**УДК 681.3**

***А.В. Анисимов, А.А. Марченко***
Киевский национальный университет им. Тараса Шевченко, Украина

# Система обработки текстов на естественном языке

В работе описывается система, которая была создана для решения таких задач обработки текстов на естественном языке (Natural Language Processing, NLP), как анализ текстов на естественном языке, синтез текстов, автоматическая генерация реферата текста, автоматическая индексация (определение тематики) текста, автоматический перевод текста с одного языка на другой, поддержка диалога на естественном языке и т.д.

## Введение

В последнее время математическая лингвистика стала одним из наиболее популярных и важных направлений в искусственном интеллекте. Достигнуты значительные результаты: разработаны большие лексикографические системы, электронные словари, мощные системы машинного перевода, системы автоматического реферирования. Но существуют проблемы, которые до сих пор остались открытыми. Эти проблемы порождены природой самого языка. Это явления омонимии, ононимии, полисемии и много других. Особенно влияние этих явлений ощутимо при создании систем автоматического перевода. Машинный перевод по качеству существенно уступает человеческому. В других системах интеллектуальной обработки текстов на естественном языке ситуация не намного лучше. Проблема состоит в сложности установления корректного отображения действительной семантико-синтаксической структуры предложения в его внутренне логическое представление, которое генерируется системой автоматически.

Одна из наиболее негативных тенденций исследований – это попытки анализировать и моделировать язык на разных функциональных уровнях изолированно, хотя большинство ученых утверждают, что эти уровни тесно связаны друг с другом и взаимодействуют во время синтеза и анализа текста. Большинство современных моделей, которые используются сегодня в системах обработки языка, являются «изолированными» по структуре. Разные уровни обработки языка объединяются только алгоритмически. Примерами моделей семантики предложения на естественном языке являются семантические сети, фреймы. Синтаксис предложений на естественном языке выражается с помощью моделей грамматик Хомского, системной грамматики Холлидея, расширенной сети переходов. Существует много лексико-морфологических моделей. Эти модели являются структурно изолированными на своих уровнях. Объединить их позволяют эвристические алгоритмы, что иногда выходит не очень эффективно.

Все уровни языка связаны друг с другом не только функционально, но и структурно.

На кафедре математической информатики Киевского национального университета имени Тараса Шевченко была разработана система обработки текстов на естественном языке, в основу которой были положены новые принципы структурирования данных для обработки текстов на естественном языке. В качестве базы знаний используется онтологическо-семантическая сеть, которая системно и иерархически описывает объекты, свойства и отношения окружающего мира или некоторой предметной области. Отдельного рассмотрения заслуживает подраздел онтологии – *дерево действий и отношений.* Оно системно и интенционально описывает иерархию действий – от абстрактных действий до более конкретных представителей-подклассов, что позволяет использовать механизм наследования при описании действий и их свойств. В онтологии концепты типа *действие* рассматриваются как функции, имеющие набор аргументов, которые соответствуют различным аспектам действия. Для каждого аргумента можно задать область определения. Для каждого аспекта действия можно указать, концепт какого типа может быть использован, и тут же можно задать, какую синтаксическую структуру следует употребить. Таким образом, получаем прямую структурную связь между семантикой и синтаксисом. Так как в онтологии семантически описываются объекты, свойства и отношения мира, то можно с помощью узлов специального типа сохранять прямо в онтологии лексические единицы, которые будут соответствовать этим объектам, свойствам и отношениям, то есть будут их названиями в некотором естественном языке.

Таким образом, получаем структуру, в которой структурно связаны семантика, синтаксис и лексика языка. Это позволяет значительно упростить процедуры синтеза и анализа текстов на естественном языке и повысить их эффективность.

Базисными функциями системы являются анализ, который переводит текст на естественном языке в формальное логическое представление его смысла, и синтез, который по логическому представлению информации генерирует текст.

В отличие от многих подобных программ, система не имеет узкой тематической направленности применения, она разработана как универсальная, то есть при необходимости на ядро системы инсталлируются модули нужных пользователю тематик и она начинает поддерживать работу в этих областях знаний. Среди позитивных свойств нужно также отметить возможность относительно легкого добавления модулей новых языков в систему.

Система разрабатывалась как ядро, которое исполняет базовые функции обработки текста и на основе которого можно реализовать ряд прикладных программ. На основе разработанного ядра были созданы программы автоматической индексации и автоматического реферирования текстов.

## Архитектура системы

На рис. 1 представлена архитектура системы.

Наиболее важной частью ядра является база знаний «Семантическая онтология». Она интенсивно используется почти на всех этапах анализа, и результаты анализа текста представляются также в виде семантической сети с

возможностью добавления в головную базу знаний непротиворечивых фактов, полученных из входного текста.

