Министерство образования Республики Беларусь
Учреждение образования БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет компьютерных систем и сетей

Кафедра информатики

РЕФЕРАТ

на тему

Обзор задач обработки естественного языка (Natural Language Processing)

Магистрант:

А.С. Долматович

МИНСК 2020

Содержание

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 NIL). Характерными чертами таких исследований являются:

• использование эмпирических методов с точными критериями оценок;

• тезаурусы,

• корпуса текстов

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 для лексических онтологий или стандарт PcnnTrceBank 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 NIL. 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 NIL 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 NIL 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 н 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 целыо найти правила классификации, описывающие полученное разбиение объектов на группы (классы) в пространстве признаков.

Ключевыми моментами машинного обучения являются:

1) выбор и формирование признакового пространства:

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.

Синтаксические зависимости, в частности, между глаголом и его аргументами, используются довольно часто в качестве признаков при выделении семантических отношений из текста. 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. Другая задача выделить глаголы, обозначающие одно и то же онтологическое отношение (семантическая кластеризация глаголов), решается аналогично, используя ограничения на аргументы глагола.

Латентный семантический анализ 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 некоторых системах, например, 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 н 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, «P1 взаимодействует с P2 », «P1 активируется через P2 », признаками могут быть множество слов, множество тегов (частей речи), слова, стоящие перед P1 и P2 и т. д. Тогда с каждой пустой позицией шаблона ассоциируется вектор признаков из множества 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.

Эти отношения рассматриваются как онтологические структуры, так как имена собственные соединяются с классами или концептами.

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.

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 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.

Целыо 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.

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 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 н 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. 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. В более общей постановке это выделение семантических отношений между заданными понятиями, например, между биологическими понятиями «клетка», «ген», «белок».

В задачах извлечения знаний из ЕЯ-текстов часто используется комбинация как традиционных методов машинного обучения, так и новых методов. Так могут комбинировать синтаксический разбор со статистическими методами классификации. Обучение при разметке семантических ролей, называемое в этом контексте 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 не размеченных 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. Примером задачи этого класса является задача построения классификации существительных на основе предикатноаргументных структур. 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. Чаще всего, выделение семантических отношений между словами основано на использовании лексико-синтаксических паттернов или фреймов синтаксической категоризации (глагольно-объектные отношения).

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 (морфологический анализ), разметку частей речи, 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.

Нормализация заключается в выделении последовательностей дат. времени и т. п. и переводе их в стандартизованный вид. Этап может включать расшифровку сокращений.

Теггинг (tagging, part of speech disambiguation) приписывание каждому слову (токену) грамматической характеристики части речи.

Традиционно различают подход на основе правил, или трансформационный подход и статистический (вероятностный) подход на основе Марковских моделей. Этап требует полного или частичного разрешения морфологической омонимии и унификации значений грамматических характеристик. Современные методы теггинга основаны на методе обучения решающим деревьям TreeTagger и Qtag Tagger и имеют точность 95 97%.

Лемматизация (lcmmatization, stemming) приведение словоформы к нормальной форме слова, репрезентирующей лексему. В монографии 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. 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.

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. Однако в последние годы в задачах синтаксического анализа начинают применяться методы машинного обучения.

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, ASSISTAT, C4.5. 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 подходящего 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.

Работы по семантическому анализу текстов, главным образом, связаны с поверхностным семантическим анализом. Типичной задачей является маркирование семантических ролей.

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 использует индуктивное логическое программирование для обучения детерминированного анализатора, написанного на языке Prolog.

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

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

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

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

Лингвистическая аннотация рассматривается как частный случай семантической аннотации, реализованной в системе CREAM. 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»).

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 н 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.

Литература

1. Abney S. Partial Parsing via Finite-State Cascades // ESSLLF96 Workshop 011 Robust Parsing Workshop. Prague. Czech Republic. 1996. P. 71 84.

2. Abney S. Chunks and Dependencies: Bringing Processing Evidence to Bear 011 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 011 Acquisition of Lexical Knowledge from Text. 1993. P. 70 81.

8. Bisson G. , Nedellee C., Canamero L. Designing clustering methods for ontology building: The Mo!K workbench // Proc. of the ECAI Ontology Learning Workshop. 2000. P. 13

9. Brill E. Some advances in transformation-based part.-of-speech tagging // Proc. of the Nat. Conf. 011 AI (AAAI). 1994. P. 722 727.

10. Buitelaar P., Olejnik D., Sintek M. A Protege plug-in for ontology extraction from text based 011 linguistic analysis // Proc. of the 1st European Semantic Web Symposium (ESWS). 2004. P. 31 44.