Подход к созданию многоязычных параллельных корпусов веб-публикаций

The approach to creation of multilingual parallel corpora of web publications

Ландэ Д. В. (dwl@visti.net), **Жигало В. В.** (vladlen@visti.net)

Информационный центр «ЭЛВИСТИ», Киев, Украина

Описан метод построения двуязычного параллельного корпуса веб-публикаций, базирующийся на использовании частотных морфологических словарей, а также эмпирико-статистических алгоритмов. Предложен подход к преодолению омонимии в родственных флективных языках, позволяющий отбирать наиболее частотные нормальные формы. Алгоритм реализован в качестве программного комплекса и интегрирован в систему контент-мониторинга InfoStream. На основе предложенного метода был создан двуязычный русско-украинский параллельный корпус текстов веб-публикаций объемом свыше 450 000 пар документов.

Большое место в документальных информационно-поисковых системах занимают алгоритмы выделения ключевых слов, с помощью которых выполняются многие процедуры, охватываемые концепцией Text Mining, например, поиск подобных документов, выявление дубликатов, построение сниппетов, информационных портретов, дайджестов и т. д.

Заметим что проблема поиска подобных документов — одна из важнейших проблем современного информационного поиска, так как важные сообщения многократно дублируются. В данной статье описан метод, с помощью которого реализуется выявление информационных дубликатов, представленных на разных языках (русском и украинском). В результате применения этого метода авторами построен параллельный по информационному содержанию документальный корпус, который можно назвать «квазипараллельным», однако, он может также считаться параллельным в понимании [8], так как оснащен некоторыми автоматически сформированными тегами и переводами выделенных лексем на 2 языка. Выравнивание данного корпуса по предложениям или словам, а также морфологическая разметка корпуса отнесена к перспективам выполненной работы и выходит за рамки данной публикации.

На сегодняшний день существуют алгоритмы создания параллельных корпусов документов, которые можно условно разделить на две группы: традиционные и статистические.

К первой группе можно отнести алгоритмы, с помощью которых создавались такие параллельные корпусы, как Корпус CRATER [1]; Параллельный корпус переводов «Слова о полку Игореве» [2]; параллельный русско-английский корпус входящий в состав Национального корпуса русского языка [3]; параллельный русско-словацкий корпус[4] и т. д. Создание данных корпусов связано с тем, что исходные данные заведомо параллельные.

Ко второй группе можно отнести параллельные корпусы, созданные с помощью статистических алгоритмов, такие как [5–8], основанные на анализе страниц многоязычных веб-сайтов, объединении заранее подготовленных фрагментарных массивов и т. д.

Авторами предлагается новый подход к созданию параллельных корпусов документов, основанный на алгоритме поиска дубликатов документов на разных языках. Подход дает возможность отыскать похожие документы на разных языках в большом массиве документов. В результате можно убедится в том что в корпус попали параллельные документы из разных источников. Методы, основанные на анализе сайтов со страницами на разных языках, не позволяют определить дубликаты на разных источниках (сайтах), не указав специально параллельность этих источников. Традиционные же методы построения параллельных корпусов используют заведомо параллельные данные, что делает их в данном случае непригодными для использования.

Предложенный подход позволил создать двуязычный украинско-русский параллельный корпус текстов из веб-публикаций на русском и украинском языках объемом свыше 450 000 пар документов. Оцененная экспертами точность предложенного алгоритма составляет 98%.

Одной из основных проблем при автоматическом анализе текста является омонимия. Существующие подходы разрешения омонимии можно разделить на два основные типа: детерминированные и вероятностные. К детерминированным можно отнести методы, применяемые, например, в системе «ЭТАП» [9], где используется «фильтровый метод» синтаксического анализа, система «Диалинг» [10], или морфологический анализатор английского языка ENGTWOL [11], которые основаны на правилах снятия неоднозначности на основе контекстных правил. Вероятностный подход к преодолению омонимии широко обсуждался в работах российских исследователей [12–14], в применялся еще в 80-х годах XX века в системе М. Харста [15] для снятия неоднозначности у существительных путем использования размеченных вручную текстовых корпусов и выбора лексических и грамматических ключей.

Предложенный авторами подход к вычислению опорных слов документов (именно так будем обозначать ключевые слова, имея в виду, возможно, более узкую сферу применения) основаны на векторном представлении текста и используют статистические свойства текстов.

