А. В. Зубов (Минск)

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ИССЛЕДОВАНИИ ТЕКСТА И РЕЧИ В МГЛУ

Применение информационных технологий при изучении **текстов** в МГЛУ проводится в следующих направлениях:

1. Создание корпусов параллельных текстов и их использование в изучении текстов и их единиц.
2. Изучение текстов, функционирующих в Internet.
3. Логико-семантическое моделирование паремий.
4. Логико-семантическое моделирование и порождение связных текстов.

В рамках первого направления решаются следующие задачи:

1.1. Создание большого корпуса текстов белорусского языка и его использование для изучения белорусского языка и его связи с другими языками Европы.

Работа проводится совместно с Институтом языкознания НАН РБ. Предполагается создание тэггированного корпуса текстов белорусского языка в 1 млн. словоупотреблений (художественные тексты, публицистика) и 3 параллельных тэггированных подкорпусов: русско-белорусского, англо-белорусского и немецко-белорусского (художественные тексты, публицистика). Тэггирование слов белорусского корпуса будет осуществляться в автоматическом режиме, а тэггирование слов подкорпусов – в полуавтоматическом режиме.

1.2. Создание параллельного тэггированного корпуса текстов и разработка методики его использования для совершенствования учебного процесса и научной деятельности.

Предполагается создание трех параллельных подкорпусов – англо-русского, немецко-русского и франко-русского, каждый объемом в 300 000 словоупотреблений. Тэггирование слов будет осуществляться в полуавтоматическом режиме. Будут использованы оригинальные тексты трех иностранных языков (научные, публицистика и тексты художественные (проза и поэзия)) и их переводы на русский язык.

1.3. Разработка алгоритмов поиска переводных эквивалентов многозначных существительных с использованием параллельного корпуса текстов.

Для проведения исследования отобраны 30 многозначных английских существительных, зафиксированных в переводах на английский язык годового Отчета Государственного патентного комитета Беларуси. В нем дается краткая характеристика основной деятельности этого комитета за пять лет его существования.

1.4. Разработка алгоритма перевода терминологических словосочетаний с английского языка на русский с использованием параллельного корпуса текстов.

Исследование проводится на материале параллельных текстов по информатике и вычислительной технике. По материалам исследования

подготовлена кандидатская диссертация. Она успешно прошла предварительную экспертизу на нашем объединении ученых по проблемам общего и прикладного языкознания.

1.5. Лингвистические особенности переводов стихотворений Дж. Байрона человеком и компьютером.

В качестве компьютерных программ перевода используется PROMT XT (6-ая версия) и "Сократ профессионал".

Второе направление связано с изучением текстов, функционирующих в Internet. Здесь решаются такие задачи:

2.1. Формализация процесса перевода сообщений электронной почты. Строится система автоматического перевода сообщений электронной почты, сообщающих о научных конференциях и приглашающих принять участие в таких конференциях, с французского и английского языков на русский. Для создания алгоритма перевода изучается по 50 текстов каждого языка.

По результатам исследования подготовлена кандидатская диссертация. Но она пока не проходила обсуждений.

2.2. Автоматическое построение WEB-сайтов на определенную тему (один из них – сервисный сайт на тему "WEB-дизайн").

Сайт должен содержать не более 600 слов и при его поиске должен находиться на 1-ой или 2-ой странице результат его поиска.

И здесь по результатам исследования подготовлена кандидатская диссертация. Успешно прошедшая обсуждение на нашей кафедре информатики и прикладной лингвистики.

К задачам, решаемым *третьим направлением*, относятся.

3.1. Моделирование процесса порождения текста пословицы.

В процессе анализа 300 французских пословиц и поговорок были выделены две группы слов (содержательные и формальные) и типы связывающих их отношений ("причина–следствие", "отрицание", "противопоставление" и др. – всего 10 типов).

На основании этих данных были построены логико-семантические формулы пословиц. С опорой на эти формулы был построен алгоритм порождения французских пословиц и написана программа на языке BASIC.

Компьютер породил 249 пословиц, из них 28 (11%) совпадают с исходными, 61 пословица (24%) была совершенно новая.

Аналогичная работа проделана и для английских пословиц.

3.2. Порождение текста загадки.

Для создания программы порождения французских загадок была изучена логико-семантическая структура более 400 загадок. По загаданным объектам ("язык", "ключ", "яйцо" и др.) загадки были разделены на 67 групп. Для построения алгоритма порождения загадок были отобраны 57 загадок, представленные самыми частыми загаданными объектами ("колокол", "яйцо", "свеча", "язык", "метла", "ключ", "лук", "дверь").

