются входными сигналами от возбужденных дизконъюнкторов предыдущего фронта волны и возбужденными рецепторами текущего кванта времени, обеспечивая незатухающую волну обработки. В процессе программирования синхронизированного линейного дерева подсистемой синтеза нейронной структуры производится обработка спецсимволов "***()*.*$%_-***" [4], [5], обеспечивая управление формированием структуры нейронной сети.

Структура линии времени

Процесс разбора текста в синхронизированном линейном дереве протекает как некоторый процесс, распределенный во времени. Каждому кванту времени в

синхронизированном линейном дереве соответствует фронт волны обработки. Время представляет собой последовательность квантов. Последовательность квантов организована в виде линии, в которой каждый квант времени связан с предшествующим и последующим квантами. Результатом разбора, извлеченным из обработанной части текста в течение одного кванта времени, является мгновенное состояние синхронизированного линейного дерева. Мгновенное состояние включает в себя мгновенный снимок множества нейронов, множества связей между нейронами и множества внутренних состояний нейронов.

Введем понятие линии времени. Линия времени представляет собой группу нейронов. Каждому кванту времени соответствует некоторый фронт волны обработки и некоторый нейрон линии времени. Нейроны линии времени связаны друг с другом и образуют линию, в которой каждый нейрон имеет связи с двумя другими нейронами линии времени: с одним нейроном предыдущего кванта времени и с одним нейроном следующего кванта времени. Каждый нейрон линии времени образует семантические связи со всеми дизконъюнкторами синхронизированного линейного дерева, возбужденными в соответствующий квант времени. Это множество возбужденных дизконъюнкторов образует волну обработки, соответствующую нейрону линии времени, с которым образованы связи. Линия времени является кратковременной памятью, отличающейся от долговременной памяти [6, с. 245], обеспечивая память о событиях, происходящих в процессе анализа текста. Она может применяться для обеспечения анализа текста на уровне дискурса, позволяя восстановить состояние сети в предшествующие моменты времени и для обеспечения процесса программирования структуры нейронной сети.

Линия времени может быть применена и для реализации долговременной памяти. Некоторые данные нейрофизиологии говорят о том, что существует вероятность наличия у человека долговременной памяти, которая не забывает ни одного восприятия, попавшего на органы чувств человека. Независимо от того, имеет ли место такой процесс в действительности, в искусственной вычислительной системе можно хранить всю линию времени, начиная с момента первого включения. Однако целесообразность такого подхода вызывает сомнения. Поэтому в имеющихся реализациях линия времени сохранялась только на время одного сеанса общения пользователя с системой. Наличие линии времени позволяет прослеживать историю беседы и поддерживать диалог, связанный с воспоминаниями о произнесенных фразах. Это возможно благодаря установке обратных связей от нейронов линейного дерева к нейронам квантов времени линии времени, в момент которых были распознаны соответствующие понятия. При необходимости по этим обратным связям возможно восстановление понятий и времени квантов, в которые были восприняты эти понятия. В результате появляется возможность обрабатывать связи между элементами текста, сильно разнесенными во времени поступления в систему, но с соответствующими друг другу понятиями. Таким образом, можно говорить, что линия времени обеспечивает на уровне дискурса обработку некоторого подмножества непроективных связей между элементами обрабатываемого текста.

Наличие линии времени позволяет дополнительно упростить реализацию синхронизированного линейного дерева и увеличить эффективность его обработки. Все дизконъюнкторы, которые находятся в возбужденном состоянии в некотором

кванте времени, связаны семантическими связями с соответствующим этому кванту нейроном линии времени. Поэтому возможно полностью исключить обмен градиентными данными между дизконъюнкторами и полностью лишить нейрон своего внутреннего состояния, в том числе и возбуждения. В результате нейронная сеть линейного дерева не изменяется во время обработки текста, что позволяет размещать эту сеть на Read-Only устройствах, обрабатывать многопользовательские запросы к нейронной сети, управлять транзакциями и нитями потоков. Состояние дизконъюнктора «возбужден / пассивен» в данный момент времени определяется по тому, имеет ли этот дизконъюнктор связи с соответствующим нейроном линии времени (рис. 2). Поэтому синхронизированное линейное дерево может быть представлено в памяти машины в виде однонаправленного графа, в котором узлы и связи нагружены дополнительными атрибутами, а нейроны (узлы) не имеют внутреннего состояния.