*Стадии анализа являются стандартными: морфологическо-лексический анализ, синтаксический анализ, семантический анализ.*

Программы автоматической индексации и автоматического реферирования обрабатывают полученную в результате анализа семантическую сеть входного текста и используют факты из главной базы знаний, когда требуется знание контекста.

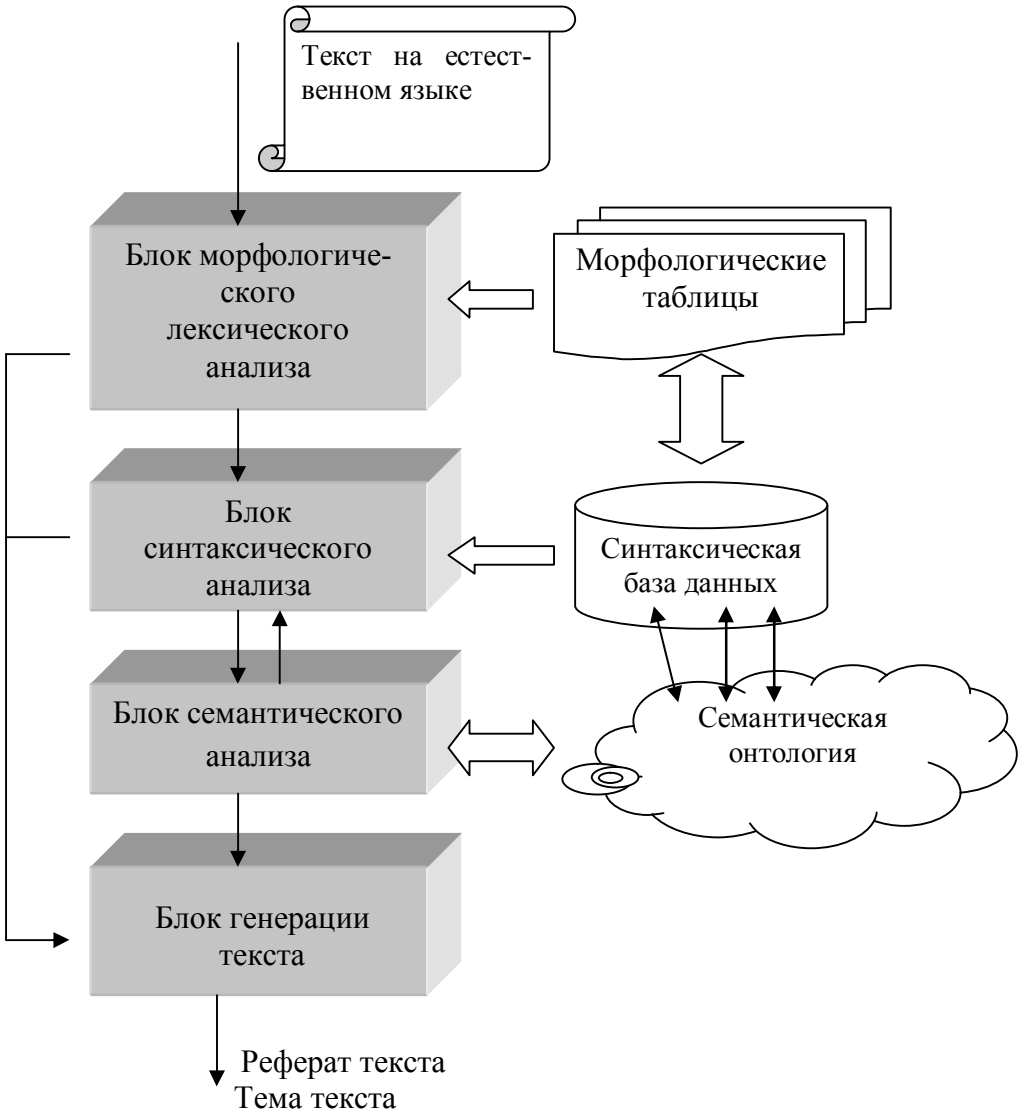


Рис. 1. Архитектура системы

Блок синтеза текста, используя правила из онтологии и шаблоны блока синтаксического анализа, строит по семантической сети текста линейные структуры предложений, которые заполняются соответствующими лексемами в нужной форме с помощью блока лексического анализа. На выходе мы получаем текст на естественном языке.

# Семантическая онтология

Семантическая онтология – это направленный гиперграф. Каждый узел представляет концепт и имеет набор связей-отношений, которые соединяют этот узел-концепт с другими узлами-концептами. То есть смысл концепта отражается его релятивной позицией. Каждый узел имеет имя – слово, которое характеризует значение смысла узла.