Формализация структуры и содержания каждой загадки проводились в 3 шага:

1. создание полных и формализованных описаний загаданных объектов;

2. построение на их основе логико-семантических формул загадок;

3. преобразование логико-семантических формул в лексико-семантические формулы.

Программой было порождено 150 загадок. Около 60% из них совпадают с теми, которые были использованы для первоначального анализа. Около 10% из общего числа синтезированных оказались новыми и семантически и грамматически верными.

Был также составлен и другой алгоритм порождения загадки, на основе **структурно-вероятностного метода** на его основе компьютер породил 150 загадок. Только 5 из них совпали с проанализированными. 70% из вновь порожденных оказались новыми, семантически и грамматически правильными.

3.3. Моделирование логико-семантической организации анекдотов.

В качестве исходного материала выступали 250 англоязычных анекдотов типа "реплика А – контр реплика В".

Например:

A: "You *married* me for my money!"

*B: "*And I *earned* it*"*

(A: "Вы *женились* на мне из-за денег!"

(B: "И я *заработал* их!"

Исследуемый массив анекдотов содержал определенное число анекдотов с одним и тем же ключевым словом в репликах А. Например, встретилось 7 анекдотов с ключевым словом *marry (married)* – "жениться".

Пока составлен алгоритм порождения английских коротких анекдотов.

К задачам *4-го направления* можно отнести следующее:

4.1. Моделирование структурно-семантической организации англоязычных рекламных объявлений на тему "Знакомство".

Исследовано более 10 рекламных объявлений этого типа из газеты "Из рук в руки " и из Internet. В процессе анализа выяснилось, что основной текст таких объявлений может быть представлен тремя семантическим блоками:

1. Характеристика автора объявления.
2. Характеристика будущего партнера (друга, спутника жизни).
3. Цель знакомства.

Каждый такой блок был представлен в виде определенного набора дифференциальных семантических признаков ("возраст", "описание внешности", "образование" и др. – всего 17).

Каждый из 3-х семантических блоков представлялся в виде цепочек ДСП (например: "возраст" + "рост" + "вес" + "личные качества" + "семейное положение").

Был составлен алгоритм порождения таких объявлений в режиме диалога с пользователем.

4.2. Формализация взаимосвязи вербальных и невербальных составляющих английского рекламного объявления.

Было изучено более 200 рекламных объявлений на тему "Косметика и парфюмерия", взятых из русских журналов.

В итоге была установлена взаимосвязь этих составляющих на уровне опорных слов текста и дескрипторов, описывающих иллюстрацию. На основе этих данных была написана компьютерная программа, которая для

вербальной части рекламного сообщения автоматически находила подходящую иллюстрацию из некоторого множества иллюстраций, находящихся в памяти компьютера.

Это исследование вылилось в кандидатскую диссертацию Н.Г. Швец, которую кафедра информатики БГУ будет оценивать как внешний оппонент.

4.3. Порождение с помощью компьютера французской волшебной сказки.

По известной методе В.Я. Проппа было проанализировано 14 французских сказок. На основе их анализа была построена база знаний, имеющая фреймовую структуру.

Были составлены три алгоритма порождения сказки, в каждом из которых уменьшался элемент случайности выбора составляющих фрейма.

С помощью компьютера было получено 50 сказок.

4.4. Порождение с помощью компьютера русских стихотворений.

В качестве исходного материала взяты более 100 стихотворений одного автора. Составлена база знаний и алгоритм порождения.

Детально принципы решения двух последних задач описаны в только что вышедшей в Москве книге Зубова А.В. и Зубовой И.И. "Основы искусственного интеллекта для лингвистов" (М.: ЛОГОС, 2006).

II. Информационные технологии в исследовании **речи.**

Кафедра совместно с Объединенным институтом проблем информатики Национальной Академии наук Беларуси выполнила тему "Разработка алгоритмов автоматического выявления ошибок произношения и создание компьютерной системы для обучения произношению".

Система разрабатывается для обучения произношению английских фраз.

Для реализации сказанного была создана база данных в виде списка фраз, отобранных для проведения обучения ученика. По данной базе была составлена вторая база – словарь всех слов, имеющихся в составе отобранных фраз. Всего было отобрано 150 английских фраз.

Для дальнейшей коррекции было отобрано 16 возможных ошибок произнесения фразы.

Построенная компьютерная система обучения английской речи выдает обучаемому после произнесения им конкретной фразы место ошибки в слове и дает подсказку, как эту ошибку исправить. Например, так: "Опять ошибка. Первое слово. Второй слог. Длительность гласного нужно значительно уменьшить".

Компьютерная система прошла испытание в одной из гимназий г. Минска и в экологическом университете им. В.Д. Сахарова в Минске.