УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
Том 150, кн. 4 Физико-математические науки *2008*

УДК 519.711.2

МАШИННОЕ ОБУЧЕНИЕ В ЗАДАЧАХ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА: ОБЗОР СОВРЕМЕННОГО СОСТОЯНИЯ ИССЛЕДОВАНИЙ

К.А. Найденова, О.А. Невзорова

Аннотация

В статье рассматриваются современные методы машинного обучения, применяемые в задачах обработки естественного языка. Обсуждаются классические задачи, в том числе задачи морфологического и синтаксического анализа текстов. Особое внимание уделено методам машинного обучения, применяемым для построения онтологических моделей.

Ключевые слова: методы машинного обучения, анализ естественного языка, морфология, синтаксис, семантика, онтологии.

Введение

Исследования в области разработки программного обеспечения для задач обработки естественного языка (Natural Language Processing – NLP, Language Engineering – LE) активно развиваются в различных исследовательских парадигмах. Устойчивые тенденции последнего десятилетия в области LE связаны с широкомасштабными исследованиями в области разработки и применения статистических методов и методов машинного обучения (Machine Learning – ML). Характерными чертами таких исследований являются:

- использование эмпирических методов с точными критериями оценок;
- расширение сферы применения статистических методов;
- использование больших ресурсов данных (текстовые базы данных, онтологии, тезаурусы, корпуса текстов);
- применение NLP-технологий в реальных областях.

Можно выделить ряд ключевых проблем данного подхода. Эффективность разработки напрямую связана с наличием больших и сверхбольших ресурсов – размеченных корпусов текстов, онтологий и тезаурусов. Весьма важным является аспект стандартизации разработки, и в настоящее время де-факто сложился ряд стандартов, например, стандарт WordNet [77] для лексических онтологий или стандарт PennTreeBank [51, 60] для синтаксически размеченных корпусов текстов и др.

Другой проблемой является оценка эффективности используемых эмпирических критериев. Метрики числовых оценок в LE подобны хорошо известным в системах извлечения информации понятиям «точность» (precision) и «полнота» (recall). В основе получения оценок лежит сравнение результатов работы человека-аналитика и компьютерной программы при решении определенной задачи. Следует отметить, что области применения сравнительных оценок в LE постоянно расширяются.

Возрастающее использование статистических методов в задачах LE порождает некоторый отход от методов исследования и моделирования глубинных механизмов, лежащих в основе мышления и языка человека. Статистические методы в NLP

позволяют достигнуть определенных результатов в решении ряда задач (распознавание речи, разрешение многозначности, аннотирование текстов и др.), однако, представляется перспективным использование гибридных моделей, в которых используется различная техника, в том числе интроспективные методы.

Одним из перспективных направлений исследований в области извлечения информации (Information Extraction – IE) является направление «машинного обучения». Компьютерные системы, реализующие методы ML, ориентированы на получение новых знаний в результате автоматизации процесса обучения. Методы автоматического получения новых знаний на основе эмпирических данных можно успешно применять для формирования баз знаний. Это обстоятельство делает актуальными исследования в области обучения языку (Language Learning), результаты которых применимы в практических приложениях NLP-систем. Можно указать несколько причин, по которым исследования по ML становятся полезными в разработках NLP.

1. Сложность задач. Язык является сложноорганизованным объектом. Полная модель языка представляет сложное взаимодействие регулярностей, нерегулярностей, зон исключений и других явлений. Разработка такой модели может быть начата с разработки моделей отдельных подъязыков, описывающих относительно простые семантические области (например, медицинская диагностика и т. п.).

2. Реальные приложения. В настоящее время существует огромный рынок NLP-приложений (машинный перевод, реферирование и др.). Методы ML несомненно могут быть полезны в решении ряда важных проблем NLP-систем.

3. Доступность больших ресурсов данных. Стандартизация и открытость многих важных ресурсов обеспечивает необходимую ресурсную составляющую методов ML.

1. Классификация методов машинного обучения

Методы машинного обучения являются методами обучения классификациям объектов, представленных описаниями в признаковых пространствах. Цель обучения есть получение необходимых и достаточных правил, с помощью которых можно произвести классификацию новых объектов, сходных с теми, которые составляли обучающую выборку (обучение с учителем – supervised learning). При этом каждый обучающий пример (описание объекта) имеет метку, показывающую, к какому классу он принадлежит. Можно сказать, что в этом случае строится классификатор (рис. 1), который предсказывает класс предъявленного объекта по аналогии с «учителем». В случае непрерывных признаков классификацию называют регрессией.

При обучении без учителя (unsupervised learning) ставится задача объединения объектов в группы, попарно не пересекающиеся, на основе заданной меры их сходства/различия. Такую задачу часто называют кластеризаций объектов. Мера сходства/различия в признаковом пространстве используется в решающем правиле при отнесении к одной из полученных групп новых объектов, не входящих в обучающую выборку. Обучение без учителя обычно применяется для анализа структуры данных, но также и для формирования обучающей выборки при последующем применении обучения с учителем с целью найти правила классификации, описывающие полученное разбиение объектов на группы (классы) в пространстве признаков.

Ключевыми моментами машинного обучения являются:

1) выбор и формирование признакового пространства;

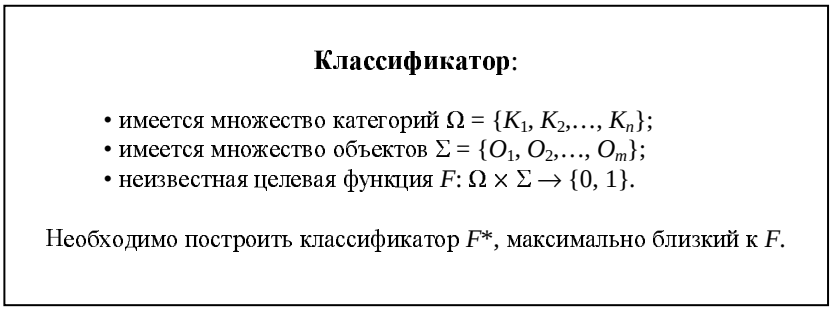

Рис. 1

2) проверка гипотез о различимости/сходстве объектов и классов объектов; задание бинарных операций сходства-различия объектов; задание мер сходства/различия для классов объектов;

3) формирование обучающей выборки;

4) формирование контрольной выборки;

5) адекватный выбор алгоритма обучения.

Если выбор признакового пространства определяет задачу обучения, главным образом, содержательно, то формирование обучающей и контрольной выборок отвечает за точность, быстроту и эффективность обучения. С помощью правильно выбранных примеров можно направлять процесс обучения. Пошаговые процедуры обучения и выбор последовательности примеров (от простого к сложному) позволяют также минимизировать число примеров, необходимых для обучения. Контрольная выборка необходима не только для проверки правильности работы классификатора, но и для целенаправленного «доучивания» классификатора, его исправления, модификации, придания ему требуемых свойств.

Достаточно естественно можно рассматривать NLP-задачи как задачи классификации ML. Действительно, в задачах лингвистических классификаций требуется построить (распознать) класс (категорию) объектов, заданных некоторым описанием. Категория обычно выбирается из некоторого множества возможных значений. Тем самым лингвистические задачи могут быть переопределены как задачи в контексте ML.

Приведем несколько примеров построения классификаций в NLP-задачах.

1) Задача классификации текстовых документов.

