УДК 002:004+81'32

Антонов С.Г., Совпель И.В.

К ЗАДАЧЕ РАЗРАБОТКИ ЛИНГВИСТИЧЕСКИХ ПРОЦЕССОРОВ

Введение. Поскольку естественный язык (ЕЯ) является универсальным средством описания действительности и коммуникации с вычислительной системой, то актуальность его автоматической обработки в составе современных информационных технологий, безусловно, очень высока. Данное направление, называемое иначе NLP (natural language processing), связано, чаще всего, с моделированием и обработкой ЕЯ в целях автоматизации его понимания. Учитывая, что ЕЯ «проявляется» как в виде текста, так и в виде речи, мы будем говорить в дальнейшем только о тексте (в самом широком смысле этого слова). Суть понимания текста состоит в представлении его содержания в терминах и отношениях некоторой заданной системы знаний, определяемой целевой задачей. Например, это может быть множество ключевых слов с указанием для них синонимических и иерархических отношений, множество объектов с указанием для них атрибутивных и функциональных отношений. В любом случае автоматизация понимания текста требует разработки процедур его лингвистического анализа (ЛА). И, таким образом, об этих процедурах, характерных для большинства важнейших приложений NLP (автоматизации инженерии знаний, информационного потока, машинного перевода, автоматического реферирования и др.), можно в совокупности говорить как о базовом лингвистическом процессоре (БЛП) [1].

Как показывают проведенные исследования и опыт разработки многих NLP-приложений, функциональность БЛП должна включать: форматирование текста, его лексический, лексико-грамматический, синтаксический и семантический анализ. Необходимость первого, вспомогательного, этапа ЛА обусловлена существованием различных форматов документов, и поэтому для упрощения процесса их обработки производится преобразование этих документов в некоторый единый формат, максимально сохраняющий стилистическую и структурную разметку документов. Кроме того, на данном этапе осуществляется разбиение текста на параграфы, выделение заголовков, подзаголовков и разделов текста, а также производится фильтрация вспомогательного текста (в случае обработки Интернет-документов к вспомогательному тексту можно отнести, например, тексты кнопок, меню и т.д.).

На этапе лексического анализа текста прежде всего распознаются границы его слов и предложений. Здесь же частично или полностью решаются задачи распознавания имен собственных, аббревиатур, электронных адресов, цифровых и других знаковых комплексов. Этот вид ЛА иначе называют сегментацией текста.

Задачей лексико-грамматического анализа текста является определение лексико-грамматической категории каждого его слова с учетом контекста. Множество всех лексико-грамматических категорий ЕЯ обычно задается заранее разработанным классификатором его лексико-грамматических свойств, основанным на разделении слов на части речи.

На этапе синтаксического анализа текста осуществляется распознавание в каждом его предложении синтаксических отношений и представление их, как правило, в виде функционального или синтаксического дерева, в котором словам предложения указывается их грамматическая функция и определяется тип синтаксической связи между ними.

Что касается семантического анализа текста, то, принимая во внимание, что текст является самым надежным и эффективным источником знаний как о предметной области/внешнем мире, так и о самом языке, есть смысл рассматривать его в контексте известных положений искусственного интеллекта, согласно которым основными типами знаний являются объекты/классы объектов, факты и правила, отображающие закономерности предметной области/внешнего мира. Поэтому упор должен быть сделан прежде всего на автоматическое распознавание в тексте семантических компонентов типа «концепт», что соответствует объекту/классу объектов, семантических отношений между ними типа «субъект-акция-объект» (САО), что соответствует факту, и «причина-следствие», что соответствует правилу, с имеющими место атрибутами и с учетом синонимических и иерархических отношений [2]. Например, предложение английского языка «Today the user can download 10,000 papers from the Web by typing the word screen» содержит:

Концепты:

- user
- 10,000 papers
- Web
- word screen

CAO1:

- Субъект: user
- Акция: download
- Объект: 10,000 papers
- Атрибут-предлог: from
- Непрямой объект: Web
- Атрибут-прилагательное: -
- Атрибут-наречие: -

CAO2:

