Лабораторная работа

«Разработка системы машинного перевода документов»

Цель работы – освоить на практике основные принципы машинного перевода документов.

Машинный перевод (МП) – процесс перевода с одного естественного языка на другой с использованием ЭВМ.

Системы машинного перевода (СМП):

- ✓ *системы прямого перевода*;

Такие системы обеспечивают результаты, близкие к пословному переводу. Операция перевода требует минимума преобразований: исходный текст постепенно превращается в текст на выходном языке путем замены всех его элементов, найденных в словаре, на переводные эквиваленты. Никакая переводная модель не используется за исключением переводных (в основном лексических) соответствий. Принимается во внимание лишь локальный контекст, он же позволяет учитывать некоторые более сложные единицы – обороты. Поэтому такой перевод называют *пословным,* или *пословно-оборотным.*

- ✓ *системы непрямого перевода*;

Они в свою очередь, подразделяются на:

- системы с *трансфером;*
- системы с *языком-посредником*.

В системах второго поколения переводные соответствия устанавливаются не «прямым» способом, а только после того, как в результате анализа для каждого предложения будет выявлена его синтаксическая и/или семантическая структура. Анализ и синтез независимы. Анализ ведется в категориях входного языка, а синтез – в категориях выходного языка; связь между этапами обеспечивается введением особого этапа *межъязыковых операций* (собственно перевода*, трансфера*).

Перевод может осуществляться с помощью семантического *языка-посредника*, универсального для разных пар естественных языков. Однако, такие системы не получили широкого распространения.

- ✓ системы, *основанные на знаниях*.

Такие СМП в качестве отдельного компонента включают *экстралингвистические знания* (знания о предметной области, ПрО), который может иметь те же формы представления, что и собственно лингвистическая информация, т.е. записываться в словарях и грамматиках. К этому классу относятся СМП, использующие при анализе *концептуальную сеть знаний*.

<u>Требования к разрабатываемой системе:</u>

- ✓ на входе – естественно-языковой текст на входном языке, подлежащий процедуре машинного перевода;
- ✓ подсчитать количество слов во входном тексте, количество переведенных слов, определить грамматическую информацию (теги частей речи и их расшифровка, для этого следует использовать функциональность лр. №3 весенний семестр).
- ✓ на выходе:

  – перевод входного текста на выходной язык;

  – упорядоченный по частоте встречаемости в тексте список слов и их переводов на выходной язык с грамматической информацией (вкладка 1) (можно использовать функциональность лр. №1 весеннего семестра);

  – построенное дерево синтаксического разбора выбранного предложения (вкладка 2).

- ✓ обеспечить наличие утилиты автоматического пополнения/корректировки полученного словаря (таблицы БД).
- ✓ обеспечить сохранение и распечатку результатов перевода и упорядоченных по частоте встречаемости в тексте списков слов и их переводов на выходной язык с грамматической информацией в файл формата txt кодировки Unicode.
- ✓ интерфейс системы должен быть простым и доступным для пользователей любого уровня, содержать понятный набор инструментов.

Варианты

№ варианта	Направление перевода	Предметная область
1	Англо-русский	Научные статьи по computer science, Сочинения по литературе
2	Англо-русский	Научные статьи по computer science, Сочинения по литературе
3	Англо-немецкий	Научные статьи по computer science, Сочинения по литературе
4	Англо-немецкий	Научные статьи по computer science, Сочинения по литературе
5	Англо-французский	Научные статьи по computer science, Сочинения по литературе
6	Англо-французский	Научные статьи по computer science, Сочинения по литературе
7	Англо-русский	Научные статьи по медицине, Критика предметов изобразительного искусства
8	Англо-русский	Научные статьи по медицине, Критика предметов изобразительного искусства
9	Англо-немецкий	Научные статьи по медицине, Критика предметов изобразительного искусства
10	Англо-немецкий	Научные статьи по медицине, Критика предметов изобразительного искусства
11	Англо-французский	Научные статьи по медицине, Критика предметов изобразительного искусства
12	Англо-французский	Научные статьи по медицине, Критика предметов изобразительного искусства

В отчет по работе необходимо включить:

- ✓ описание структуры разработанной системы;
- ✓ основные алгоритмы реализации компонентов системы;
- ✓ результаты тестирования;
- ✓ результаты анализа полученных данных, и предложения по улучшению работы системы.