**Отзыв о магистерской диссертации
Сергеевой Елены Сергеевны**

**«Использование статистических методов**

**при автоматическом извлечении информации
из новостных текстов»**

Отрадно, что в последнее время наметилось усиление интереса
к факто­графическим ИПС, которые в Советские годы разрабатывались повсеместно и в широких масштабах. Мне самому довелось в этом участвовать.

При создании таких систем, как я это себе представляю и как было принято поступать в прежние годы, разработка ФИПС включала несколько этапов: извлечение фактографической информации из текста, построение языка семантической экспликации, т. е. некоторого промежуточного языка, и конструирование на этой основе собственно информационно-поискового языка объектно-признакового типа.

Данная диссертация посвящена решению первой задачи: извлечению фактографической информации из новостных текстов, циркулирующих в сети ИНТЕРНЕТ. Поток такой информации огромен. Автор диссертации отмечает, что по данным Google, видимая индексируемая часть Интернета перешла порог в один триллион уникальных документов уже в 2008 году.

Назначение создаваемой автором системы — извлечение из текстового потока информации о высказываниях и их авторах. Представляется, что данная система представляет собой нечто вроде досье, содержащего мнение индивидуума об отдельных проблемах или персоналиях, упорядочивание и систематизацию фактов, имеющих отношение к данной персоналии. Иначе говоря, это некоторый речевой портрет активного интернетного коммуниканта. Такого рода информация представляет непреходящий интерес для определенных властных и управленческих структур. В этом заключается практическая значимость данной работы.

Елена Сергеевна так формулирует задачи своего исследования:

* поиск и анализ информации о существующих статистических методах извлечения структурированной информации из текста;
* разработка формальной структуры для описания лингвистической информации;
* создание программы, производящей лингвистическую разметку новостных текстов в данной системе;
* разработка модуля поиска фактографической информации, использующего лингвистическое метаописание и опирающегося на статистические методы при поиске информации;
* оценка функционирования разработанной системы.

Работа четко структурирована. Она состоит из четырех частей. В первой части вводится понятие фактографического поиска и описывается его классическая модульная система. Во второй части производится анализ наиболее популярных существующих статистических методов извлечения информации из текста. В третьей части содержится описание созданной автором системы извлечения фактографической информации. Четвертая часть посвящена оценкам результатов работы созданной системы.
В заключении сформулированы выводы и рекомендации по использованию данной работы.

# Остановимся на наиболее важных достижениях диссертации.

1. Проведен анализ нескольких современных статистических методов, используемых для решения задачи по извлечению фактографической информации из текста. Это методы, использующие наивный классификатор Байеса, классификацию на базе дерева решений, классификацию на базе метода опорных векторов и нейронные сети. Анализ литературы проведен главным образом на материале англоязычных источников.

2. Предложена собственная формальная система представления лингвистической и метаинформации о тексте, написана программа-разметчик, которая проводит разметку входных новостных текстов в разработанном XML формате.

3. Система представления лингвистической информации о тексте была выполнена в форме многоуровневой разметки, содержащей набор поддеревьев, каждое из которых представляет собой определенный уровень лингвистического описания.

4. Разработан статистический модуль извлечения фактографической информации, использующий лингвистическую информацию метоописания для статистической классификации предложений как содержащих цитату/не содержащих цитату и выделяющий наиболее вероятного автора цитаты из текста.

Тестирование полученной системы проводилось на информационном массиве, содержащем 50 текстов. Оценена эффективность поиска. Полнота равна 0.57, точность — 0.79. Намечены пути увеличения эффективности поиска.

Работа в целом выполнена на высоком уровне. Поставленные цели достигнуты.

Некоторые недостатки.

1. Недостаточно освещены современные разработки ФИПС в отечественной науке. Непонятно, ведутся они или не очень.
2. Во введении диссертации не сформулирована новизна исследования и ее теоретическая значимость. Представляется, что новизну можно усмотреть во всесторонней оценке популярных зарубежных систем и ознакомлении с ними отечественных ученых. Элементы новизны можно «разглядеть» и в многоуровневом древесном представлении фактографической информации.

Диссертация Сергеевой Елены Сергеевны отвечает всем требованиям, предъявляемым к сочинениям подобного рода, и заслуживает высокой оценки.

Профессор кафедры

математической лингвистики, д.ф.н. Мартыненко Г.Я.

10.05.13