Объекты: Текстовые документы D_j, $j = 1, \ldots, n$;

Задача: Получить правила для классификации текстов (обучение с учителем);

Признаки: Ключевые слова T_i, $i = 1, \ldots, t$;

Вектор признаков $\mathbf{W}_j$, описывающий документ D_j:

а) компонента вектора w_{ij} О {1,0}, где 1 означает присутствие ключевого слова T_i в тексте D_j, 0 – его отсутствие; например, $\mathbf{W}_j = [011110]$, то есть T_1, T_6 отсутствуют в D_j; T_2, T_3, T_4, T_5 присутствуют;

б) компонента вектора w_{ij} О {1,0} – показывает частоту встречаемости ключевого слова T_i в документе D_j, например, $\mathbf{W}_j = [0.00 \quad 1.00 \quad 0.10 \quad 0.75 \quad 0.90 \quad 1.00]$.

2) Задача построения предметной онтологии на основе текста (пример взят из статьи [19]).

Объекты: существительные в ролях агента или прямого объекта глагола в предложении.

Признаки: глаголы, с которыми ассоциированы существительные.

Табл. 1

Знания об услугах туристического агентства

	Book (bookable)	Rent (rentable)	Drive (driveable)	Ride (rideable)	Join (joinable)
Hotel	X				
Apartment	X	X			
Car	X	X	X		
Bike	X	X	X	X	
Excursion	X				X
Trip	X				X

Так, из рекламного текста туристического агентства, извлекается следующие отношения (табл. 1), связывающие объекты (строки таблицы) и признаки (столбцы таблицы):

На основе полученных описаний объектов (терминов предметной области) можно построить следующую классификацию (онтологию) объектов (рис. 2).

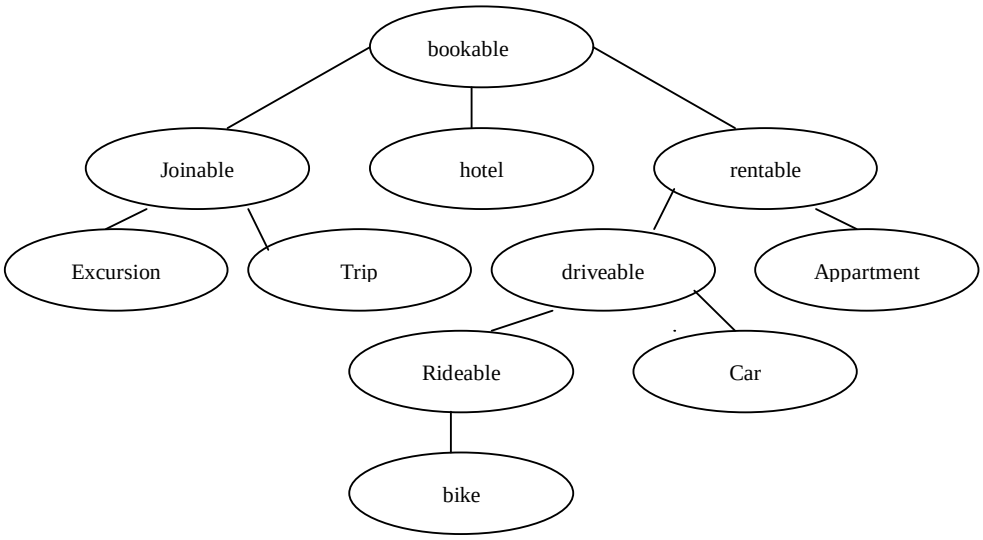

Рис. 2

Синтаксические зависимости, в частности, между глаголом и его аргументами, используются довольно часто в качестве признаков при выделении семантических отношений из текста. Эта идея развивается в работах [10, 17, 30, 41, 69]. Здесь важно найти верный уровень генерализации для глагольных аргументов по отношению к заданной концептуальной иерархии. Этому подходу уделяется много внимания в компьютерной лингвистике в контексте так называемых селективных ограничений [23, 64, 65]. Другая задача – выделить глаголы, обозначающие одно и то же онтологическое отношение (семантическая кластеризация глаголов), – решается аналогично, используя ограничения на аргументы глагола [67].

Латентный семантический анализ (LSA) [46] основан на статистической оценке сходства слов по их значению. Значения слов сходны, если они употребляются в сходных контекстах. Здесь контекст выступает в качестве признака слова. Смысловое сходство контекстов также оценивается в LSA. В системе iSTART [52], на-

пример, с помощью LSA вычисляется смысловое подобие между предложениями из разных текстов, а также между любыми текстовыми фрагментами.

В задаче сегментации предложений, где под сегментом понимается часть предложения (в частном случае целиком простое предложение), выделенная на письме знаками пунктуации и описывающая отдельную ситуацию, в качестве классификационных признаков слов используется их лексический контекст. Позиция слова называется сегментной позицией, если слово начинает сегмент или является границей между двумя сегментами. Выделяются два вида лексических контекстов: активный для целевого слова в сегментной позиции и неактивный для других слов. Лексический контекст слова, как правило, включает пять позиций: само слово, два слова справа и два слова слева от него. Лексический контекст также содержит некоторую дополнительную информацию, например, теги (метки) частей речи слов контекста.

В задаче обнаружения семантических отношений между словами путем заполнения пустых позиций в паттернах (шаблонах), например, «P_1 взаимодействует с P_2», «P_1 активируется через P_2», признаками могут быть множество слов – Σ_1, множество тегов (частей речи) – Σ_2, слова, стоящие перед P_1 и P_2 – Σ_3 и т. д. Тогда с каждой пустой позицией шаблона ассоциируется вектор признаков из множества $\Sigma = \Sigma_1 \times \Sigma_2 \cdots \times \Sigma_k$ всех возможных векторов признаков.

Теоретически любое определимое отношение может рассматриваться как признак, так что выделяемые паттерны характеризуют тексты, и сами сводятся к специфическим идиоматическим, синтаксическим, семантическим отношениям.

В табл. 2 представлены паттерны семантических отношений, которые выделяются при автоматическом аннотировании англоязычных текстов с использование Web.

Эти отношения рассматриваются как онтологические структуры, так как имена собственные соединяются с классами или концептами [15, 20, 35, 37, 49].

Методы машинного обучения подразделяются на вероятностные и логические в зависимости от природы объектов и признаков и формы представления функции или решающего правила, с помощью которых приближается заданная классификация.

Кратко рассмотрим некоторые широко известные **вероятностные методы машинного обучения**.

Метод Байеса [80, 86] определяет вероятность наступления события в условиях, когда наблюдается лишь некоторая частичная информация о событиях, например, по наблюдаемым признакам определяется принадлежность некоторого объекта к одному из заданных классов.

Метод опорных векторов (Support Vector Machine) [25] относится к линейным разделяющим методам. Каждый вектор признаков (объект) представляется точкой в многомерном пространстве признаков. При заданной классификации объектов строятся две параллельные гиперплоскости (границы), разделяющие объекты разных классов, таким образом, что расстояние между этими гиперплоскостями максимизируется. Примеры, расположенные вдоль гиперплоскостей, называют опорными векторами. Значения признаков – вещественные числа.

К вероятностным методам относится скрытая марковская модель (Hidden Markov chaines) [63]. Это статическая модель, имитирующая некоторый последовательный процесс, в котором на наблюдаемые переменные оказывают влияния скрытые состояния. Переход из одного состояния в другое происходит с некоторой вероятностью. По последовательности наблюдений можно получить информацию о последовательности состояний.

Табл. 2

Примеры паттернов при аннотировании текстов с использованием Web

Паттерн	Пример паттерна
<CONCEPT>s such as <INSTANCE>	Hotels such as Ritz
Such <CONCEPT>s as <INSTANCE>	Such hotels as Hilton
<INSTANCE> (and | or) other <CONCEPT>	The Eiffel Tower and other sights in Paris
<CONCEPT>s, (especially | including) <INSTANCE>	Presidents, especially George Washington
The <INSTANCE> <CONCEPT>	The Hilton hotel
The <CONCEPT> <INSTANCE>	The hotel Hilton
<INSTANCE>, a <CONCEPT>	Excelsior, a hotel in the center of Nancy
<INSTANCE> is a <CONCEPT>	The Excelsior is a hotel in the center of Nancy

Наиболее распространенным методом кластеризации является метод k-ближайшего соседа [83]. Этот алгоритм классифицирует объект по большинству «голосов» его соседних объектов в многомерном пространстве признаков. Объект относят к классу, к которому принадлежит наибольшее число из его k наиболее близких соседей, где k – некоторое целое число. В задаче кластеризации для двух классов число k выбирают нечетным. «Соседи» выбираются из множества правильно расклассифицированных объектов. Для вычисления расстояний задают различные меры, например евклидово расстояние или расстояние Манхеттэна.