В данной работе описываются процедуры создания частотного словаря на основе морфологического словаря (МС) с использование тестового корпуса документов, построения алгоритма вычисления опорных слов с использованием частотного МС и модификации общеизвестного подхода TF IDF [16–13], а также статистического подхода к преодолению омонимии. На основе созданного алгоритма был построен программный комплекс, который интегрирован в систему контент-мониторинга InfoStream [17].

Реализован алгоритм построения параллельного корпуса документов, который учитывает не только статистические свойства текстов, но и некоторые морфологические признаки. В соответствии с этим алгоритмом построение параллельного корпуса происходит в несколько основных этапов:

- создание морфологических словарей;
- создание частотных морфологических словарей;
- создание словарей переводов;
- создание процедуры определения опорных слов в документах;
- определение разноязычных дубликатов.

Для русского и украинского языков были использованы свободно доступные электронные словари: ispell с набором более 1 млн. словоформ и «Словники України» [18], с набором более 4 млн. словоформ, а также словарь Зализняка, который насчитывает порядка 100 тыс. слов.

Эксперты дополнили морфологические словари неологизмами, названиями известных фирм, брендов и известными фамилиями, которых не было в исходных словарях.

Для обучения частотных морфологических словарей были взяты электронные публикации новостей, полученные из Интернет с помощью системы контент-мониторинга InfoStream. Количество публикаций составило 3 млн. 700 тыс. документов, 1 млн. 300 тыс. на украинском языке и около 2 млн. 400 тыс. на русском языке, за период с 01.01.2007 по 31.12.2007.

В соответствии с предложенным методом, «обучение» словарей проводится в несколько этапов. Первый этап заключается в разделении документов на словоформы и сохранении полученных словоформ с информацией о номерах соответствующих документов.

На втором этапе, созданный файл словоформ сортируется, после чего подсчитывается количество вхождений каждой словоформы, и количество документов в которых она встретилась. Найденные частоты записываются в частотный словарь, на основании которого определяется вероятная нормальная форма каждого слова.

Для выявления омонимии, в выходной файл записываются все нормальные формы соответствующие словоформе, т. е. если одной словоформе соответствует сразу несколько нормальных форм, сохраняются подсчитанные частоты со всеми найденными нормальными формами. На третьем этапе происходит заключительный подсчет количества нормальных форм и сохранение результатов в частотный словарь.

Для решения задачи построения параллельных текстовых корпусов в результирующие словари отбираются все словоформы имен существительных.

Описанный подход предусматривает использование алгоритма разрешения контекстной неоднозначности, так как омонимия является существенной проблемой при определении опорных слов документа, например, слово «села», которое в практике русского языка может быть множественным числом от слова «село», а также производной от глагола «садиться», может некорректно переводиться и использоваться на украинском языке, так как слово «село» переводится на украинский язык как «село», а слово «садиться» — «сідати». Неправильный выбор нормальной формы может привести к тому, что в одинаковых по информационному содержанию документах на разных языках будут использованы различные опорные слова. Для решения этой проблемы использовался, как оказалось позднее, эффективный и достаточно быстрый алгоритм, что особенно важно, так как этап обучения частотных словарей и этап их использования связаны с обработкой больших объемов текстовой информации.

В Табл. 1 показан пример обучения частотного словаря для слов «садиться» и «село». Предложено правило, в соответствии с которым, если в систему поступила словоформа, которая на практике может приводить к нескольким нормальным формам (например, для словоформы «села» допустимы нормальные формы «село» и «садиться»), то так называемые «индексы нормальных форм» для этой словоформы увеличиваются на единицу. В табл. 1 показан пример, когда в текстовом корпусе словоформа «села» встретилось 20 раз, словоформа «село» — 50 раз, словоформа «сели» — 10 раз, а словоформа «селом» — 30 раз. В результате обучения, в словари попадают слова «село» с индексом нормальной формы 100 и «садиться» с индексом 80, соответственно, в дальнейшем при отборе опорных слов предпочтение будет отдано слову «село».

