MegaLing'2011
Горизонты прикладной лингвистики и лингвистических технологий

ВЫРАВНИВАНИЕ ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ СЛОВАРЕЙ N-ГРАММ

Ландэ Дмитрий Владимирович[1,2], Дармохвал Александр Теодорович[2], Жигало Владлен Викторович[2]

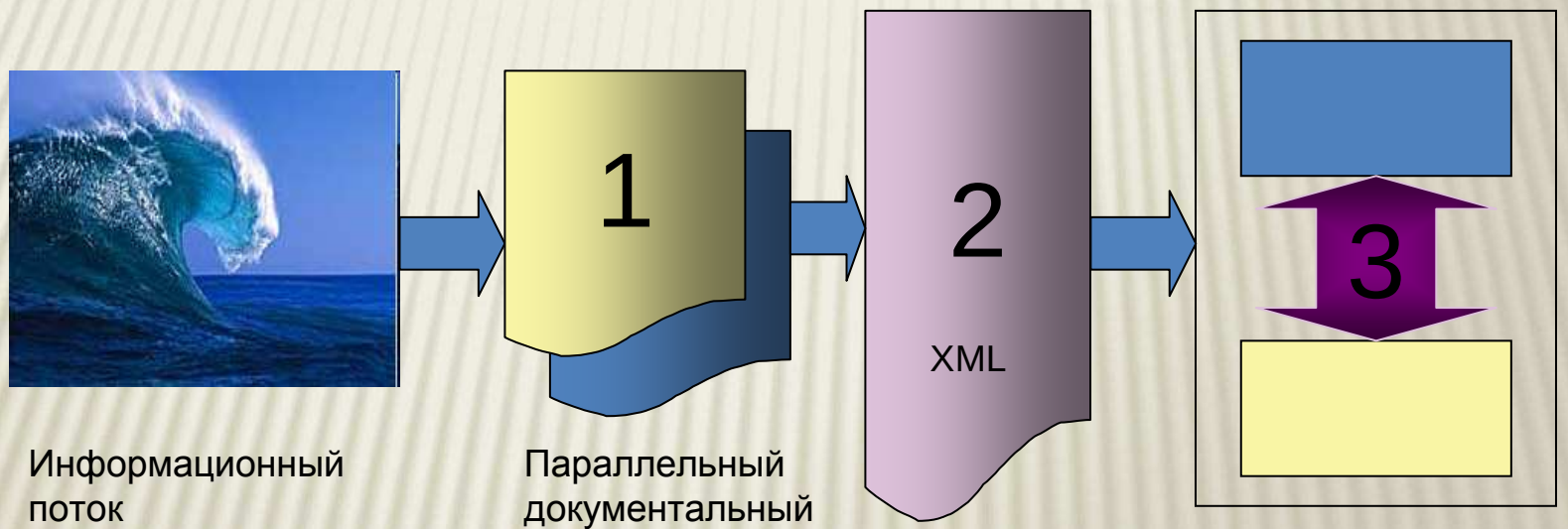
[1]ИПРИ НАН Украины,
[2]Информационный центр «ЭЛВИСТИ»
Киев, Украина

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 потоковый переводчик

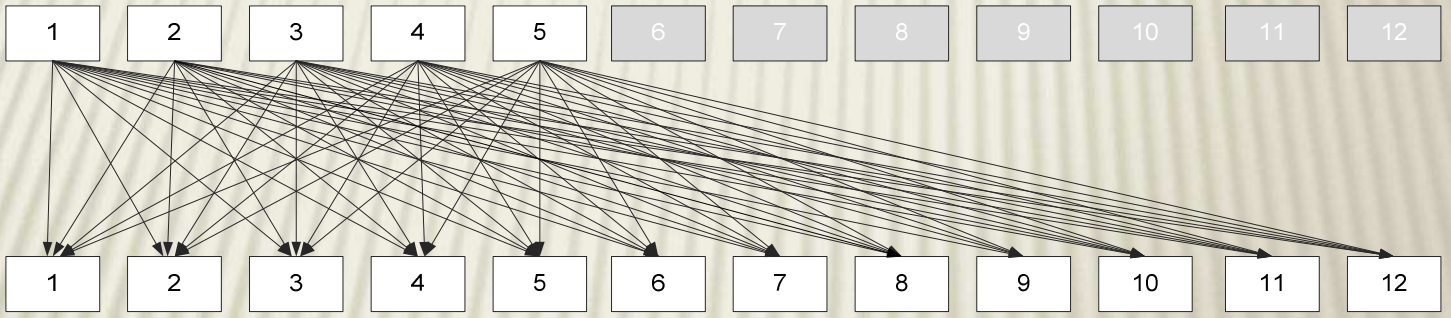
Несколько слов о технологии контент-мониторинга