Рис. 2. Общий вид бинарного синхронизированного линейного дерева с линией времени

Нейроны бинарного синхронизированного линейного дерева представляют узлы некоторого графа. Аксоны и дендриты нейронов представляют ориентированные ребра этого графа. Как было показано в [7], реализация связей между нейронами представляет собой списки записей некоторой структуры. Однотипные связи помещаются в один список, сохраняемый в теле нейрона в виде одной секции. Для реализации синхронизированного линейного дерева наиболее простой вид структуры, обеспечивающий требуемую функциональность, будет иметь вид одного поля. В этом поле будет храниться идентификатор нейрона, с которым установлена связь. Таким образом, секции будут представлять собой списки связей между нейронами.

Структура системы, обрабатывающей текст

Обработку текста естественного языка принято разделять на операции морфологического и синтаксического разбора, семантического анализа и синтеза. Операции морфологического и синтаксического разбора реализуем двумя

синхронизированными линейными деревьями. Каждое такое дерево обрабатывает понятия своего уровня абстракции. Обработка предложения на этих уровнях организована последовательно, уровень морфологического разбора готовит данные для уровня синтаксического разбора. Для каждого линейного дерева организуется своя линия времени. Линейное дерево уровня морфологического разбора выполняет выделение из текста отдельных слов, распознавание и разбиение этих слов на морфемы и определение для каждого слова на основе произведенного разбора признаков, требуемых для синтаксического разбора. Линейное дерево уровня синтаксического разбора проводит подготовительную работу, необходимую для полноценного семантического разбора. Синтаксический разбор определяет синтаксическую структуру текста, синтаксические связи между словами и синтаксические признаки слов, входящих в этот текст [8, с. 34]. Уровень синтаксического разбора текста назначает словам синтаксические признаки, находящиеся в памяти системы, соответствующие этим словам и необходимые для проведения семантического разбора. В число признаков, назначенных словам синтаксическим разбором, входят уникальные идентификаторы слов, по которым возможно сопоставление в процессе семантического разбора соответствующих этим словам семантических единиц.

Задачей уровня семантического анализа является формирование имитационной модели фрагмента реальности, описываемой обрабатываемым текстом. Такая модель не может быть сформирована на данных, содержащихся только в анализируемом предложении. Формирование модели требует наличия в памяти системы модели реальности, в которой существует эта понимающая система. Уровень семантического анализа возбуждает модели понятий, находящиеся в памяти системы в соответствии с понятиями, извлеченными из текста предыдущими уровнями разбора. Совокупность возбужденных моделей понятий образует семантически связанную модель фрагмента реальности, описываемого анализируемым текстом. Семантический анализ содержимого текста реализуем, используя модель предметной области в виде нейронной экспертной системы [9]. Операцию синтеза текста реализуем программно, путем восстановления текста по возбужденному нейрону-эффектору в синхронизированном линейном дереве.

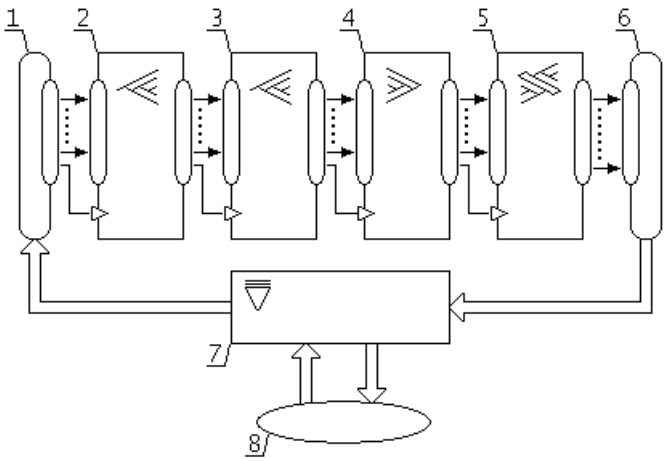

Рис. 3. Блок-схема системы, обрабатывающей естественный язык

На рис. 3 обозначены:
1. Слой рецепторов;
2. Линейное дерево морфологического разбора;
3. Линейное дерево синтаксического разбора;
4. Экспертная система с моделью предметной области;
5. Слой синтеза реакций системы (синтеза текста);
6. Слой эффекторов;
7. Коммутатор;
8. Внешняя среда.