- Субъект: user
- Акция: type
- Объект: word 'screen'
- Атрибут-предлог: -
- Непрямой объект: -
- Атрибут-прилагательное: -
- Атрибут-наречие: -

Причина-следствие:

CAO-причина (CAO2): user – type – word screen

CAO-следствие (CAO1): use – download -10,000 papers – from Web

Эти семантические компоненты и отношения являются универсальными, не зависящими от предметной области и конкретного языка, охватывают практически все информативные словоупотребления каждого предложения анализируемого текста, обеспечивают максимально возможное обобщение в виде паттернов лингвистических правил распознавания других, требуемых конкретным приложением, семантических компонентов и отношений, например, таких как время, параметр, месторасположение, состав, часть-целое и др., и в конечном счете обеспечивают распознавание не только тематического, но и логического содержания текста.

Таким образом, функциональность БЛП в направлении увеличения глубины анализа ЕЯ должна заканчиваться построением некоторой структуры, включающей в себя синтаксические деревья и соответствующие им множества объектов, фактов и правил и допускающей проектирование ее последующей обработки для конкретных приложений в режиме открытой архитектуры, обеспечивающей эффективное использование указанного выше основного ресурса решения задачи семантического анализа текста в виде совокупности эвристических алгоритмов. При этом особое внимание должно быть уделено достижению высоких показателей эффективности работы ЛП, что в свою очередь требует построения качественной технологии тестирования ЛП на всех указанных этапах лингвистического анализа текста.

Отметим, что указанные компоненты знаний могут и не фиксироваться в выходной структуре в явном виде. В этом случае, как, например, в [3], фиксируются только отношения SimpleNounPhrase, VerbPhrase, NounPhrase_additional, ComplexSentence, предопределяющие знания основных типов. Результаты всех указанных этапов обработки текста образуют его лингвистический индекс (LI), который формально может быть представлен в виде:

$$LI = <W, POS, SYN, REL>,$$

где W – множество слов текста; POS, SYN и REL – отображения слов в множества их соответственно лексико-грамматических и синтаксиче-

***Антонов Сергей Георгиевич**, д.т.н., научный консультант ООО «Крос2000», г. Москва.*

***Совпель Игорь Васильевич,** д.т.н., профессор кафедры информационных систем управления УО «Белорусский государственный университет».*

Беларусь, БГУ, 220050, г. Минск, пр. Независимости, 4.

ческих классов (меток), а также меток семантико-синтаксических отношений, предопределяющих знания основных типов.

Такая модель обладает одним очень важным свойством – она допускает естественное включение в себя новых компонентов в соответствии с разрабатываемым приложением БЛП. Так, например, при решении задачи автоматизации инженерии знаний (если текст рассматривать как их основной источник) могут быть добавлены компоненты, соответствующие отображениям слов в множества типов основных и атрибутивных знаний [4]. В этом случае происходит естественный переход от LI текста к его семантическому индексу, рассматриваемому в качестве эффективной модели представления автоматически распознаваемых в тексте знаний, которая, во-первых, в отличие от известных моделей, не ориентирована на конкретные механизмы вывода, но может быть при необходимости трансформирована в любую из этих моделей, и, во-вторых, обеспечивает пользователю ЕЯ-доступ к распознаваемым в тексте знаниям.

Отмеченная ранее функциональность базового ЛП обеспечивается лингвистической базой знаний (ЛБЗ), которая, прежде всего, включает в себя базовый аннотированный словарь ЕЯ, базовый аннотированный корпус текстов и множество лингвистических правил (паттернов) анализа текста на различных уровнях глубины ЕЯ. Такие паттерны, получаемые лингвистами-экспертами, являются основой разработки машинных алгоритмов для большинства этапов автоматического ЛА текста. С целью повышения эффективности этих алгоритмов разработан специальный язык расширенных регулярных выражений WRE [3] для формального описания самих лингвистических правил. Он максимально соотнесен с требованием его доступности для использования экспертами, возможностью обобщения разрабатываемых правил путем оперирования не только символами алфавита, морфемами, отдельными словами и их совокупностями, но и лексико-грамматическими, синтаксическими и семантическими классами лексических единиц. Ниже приводится пример описания на языке WRE правила распознавания именной группы:

