

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національна академія наук України
Національний авіаційний університет
Факультет комп'ютерних систем



МІЖНАРОДНА
НАУКОВО-ТЕХНІЧНА
КОНФЕРЕНЦІЯ
«ІНТЕЛЕКТУАЛЬНІ ТЕХНОЛОГІЇ
ЛІНГВІСТИЧНОГО АНАЛІЗУ»

26-27 жовтня 2010 року

Тези доповідей

Київ 2010

УДК 004.01(082)

Міжнародна наукова-технічна конференція «Інтелектуальні технології лінгвістичного аналізу»: Тези доповідей. – К.: НАУ, 2010. – 52 с.

Збірник містить тези доповідей, які були представлені на конференції «Інтелектуальні технології лінгвістичного аналізу».

В доповідях розглянуті наукові та методичні питання інтелектуальних технологій: методологія інтелектуальних мовно-інформаційних систем, комп'ютерна технологія порівняльного аналізу електронних текстів, технології інформаційного пошуку, методи системного моделювання. Для фахівців з комп'ютерної лінгвістики.

Редакційна колегія:

Литвиненко О.Є. – д.т.н., професор, декан факультету комп'ютерних систем НАУ

Широков В.А. – чл.-кор. НАН України, д.т.н., с.н.с., директор Українського мовно-інформаційного фонду НАН України

Денисюк В.П. – д.ф.-м.н., професор, завідувач кафедри вищої та обчислювальної математики факультету комп'ютерних систем НАУ

*Затверджено до друку вченою радою факультету комп'ютерних систем Національного авіаційного університету
(протокол № 13 від 08.11.2010 р.)*

© Національний авіаційний університет, 2010

ЗМІСТ

В.А. Широков, О.О. Сидоренко	
СЕМАНТИЧНІ СТАНИ В ДЕРЖАВНОМУ ТЕЗАУРУСІ УКРАЇНИ	8
В.А. Широков, чл.-кор. НАН України, д.т.н., с.н.с.....	9
ГІПЕРЛАНЦЮГИ В ЛЕКСИКОГРАФІЧНІЙ РЕПРЕЗЕНТАЦІЇ ЛЕКСИКО-СЕМАНТИЧНИХ ВІДНОШЕНЬ.....	9
Д.В. Ланде, д.т.н., с.н.с, В.В. Жигало.....	10
ЭТАПЫ СОЗДАНИЯ СТАТИСТИЧЕСКОГО ПЕРЕВОДЧИКА ПОТОКОВ НОВОСТЕЙ.....	10
К.А. Мацуєва, К.А. Мацуєва	11
ОБРАБОТКА ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ЗАПРОСОВ К ПОИСКОВОЙ МАШИНЕ НА ОСНОВЕ ИХ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА.....	11
О.О. Бєляков, О.О. Добріна.....	12
ЗАСТОСУВАННЯ ЛОГІКО-ЛІНГВІСТИЧНИХ МОДЕЛЕЙ У СИСТЕМАХ ОБРОБКИ ТЕКСТІВ	12
Ю.Н. Минаев, д.т.н., Е.В. Толстикова.....	13
ІНТЕЛЕКТУАЛЬНІ ТЕХНОЛОГІЇ ІДЕНТИФІКАЦІЇ АНОМАЛЬНИХ СТАНІВ КОМП'ЮТЕРНИХ МЕРЕЖ.....	13
І.М. Курченко, Є.О. Гончарова, Ю.О. Гончарова.....	14
РОЗПІЗНАННЯ ЛІТЕР АЛФАВІТУ НА ОСНОВІ МІНІМАЛЬНОЇ СИСТЕМИ ОЗНАК.....	14
В.І. Пустоваров, к.т.н., С.В. Кизима.....	15
СТВОРЕННЯ СЛОВНИКІВ ТА ДОВІДНИКІВ ДЛЯ НАКОПИЧЕННЯ ЗМІСТОВНОЇ ІНФОРМАЦІЇ В МУЛЬТИМЕДІЙНИХ СИСТЕМАХ	15
О.М. Романов	16
ПОБУДОВА МОДЕЛЕЙ НА ОСНОВІ ШТУЧНОГО ІНТЕЛЕКТУ .	16
А.І. Вавіленкова.....	17
АНАЛІЗ ОНТОЛОГІЇ ЯК МОДЕЛІ ПРЕДСТАВЛЕННЯ ЗНАНЬ ТЕКСТОВОЇ ІНФОРМАЦІЇ	17
А.С. Шевченко	18
ВИДИ ЗАПИТІВ У ВЕБ-СЕРЕДОВИЩІ.....	18
Б.Г. Масловський, к.т.н., Є.В. Тупота	19

УДК 004.012:82-995(045)

Д.В. Ланде, д.т.н., с.н.с, В.В. Жигало
Информационный центр «Электронные вести»

ЭТАПЫ СОЗДАНИЯ СТАТИСТИЧЕСКОГО ПЕРЕВОДЧИКА ПОТОКОВ НОВОСТЕЙ

При создании статистических русско-украинского и украинско-русского переводчиков решается ряд задач, имеющих важное значение для таких приложений, как автоматическое выявление опорных (ключевых) слов в документах, создание словарей опорных слов [1], выявление дубликатов документов (плагиата), представленных на различных языках, построение корпусов параллельных текстовых документов, предложений, n -грамм, и, наконец, создания автоматического переводчика.

В докладе описана методология создания «самообучаемого» статистического переводчика, ориентированного на массовый перевод текстовой информации из информационных потоков. Данная методология охватывает такие основные этапы: 1) создание корпуса параллельных документов; 2) создание корпуса параллельных предложений; 3) создание массивов параллельных n -грамм; 4) непосредственное создание модулей статистического переводчика.

При построении первичного параллельного текста корпуса авторами использовались лингво-статистические алгоритмы, применяемые к результатам контент-мониторинга сетевых СМИ [2]. При этом использовались электронные двуязычные словари. В результате был сформирован выровненный украинско-русский корпус объемом свыше 2,5 млн. пар предложений, доступный для работы в онлайн-режиме на сайте <http://ling.infostream.ua/>. Часть корпуса объемом 500 тысяч пар предложений, представленная в формате XML, общедоступна для некоммерческого использования.

Разработанное ПО сейчас находится в стадии комплексной отладки и тестирования. Подобный подход оказался эффективным для текстов на близких по статистическим параметрам языках.

Список литературы

1. *Потемкин С.Б., Кедрова Г.Е.* Выравнивание неразмеченного корпуса параллельных текстов // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог 2008». – Вып. 7 (14). – М.: РГГУ, 2008. – С. 431–436.

2. *Ланде Д.В., Жигало В.В.* Константы. Подход к созданию многоязычных параллельных корпусов веб-публикаций // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог 2009». – Вып. 8. – М.: РГГУ, 2009. – С. 278 – 283.