Описываемая система, обрабатывающая естественный язык, содержит рецепторы, слои морфологического и синтаксического разбора; слой семантического анализа в виде экспертной системы с моделью предметной области; слой синтеза символьных последовательностей; эффекторы и коммутатор символьных потоков. Коммутатор 7 содержит в себе очередь символов текста, обрабатываемых системой. Входной текст из внешней среды 8 попадает в коммутатор 7, и пройдя через него, на слой рецепторов 1. Два синхронизированных дерева 2 и 3 соответственно, проводят морфологический и синтаксический разбор текста [4]. Результаты синтаксического анализа текста поступают на вход экспертной системы 4, в которой содержится модель предметной области, сопоставляемая с текстом естественного языка. Экспертная система [9] проводит семантический анализ понятий, содержащихся в тексте. Результаты семантического анализа попадают в слой синтеза символьных последовательностей 5 и далее на слой эффекторов 6. Со слоя эффекторов символьная последовательность попадает в коммутатор 7. Символьные последовательности, снимаемые со слоя эффекторов, могут быть двух видов: внешняя реакция системы и внутренний промежуточный результат. Тип последовательности определяется по результату работы блоков 2, 3, 4 и 5 как наличие возбуждения эффекторов соответствующего спецсимволам переключения режимов коммутатора. Если символьная последовательность является внешней реакцией системы, то коммутатор 7 передает ее во внешнюю среду 8. Если символьная последовательность является промежуточным результатом, то коммутатор 7 подает ее на слой рецепторов 1. В результате в системе возникает семиотическая обратная связь. Система получает возможность оценить и скорректировать промежуточные результаты и только после окончательного формирования результата передать его во внешнюю среду.

Коммутатор 7 был реализован программно, поэтому на рис. 3 он находится между слоями эффекторов 6 и рецепторов 1. Результаты экспериментов показали, что наличие семиотической обратной связи и повторной обработки символьных последовательностей, синтезированных в блоке 5, значительно облегчает реализацию словообразования и словоизменения, так как в этом случае эти операции производятся в блоках 2 и 3. При этом перевод словарных статьей из одного субсостояния в другое осуществляется путем подачи на слой рецепторов синхронизированного линейного дерева символьных последовательностей, синтезируемых в блоке 5. Также было обнаружено, что наличие внутренних спецсимволов, циркулирующих по контуру семиотической обратной связи, позволяет на примитивном уровне реализовать операции планирования беседы и управления вниманием в пределах нескольких

обсуждаемых тем. Текущий контекст беседы сохраняется в контуре семиотической обратной связи как состояние циркулирующих по этому контуру спецсимволов. Эти спецсимволы управляют субсостояниями нейронных субавтоматов синхронизированного линейного дерева, определяя возбуждение словарных статей, относящихся к текущему контексту.

Имитационная модель позволяет прогнозировать развитие событий в моделируемом фрагменте реальности. Анализируя состояние имитационной модели, система в состоянии смоделировать влияние предполагаемых событий на моделируемый фрагмент реальности. На основе модели возможно создание нескольких вариантов моделей воздействия событий на имитационную модель и прослеживание различных вариантов развития этой модели в различных контекстах. После проведения моделирования воздействий на имитационную модель система в состоянии выбрать один из вариантов развития модели и продолжить моделирование с учетом выбранного варианта. Этот процесс моделирования происходит циклически. Можно сказать, что в понимающей системе возникает семиотическая обратная связь. Внутренняя обратная связь появляется в результате применения к имитационной модели созданных на ее основе воздействий. Эта внутренняя обратная связь может привести к известным в психологии внутреннему монологу и самосознанию [10].