AT|ATI|DT|DTI|DTS ? (RB * JJ|VBN)* NN|NNS +

Правило задает опциональный артикль или детерминатив, за которым следует произвольное количество наречий и прилагательных, за которыми в свою очередь следуют существительные. Заданному правилу в предложении на английском языке

Transmission_NN for_IN electrically_RB driven_VBN tool_NN

соответствуют последовательности слов Transmission_NN, electrically_RB driven_VBN tool_NN, driven_VBN tool_NN и tool_NN. Если из всех возможных соответствий выбирать «самый левый самый длинный», то найденные последовательности слов и будут соответствовать именным группам в предложении, т.е. в нашем примере Transmission_NN и electrically_RB driven_VBN tool_NN.

Еще одним важным достоинством WRE оказалось то, что он позволил, используя теорию конечных автоматов, значительно оптимизировать алгоритмическое обеспечение базового ЛП, основу которого составляет очень трудоемкая процедура сопоставления входной цепочки обрабатываемого текста с множеством описанных в ЛБЗ на языке WRE лингвистических правил. Это, в частности, позволило значительно превзойти по скорости обработки текста даже такого известного производителя промышленных лингвистических ресурсов, как Connexor [5].

Учитывая, что множество паттернов является одним из основных лингвистических ресурсов ЛБЗ, должна быть разработана эффективная технология построения паттернов, которая в нашем случае включает следующие основные этапы:

1. Задание лексических единиц, которые заведомо являются носителями объекта или отношения, для распознавания которого строится паттерн (например, для отношения типа «isA» в английском языке это может быть пара Asian country и Japan).
2. Поиск в корпусе текстов предложений, в которых одновременно встречаются все заданные на этапе 1 лексические единицы (с учетом их словоизменения и синонимии).
3. Экспертный анализ множества найденных на этапе 2 предложений и формулирование паттерна.
4. Тестирование сформулированного паттерна на эталонном корпусе текстов (предложений) с целью определения его качественных характеристик (точности и полноты), определяющих возможность включения паттерна в состав ЛБЗ.

Относительно перечисленных процедур отметим следующее. Желательно, чтобы те лексические единицы, о которых идет речь на этапе 1, имели высокую частотность, обеспечивающую, очевидно, получение на этапе 2 максимально возможного числа предложений, в которых планируемый к распознаванию лингвистический объект или отношение будут выражены разными лексическими и грамматическими средствами языка, что очень важно для построения паттерна в целом и его обобщения в частности. На этапе 2 и 3 мы говорим о предложениях только для простоты изложения. Понятно, что на входе могут быть заданы объект или отношение, носителями которых являются два и более предложений. Что касается собственно процедуры формулирования паттерна (этап 3), то она осуществляется исходя из определения этого понятия как формальной спецификации свойства набора примеров, определенной в терминах некоторого формального языка [6], а также с учетом функциональности используемого ЛП. Так, например, обобщение паттерна до уровня синтаксических отношений имеет смысл только в том случае, если используемый в дальнейшем при решении задачи ЛП на этапе анализа текста сможет довести его до этого уровня. То есть, обобщение паттерна есть смысл доводить до максимально возможного уровня глубины языка (от лексического до семантического), но не выше того, до которого сможет доводить анализ текста, используемый для обработки этого паттерна ЛП. Ранее нами был приведен пример паттерна для распознавания именной группы, в котором лексические единицы (артикли, прилагательные, существительные и др.) обобщены до уровня лексико-грамматических классов слов. Ниже приводится еще один пример паттерна для английского языка, предназначенного для распознавания в тексте причинно-следственных отношений.

Если в предложении текста присутствует семантическое отношение Субъект-Акция-Объект, причем все три компонента этого отношения не пусты, а Акция имеет семантический класс CAUSE (CAUSE :: = cause|result in|create|activate|generate… (более 40 конкретных акций)), то в данном предложении присутствует причинно-следственное отношение, в котором Причина выражена Субъектом, а Эффект – Объектом.