К методам классификации с учителем можно отнести искусственные нейронные сети [82]. Нейронная сеть состоит из нескольких слоев нейронов, связанных между собой. Веса связей при обучении изменяются таким образом, чтобы выходной элемент (элементы) сети давал(и) бы правильный классификационный ответ на входной сигнал. Математически в основе обучения лежит построение разделяющих нелинейных функций в пространстве признаков, что дает хорошую точность предсказания, но полученное решение не поддается объяснению в терминах значений признаков.

К логическим методам относятся методы, при которых в процессе обучения строятся логические правила в форме продукций или в форме решающих деревьев, в узлах которых проверяются значения отобранных при обучении признаков и принимается решение о разбиении объектов на подклассы. Логические методы работают на описаниях, которые имеют как символьную, так и целочисленную природу, или в пространстве булевых признаков. С этой точки зрения эти методы можно отнести к методам концептуальных классификаций или методам построения правил рассуждений на уровне концептов. Но формирование правил в практических задачах, в особенности в NLP-задачах, часто влечет за собой громадную ручную работу и заключает в себе творческую и неавтоматизированную компоненту методов машинного обучения.

К методам концептуального обучения (извлечение концептов и их иерархических отношений) относится направление машинного обучения, называемое формальным концептуальным анализом [31]. Часто в литературе формальный концептуальный анализ называется концептуальной кластеризацией. Примером применения концептуального анализа может служить моделирование лексической базы данных [59].

Концептуальное обучение представляет собой особый класс методов, основанных на порождении и использовании концептуальных знаний, элементами которых

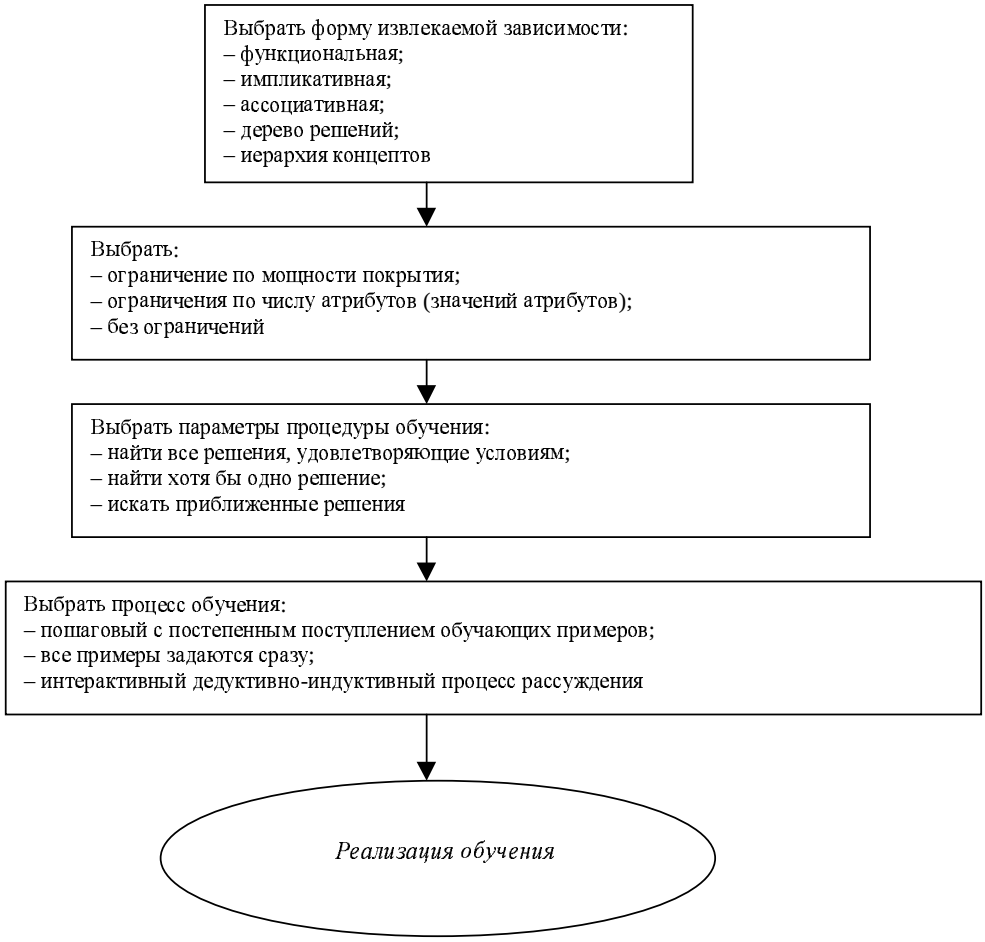

Рис. 3

являются объекты, атрибуты (значения атрибутов), классификации (разбиения объектов на классы) и связи между ними. Эти связи выражаются через импликативные отношения вида «объект → класс», «класс → объект», «объект → значение атрибута (свойство)», «значение атрибута (свойство) → объект», «значение атрибута → класс», «класс → значение атрибута», «подкласс → класс», «класс → подкласс». В работе [87] показано, что индуктивный вывод концептуальных знаний сводится к задаче вывода хороших диагностических тестов для заданных классификаций объектов. Хороший диагностический тест определяется как совокупность атрибутов (значений атрибутов), которая порождает наилучшую аппроксимацию заданной классификации (разбиения) некоторого заданного множества объектов. Критерий поиска наилучшей аппроксимации заданной классификации связан с выбором поднабора атрибутов, который порождает разбиение заданного множества объектов, вложенное в заданную классификацию и имеющее наименьшее возможное число классов среди всех возможных разбиений, вложенных в заданную классификацию и порождаемых поднаборами заданного множества атрибутов. К поиску хороших тестов сводятся почти все хорошо известные логические методы машинного обучения. Эти методы охватывают вывод импликативных и функциональных зависимостей из данных, вывод логических правил (правила «если-то»,

«грубые» множества), вывод ассоциативных правил, конструирование решающих деревьев, извлечение иерархических концептуальных классификаций и ряд других. Все перечисленные задачи отличаются по форме представления выделяемых отношений включения на множестве всех подмножеств объектов и решаются с использованием одной и той же структуры данных и одних и тех же алгоритмов, которые можно найти в [55, 56, 57].

На рис. 3 показано, что требуемая конфигурация алгоритма машинного обучения порождается путем задания параметров, определяющих форму представления результата, количество и некоторые свойства искомых концептов.

2. Применение методов машинного обучения в NLP-задачах

Основные методы в области машинного обучения задачам NLP можно отнести к двум классам: «ленивое» обучение (lazy learning) и «жадное» обучение (greedy learning). Существенное различие между этими подходами заключается в том, что при «ленивом» обучении извлекаемая информация не обобщается, в то время как при «жадном» обучении извлекаемая информация обобщается посредством реструктурирования и удаления избыточных и несущественных частей.

