

MegaLing'2011

Горизонты прикладной лингвистики и
лингвистических технологий



ВЫРАВНИВАНИЕ ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ СЛОВАРЕЙ N-ГРАММ

Ландэ Дмитрий Владимирович^{1,2}, Дармохвал Александр Теодорович², Жигало Владлен
Викторович²

¹ИПРИ НАН Украины,

²Информационный центр «ЭЛВИСТИ»

Киев, Украина

dwl@visti.net, hval@visti.net, vladlen@visti.net

12 - 16 мая 2011 г.
Украина, Крым, Партенит

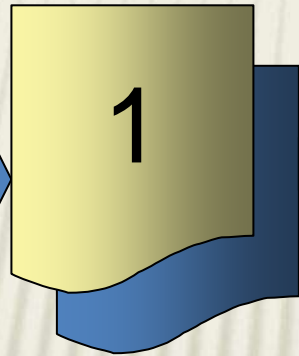


Три задачи – три этапа

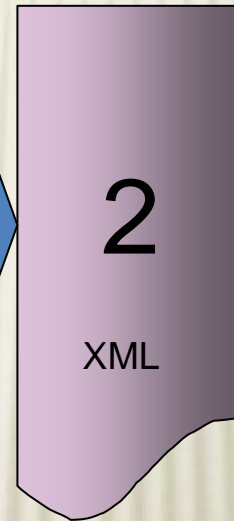
СОЗДАНИЯ СТАТИСТИЧЕСКОГО ПЕРЕВОДЧИКА ПОТОКОВ НОВОСТЕЙ



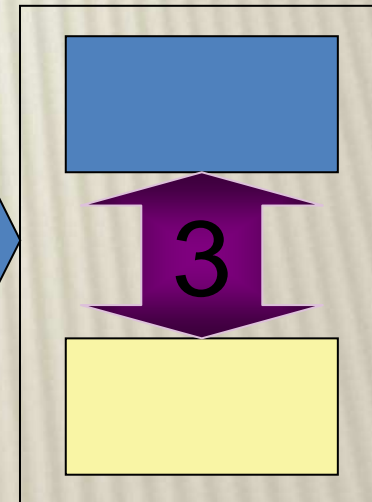
Информационный
поток



Параллельный
документальный
корпус



Параллельный
корпус
предложений



Статистический
поточный
переводчик



Несколько слов о технологии контент-мониторинга



InfoStream

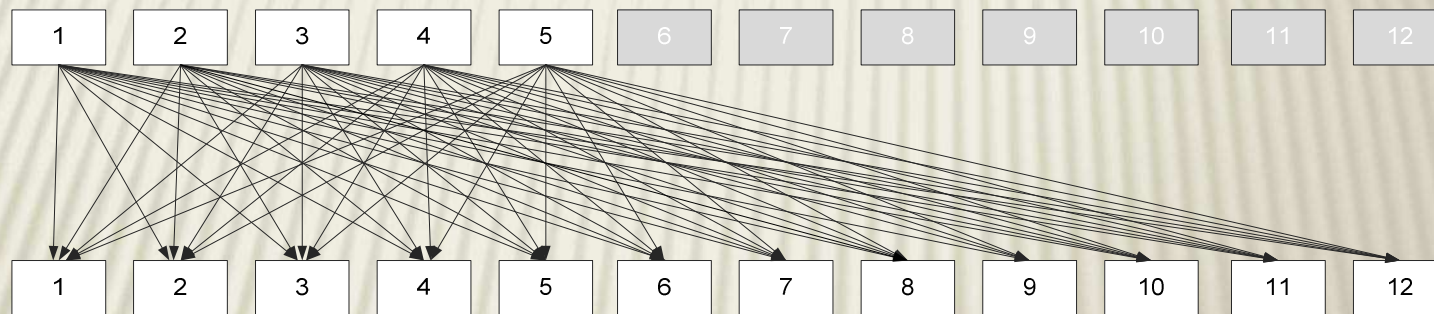
В Информационном центре "ЭЛВИСТИ" (Киев) создана система InfoStream, с помощью которой охватываются новости из более 5 тысяч отечественных и зарубежных веб-сайтов, осуществляется их обработка и обобщение.





Выявление дубликатов в InfoStream

В системе InfoStream используется механизм поиска дубликатов, в котором 5 опорных слов исследуемого документа, сравниваются с 12-ю опорными словами каждого из документов корпуса.



Процедура сравнения была дополнена рядом эвристических критериев, например:

- общее количество слов в переведенном варианте не должно отличаться от оригинала более чем на 10%;
- количество чисел в документах не должно отличаться больше чем на два.