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 отличатся 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 отличатся больше чем на два числа;
   - найденные числа в документах не должны отличаться более чем на 15 %.

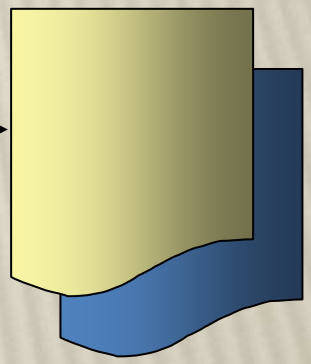
Процедура формирования параллельного корпуса документов

Определение «опорных слов»

Перевод «опорных слов»

Определение дубликатов на разных языках

Оконча-тельная фильтрация



рус

укр

Морфологические частотные словари

рус / укр

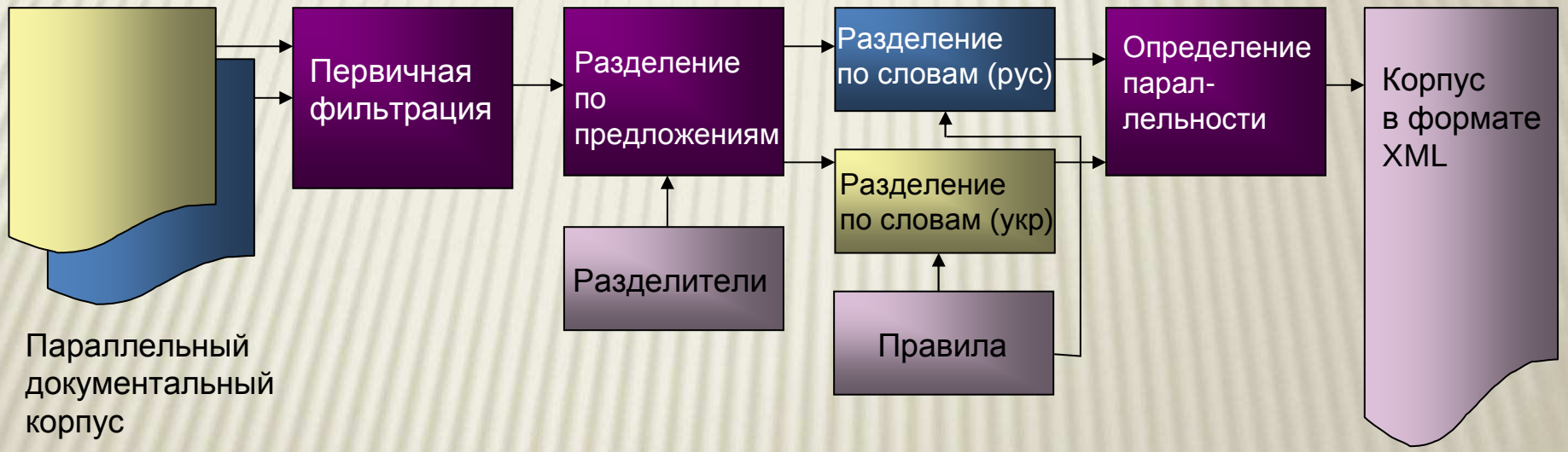
укр / рус

Словари переводов

InfoStream

Параллельный документальный корпус

Процедура формирования корпуса параллельнных предложений

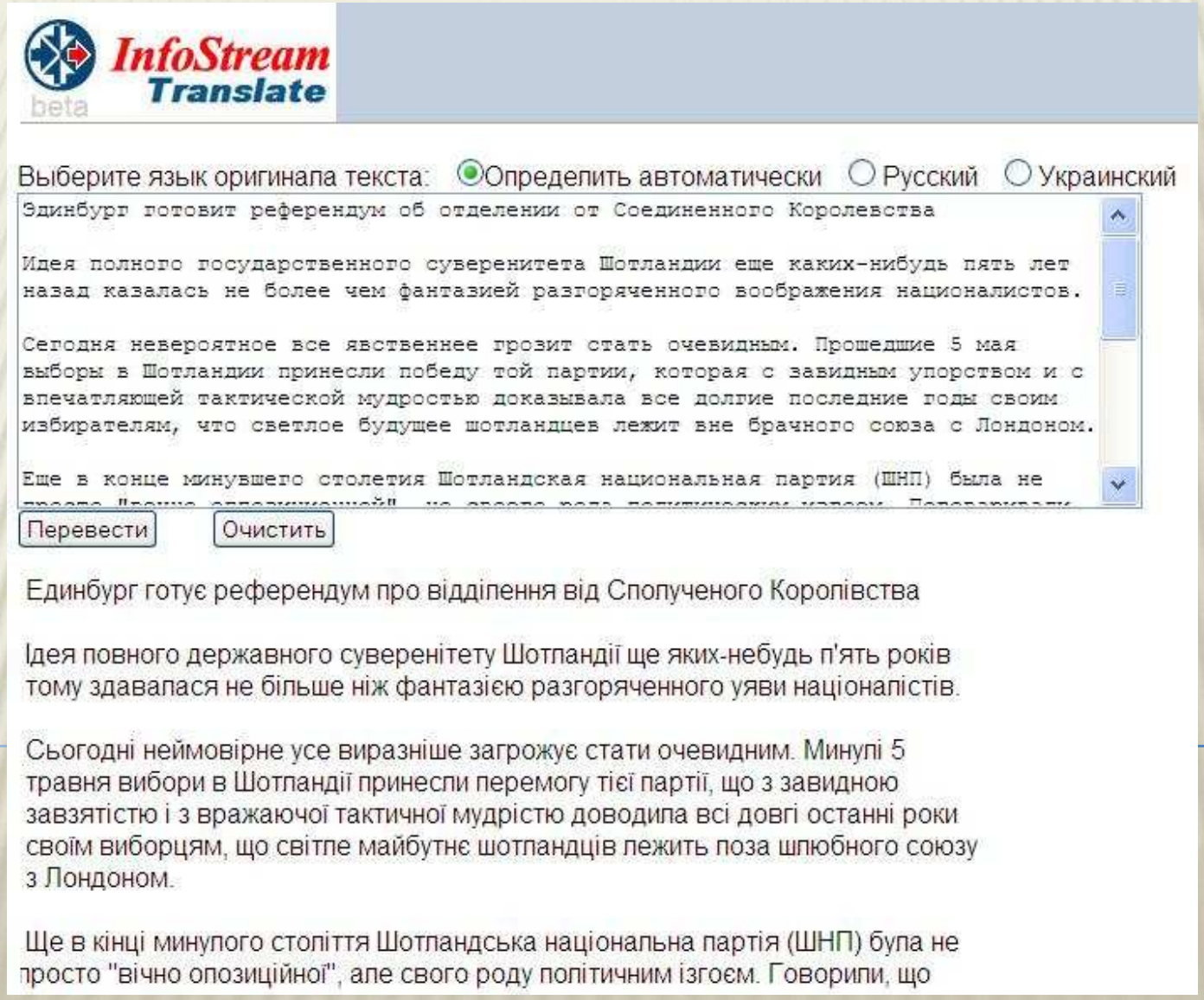