Подход lazy learning [4] основывается на гипотезе, что решение когнитивных задач (обучение языку, в частности) базируется на построении выводов на основе аналогий, а не на основе абстрактных правил, полученных из экспериментов. Этот подход используется в различных дисциплинах искусственного интеллекта и лежит в основе таких методов, как вывод на основе сходства, вывод на основе примеров, вывод на основе аналогии, вывод на основе прецедентов (case-based reasoning) и пр. При «ленивом» обучении обучающие примеры добавляются в память без обобщений и реструктурирования. Сходство нового примера с остальными вычисляется по метрике сходства и категория большинства сходных примеров используется как базовая для предсказания категории нового примера. Данный подход применяется в фонологических и морфологических задачах, задачах распознавания речи, морфологического и синтаксического анализа, в задачах разрешения морфосинтаксической и семантической многозначности.

Основными методами подхода greedy learning являются обучение на основе деревьев решений, индуктивного вывода, обучение на нейронных сетях и индуктивное логическое программирование. Обучение на основе деревьев решений основывается на предположении, что сходство примеров может быть использовано для автоматического построения деревьев решений, на базе которых порождаются обобщения и объяснения.

Целью индуктивного вывода является построение ограниченного множества интерпретируемых правил на основе обучающих примеров или деревьев решений. На основе алгоритмов индуктивного логического программирования формируются гипотезы логики первого порядка на основе примеров. Наиболее интересные результаты в этом направлении получены В.К. Финном [88].

Анализируя круг решаемых NLP-задач, можно сделать предварительные выводы об эффективности применения рассмотренных выше подходов. Выбор метода существенно зависит от целей системы. Если цель – точность, то метод lazy learning является предпочтительным. Алгоритмы lazy learning, дополненные методами взвешивания признаков, вероятностными правилами, дают хорошие результаты для большого класса лингвистических задач. Если цель машинного обучения – создание проверяемых, объясняющих обобщений данных, предпочтительны методы greedy learning.

Рассмотрим типовые NLP-задачи, для которых активно применяются методы ML.

В обработке естественно-языковых (ЕЯ) текстов можно выделить два главных направления: извлечение информации из текстов (Information Extraction) и извлечение знаний из текстов (Text Mining). В первом случае речь идет о выделении явных сведений, имеющихся в текстах, например, ключевых слов, дат, названий организаций, имен, описок, оговорок и т. д. Извлечение информации можно рассматривать как неизбежный предварительный этап более серьезных задач извлечения знаний из текстов. Этот этап использует различные системы категоризации и классификации текстов на основе методов машинного обучения, главным образом, метод опорных векторов и логические методы (решающие деревья, извлечение правил «если-то») со всеми вытекающими отсюда требованиями к предварительному формированию классификационных признаков текстов, фрагментов текстов и их классов.

К задачам извлечения знаний из текстов относятся задачи понимания текстов [5]. В частном случае, это может быть проблема выделения в текстах мнений людей о тех или иных продуктах (товарах) с их оценочным содержанием – положительным или отрицательным [58]. В более общей постановке это выделение семантических отношений между заданными понятиями, например, между биологическими понятиями «клетка», «ген», «белок» [11], семантических ролей [54].

В задачах извлечения знаний из ЕЯ-текстов часто используется комбинация как традиционных методов машинного обучения, так и новых методов. Так, в [54] авторы комбинирует синтаксический разбор со статистическими методами классификации. Обучение при разметке семантических ролей, называемое в этой работе активным обучением, состоит из следующих этапов:

1) объединяются предложения с одним и тем же целевым глаголом;
2) группируются предложения с одинаковым деревом синтаксического разбора;
3) ручная разметка применяется к группе с одним и тем же деревом разбора;
4) на примерах разметки обучается классификатор;
5) классификатор работает на неразмеченных предложениях.

При этом на шаге 4 используются три различных классификатора и классификационные признаки выделяются на дереве синтаксического разбора.

В последние годы сформировалось новое направление извлечения знаний из текстов, связанное с построением онтологий [18]. Примером задачи этого класса является задача построения классификации существительных на основе предикатно-аргументных структур [39]. В основе этой работы лежит дистрибутивная гипотеза, в которой сходство имен существительных устанавливается на основе сходства их синтаксических контекстов употребления. Таким образом, существительные группируются, если они появляются в сходных глагольных фреймах как подлежащее и/или прямое дополнение. В работе [21] дополнительным шагом является построение иерархий полученных таксонов. Аналогичный подход дается в [27]. Чаще всего, выделение семантических отношений между словами основано на использовании лексико-синтаксических паттернов или фреймов синтаксической категоризации (глагольно-объектные отношения) [67]. В работе [7] при кластеризации глаголов с помощью пошагового алгоритма учитывается полисемия глаголов, но вместо синтаксических отношений используются селективные ограничения, которые вводятся вручную. В литературе описываются инструментальные системы построения онтологий, интегрирующие различные методы кластеризации, например система Mo'K [8], система OntoLearn [74].

Система выделения онтологий из текстов OntoLearn [74] обеспечивает:

1) выделение релевантной терминологии;
2) именование построенных классов;
3) построение таксономии по отношению “is_a”;

4) выделение других типов синтаксических отношений.

Извлечение знаний из текстов в частных случаях может не требовать полного синтаксического и семантического разбора предложений. Но глобально, конечно, задача извлечения знаний из текстов не может быть решена без полного синтактико-семантического анализа предложений. Синтаксический анализ требуется для таких задач, как извлечение семантических отношений между словами (понятиями), построение онтологий, машинный перевод, исправление грамматических ошибок, реферирование и др.

Традиционно процесс анализа и понимания ЕЯ-текстов является последовательностью следующих этапов: предобработка (предпроцессы), синтаксический анализ, семантический анализ, контекстуальная интерпретация. Предпроцессы включают в себя: выделение токенов (tokenization), нормализацию, лемматизацию (lemmatization or stemming) (морфологический анализ), разметку частей речи (part-of-speech tagging (POS)), распознавание имен собственных (уникальных объектов в мире – персон, организаций, мест и дат и т. д.), разрешение кореференций.

Токенизация (tokenization, lexical analysis, графематический анализ, лексический анализ) – выделение в тексте слов, чисел, и иных токенов, а также границ предложений. Общеизвестны проблемы многозначности, связанные со знаками пунктуации [26].

Нормализация заключается в выделении последовательностей дат, времени и т. п. и переводе их в стандартизованный вид. Этап может включать расшифровку сокращений. Более детально этапы токенизации и нормализации рассматриваются в [50].

Теггинг (tagging, part of speech disambiguation) – приписывание каждому слову (токену) грамматической характеристики части речи. Традиционно различают подход на основе правил, или трансформационный подход [9], и статистический (вероятностный) подход на основе Марковских моделей [14]. Этап требует полного или частичного разрешения морфологической омонимии и унификации значений грамматических характеристик [48, 50]. Современные методы теггинга основаны на методе обучения решающим деревьям TreeTagger [66] и Qtag Tagger [73] и имеют точность 95–97%.

Лемматизация (lemmatization, stemming) – приведение словоформы к нормальной форме слова, репрезентирующей лексему. В монографии [18] описывается специальный анализатор для проведения лемматизации – LogPar-анализатор. Лемматизация может быть неполной (stemming) и полной, требующей глубокого морфологического анализа, выявляющего внутреннюю структуру слова.

Распознавание уникальных имен, если они не ограничены заданным списком, требует использование процедур машинного обучения [18]. Наряду с задачей распознавания уникальных имен рассматривается задача кореференции имен. Только некоторые частные задачи кореференции относятся к предпроцессам, например, распознавание разных имен одной и той же персоны (профессор Иванов, Николай Иванов, Н.Иванов). В общем случае разрешение анафоры и более сложных отношений референции не относится к предпроцессам.

Синтаксический анализ подразделяется на поверхностный (shallow) и полный (deep). Поверхностный анализ (chunking) предназначен для выделения смысловых составляющих (chunks), таких, как именная группа (Noun Phrase (NP)), глагольная группа (Verb Phrase (VP)), предложная группа (Prepositional Phrase (PP)). Эти семантические единицы не пересекаются, не рекурсивны и не избыточны. Одна из систем SMES поверхностного синтаксического анализа для немецкого языка детально описана в [48].

