ВЫРАВНИВАНИЯ ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ СЛОВАРЕЙ N-ГРАММ

***Ландэ** Дмитрий Владимирович, **Дармохвал** Александр Теодорович, **Жигало** Владлен Викторович*

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

В докладе описывается алгоритм первичного выравнивания параллельного русско-украинского корпуса документов [1] и метод улучшения результата работы этого алгоритма на базе использования словарей N-грамм.

Первичное выравнивание параллельного корпуса документов, на уровне предложений, выполняется на основе принципа монотонности. Предполагается, что различные языковые версии одного и того же документа содержат предложения, размещенные в одинаковом порядке. При построении параллельных выровненных документальных корпусов используются многочисленные эмпирические критерии, например, подсчитывается количество предложений, цифр, имен собственных, длины фрагментов текстов и т.п. Затем происходит построение параллельных словарей N-грамм.

Усовершенствование данного алгоритма заключается в том, что используя полученные ранее N-граммы на разных языках, производится повторное выравнивание параллельного корпуса документов, которые в свою очередь, выступают источником для параллельного корпуса предложений.

В результате работы алгоритма формируется выровненный параллельный корпус с большей точностью и полнотой.

ЛИТЕРАТУРА

1. Ландэ Д.В., Жигало В.В. Подход к созданию многоязычных параллельных корпусов веб-публикаций // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Междунар. конф. "Диалог 2009". – Вып. 8 (15). – М.: РГГУ, 2009. - С. 278-283.