В рамках данного исследования использовались словари переводов с русского на украинский и с украинского на русский язык. Данные словари были получены путем перевода наиболее частотных нормальных форм имен существительных с помощью бесплатных онлайн-словарей переводов в Интернет [19–21]. В случае, если одной словоформе соответствовало несколько переводов, то выбиралось наиболее употребляемые словоформы языка перевода в соответствии с частотным словарем. Полученный таким образом русско-украинский словарь насчитывал 80 тыс. наиболее частотных нормальных форм имен существительных, украино–русский — 90 тыс. наиболее частотных нормальных форм имен существительных.

Табл. 1. Пример обучения системы

Словоформа	Количество	Индекс нормальных форм
села	20	садиться → +20 село →+20
село	50	садиться → +50 село → +50
сели	10	садиться → +10
селом	30	село → +30
		село = 100 садиться = 80

Одним из эффективных подходов к выделению опорных слов из текста является векторная модель, в рамках которой, каждому слову документа присваивается его весовой коэффициент. Чем больше коэффициент слова, тем больше это слово характеризует документ. Для выявления опорных слов в тексте была использована модификация метода TF IDF — формула Okapi BM25 [22], которая в отличии от общепринятого подхода TF IDF позволяет учитывать среднюю длину документа в корпусе.

При использовании морфологических словарей предусмотрено, что отсеиваются все нормальные формы, соответствующие словам, находящихся в «стоп-словарях».

Для создания параллельного корпуса были взяты электронные публикации из Интернет, полученные с помощью системы InfoStream, за период с 01.01.96 по 28.02.2009, с общим количеством документов 60 млн., по всем политематическим источникам.

При реализации алгоритма происходит считывание текстового документа из входного потока, после чего выполняется выделение словоформ и поиск нормальной формы для каждой из них. В случае омонимии, выбирается наиболее частотная (с наибольшим индексом) по словарю нормальная форма словоформы. После вычисления соответствующих весовых коэффициентов с помощью формулы Okapi BM25 происходит ранжирование нормальных слов и выбирается двенадцать наиболее «весомых». Полученные двенадцать опорных слов переводятся на другой язык с помощью словарей переводов. Все опорные слова и слова-переводы приписываются к документу и выдаются в выходной поток.

Уже несколько лет в системе InfoStream используется механизм поиска дубликатов, который позволяет с помощью опорных слов находить подобные документы, представленные на одном языке. В этом механизме 6 опорных (наиболее весомых) слов исследуемого документа, сравниваются с 12-ю опорными словами каждого из документов корпуса вебпубликаций (рис. 1).

Именно таким же путем проводился поиск разноязычных дубликатов. Кроме того, данная процедура была дополнена рядом эвристических критериев, например:

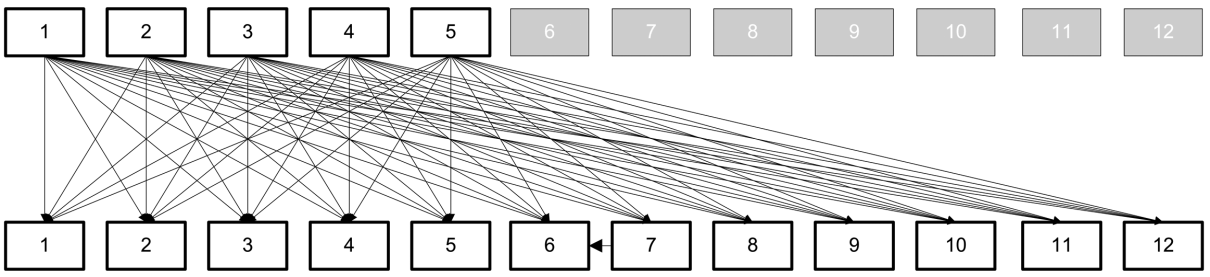

Рис. 1. Сравнение опорных слов

- 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.

Анализа полученных результатов проводился путем изучения экспертами случайных выборок документов, определенных как разноязычные дубликаты. «параллельных» документов. Проведенный таким образом анализ показал, что в среднем 98% содержания каждого документа имеют разные дополнения: например, ссылки на другое издательство, или же название издательства издавшего документ.

На базе системы InfoStream был разработан программный комплекс для работы с параллельным корпусом в поисковом режиме [23]. Данный программный комплекс позволяет производить поиск по корпусу документов как на русском так и на украинском языках, а также поддерживает одновременный вывод параллельных текстов, релевантных запросам пользователей. На рис. 3 приведен интерфейс, на котором представлены результаты поиска по «экономический кризис» (в результате было выбрано 157 параллельных текстов, релевантных данному запросу).