Полный синтаксический анализ (parsing) представляет структуру предложения в виде синтаксического дерева. В настоящее время разработаны различные формальные грамматические теории синтаксиса: грамматика зависимостей, грамматика непосредственных составляющих, категориальная грамматика, лексико-функциональная грамматика (Lexical Functional Grammar – LFG), вершинная грамматика составляющих (Head-driven Phrase Structure Grammar – HPSG), вероятностная контекстно-свободная грамматика (Probabalistik Context-Free Grammar – PCFG) и др.

Отмечается, что наилучшие результаты дают статистические методы синтаксического анализа. Синтаксический анализ приводит к сложностному взрыву и порождает множество вариантов синтаксического разбора предложения, при этом основные трудности связаны именно с разрешением возникающих многозначностей.

Одной из первых задач синтаксического анализа является задача сегментации предложения. Под сегментом понимается часть предложения (в частном случае – простое предложение), выделенная знаками пунктуации и описывающая отдельную ситуацию. В сегменте выделяется его предикативная вершина (head), выраженная в большинстве случаев финитной формой глагола или другим предикативным словом (деепричастием, причастием, именем с семантической характеристикой действия).

В западной лингвистической традиции понятие «сегмент» эквивалентно термину «клауза»: «клаузой» называется любая группа, в том числе и непредикативная, вершиной которой является глагол, а при отсутствии полнозначного глагола – связка или грамматический элемент, играющий роль связки. На следующем этапе синтаксического анализа решается задача установления внутрисегментных связей.

Задача синтаксического анализа решается на основе различных методов формальных грамматик, устанавливающих определенные правила композиции синтаксических структур. Для моделирования русского синтаксиса чаще всего применяются правила грамматики зависимостей [81, 84, 85]. Однако в последние годы в задачах синтаксического анализа начинают применяться методы машинного обучения.

В статье [42] описан метод сегментации длинных предложений английского языка, основанный на методах машинного обучения (используются решающие деревья и генетические алгоритмы). Сегментация на основе обучения происходит в два этапа: определение возможных позиций сегментов и определение действительных границ сегментов. Процесс длится до тех пор, пока каждый сегмент не достигнет пороговой длины. Порог выбирается, исходя из оценок сложности синтаксического разбора.

На первом этапе потенциальные позиции сегментов определяются с помощью классификации: каждое слово предложения относится к одному из двух классов: «может быть границей сегмента» (segmentable) и «не может быть границей сегмента» (nonsegmentable). Функция, которая классифицирует слова, определяется с помощью обучения. Функция классификации представляет собой множество правил в виде дерева решений. Используются хорошо известные обучающие алгоритмы ID3 [61], ASSISTAT [13], C4.5 [62]. Обучающие примеры представлены как пары «атрибут – значение» и включают атрибуты, которые специфицируют слова. Число значений атрибутов определяется числом входов в словаре. Чтобы уменьшить сложность дерева решений для связки значений атрибутов, используется только функция конъюнкции. В обучающем тексте позиции сегментов расставляются вручную.

На втором этапе происходит выделение действительных позиций сегментов из полученного множества возможных границ на основе функции «наиболее подхо-

дящего сегментирования». Такая функция есть линейная сумма взвешенных переменных, отображающих факторы, которые влияют на выбор. Веса переменных выбираются на основе некоторых критериев оптимальности функции. Для поиска весов используются генетические алгоритмы машинного обучения.

Предложение рассматривается как конкатенация последовательных сегментов. Сегмент соответствует элементарной фразе, или «клаузе», в предложении, отношения между сегментами описываются с помощью контекстно-свободной грамматики. Контекстно-свободный синтаксический анализатор обозревает сегменты справа налево, подобно «chuck by chunk» стратегии [1–3, 38].

Работы по семантическому анализу текстов, главным образом, связаны с поверхностным семантическим анализом. Типичной задачей является маркирование семантических ролей [12, 34, 45, 71, 72]. Более амбициозная задача семантического разбора, который конструирует полное формальное представление смысла предложения, решается в [75]. Идея построения синтаксически ориентированного обучаемого семантического анализатора рассматривается в [53].

В последнее десятилетие разработаны методы машинного обучения для индуктивного построения семантических анализаторов по примерам предложений, смысл которых представлен на специальном формальном языке.

Наиболее ранняя система CHILL для обучаемого семантического анализатора [78] использует индуктивное логическое программирование для обучения детерминированного анализатора, написанного на языке Prolog. Система CHILL позже расширена в [70] до системы COCKTAIL, которая использует конструктор множества «клауз» в развитие идей системы CHILL.

Сравнительно недавно разработаны подходы к обучению статистических семантических анализаторов, работающих на больших обучающих выборках. Они используют разные технологии статистической обработки ЕЯ-текстов. Система SCISSOR [32, 33] добавляет подробную семантику к современному статистическому синтаксическому анализатору Collins parser [24]. Обучаемый семантический анализатор WASP [76] адаптирует метод статистического машинного перевода для задачи семантического анализа. Семантический анализатор KRISP [40] использует метод опорных векторов [25] с последовательными ядрами [47].

В [79] предложена система обучаемого семантического анализатора на основе вероятностной комбинаторной категориальной грамматики (Probabilistic Combinatorial Categorial Grammars – CCG).

Области применения семантических анализаторов ограничены ЕЯ-запросами к специфическим базам данных и некоторыми специальными приложениями, например интерпретацией инструкций футбольного тренера.

В [53] обсуждается возможность использования контекстов для разрешения многозначности. Автор предлагает объединить синтаксически управляемый обучаемый семантический анализ с формальным представлением смысла предложений и описанием контекста. В качестве области применения такого подхода предлагается выбирать динамические задачи с ограниченным контекстом. В частности, можно использовать систему [16], которая обеспечивает детальное физическое моделирование роботов – футбольных болельщиков или систему комментаторов чемпионата роботов (Robocup “commentator” systems) [6].

Можно указать и другие задачи, в которых используется контекстное обучение языку [29]. К таким задачам относится обучение языку по картинкам, обучение значению отдельных слов по символьным описаниям контекста [68] и обучение описаниям объектов и действий, извлекаемых из видео изображений [28].

Весьма актуальная задача аннотирования текстов также решается на основе методов машинного обучения. При этом методы машинного обучения требуют суще-

ственных затрат на создание обучающих выборок. Преодоление этого затруднения ведется двумя путями. Первый путь приводит к включению в процесс аннотирования автоматизированного этапа сборки результатов аннотирования для создания обучающей выборки предложений текста с таким же деревом разбора (аннотирование с обучением). Второй путь ведет к исключению (полному или частичному) ручного аннотирования путем использования ресурсов интернета (аннотирование без обучения). При этом различают аннотирование лингвистическое (синтаксическое) и семантическое. К первому относят разметку морфологических и синтаксических признаков слов, построение аннотации на основе дерева разбора. Ко второму – приписывание словам или выражениям семантических категорий или онтологических меток. Примером системы второго типа является система полуавтоматического аннотирования PANKOW [20], в которой именам собственным приписываются онтологические категории на основе обращения к ресурсам интернета, содержащим уже готовые онтологические схемы. Сначала через запрос к этим ресурсам генерируются гипотезы о возможных онтологических отношениях для выделенного имени собственного, затем выбирается одна из гипотез на основе статистического правила максимального правдоподобия.

