УДК 808.2: 159.937

М. В. МАКАРИЧ

РЕАЛИЗАЦИЯ ОБУЧАЮЩЕЙ КОМПЬЮТЕРНОЙ ПРОГРАММЫ НА ОСНОВЕ ЛИНГВИСТИЧЕСКОЙ БАЗЫ ДАННЫХ АВТОМАТИЧЕСКОЙ СИСТЕМЫ ОБРАБОТКИ ТЕКСТА

Белорусский национальный технический университет

Вследствие постоянного увеличения объема электронной текстовой информации современное общество испытывает острую потребность в автоматизированной обработке естественного языка (ЕЯ). Главное предназначение систем автоматической обработки ЕЯ – анализировать и синтезировать тексты, то есть преобразовывать тексты из их исходного вида в некоторое абстрактное представление, приближенное к представлению смысла, и обратно. Целью предлагаемой работы является разработка лингвистического и информационного обеспечения автоматической системы для обработки англоязычных публицистических текстов и ее последующая реализация в виде обучающей компьютерной программы. В статье рассматриваются примеры различных подходов к технологии создания лингвистической базы данных для систем обработки текста естественного языка. Автор дает подробное описание основных структурных блоков принципиально нового лингвистического процессора: лексико-семантического, синтаксического и семантико-синтаксического. Основным преимуществом данного процессора является использование в алфавитном словаре специальных семантических кодов в соответствии с разработанной лексико-семантической классификацией. Это позволяет автоматической системе точно определять семантические функции ключевых слов, входящих в выделенные в процессе синтаксического анализа группы. Что, в свою очередь, помогает избежать ошибок, характерных для такого рода систем.

Ключевые слова: *система автоматической обработки ЕЯ, лингвистическое и информационное обеспечение, лингвистический процессор, обучающая компьютерная программа.*

Введение

Естественный язык (ЕЯ) является неотъемлемой частью нашей повседневной жизни. Однако вследствие постоянного увеличения объема электронной текстовой информации современное общество испытывает острую потребность в автоматизированной обработке ЕЯ. Главное предназначение систем автоматической обработки ЕЯ – анализировать и синтезировать тексты, то есть преобразовывать тексты из их исходного вида в некоторое абстрактное представление, приближенное к представлению смысла, и обратно. Помимо этого главного предназначения, направленного на решение фундаментальной задачи моделирования ЕЯ, имеется и прикладной аспект. Способность автоматической системы понимать и строить тексты может быть использована в приложениях, способных принести конкретную пользу. Предлагаемая работа посвящена разработке лингвистического и информационного обеспечения автоматической системы для обработки англоязычных текстов и ее последующей реализации в виде обучающей компьютерной программы.

Общее описание и структура базы данных автоматической системы

Как уже говорилось выше, целью данной работы является разработка лингвистического и информационного обеспечения, на основе которого может быть создана система или, другими словами, построена функциональная модель, позволяющая раскрыть суть механизма создания такого текста, который был бы максимально удобным для восприятия человеком. Анализ литературы, содержащей общие принципы процесса автоматической обработки текстов ЕЯ, показывает, что такого рода системы должны содержать три последовательных блока: лексико-грамматический, синтаксический и блок семантического анализа текстов [1].

Целью *лексико-грамматического анализа* является разбор входного потока слов с распознаванием частей речи: существительное, прилагательное, глагол, наречие и т. д., а также других морфологических параметров, таких как род, число, падеж и др. Основой данного блока является автоматический словарь. В настоящее время наиболее используемыми являются словари словоформ, где для любого слова хранятся все его возможные формы с указанием возможных лексико-грамматических классов. Для распознания неизвестных, т. е. не содержащихся в словаре слов могут быть реализованы алгоритмы идентификации, использующие формальные признаки частей речи (окончания, суффиксы и т. п.) [2].