Процедура формирования параллельного корпуса документов

1. Создание частотных морфологических словарей;
2. Выделение с их помощью опорных слов из документов;
3. Перевод опорных слов, с помощью словарей переводов;
4. Определение дублей документов на разных языках (сравнение 5-и переведенных опорных слов с 12 опорными словами др. документа);
5. Отсеивание с полученного множества документов «неполных дублей». Были использованы такие дополнительные критерии:
 - общее количество слов в переведенном варианте не должно отличаться больше чем на 10%;
 - количество слов начинающихся с большой буквы не должно отличаться больше чем на 3 слова;
 - количество чисел в документах не должно отличаться больше чем на два числа;
 - найденные числа в документах не должны отличаться более чем на 15 %.



Процедура формирования параллельного корпуса документов



Определение «опорных слов»

рус

укр

Морфологические частотные словари

Перевод «опорных слов»

рус / укр

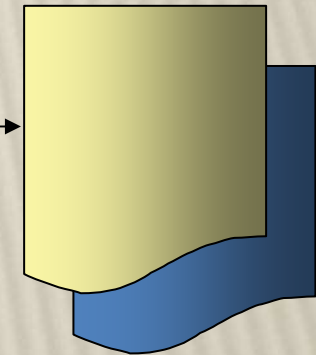
укр / рус

Словари переводов

Определение дубликатов на разных языках



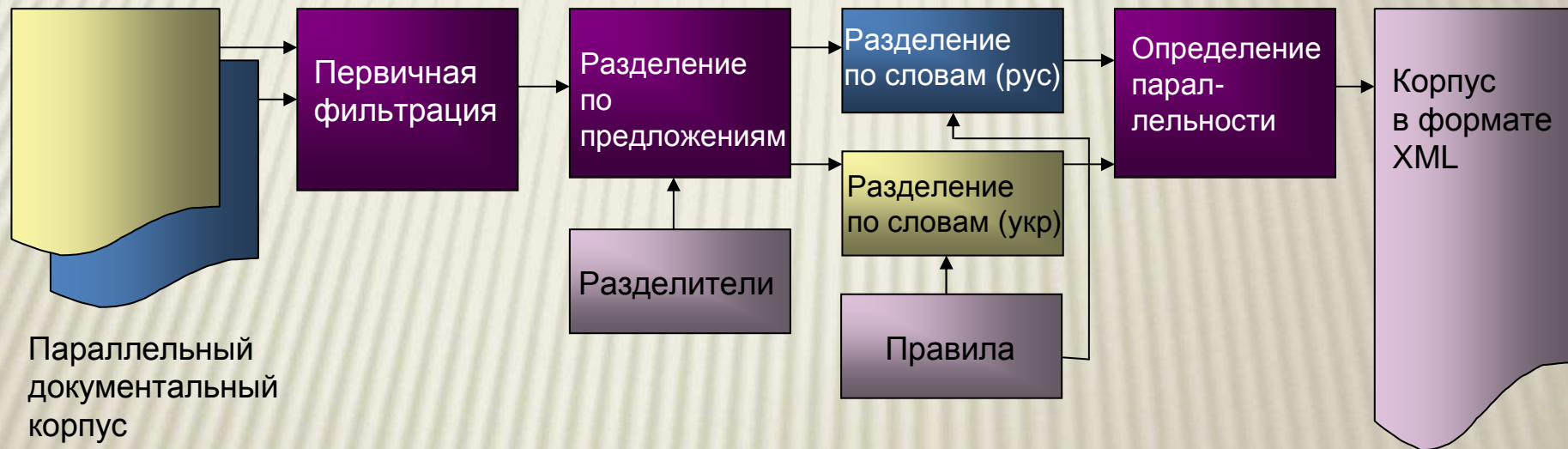
Окончательная фильтрация



Параллельный документальный корпус



Процедура формирования корпуса параллельных предложений





Алгоритм работы автоматического переводчика

1. Разделение документов на предложения
 2. Построение массивов триграмм, биграмм, слов
 3. Поиск триграмм, биграмм и слов в словарях
 4. Перевод предложений документа с использованием словарей n-грамм переводов и правил
 5. Форматирование документа
-



Пример русско-украинского перевода

InfoStrim Translate (<http://ling.infostream.ua/translate.php>)



Выберите язык оригинала текста: Определить автоматически Русский Украинский

Эдинбург готовит референдум об отделении от Соединенного Королевства

Идея полного государственного суверенитета Шотландии еще каких-нибудь пять лет назад казалась не более чем фантазией разгоряченного воображения националистов.