Алгоритм работы автоматического переводчика

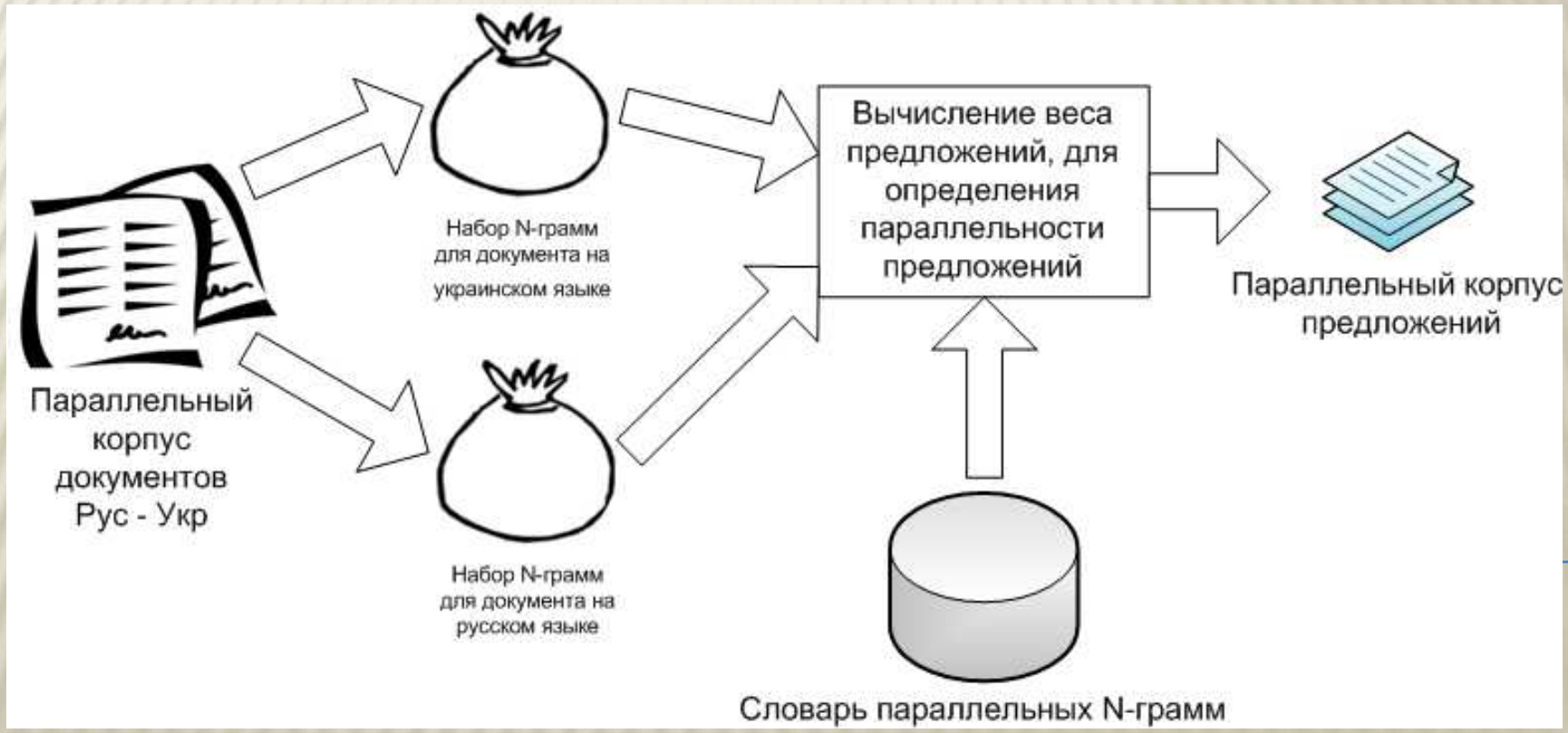
1. Разделение документов на предложения
2. Построение массивов триграмм, биграмм, слов
3. Поиск триграмм, биграмм и слов в словарях
4. Перевод предложений документа с использованием словарей n-грамм переводов и правил
5. Форматирование документа