На этапе лексико-грамматического анализа решается и проблема устранения лексической многозначности слов. Большое количество слов имеют одинаковое написание, но являются различными частями речи. Любому входному слову текста, учитывая контекст, необходимо поставить в соответствие единственный лексико-грамматический код. Данная задача является легко выполнимой применительно к русскому языку. Его развитая морфология позволяет осуществить это практически со стопроцентной точностью. В английском языке простой алгоритм, присваивающий каждому слову в тексте наиболее вероятный лексико-грамматический код, работает с точностью 95–96% [3].

Для улучшения точности анализа используются два подхода: вероятностно-статистический и основанный на продукционных правилах, оперирующих словами и кодами. Вероятностно-статистические методы используют статистику встречаемости различных кодов и слов, снятой с некоторого эталонного текста. В результате каждому слову в предложении ставится в соответствие наиболее вероятный лексико-грамматический класс. Алгоритмы, основанные на продукционных правилах, используют правила собранные автоматически, либо подготовленные квалифицированными экспертами-лингвистами. Оба подхода дают примерно одинаковый результат. При их использовании раздельно либо в различных комбинациях точность лексико-грамматического анализа английского языка улучшается до 96–98%, что сравнимо с точностью ручной обработки.

Синтаксический анализ подразумевает сегментирование (фрагментацию) текста на предложения или близкие к ним фрагменты для построения синтаксических структур. Процедура автоматического синтаксического анализа позволяет получить при помощи алгоритмов формализованную синтаксическую структуру предложения. Результатом работы автоматической системы синтаксического анализа является представление синтаксической структуры входного предложения обрабатываемого текста в виде синтаксического дерева. Исходной информацией для работы такой системы служит морфологическое представление слов в виде цепочки кодов, репрезентующих грамматический класс слова и его словоизменительные характеристики. Таким образом, морфологический анализ путем кодификации слов текста обеспечивает доступ к денотативной информации. На следующем этапе создается семантическая структура текста с помощью синтаксических, семантических и пунктуационных средств. Существует определенный параллелизм между синтаксической и семантической структурами, который проявляется в соответствии структурных связей семантическим. Исследователями замечено сходство результатов системы автоматического анализа текста и, например, лингвиста, изучающего язык и действующего в противоположном направлении. Данный факт подтверждает гипотезу о том, что язык хорошо коррелирован на самых различных уровнях. Это значит, что синтаксическая структура, полученная на основе только грамматических критериев, очень близка синтаксической структуре, полученной на основе семантических критериев, в пользу чего говорят многие данные, накопленные лингвистикой за последние годы.

Семантический анализ заключается в выделении из документа основных смысловых единиц (слов, словосочетаний) и определении ассоциативных, причинно-следственных и др. связей между ними [4]. Основным средством для этого является некоторая система правил. Семантический этап – базовая составляющая систем автоматического понимания текста. Он выступает в роли посредника и должен согласовать три разных «языка» [5]:

- язык построенных системой лингвистических структур (плюс другие лингвистические знания), получаемые им на входе;

• язык той предметной области, к которой относится текст и термины которой желательно использовать при построении выходной структуры;

• язык пользователя, для которого система автоматической обработки текста должна предоставить информацию.

Информация, которую автоматическая система обработки текста получает из текста, должна быть изложена на языке, понятном пользователю как с естественно-языковой точки зрения, так и с точки зрения той предметной области, которой он владеет как специалист. Иначе результат работы системы не может быть назван информацией для этого конкретного пользователя – адресата информации.

В зависимости от конкретных задач обработки текста автоматическая система может не содержать всех трех блоков: лексико-грамматического, синтаксического и семантического, или содержать дополнительные блоки, необходимые для конкретной системы. Примером может служить система, разработанная для распознавания в тексте концептов и семантических отношений между ними типа «субъект-акция-объект», отношений типа «причина-следствие» и «часть-целое» [6]. Данная система реализует следующие этапы обработки поступившего на ее вход текста: переформатирование, лексический (распознавание границ слов и предложений), лексико-грамматический, синтаксический и семантический анализ.

Таким образом, выбор подхода к разработке систем автоматического экстрагирования информации связан с конкретной задачей, поставленной перед системой автоматической обработки.