Для такого большого полученного корпуса возникает проблема ручной проверки, в таком случает было решено использовать метод случайной выборки документов, по которым эксперты смогли определить точность соответствия документов в 98%.

Был произведен детальный анализ корпуса параллельных документов и получены такие результаты:

Общее количество слов в корпусе составляет более 192,7 млн., из которых 96 млн. в украинских документах, 96.7 млн. — в русских документах.

Средняя длина документа в корпусе составляет 195 слов для украинского и 196 слов для русского.

Количество источников документов на украинском языке содержащихся в корпусе — 997. Количество источников документов на русском языке — 1768. Наиболее частотные источники приведены в Табл. 2.

На рис. 2 представлен пример вывода заголовков и аннотаций параллельных документов, содержащихся в корпусе, найденных по ключевым словам «экономический кризис». Полный текст пары параллельных документов приведен на рис. 3.

Указанный параллельный корпус расположен по адресу http://ling.infostream.ua и свободно доступен через поисковую систему. Корпус постоянно расширяется (по мере мониторинга новостей из Интернет) и в данный момент уже содержит более 450 тыс. пар документов на русском и украинском языках. Также выложен для скачивания и использования в научных и учебных целях параллельный корпус объемом около 30 тыс. пар документов.

Используя приведенный подход можно создавать не только русско-украинский параллельный корпус, но и, вероятно, подобные корпусы для любых языков входящих в славянскую группу языков. Авторами планируется построение корпуса параллельных украинско-английских, русско-английских корпусов и украино-русско-анлийских корпусов, однако, для перехода к работе с нефлективными языками необходим пересмотр некоторых из приведенных алгоритмов.

К перспективам данной работы также можно отнести:

- расширение разнообразия много языковых корпусов;
- расширение украинско-русского параллельного корпуса;
- совершенствование программной оболочки для просмотра параллельных корпусов, а также выравнивание данных корпусов по предложениям;
- создание автоматических переводчиков на основании построенных корпусов.

Табл. 2. Наиболее частотные источники

№ п.п.	Украино-язычные источники	Кол-во публикаций	Русскоязычные источники	Кол-во публикаций
1.	ForUm	33547	ForUm	30903
2.	УНІАН	31573	УНИАН	26509
3.	РБК-Екраїна	21517	УКРИНФОРМ	25838
4.	УТРО-Україна	20019	РБК-Украина	21849
5.	УКРІНФОРМ	19031	Корреспондент.net	21646
6.	Оглядач	18460	УТРО-Украина	19769
7.	ProUa	14090	ICTV	19719
8.	Кореспондент.net	13505	ProUa	15189
9.	Укроп	12346	Обозреватель	14844
10.	ГлавРед	8905	ГлавРед	10475
11.	Новинар	8159	NewsRu.ua	6284
12.	NewsRu.ua	7377	Форпост	5621
13.	УКРИНФОРМ	6518	Подробности	4204
14.	Форпост	6017	Київ-Прес-Інформ	3385
15.	Вголос	5535	Zaxid.net	3081