Сегодня невероятное все явственнее грозит стать очевидным. Прошедшие 5 мая выборы в Шотландии принесли победу той партии, которая с завидным упорством и с впечатляющей тактической мудростью доказывала все долгие последние годы своим избирателям, что светлое будущее шотландцев лежит вне брачного союза с Лондоном.

Еще в конце минувшего столетия Шотландская национальная партия (ШНП) была не просто «вечно оппозиционной», но своего рода политическим изгоем. Говорили, что

Перевести

Очистить

Единбург готує референдум про відділення від Сполученого Королівства

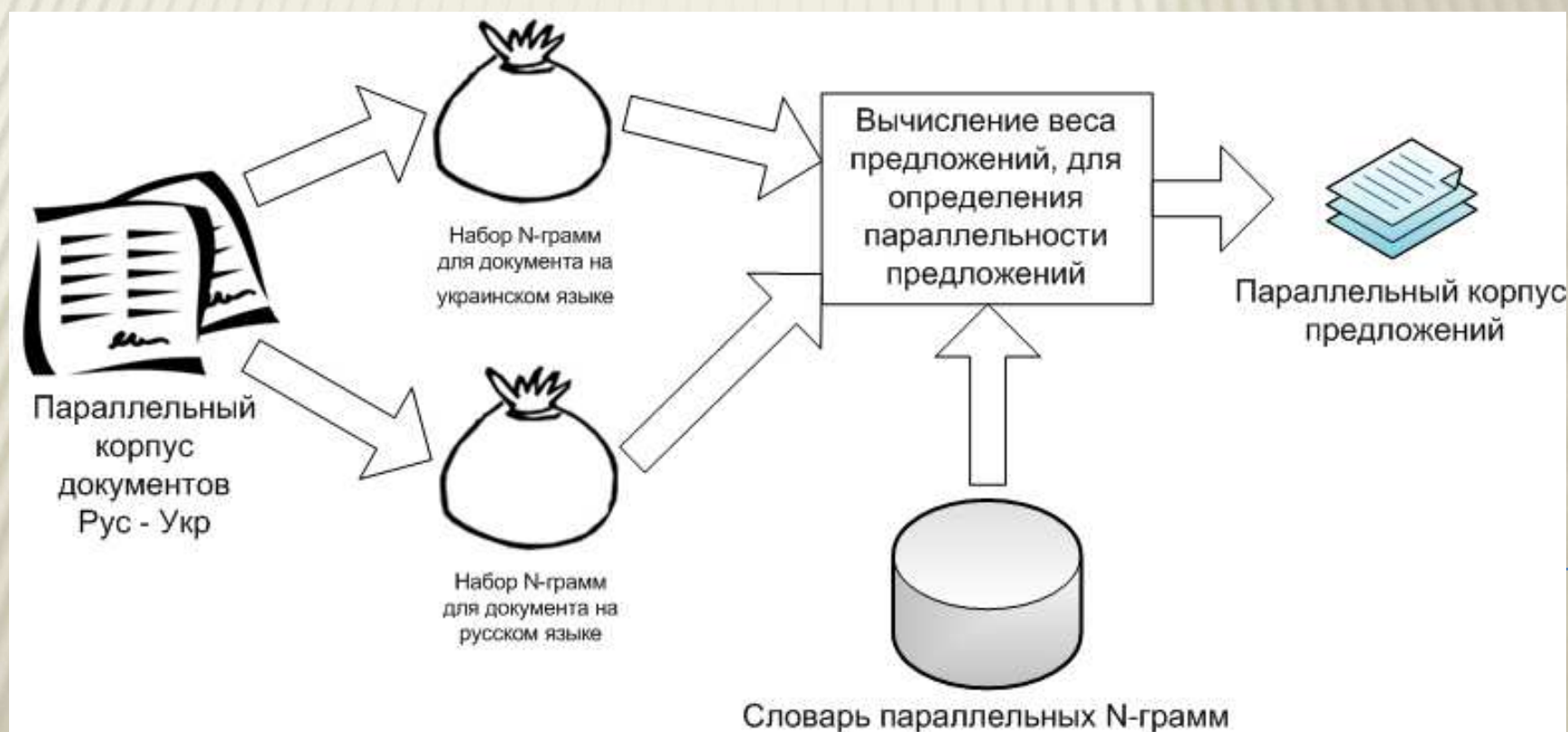
Ідея повного державного суверенітету Шотландії ще яких-небудь п'ять років тому здавалася не більше ніж фантазією розгоряченого уяви націоналістів.