С учетом того, что целью нашего исследования является выделение главных субъектов, объектов и их действий из англоязычных публицистических текстов, в общую структуру системы автоматической обработки включены следующие блоки:

1. Блок лексико-семантического анализа слов входящего предложения, опирающийся на алфавитный словарь исследуемых текстов, содержащий семантические коды в соответствии с разработанной лексико-семантической классификацией.

2. Блок синтаксического анализа, основой которого являются списки граничных сигналов для синтаксического анализа английского предложения и выявления в нем членов предложений (группы подлежащего, группы сказуемого и т. п.).

3. Блок семантико-синтаксического анализа, определяющий семантические функции ключевых слов, входящих в выделенные в процессе синтаксического анализа группы слов. Работа данного блока построена с использованием семантических функций формального языка TABLING [7] и позволяют компьютеру ответить на ряд самых важных вопросов, касающихся содержания текста, путем построения таблицы его основного содержания. Здесь, в соответствии со значением семантических функций языка TABLING, составляющие основного содержания текста означают:

(AGA) – субъект 1, активный одушевленный инициатор некоторого действия или события.

(AGN) – субъект 2, активный неодушевленный инициатор некоторого действия или события.

(ONG) – предмет, главный объект некоторого события текста.

(R) – действие, репрезентация конкретного действия активного одушевленного инициатора.

(PRP) – свойство (признак) – свойство предмета, процесса, материала, зафиксированного в тексте.

(LOC) – место совершения некоторого события текста.

(TIM) – время совершения некоторого события текста.

Реализация созданной базы данных в обучающей компьютерной программе

Между любой теоретической моделью и ее компьютерной реализацией имеется двусторонняя зависимость. С одной стороны, компьютерная система должна как можно более точно воплотить разработанные теоретические принципы. Это, в первую очередь, представление каждого предложения обрабатываемого текста на нескольких уровнях (морфологическом, поверхностно-синтаксическом и глубинно-синтаксическом), изображение синтаксического строения предложения в виде дерева зависимостей между словами и признание словаря, наряду с грамматикой, важнейшим ком-