Описанная ситуация имеет, например, место в предложении:

The vacuum knife causes a shearing air flow

Здесь действительно имеется причинно-следственное отношение, в котором Причиной является Субъект «vacuum knife», а Эффектом - Объект «shearing air flow».

Что касается тестирования паттерна (этап 4), то оно осуществляется экспертом с использованием эталонного корпуса текстов K, заранее аннотированного лексико-грамматическими, синтаксическими и другими классами, в том числе и метками того лингвистического объекта или отношения, для распознавания которого сформулирован данный паттерн. В качестве показателей эффективности паттерна предлагается использовать широко применяемые в теории информационных систем понятия точности P и полноты R [7]. Адаптация к нашей задаче позволяет дать следующее их определение. Обозначим:

$P_i(K)$ – множество лингвистических объектов или отношений данного типа, полученных ЛП на основе паттерна P_i при обработке эталонного корпуса K;

$P_i^t(K)$ – подмножество тех элементов из $P_i(K)$, которые, по оценкам эксперта, распознаны ЛП корректно;

$E(K)$ – множество лингвистических объектов или отношений данного типа, имеющих место, по оценкам эксперта, в K.

Тогда: $P = \dfrac{\left|P_i^t(K)\right|}{\left|P_i(K)\right|}$; $R = \dfrac{\left|P_i^t(K)\right|}{\left|E(K)\right|}$.

Заметим, что попытка оценить здесь полноту R по классической схеме привела бы к решению задачи об объективной оценке количе-

ства элементов данного типа в корпусе K в соответствии с субъективным паттерном P_i, постановку которой нельзя считать корректной. Поэтому в данном случае полнота R паттерна P_i оценивается по отношению к количеству всех элементов данного типа в K, независимо от тестируемого паттерна. Поскольку решение задачи в общем требует разработки целого множества паттернов (например, в силу такой особенности ЕЯ как неоднородность его правил [8]), то очевидно, что на начальном этапе показатель полноты первых формулируемых паттернов будет, как правило, невысоким, что означает необходимость, с одной стороны, возможной их корректировки, а с другой – разработки в дополнение к сформулированным новых паттернов. Это, собственно, и характеризует процесс разработки системы паттернов как итерационный процесс.

Заключение. Изложенные аспекты решения задачи разработки эффективных лингвистических процессоров, в том числе и базовых, позволяют переходить к построению инструментальных программных средств интеллектуализации информационных систем, прежде всего, ЕЯ-интерфейса пользователя, семантического поиска, автоматизации инженерии знаний, а это путь к построению компьютерной системы знаний, как компонента кратко- и долговременной памяти искусственного интеллекта.

СПИСОК ЦИТИРОВАННЫХ ИСТОЧНИКОВ

1. Совпель, И.В. Базовые лингвистические процессоры: назначение, принципы построения, функциональность, состав, приложения / И.В. Совпель, А.В. Чеусов // Актуальные проблемы теоретической и прикладной лингвистики: материалы международной конференции, Минск, 15-16 июня 2010 г.; в 2 т.; редкол.: А.В. Зубов [и др.]. – Минск: МГЛУ, 2010. – Т. 2. – С. 122–126.
2. Совпель, И.В. Система автоматического извлечения знаний из текста и ее приложения // Искусственный интеллект. – 2004. – № 3. – С. 668–677.
3. Чеусов, А.В. Разработка алгоритмов и технологии построения многоязычного базового лингвистического процессора: диссертация на соискание ученой степени к-та технических наук: 05.13.17. – Минск, 2013. – 116 с.
4. Постаногов, Д.Ю. Автоматическая обработка естественного языка в задаче инженерии знаний и доступа к ним: диссертация на соискание ученой степени к-та технических наук: 05.13.17. – Минск, 2012. – 134 с.
5. Режим доступа: http://connexor.com.
6. Городецкий, В.И. Современное состояние технологии извлечения знаний из баз и хранилищ данных (часть I) // Новости искусственного интеллекта. – 2002. – № 3. – С. 3–12.
7. Солтон, Д. Динамические библиотечно-информационные системы. – М.: Мир, 1979. – 557 с.
8. Апресян, Ю.Д. Лингвистические процессоры для машинного фонда русского языка / Ю.Д. Апресян, О.С. Кулагина // Доклады второй всесоюзной конференции по созданию Машинного фонда русского языка / Институт русского языка АН СССР; ред.: Ю.Н. Караулов. – М., 1987. – С. 27–40.