Наиболее важным видом связей в графе является отношение «быть». Эти связи образуют онтологический иерархический граф концептов естественного языка. В корне онтологического дерева находится наиболее абстрактный объект «все», от него идут сыновья «действие», «объект», «свойство» и «отношение», которые являются выделенными в онтологии основными категориями естественного языка. Эти узлы имеют, в свою очередь, несколько менее абстрактных детей. Один узел может иметь несколько отцов, то есть может наследовать семантические отношения и атрибуты всех своих отцов. На втором уровне сети, как уже упоминалось, идет разделение на изолированные подсети «действие», «объект», «свойство» и «отношение» (рис. 2).

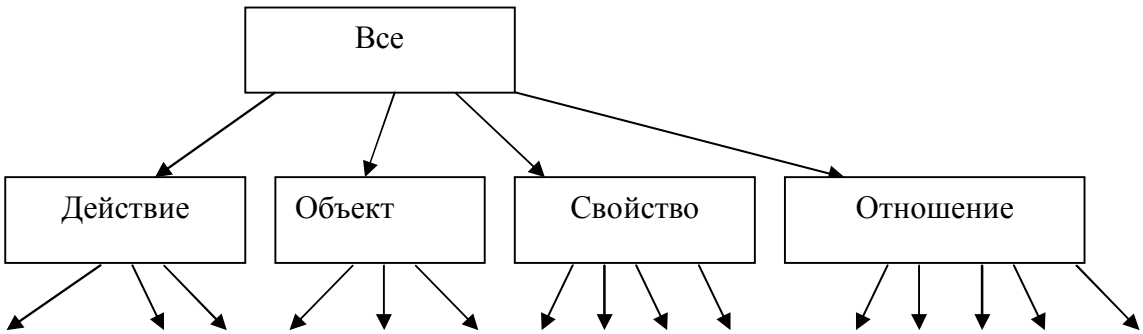


Рис. 2

Таким образом, база знаний представляет собой онтологическую иерархическую сеть [1], которая содержит множество концептов языка. Кроме вертикальных связей «быть», содержится набор горизонтальных отношений, таких, как «имеет свойство», «делает» и другие, которые описывают свойства объектов, отношения, в которые они вступают и с кем, действия и с кем они происходят и другие факты окружающего мира.

Иерархичность обеспечивает эффективное использование механизма наследования, что помогает избежать избыточности. Приведем пример описания концепта в сети. Возьмем понятие «Родина». Толковый словарь дает определение, что это – страна, где родился данный человек. Концепт «Родина» будет иметь позицию в сети как показано на рис. 3.

При добавлении в систему новой проблемной области добавляется ее иерархическая сеть, которая описывает ее концепты и связи между ними. В системе вместе с главной сетью может быть неограниченное количество дополнительных. Этим объясняется относительная легкость добавления в систему новых тематик.

Лексикон привязывается к семантической онтологии – конкретное слово к соответствующему концепту. В случае синонимии одному концепту соответствует несколько слов лексикона, в случае омонимии – одному слову соответствует несколько концептов.

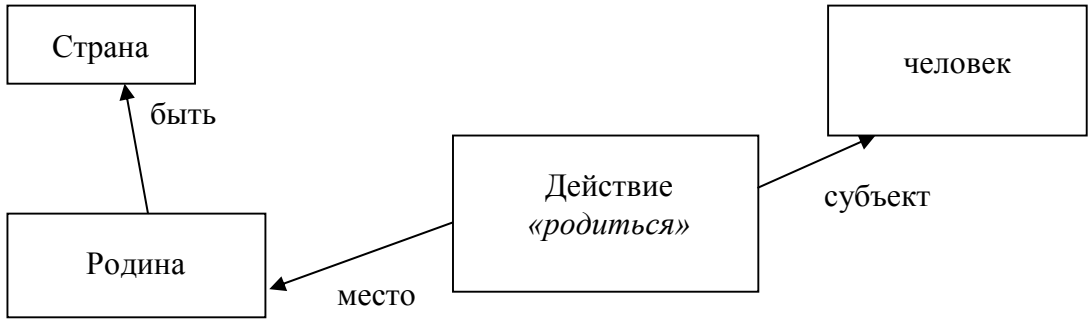


Рис. 3.

При создании сети использовались данные WordNet [2], базы данных, которые содержат лексическую и семантическую информацию словоформ английского языка.

## Блок морфологического лексического анализа

На вход системы подается текст в виде последовательности предложений на естественном языке. Задача морфологическо-лексического анализатора состоит в том, чтобы для каждого слова входных предложений был найден вход в лексиконе онтологии системы, а также полностью определены морфолексические характеристики входных слов (род, число, падеж и т.д.) Для решения этой проблемы были разработаны словотворные модели для английского и украинского языков.