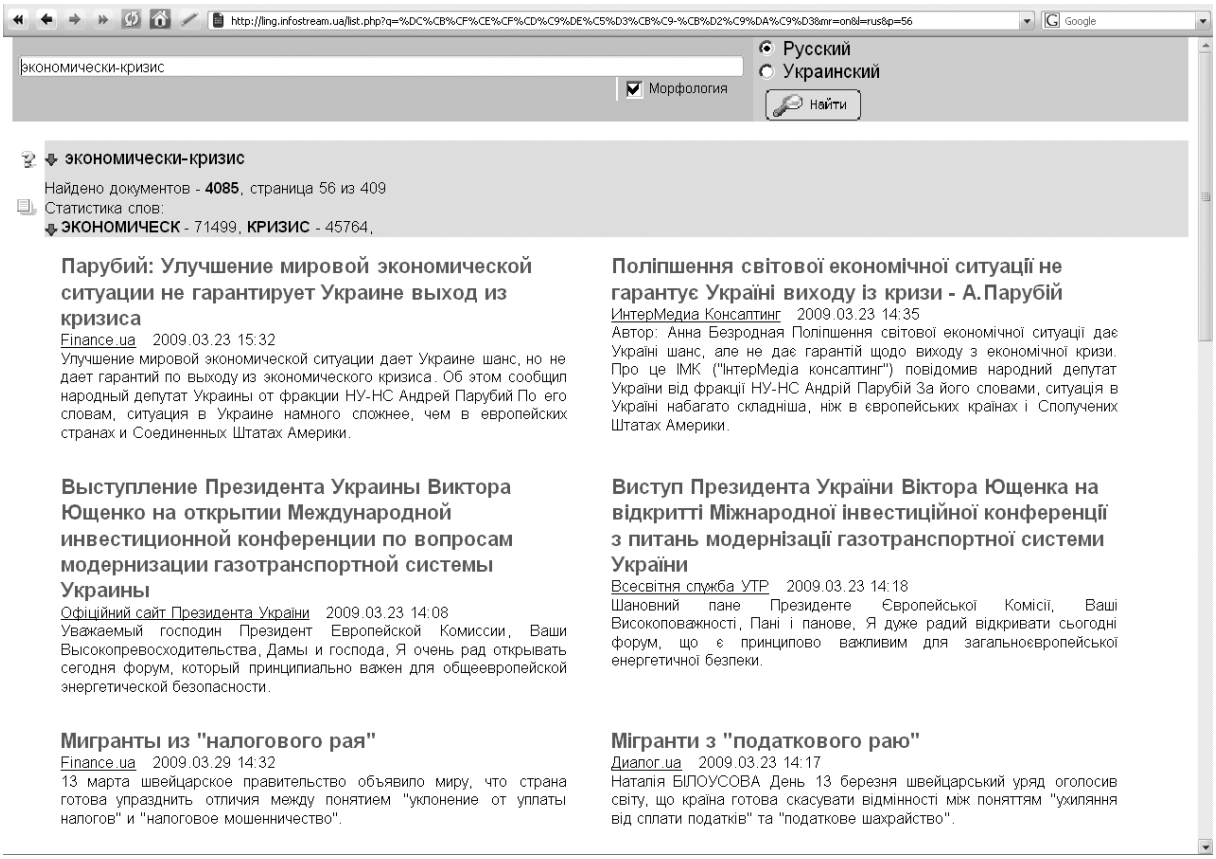

Рис. 2. Поисковый интерфейс для параллельного корпуса

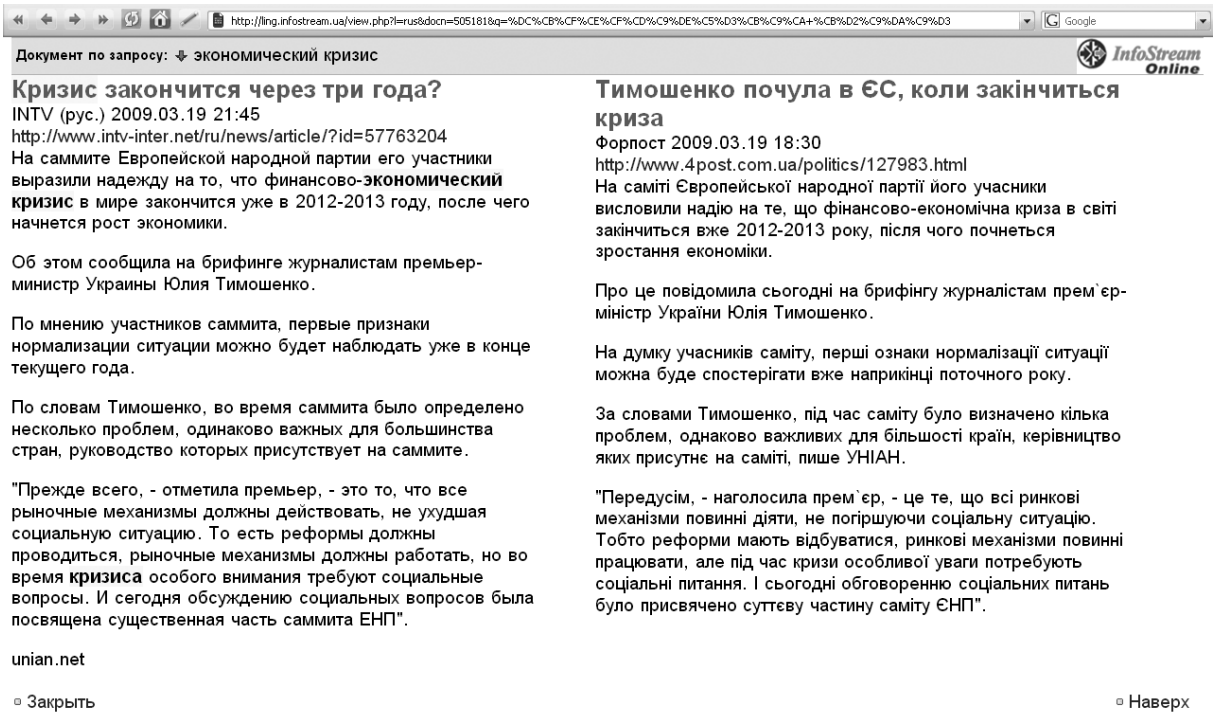

Рис. 3. Пример параллельных документов

Литература

1. *В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна.* Корпусна Лінгвістика — К.: Довіра, 2005. — 471 с.
2. http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html(сайт CRATER Multilingual Aligned Annotated Corpus)
3. http://nevmenandr.net/slovo/ (сайт Параллельного корпуса переводов «Слова о полку Игореве»).
4. http://www.ruscorpora.ru/corpora-biblio.html (сайт Национального корпуса русского языка).
5. *Гарабик Р., Захаров В.* Параллельный русско-словацкий корпус//Труды международной конференции Корпусная лингвистика — 2006. Sankt-Petersburg: St. Petersburg University Press 2006, s. 81–87.
6. *Resnik P.* Parallel strands: a preliminary investigation into mining the web for bilingual text. In D. Farwell, L. Gerber and E. Hovy (eds) Machine Translation and the Information Soup, Springer, Berlin, pp. 72–82.
7. *Resnik, P. and Smith, N. A.* 2003. The Web as a parallel corpus. Comput. Linguist. 29, 3 (Sep. 2003), pp. 349–380.
8. *Xiaoyi Ma, Mark Y. Liberman.* BITS: A Method for Bilingual Text Search over the Web// http://papers.ldc.upenn.edu/MTSVII1999/BITS.pdf
9. *Цинман Л. Л. , Сизов В. Г.* Лингвистический процессор ЭТАП: дескрипторное соответствие и обработка метафор // Труды межд. семинара Диалог‘2000. — М.: Изд-во РГГУ, 2000. — С. 366–369.