Пример русско-украинского перевода

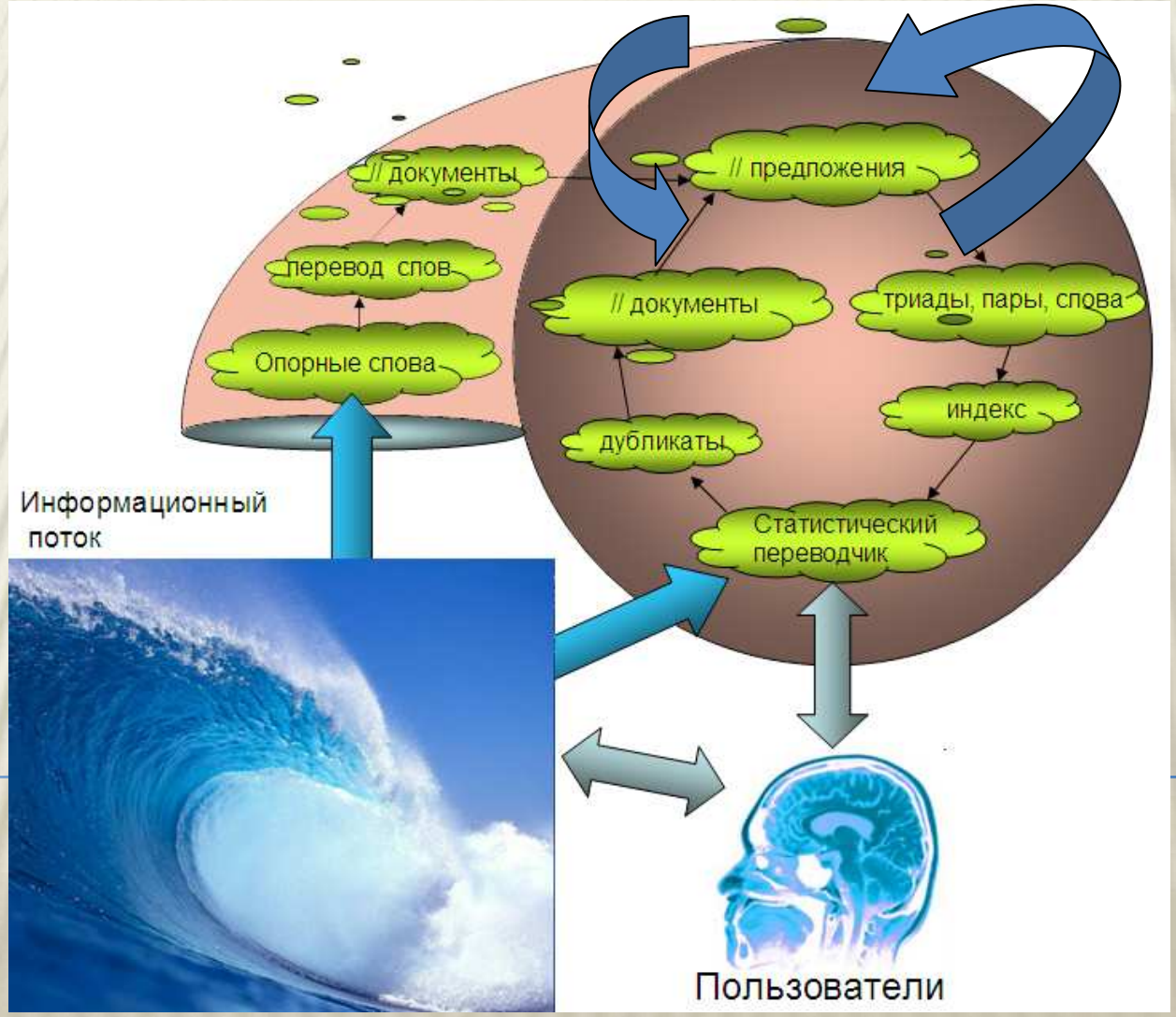
InfoStrim Translate (http://ling.infostream.ua/translate.php)



Формирование уточненного параллельного корпуса предложений



Место в технологии перевода потоков новостей



Спасибо за внимание!

Ландэ Д.В.

dwl@visti.net

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