Интересный подход, сводящий синтаксическое аннотирование к семантическому полуавтоматическому аннотированию (без ручного аннотирования), приводится в работе [22]. Лингвистическая аннотация рассматривается как частный случай семантической аннотации, реализованной в системе CREAM [36]. Например, теггинг сводится к задаче выбора соответствующего тега для слова из онтологии категорий слов. Задача понимания смысла слов и разрешения неоднозначностей сводится к выбору правильного семантического класса или концепта для слова из соответствующей онтологии в WordNet. Заполнение шаблонов в системах извлечения информации сводится к задаче нахождения и разметки всех атрибутов заданного онтологического концепта (например, для концепта «персона» задаются атрибуты «имя», «место работы», «должность»). Таким образом, грамматическая информация вводится через атрибуты концепта, и схемы аннотации представляются в онтологических структурах.

Рассматриваются три пути аннотирования:

1) лингвистическое выражение может аннотироваться как пример некоторого онтологического концепта;

2) лингвистическое выражение может аннотироваться как пример атрибута некоторого лингвистического концепта, предварительно аннотированного как некоторый концепт;

3) семантическое отношение между двумя лингвистическими выражениями, аннотированными как примеры двух концептов, может аннотироваться как пример отношения.

Если использовать язык OWL как стандартный формализм для записи онтологий, то можно вычислять согласованность между различными аннотациями, определив меры близости между иерархиями концептов. Описанный в работе подход применяется для аннотирования анафорических отношений. Аннотирование анафорического отношения между двумя выражениями может соответствовать более общему отношению в онтологической иерархии. Идентичность и кореференция рассматриваются как специальные случаи анафоры. На основе декларированного подхода возможно разрешение неявно выраженного отношения идентичности («Джон купил вчера машину»; «“Тачка” в хорошем состоянии»).

Полуавтоматическое семантическое аннотирование документов через интернет-онтологии рассматривается также в [43]. Первый шаг – семантическая аннотация с опорой на семантическую модель: онтологию или концептуальную схему. Второй

шаг – ручная коррекция результатов первого шага. Оригинальность подхода, предложенного в этой работе, состоит в том, что для семантического аннотирования на первом шаге применяется модифицированный процесс анализа программного кода [44]. На первой стадии используется приближенная контекстно-свободная неоднозначная грамматика для получения приближенной фразовой структуры. При этом используется быстрый детерминированный парсинг, по сложности линейно зависимый от входа. Результаты аннотирования переводятся в XML и помещаются в базу данных для дальнейшего использования для ответов на ЕЯ-запросы к базе данных.

Заключение

Методы машинного обучения, хотя и находят все большее применение для различных задач обработки ЕЯ-текстов, пока ещё остаются чрезвычайно сложными и трудоемкими для реального применения. Это объясняется не столько сложностью алгоритмов обучения, сколько, возможно, неудачными методологическими подходами к обучению. Задачи обучения применяются фрагментарно, к какому-либо отдельному этапу последовательного процесса обработки текста. Именно поэтому приходится заниматься ручной разметкой, а не использовать результаты предыдущего обучения системы на предшествующих и взаимосвязанных этапах обработки.

Например, успешные результаты морфологического разбора можно было бы использовать при обучении системы распознаванию синтаксических зависимостей между словами. По-видимому, целесообразно моделировать поведение «обучаемого лингвистического агента», который накапливает знания о том, как взаимосвязаны синтаксические составляющие между собой и как они связаны со смыслом предложений. Таким образом, нужна программа постепенного обучения лингвистического агента от «простого к сложному», причем обучение должно управляться семантической компонентой анализатора. В таком процессе признаки для каждого подпроцесса могли бы формироваться автоматически на предыдущих этапах.

Существует важная проблема проверки правильности работы обученной программы. Весьма важно, чтобы сама программа могла «понимать», что она не может справиться с задачей. Такое «понимание» может базироваться на том обстоятельстве, что для какого-либо шага нет однозначного решения или имеет место противоречие, конфликт некоторых правил. В этом случае программа должна запрашивать новые примеры или дополнительные знания экспертов-лингвистов.

Машинные методы обучения концептуальным знаниям представляют собой модель правдоподобных индуктивных и дедуктивных рассуждений, в которых вывод знаний и их использование не отделимы друг от друга. Реализация обучения в режиме правдоподобных рассуждений позволит организовать взаимодействие не только данных и знаний в процессах обработки текстов, но и моделировать процесс взаимодействий учителя и ученика в процессе приобретения знаний в схемах многоагентных взаимодействий.

Summary

X.A. Naidenova, O.A. Nevzorova. Machine Learning for Natural Language Processing: Contemporary State.

Contemporary methods of machine learning used for natural language processing tasks are discussed in the paper. Various classic tasks of natural language processing are considered, including the tasks of morphological and syntactical analysis. Special attention is given to the methods of machine learning used for construction of ontological models.

Key words: methods of machine learning, natural language processing, morphology, syntax, semantics, ontology.

Литература

1. *Abney S.* Partial Parsing via Finite-State Cascades // ESSLLI'96 Workshop on Robust Parsing Workshop. – Prague, Czech Republic, 1996. – P. 71–84.
2. *Abney S.* Chunks and Dependencies: Bringing Processing Evidence to Bear on Syntax // Computational Linguistics and the Foundations of Linguistic Theory / J. Cole, G.M. Green, J.L. Morgan (eds.). – Stanford, CA: CSLI Publications, 1995. – P. 145–164.
3. *Abney S.* Parsing by Chunks // Principle-Based Parsing / R. Berwick, S. Abney, C. Tenny (eds.). – Dordrecht, The Netherlands: Kluwer Acad. Publ., 1991. – P. 257–278.
4. Lazy Learning / D.W. Aha (ed.). – Dordrecht, The Netherlands: Kluwer Acad. Publ., 1997. – 625 p.
5. *Allen J.* Natural Language Understanding. – Menlo Park, CA: Benjamin/Cummings Publishing Company, 1995. – 625 p.
6. *Andr'e E., Binsted K., Tanaka-Ishii K., Luke S., Herzog G., Rist T.* Three RoboCup simulation league commentator systems // AI Magazine. – 2000. – V. 21, No 1. – P. 57–66.
7. *Basili R., Pazienza M.T., Velardi P.* Hierarchical clustering of verbs // Proc. of the Workshop on Acquisition of Lexical Knowledge from Text. – 1993. – P. 70–81.
8. *Bisson G. , Nedellec C., Canamero L.* Designing clustering methods for ontology building: The Mo'K workbench // Proc. of the ECAI Ontology Learning Workshop. – 2000. – P. 13–18.
9. *Brill E.* Some advances in transformation-based part-of-speech tagging // Proc. of the Nat. Conf. on AI (AAAI). – 1994. – P. 722–727.
10. *Buitelaar P., Olejnik D., Sintek M.* A Protégé plug-in for ontology extraction from text based on linguistic analysis // Proc. of the 1st European Semantic Web Symposium (ESWS). – 2004. – P. 31–44.
11. *Bunescu C., Mooney J.* Extracting Relations from Text: From Word Sequences to Dependency Paths // Natural Language Processing and Text Mining / A. Rao, S.R. Poteet (Eds). – Springer, 2007. – P. 29–44.
12. *Carreras X., M'arquez L.* Introduction to the CoNLL-2005 shared task: Semantic role labeling // Proc. of the 9th Conf. on Natural Language Learning (CoNLL-2005). – 2005. – P. 152–164.
13. *Cestnik B., Kononenko I., Bratko I.* ASSISTANT-86: A Knowledge-Elicitation Tool for Sophisticated Users // Progress in Machine Learning / I. Bratko, N. Lavrac (eds.). – Wilmslow: Sigma Press, 1987.
14. *Charniak E., Hendrickson C., Jacobson N., Perkowitz M.* Equations for part-of-speech tagging // Proc. of the 11th Nat. Conf. on AI (AAAI). – 1993. – P. 784–789.
15. *Charniak E., Berland M.* Finding parts in very large corpora // Proc. of the 37th Annual Meeting of the ACL. – 1999. – P. 57–64.
16. *Chen M., Foroughi E., Heintz F., Kapetanakis S., Kostiadis K., Kummeneje J., Noda I., Obst O., Riley P., Steffens T., Wang Y., Yin X.* Users manual: RoboCup soccer server manual for soccer server version 7.07 and later. – 2003. – URL: http://sourceforge.net/projects/sserver/.
17. *Ciaramita M., Gangemi A., Ratsch E., Šarić J., Rojas I.* Unsupervised learning of semantic relation between concepts of a molecular biology ontology // Proc. of the 19th Int. Joint Conf. on AI (IJCAI). – 2005. – P. 659–664.