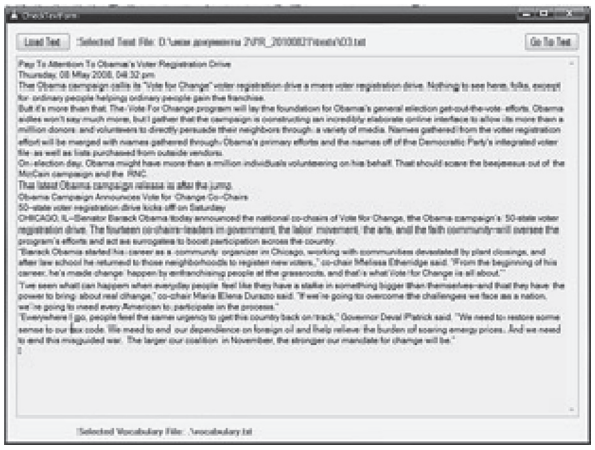

Рис. 1. Структура окна в обычном режиме

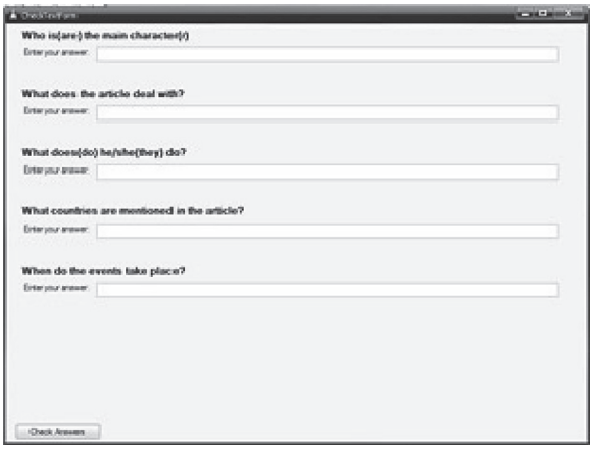

Рис. 2. Структура окна в режиме тестирования

Рис. 3. Структура окна в режиме контроля результатов

понентом лингвистической модели. С другой стороны, компьютерная модель имеет неоценимое значение для развития теории, поскольку служит объективным и надежным полигоном для проверки и отладки теоретических положений и конкретных решений. Дело в том, что как только мы покидаем область исследуемой группы текстов и переходим на более абстрактные уровни представления, мы лишаемся возможности непосредственно оценивать результаты. Лишь компьютерное моделирование предоставляет исследователю возможность наглядно увидеть, насколько адекватна действительности разработанная им теоретическая схема.

Предложенная нами теоретическая модель автоматической системы для обработки текстов ЕЯ ранее уже была реализована в компьютерной программе автоматического реферирования TRT [8]. Другая прикладная система, в составе которой описанная теоретическая модель была опробована – обучающая компьютерная программа для совершенствования навыка просмотрового чтения текстов на английском языке. Программа реализована на языке программирования C# (Sharp). Структура окна программы в обычном режиме приведена на рис. 1.

Как видно на рис. 1, основным элементом окна является рабочая область для загрузки одного из предложенных англоязычных текстов. Работа с элементами окна производится с помощью выбора необходимого файла в строке заголовка. После ознакомления с содержанием текста обучаемый переходит в режим контроля знаний при помощи кнопки «Go To Test», находящейся в правом верхнем углу. Контрольные вопросы сформулированы таким образом, чтобы проверить правильность понимания предложенного текста: *Кто главное действующее лицо текста? О чем говорится в тексте? Какие основные действия совершают главные действующие лица? И т. д.* (рис. 2).

После заполнения всех соответствующих рабочих областей появляется возможность проверить правильность данных ответов при помощи кнопки «Check Answers» в левом нижнем углу. На рис. 3 приведен пример исправленных ответов на поставленные вопросы: выбранные обучаемым ответы приведены в строке «Your Answer», а аналогичные правильные – в строке «Correct Answer».

Заключение

Созданная нами на основе разработанной лингвистической базы данных компьютерная программа может быть использована в качестве тренажера на практических занятиях по английскому языку для формирования навыка

восприятия и интерпретации англоязычного публицистического текста. Структура лингвистической базы созданной программы содержит три блока: блок лексико-семантического анализа, блок синтаксического анализа и блок семантико-синтаксического анализа. Программа открыта для доработки и может быть использована для работы с англоязычными текстами другой предметной области при внесении дополнительного материала в автоматический алфавитный словарь. Это является основным преимуществом представленной программы в процессе ее использования на практических занятиях по английскому языку, так как преподаватель может значительно сэкономить время в поиске необходимого и эффективного учебного материала в зависимости от того, какую методическую цель он ставит перед собой. С помощью интернет-ресурсов можно ввести новый лексический, страноведческий материал, сделать занятие более наглядным, закрепить учебный материал, вооружить студентов стратегиями саморазвития.

Литература

1. Пиотровский, Р. Г. Методы автоматического анализа и синтеза текста / Р. Г. Пиотровский [и др.]. – Минск: Вышэйшая школа, 1985.– 222 с.

2. Воронцов, А. В. Промышленная реализация системы лексико-грамматического анализа текстовых документов / А. В. Воронцов // Вестн. МГЛУ. Сер. 1, Филология. – 2007. – № 1(26). – С. 189–203.

3. Jurafsky, D. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition / D. Jarafsky, J. Martin. – New Jersey: Prentice Hall, 2000. – 934 p.

4. Todhunter, J. Semantic processor for recognition of whole-part relations in natural language documents: US Patent Appl. 20070156393 / J. Todhunter, I. Sovpel, D. Pastanohau and others; Intention Machine Corp. – Serial no. 686660; Series code 11; Filed 15.03.2007. – 2007.

5. Леонтьева, Н. Н. Автоматическое понимание текстов: системы, модели, ресурсы: учеб. Пособие для студентов лингвистических факультетов вузов / Н. Н. Леонтьева. – Москва: Академия, 2006–304 с.