## Блок синтаксического анализа

На вход этого блока поступают линейные последовательности предложений текста, которые в явном виде содержат все морфолексические характеристики. На основе синтаксических шаблонов, привязанных к онтологии, с помощью морфолексических характеристик слов предложения преобразуются в синтаксические структуры. При первом проходе синтаксического анализатора образуются синтаксические группы (глагольные группы, группы существительного и т.д.). На втором проходе идет сборка отдельных групп в единую синтаксическую структуру на основе заполнения аспектных полей глагольной группы согласно синтаксическим шаблонам, прикрепленным к онтологии. Иногда синтаксический анализ не в состоянии однозначно определить корректную синтаксическую структуру предложения с помощью правил синтаксиса. Например, предложение «Девушка ехала в автобусе в шляпе». Правила синтаксиса не могут нам ответить, с чем связана лексема «шляпа» – с лексемой «девушка» или с лексемой «автобус». Это вопрос семантики. Поэтому при сборке синтаксической структуры подключается блок семантического анализа. Блок семантического анализа вычисляет длину путей в семантической онтологии между одной парой концептов и между другой (от «девушки» до

«шляпы» и от «автобуса» до «шляпы»), а потом после сравнения делается вывод, который возвращается блоку синтаксического анализа, который заканчивает формирование синтаксической структуры предложения.

## Блок семантического анализа

В системе семантический анализ работает параллельно с синтаксическим**.** Блок семантического анализа заменяет в собранных структурах слова на концепты. При этом он добавляет из семантической сети системы вместе с концептом некоторый семантический контекст – набор наиболее специфических атрибутов и отношений концепта. Дальше идет проверка на непротиворечивость объединенных в структуре концептов. Идет логическая проверка – насколько естественно объединены в структуре объекты и отношения, которые их связывают; насколько образованная сеть предложения «квазиизоморфна» соответствующей подсети семантической онтологии системы. Одной из первых решается проблема с заменой местоимений на концепты, на которые они ссылаются в тексте. Тут алгоритм ориентируется на морфолексические характеристики местоимений (род, число, падеж) и слов-концептов, которые встречались в тексте (они должны совпадать), а также семантические свойства – реляционная позиция местоимения в семантической сети предложения должна быть подобной той, которую занимает концепт-кандидат на ее место в семантической сети системы. Таким образом, постепенно из сетей предложений мы получаем семантическую сеть текста.

## Процедура генерации реферата

Дальше мы можем «взвесить» вершины семантической сети текста. Наиболее важными узлами сети считаются вершины, которые имеют наибольшее количество связей с другими. Таким образом, взвесив вершины графа и откинув наиболее легкие «маргинальные», система получает семантический образ будущего реферата. Проведя сравнительный анализ концептов и связей полученного образа с сетями проблемных областей, которые содержатся в семантической онтологии системы, программа автоматической индексации может сделать вывод о тематике входного текста. Также полученный семантический образ текста можно использовать как поисковый образ при создании новых поисковых систем.

В полученном оптимизированном графе вершины и связи имеют свою временную оценку. В простейшем случае она соответствует порядку, в котором соответствующие этим концептам предложения появляются в тексте. Генератор текста обрабатывают последовательно подграфы сети, вершины и связи которых имеют одинаковую временную оценку по возрастанию – от наименьшей до самой большой. Генератор с помощью блока синтаксического анализа находит соответствие между структурой подграфа и некоторым синтаксическим шаблоном. Дальше, согласно найденным синтаксическим шаблонам, генератор перестраивает структуру подграфа в линейную. После этого с помощью блока морфолексического анализа в позиции полученной линейной структуры

вставляются соответствующие концептам лексемы в нужной форме. Таким образом, генерируется предложение на естественном языке. Обработав последовательно все подграфы оптимизированной сети текста, система порождает реферат текста. Отметим, что для повышения качества текста были использованы механизмы синонимов, местоимений и другие стилистические приемы.

# Заключение

При разработке системы обработки текстов на естественном языке были практически проверены и реализованы новые идеи и принципы анализа и синтеза текстов, заключающиеся в системном объединении семантики, синтаксиса и морфологии языка. На сегодняшний день на базе ядра созданы системы автоматического реферирования и индексирования, подходит к завершению создание диалоговой системы на английском языке.

# Литература

1. Nirenburg Sergei and Raskin Victor. Ontological Semantics, 2001 // crl.nmsu.edu/stuff.pages/Techial/book/index-book.html
2. Miller G. Wordnet: An online lexical database // International Journal of Lexicography. – 1990. – № 3 (4).

In this article Natural Language Text Processing system is described. It has been created as a general purpose Natural Language Processing system (NLP) for such tasks as analysis and synthesis of natural language texts, reviewing of texts, translation systems, natural language dialogue systems.

*Статья поступила в редакцию 17.07.02.*