18. *Cimiano P.* Ontology Learning and Population from Text. Algorithms, Evaluation and Applications. – Springer, 2006. – 347 p.

19. *Cimiano P., Holto A., Staab S.* Learning Concept Hierarchies from Text Corpora Using Formal Concept Analysis // J. Artif. Intellig. Res. – 2005. – V. 24. – P. 305–339.

20. *Cimiano R., Handschuh S., Staab S.* Towards the self-annotating web // Proc. of the 13th World Wide Web Conf. – 2004. – P. 462–471.

21. *Cimiano P., Staab S., Tane J.* Automatic acquisition of taxonomies from text: FCA meet NLP // Proc. of the PKDD/ECML'03 Int. Workshop on Adaptive Text Extraction and Mining (ATRM). – 2003. – P. 10–17.

22. *Cimiano P., Handschuh S.* Ontology-based linguistics annotation // Proc. of the ACL 2003 workshop on Linguistic annotation: getting the model right. – 2003. – V. 19. – P. 14–21.

23. *Clark S., Weir D.* Class-based probability estimation using a semantic hierarchy // Comput. Linguist. – 2002. – V. 28, No 2. – P. 187–206.

24. *Collins M.J.* Three generative, lexicalized models for statistical parsing // Proc. of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97). – 1997. – P. 16–23.

25. *Cristianini N., Shawe-Taylor J.* An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. – Cambridge: Cambridge Univ. Press, 2000. – 204 p.

26. *van Delden S., Gomez F.* Combining Finite State Automata and Greedy Learning Algorithm tj Determine the Syntactic Roles of Commas // Proc. of 14th IEEE Int. Conf. on Tools and AI (ICTAI'02). – 2002. – P. 293.

27. *Faure D., Nedellec C.* A corpus-based conceptual clustering method for verb frames and ontology // Proc. of the 1st LREC Workshop on Adapting lexical and corpus resources to sublanguages and applications. – Grenada, Spain, 1998. – P. 1–8.

28. *Fern A., Givan R., Siskind J.M.* Specific-to-general learning for temporal events with application to learning event definitions from video // J. Artif. Intellig. Res. – 2002. – V. 17. – P. 379–449.

29. *Fleischman M., Roy D.* Intentional context in situated natural language learning // Proc. of 9th Conf. on Computational Natural Language Learning (CoNLL-2005). – 2005. – P. 104–111.

30. *Gamallo P., Gonzalez M., Agustini A., Lopes G., de Lima V.S.* Mapping syntactic dependencies onto semantic relations // Proc. of the ECAI Workshop on Machine Learning and NLP for Ontology Engineering. – 2002. – P. 15–22.

31. *Ganter B., Wille R.* Formal Concept Analysis: Mathematical Foundations. – N. Y.: Springer-Verlag, 1999. – 294 p.

32. *Ge R., Mooney R.J.* Discriminative reranking for semantic parsing // Proc. of Joint Conf. of the Int. Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL-06). – Sydney, Australia, 2006. – P. 263–270.

33. *Ge R., Mooney R.J.* A statistical semantic parser that integrates syntax and semantics // Proc. of the Ninth Conf. on Computational Natural Language Learning (CoNLL-2005). – 2005. – P. 9–16.

34. *Gildea D., Palmer M.* Automatic labeling of semantic roles // Comput. Linguist. – 2002. – V. 23. – P. 245–248.

35. *Hahn U., Schnattinger K.* Towards text knowledge engineering // Proc. of the 15th Nat. Conf. on Artificial Intelligence and the 10th Conf. on Innovative Applications of Artificial Intelligence (AAAI'98/IAAI'98). – 1998. – P. 524–531.

36. *Handschuh S., Staab S.* Authoring and annotation of web pages in CREAM // Proc. of the 11th Int. World Wide Web Conf., WWW 2002. – Honolulu, Hawaii, 2002. – P. 462–473.

37. *Hearst M.A.* Automatic acquisition of hyponyms from large text corpora // Proc. of 14th Int. Conf. on Computational Linguistics. – 1992. – P. 539–545.

38. *Haruno M., Shirai S., Ooyama Y.* Using Decision Trees to Construct a Practical Parser // Machine Learning. – 1999. – V. 34. – P. 131–149.

39. *Hindle D.* Noun classification from predicate-argument structures // Proc. of the Annual Meeting of the Association for Computational Linguistics. – 1990. – P. 268–275.

40. *Kate R., Mooney R.J.* Using string-kernels for learning semantic parsers // Proc. of Joint Conf. of the Int. Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL-06). – Sydney, Australia, 2006.

41. *Kavalec M., Svátek V.* A study on automated relation labeling in ontology learning // Ontology Learning from Text: Methods, Evaluation and Applications / P. Buitelaar, P. Cimiano, B. Magnini (eds.). – 2005. – P. 44–58.

42. *Kim S.-D., Zhang B.-T., Kim Y.T.* Learning-based Intrasentence Segmentation for Efficient Translation of Long Sentences // Machine Learning. – 2001. – V. 16. – P. 151–174.

43. *Kiyavitskaya N., Zeni N., Mich L., Cordy J.R., Mylopoulos J.* Text Mining through Semi Automatic Semantic Annotation // Proc of PAKM'2006. – 2006. – URL: htpp://citeseer.ist.psu.edu/Kiyavitskaya06text.html.

44. *Kiyavitskaya N., Zeni N., Mich L., Cordy J.R., Mylopoulos J.* Applying Software Analysis Technology to Lightweight Semantic Markup of Document Text // Proc. of Int. Conf. on Advances in Pattern Recognition (ICAPR-2005). – Bath, UK, 2005. – P. 590–600.

45. *Koomen P., Punyakanok V., Roth D., Yih W.* Generalized inference with multiple semantic role labeling systems // Proc. of the 19th Conf. on Computational Natural Language Learning (CoNLL-2005). – 2005. – P. 181–184.

46. *Landauer T.K., Foltz P.W., Laham D.* Introduction to Latent Semantic Analysis // Discourse Processes. – 1998. – V. 25. – P. 259–284.

47. *Lodhi H., Saunders C., Shawe-Taylor J., Cristianini N., Watkins C.* Text classification using string kernels // J. Machine Learning Res. – 2002. – V. 2. – P. 419–444.

48. *Maedche A., Staab S.* Discovering Conceptual Relations from Text. Technical Report 399. – Institute AIFB, Karlsruhe University, 2000. – URL: http://citeseer.ist.psu.edu/maedche00discovering.html.

49. *Maedche A., Pekar V., Staab S.* Ontology learning. Part one – on discovering taxonomic relations from the web // Web Intelligence. – Springer, 2002.

50. *Manning C., Schutze H.* Foundations of Statistical Natural Language Processing. – Cambridge, MA: The MIT Press, 1999. – 620 p.

51. *Marcus M., Santorini B., Marcinkiewicz M.A.* Building a large annotated corpus of English: The Penn Treebank // Comput. Linguist. – 1993. – V. 19, No 2. – P. 313–330.