Материал поступил в редакцию 05.12.13

ANTONOV S.G., SOVPEL I.V. **To a problem of development of linguistic processors**

The definition, functionality and linguistic knowledge base components of basic linguistic processor are presented. The model of results of linguistic text analysis in view of linguistic index and technology of linguistic patterns development are described.

УДК 004.8

Савицкий Ю.В., Давидюк Ю.И.

НЕКОТОРЫЕ АСПЕКТЫ ПРИМЕНЕНИЯ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ В ЗАДАЧЕ АНАЛИЗА СИГНАЛОВ ЭЭГ И ЭКГ

Нейросетевые методы анализа хаотических сигналов находят все большее применение в различных областях благодаря ряду преимуществ по сравнению с традиционными методами: возможностью исследования систем, математическая модель которых неизвестна (неизвестны математические соотношения, характеризующие поведение динамической системы); использованием для исследований выборки данных ограниченного объёма [1]. Высокая актуальность данного направления объясняется всё возрастающей потребностью в наличии эффективных средств для решения сложных нетривиальных задач в плохо формализуемых областях обработки информации.

Хаос в динамике означает чувствительность динамической эволюции к изменениям начальных условий. Старший показатель Ляпунова характеризует степень экспоненциального расхождения близких траекторий. Наличие у системы положительной экспоненты Ляпунова свидетельствует о том, что любые две близкие траектории быстро расходятся с течением времени, то есть имеет место чувствительность к значениям начальных условий.

В результате экспериментов установлено, что наиболее приемлемой для цели данного исследования является модель гетерогенной многослойной нейронной сети (НС) с нейронами сигмоидального типа в скрытом слое и линейными нейронами выходного слоя сети [2].

Для обучения НС применяется алгоритм обратного распространения ошибки (и его более быстродействующие модификации), использующий метод градиентного спуска для минимизации функции среднеквадратичной погрешности [2, 3]. Благодаря высокой точности алгоритм позволяет достигать малой погрешности обучения, что является крайне важным фактором для решения большинства практических задач в нейросетевом базисе.

В общем виде алгоритм обработки хаотических сигналов состоит из следующих этапов: 1) нормализация исходного временного ряда, состоящего из N точек, выбранных с учетом задержки τ; 2) сегментация исходного временного ряда методом фиксированных отрезков; 3) обучение нейронной сети прогнозированию по методу скользящего окна; 4) расчет старшего показателя Ляпунова на базе сформированной нейросетевой прогнозной модели по методу отклонений траекторий прогнозов [5].

Существует проблема в выборе метода сегментации исходной выборки [4]. Для решения подобных задач применяются: метод фиксированных отрезков; метод наложения отрезков друг на друга; адаптивный метод сегментации при помощи НС.

Наиболее приемлемым для решения поставленной в работе задачи авторы определили метод фиксированных отрезков.

Были проведены 2 группы вычислительных экспериментов на базе вышеописанной архитектуры НС, результаты которых представлены ниже.

1. Исследование наборов сигналов электроэнцефалограмм (ЭЭГ) человека (A,D,E) [6]. Каждый набор содержит в себе 100 сигналов определенной группы в зависимости от эпилептической активности. Результаты анализа сведены в таблицу 1.

Савицкий Юрий Викторович, *к.т.н., доцент кафедры интеллектуальных информационных технологий факультета электронных информационных систем Брестского государственного технического университета.*

Давидюк Юлия Ивановна, *магистр технических наук, ассистент кафедры интеллектуальных информационных технологий факультета электронных информационных систем Брестского государственного технического университета.*

Беларусь, БрГТУ, 224017, г. Брест, ул. Московская, 267.