ВЫРАВНИВАНИЕ УКРАИНСКО-РУССКОГО КОРПУСА ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ ИЗ СЕТЕВЫХ СМИ

***Ландэ** Дмитрий Владимирович, **Жигало** Владлен Викторович*

Информационный центр «ЭЛВИСТИ»
Киев, Украина
dwl@visti.net, vladlen@visti.net

Выравнивание параллельных текстов является важным этапом обработки для таких приложений, как машинный перевод, поиск по текстам на различных языках, составление словарей и т.д. [1]. В докладе описана методология выравнивания неразмеченного украинско-русского корпуса параллельных текстов на уровне предложений.

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, что позволило при достаточном уровне точности и полноты избежать дорогостоящей ручной обработки. Благодаря этому при выравнивании корпуса авторы смогли ограничиться преимущественно статистическими методами для получения удовлетворительных результатов.

Так как при построении выровненного корпуса во главу угла ставилась задача достижения точности (даже в ущерб полноте), метод выравнивания был дополнен фильтрацией аномально несовпадающих фрагментов текстов (было отсеяно около 14% пар предложений). Подобный подход оказался эффективным только для текстов на близких по статистическим параметрам языках.

В результате был сформирован выровненный украинско-русский корпус объемом свыше 5 млн. пар предложений, доступный для работы в онлайн-режиме на сайте http://ling.infostream.ua/. Часть корпуса объемом 500 тыс. пар предложений, представленная в формате XML, общедоступна для некоммерческого использования.

ЛИТЕРАТУРА

[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] Ландэ 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 (15). – М.: РГГУ, 2009. - С. 278-283.