52. *McNamara D.C., Levinstein I.B., Boonthum C.* iSTART: Interactive Strategy Training for Active Reading and Thinking // Behavior Research Methods, Instrument, and Computers. – 2004. – V. 36.– P. 222–233.

53. *Mooney R.J.* Learning Language from Perceptual Context: A Challenge Problem for AI // Proc. of the 2006 AAAI Fellows Symposium. – Boston, MA, 2006. – URL: http://www.cs.utexas.edu/ ml/publication/nl.html.

54. *Mustafaraj E., Hoof M., Freisleben B.* Mining Diagnostic Text Reports by Learning to Annotate Knowledge Roles // Natural Language Processing and Text Mining / A. Rao, S.R. Poteet (eds.). – Springer, 2007. – P. 45–67.

55. *Naidenova X.A.* Reducing a Class of Machine Learning Algorithms to Logical Commonsense Reasoning Operations // Mathematical Methods for Knowledge Discovery and Data Mining. / G. Felici, C. Vercellis (eds.). Information Science Reference, Hershey, New York, 2007. – Chapter III. – P. 41–64.

56. *Naidenova X.* An Incremental Learning Algorithm for Inferring Logical Rules from Examples in the Framework of Common Reasoning Process // Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques / E. Triantaphyllou, G. Felici (eds). – Springer, 2006. P. 89–146.

57. *Naidenova X.A., Shagalov V.L., Plaksin M.V.* Inductive Inferring All Good Classification Tests // Proc. of the Int. Conf. "Knowledge – Dialog - Solution" (KDS-95). – 1995. – V. l. – P. 79–84.

58. *Popescu A.-M., Etzioni O.* Extracting Product Features and Opinions from reviews // Natural Language Processing and Text Mining / A. Rao, S.R. Poteet (eds.). – Springer, 2007. – P. 9–28.

59. *Priss U., Old L.J.* Modelling Lexical Databases with Formal Concept Analysis // J. Univer. Comput. Sci. – 2004. V. 10, No 8. – P. 967–984.

60. The Penn Treebank. – URL: http://www.cis.upenn.edu/∼treebank/.

61. *Quinlan J.R.* Induction of Decision Trees // Machine Learning. – 1986. – V. 1. – P. 81–106.

62. *Quinlan J.R.* C4.5: Programs for Machine Learning. – San Mateo, CA: Morgan Kaufmann Publ., 1993. – 302 p.

63. *Rabiner L.R.* A tutorial on hidden Markov models and selected application in speech recognition // Proc. IEEE. – 1989. – V. 77, No 2. – P. 257–286.

64. *Resnik P.* Selectional preference and sense disambiguation // Proc. of the ACL SIGLEX WORKSHOP on Tagging Test with Lexical Semantics: Why? What? And How? – 1997. – P. 52–57.

65. *Ribas F.* On learning more appropriate selectional restrictions // Proc. of the 7th Conf. of the European chapter of the Association for Computational Linguistics (EACL). – 1995. – P. 112–118.

66. *Schmid H.* Probabilistic part-of-speech tagging using decision trees // Proc. of Int. Conf. on New Methods in Language Processing. – Manchester, UK, 1994. – P. 44–49.

67. *Shulte im W.* Clustering verbs semantically according to their alternation behavior // Proc. of the 18th Int. Conf. on Computational Linguistics (COLING-00). – 2000. – P. 747–753.

68. *Siskind J.M.* A computational study of cross-situational techniques for learning word-to-meaning mappings // Cognition. – 1996. – V. 61, No 1. – P. 39–91.

69. *Shutz A., Buitelaar P.* RelExt: A tool for relation extraction from text in ontology extension // Proc. of the Int. Semantic Web Conf. – 2005. – P. 593–606.

70. *Tang L.R., Mooney R.J.* Using multiple clause constructors in inductive logic programming for semantic parsing // Proc. of the 12th Europ. Conf. on Machine Learnin. – Freiburg, Germany, 2001. – P. 466–477.

71. *Thompson C.A., Mooney R.J.* Acquiring word-meaning mappings for natural language interfaces // J. Artif. Intellig. Res. – 2003. – V. 18. – P. 1–44.

72. *Toutanova K., Haghighi A., Manning C.D.* Joint learning improves semantic role labeling // Proc. of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL-05). – 2005. – P. 589–596.

73. *Tufis D., Mason O.* Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger // Proc. of the 1st Int. Conf. on Language Resources and Evaluation (LREC). – 1998. – P. 589–596.

74. *Velardi P., Fabriani P., Missikoff M.* Using text processing techniques to automatically enrich a domain ontology // Proc. of the ACM Int. Conf. on Formal Ontology in Information Systems. – 2001. – P. 270–274.

75. *Wong Y.W.* Learning for semantic parsing using statistical machine translation techniques. Doctoral Dissertation Proposal. University of Texas at Austin. Also appears as Technical Report UT-AI-05-323. – Artificial Intelligence Lab, University of Texas at Austin, October 2005. – 53 p.

76. *Wong Y.W., Mooney R.J.* Learning for semantic parsing with statistical machine translation // Proc. of the Human Language Technology Conference - North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL-06). – N. Y., 2006. – P. 439–446.

77. Wordnet. – URL: http://wordnet.princeton.edu/.

78. *Zelle J.M., Mooney R.J.* Learning to parse database queries using inductive logic programming // Proc. of the 13th Nat. Conf. on Artificial Intelligence (AAAI-96). – 1996. P. 1050–1055.

79. *Luke S.* Zettlemoyer and Michael Collins. Learning to map sentences to logical form: Structured classification with Probabilistic Categorial Grammars // Proc. of 21th Conf. on Uncertainty in Artificial Intelligence (UAI-2005). – Edinburgh, Scotland, 2005. – P. 658–666.

80. *Афифи А., Эйзен С.* Статистический анализ. Подход с использованием ЭВМ / Пер. с англ. – М.: Мир, 1982. – 488 с.

81. *Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л., Лазурский А.В., Митюшин Л.Г., Санников В.З., Цинман Л.Л.* Лингвистический процессор для сложных информационных систем. – М.: Наука, 1992. – 256 с.

82. *Галушкин А.И.* Теория нейронных сетей. Кн. 1. – М.: ИПРЖР, 2000. – 416 с.

83. *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во ИМ, 1999. – 270 с.

84. *Иомдин Л.Л., Сизов В.Г., Цинман Л.Л.* Использование эмпирических весов при синтаксическом анализе // Обработка текста и когнитивные технологии. – Казань: Отечество, 2001. – С. 64–72.

85. *Кобзарева Т.Ю., Лахути Д.Г., Ножов И.М.* Модель сегментации русского предложения // Труды конференции Диалог'2001. – 2001. – Т. 2. – С. 185–194.

86. *Ледли Р., Ластед Л.* Медицинская диагностика и современные методы выбора решения // Медицинские проблемы в биологии / Под ред. Р. Беллмана. – М.: Мир, 1966. – С. 141–198.

87. *Найденова К.* Редукция задач машинного обучения к аппроксимации заданной классификации на множестве примеров // Труды 5-й нац. конф. по искусственному интеллекту. – Казань, 1996. – Т. 1. – С. 275–279.

88. *Финн В.К.* Об Интеллектуальном анализе данных // Новости искусственного интеллекта. – М., 2004. – № 3. – С. 3–18.

Поступила в редакцию
27.03.08

Найденова Ксения Александровна – кандидат технических наук, старший научный сотрудник Военно-медицинской Академии, г. Санкт-Петербург.

E-mail: *naidenovaxen@gmail.com*

Невзорова Ольга Авенировна – кандидат технических наук, ведущий научный сотрудник НИИ математики и механики им. Н.Г. Чеботарева Казанского государственного университета, доцент Татарского государственного гуманитарно-педагогического университета.

E-mail: *olga.nevzorova@ksu.ru*