6. Совпель, И. В. Автоматическое распознавание основных типов знаний в текстовых документах / И. В. Совпель // Искусственный интеллект ISSN1561–5359 – НАН Украины «Наука и просвещение» – 2007. – № 3 – С. 328–332.

7. Зубов, А. В. Семантико-синтаксический язык для записи текстов в памяти ЭВМ / А. В. Зубов // Функционирование и развитие языковых систем. Сборник научных трудов.– Минск: Вышэйшая школа, 1990. – С. 110–117.

8. Макарич, М. В. Автоматическая система для создания табличного реферата группы текстов / М. В. Макарич. – Germany: LAP LAMBERT Academic Publishing, 2012–145с.

References

1. R. G. Piotrovsky, Automatic text analysis and synthesis methods. Minsk: Vyshejshaya shkola, 1985.– 222 p.

2. A. V. Vorontsov, Industrial implementation of a system for lexical and grammatical analysis of text documents. J. Vestn. MSLU, Vol. 1(26), pp. 189–203, 2007.

3. D. Jurafsky, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. New Jersey: Prentice Hall, 2000. – 934 p.

4. J. Todhunter, I. Sovpel, D. Pastanohau, Semantic processor for recognition of whole-part relations in natural language documents: US Patent Appl. 20070156393; Intention Machine Corp. – Serial no. 686660; Series code 11; Filed 15.03.2007.

5. N. N. Leonteva, Automatic text interpretation: Systems, models, resources: Handbook for students of linguistic faculties. Moscow: Akademy, 2006–304 с.

6. I. V. Sovpel, Automatic recognition of basic knowledge in text. J. Artificial intelligence, Vol. 3, pp. 328–332, 2007.

7. A. V. Zubov, A semantic and syntactic language for text entry in computer memory. Functioning and development of language systems: Collection of scientific papers. Minsk: Vyshejshaya shkola, pp. 110–117,1985.

8. M. V. Makarych, Automatic system for creating a table abstract of texts. Germany: LAP LAMBERT Academic Publishing, 2012–145 p.

Поступила 01.03.2016

M. V. MAKARYCH

REALIZATION OF TRAINING PROGRAMME ON THE BASIS OF LINGUISTIC DATABASE FOR AUTOMATIC TEXTS PROCESSING SYSTEM

Belarusian National Technical University

Due to the constant increasing of electronic textual information, modern society needs for the automatic processing of natural language (NL). The main purpose of NL automatic text processing systems is to analyze and create texts and represent their content. The purpose of the paper is the development of linguistic and software bases of an automatic system for process-

ing English publicistic texts. This article discusses the examples of different approaches to the creation of linguistic databases for processing systems. The author gives a detailed description of basic building blocks for a new linguistic processor: lexical-semantic, syntactical and semantic-syntactical. The main advantage of the processor is using special semantic codes in the alphabetical dictionary. The semantic codes have been developed in accordance with a lexical-semantic classification. It helps to precisely define semantic functions of the keywords that are situated in parsing groups and allows the automatic system to avoid typical mistakes. The author also represents the realization of a developed linguistic database in the form of a training computer program.

Keywords: *NL automatic text processing systems, linguistic and software bases, linguistic processor, training computer program.*

Макарич Марина Васильевна
E-mail: 2348843@tut.by.
Доцент кафедры английского языка № 2 БНТУ, кандидат филологических наук (специальность 10.02.21 – прикладная и математическая лингвистика), доцент. Также имеет диплом БНТУ (БПИ, факультет роботов) по специальности – автоматизация и комплексная механизация машиностроения. Научные интересы: педагогический аспект технического образования и лингвистическое обеспечение информационных систем.

Makarych Marina, Associate Professor of the 2nd English Department of the Belarusian National Technical University, PhD in Applied and mathematical linguistics. In addition she has B. Sc. in robotics (e-mail: 2348843@tut. by).

Her scientific interests focus on engineering education and interdisciplinary education that combines linguistics and computer science.