10. *Сокирко А. В., Ножов И. М.* Описание МаПоста // АОТ :: Технологии :: Описание МаПоста: http://www.aot.ru/docs/mapost.html (17 октября 2005 г.)
11. *Jurafsky D., Martin J. H.* Speech and Language Processing: An ntroduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall PTR, Upper Saddle River, NJ, 2000.
12. *Зеленков Ю. Г., Сегалович И. В., Титов В. А.* Вероятностная модель снятия морфологической омонимии на основе нормализубющих подстановок //Труды межд. конф. Диалог'2005. — М.: Наука, 2005.
13. *Баглей С. Г., Антонов А. В., Мешков В. С., Титов А. В.* Вероятностный подход к задаче разрешения омонимии слов и словарных пар // Труды межд. конф. Диалог'2007. 2007. С. 23–28.
14. *Зинькина Ю. В., Пяткин Н. В., Невзорова О. А.* Разрешение функциональной омонимии в русском языке на основе контекстных правил // Труды межд. конф. Диалог'2005. — М.: Наука, 2005. С. 198–202.
15. *Hearst M. A.* Noun homograph disambiguation using local context in large text corpora// Processing of the 7th conference on Research and Development in Information Retrieval ACM/SIGIR, pp. 36–47. — UW Centre for the New OED & Text Research Using Corpora, Pittsburgh, PA., 1991.
16. *Salton, G., Buckley, C.* Term-Weighting Approaches // Automatic Text Retrieval. Information Processing and Management, 24(5), pp. 513–523, 1988.
17. http://www.infostream.ua
18. http://www1.ulif.org.ua/ulif/
19. http://perevod.uaportal.com/
20. http://www.trident.com.ua/rus/online.php
21. http://translate.google.com/
22. http://www.xapian.org/docs/bm25.html
23. http://ling.infostream.ua