Сьогодні неймовірне усе виразніше загрожує стати очевидним. Минулі 5 травня вибори в Шотландії принесли перемогу тієї партії, що з завидною завзятістю і з вражаючої тактичної мудрістю доводила всі довгі останні роки своїм виборцям, що світле майбутнє шотландців лежить поза шлюбного союзу з Лондоном.

Ще в кінці минулого століття Шотландська національна партія (ШНП) була не просто «вічно опозиційної», але свого роду політичним ізгоем. Говорили, що

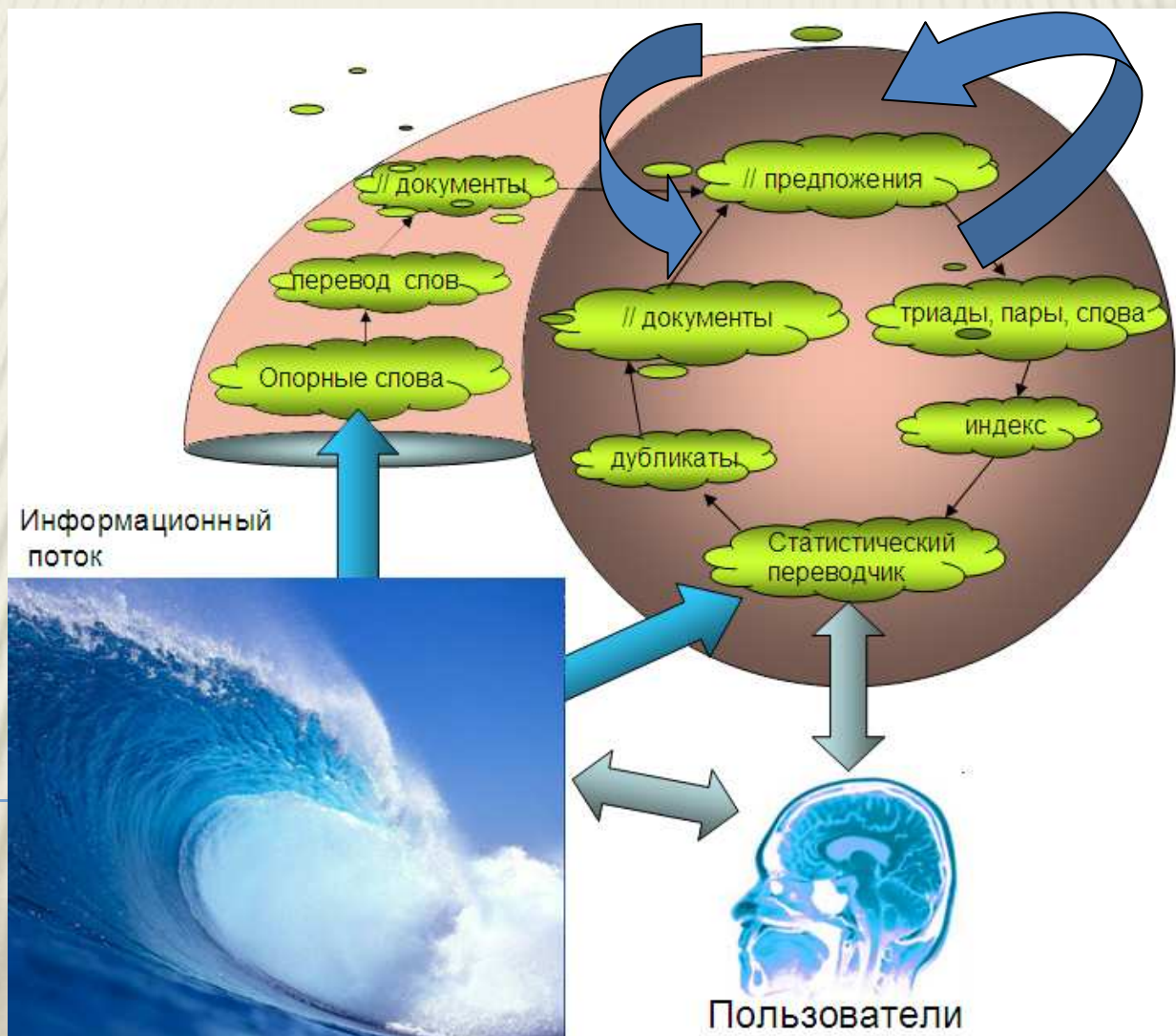


Формирование уточненного параллельного корпуса предложений





Место в технологии перевода ПОТОКОВ НОВОСТЕЙ



MegaLing'2011
Горизонты прикладной лингвистики и
лингвистических технологий



Спасибо за внимание!

Ландэ Д.В.

dwl@visti.net

12 - 16 мая 2011 г.
Украина, Крым, Партенит