Благодаря наличию внешней обратной связи системе становится известно о результатах ее действий в окружающей среде. Синтезированные реакции системы, будучи спроецированными во внешний мир, возвращаются по принципу обратной связи в виде восприятия фрагмента реальности, в которой существует система. В случае несовпадения реальных и моделируемых восприятий система производит коррекцию имитационной модели. Модель в процессе своего функционирования влияет сама на себя, используя внутреннюю и внешнюю обратные связи. Этим механизмом обеспечивается самообучение системы и построение внутренней модели внешнего мира. Есть основания полагать, что описанный механизм является основным в обучении. Только благодаря наличию множества обратных связей человек обучается координировано управлять мышцами своего тела. Можно позволить сознанию человека получать информацию о состоянии внутренних органов, которыми человек обычно не может управлять сознательно. Тогда через некоторое время благодаря такой искусственной обратной связи человек получает возможность сознательно управлять состоянием этих внутренних органов. Например, используя активное биоадаптивное управление [11], можно обучить испытуемого сознательно управлять частотой сокращений своего сердца.

Заключение

В разработанной системе механизм семиотической обратной связи напоминает феномен внутреннего монолога, циркуляция по контуру этой обратной связи символов обрабатываемого текста и управляющих спецсимволов напоминает процесс мышления. Экспертная система может быть поставлена в соответствие подсознанию. В реализованной системе экспертная система 4 не была интегрирована в степени, достаточной для того, чтобы определить степень влияния наличия или отсутствия модели предметной области на процесс общения.

Также не был разработан механизм автоматического изменения имитационной модели на основе работы внутренней и внешней обратных связей. Представляет интерес поставить эти эксперименты с наличием в системе достаточно развитой модели предметной области.

Литература

1. Дударь З.В., Шуклин Д.Е. Семантическая нейронная сеть как формальный язык описания и обработки смысла текстов на естественном языке // Радиоэлектроника и информатика. – 2000. – № 3. – С. 72-76.
2. Дж. фон Нейман. Теория самовоспроизводящихся автоматов / Закончено и отредактировано А. Бёрксом. – М.: Мир, 1971. – 384 с.
3. Шуклин Д.Е. Структура семантической нейронной сети, извлекающей в реальном времени смысл из текста // Кибернетика и системный анализ. – 2001. – № 2. – С. 43-48.
4. Шуклин Д.Е. Структура семантической нейронной сети, реализующей морфологический и синтаксический разбор текста // Кибернетика и системный анализ. – 2001. – № 5. – С. 172-179.
5. Шуклин Д.Е. Морфологический и синтаксический разбор текстов как конечный автомат, реализованный семантической нейронной сетью, имеющей структуру синхронизированного линейного дерева // Мат-лы V науч.-практ. семинара «Новые информационные технологии». – М.: МГИЭМ, 2002. – С. 74-85
6. Хофман И. Активная память: Эксперимент. исслед. и теории чловеч. памяти: Пер. с нем. / Под ред. Б.М. Величковского, Н.К. Корсаковой. – М.: Прогресс, 1986. – 312 с.
7. Дударь З.В., Шуклин Д.Е. Реализация нейронов в семантических нейронных сетях // Радиоэлектроника и информатика. – 2000. – № 4. – С. 89-96.
8. Виноград Т. Программа, понимающая естественный язык. – М.: Мир, 1976. – 296 с.
9. Шуклин Д.Е. Применение семантической нейронной сети в экспертной системе, преобразующей смысл текста на естественном языке // Радиоэлектроника и информатика. – 2001. – № 2. – С. 61-65.
10. Кучинский Г.М. Психология внутреннего диалога. – Мн.: Университетское, 1988. – 206 с.
11. Пат. 20874А України, МПК А 61 В 5/16. Спосіб активного біоадаптивного регулювання психофізіологічного стану та пристрій для його здійснення / А.І. Тесленко, Д.Є. Шуклін (Україна). – № 96052105; Заявлено 28.05.96; Опубл. 27.02.98, Бюл. № 1.-3.1. – 40с. ил.

A connection structure of semantic neural network as the synchronised linear tree and the line of time are described. The multilevel structure of the natural language text processing system is offered. The semiotics feedback serving for correction of the entrance text processing result is described. It is revealed, that presence of the entrance text symbols and the managing subsymbols circulating in the semiotics feedback contour reminds a process of thinking. It has been assumed that the semiotic feedback can cause an internal monologue and consciousness to occur in the system.

Статья поступила в редакцию 26.07.02.