ПОСТРОЕНИЕ МАШИННО-ЧИТАЕМОГО СЛОВАРЯ НА ОСНОВЕ РУССКОГО ВИКИСЛОВАРЯ

КРИЖАНОВСКИЙ А.А.

УДК 004.912

Крижановский А.А. **Построение машинно-читаемого словаря на основе русского викисловаря.**

Аннотация. Сформулированы и решены практические вопросы извлечения данных из викисловаря, представляющего собой тезаурус и многофункциональный многоязычный словарь (только в русском викисловаре представлено более 300 языков). Для хранения лексикографической информации, извлеченной из русского викисловаря, разработаны структура базы данных машинно-читаемого словаря, а также интерфейс к этой базе данных который позволяет выводить на экран карточки словарных статей. В работе рассказывается о создании машинно-читаемого словаря на основе данных русского викисловаря.

Ключевые слова: машинно-читаемый словарь, лексикография, автоматическая обработка текста, вики.

Krizhanovsky A.A. **Constructing maching-readable dictionary based on Russian Wiktionary.**

Abstract. The practical questions of data extraction from Wiktionary are elaborated. Wiktionary is a multilingual free content dictionary (and in Russian Wiktionary there are more than 300 languages). In order to store the lexicographic data extracted from Russian Wiktionary (1) a database structure (tables and relations) was designed, (2) an application programming interface to this database was developed. The graphical user interface was implemented, which allows present the word-cards to the user. The paper is devoted of the creation of a machine-readable dictionary based on data from Russian Wiktionary.

Keywords: machine-readable dictionary, lexicography, information retrieval, wiki.

1. Введение. Викисловарь, как лингвистический ресурс — по многогранности и объему содержащихся в нем лексикографических данных является важным источником данных для систем автоматической обработки текста. Но только после преобразования текстов викисловаря в машинный формат, после создания программного интерфейса приложения (API) с богатой функциональностью, т. е. такого набора функций, который позволит выполнять разнообразные запросы к извлеченным структурированным данным.

Викисловарь по определению[1] является «многофункциональным многоязычным словарем и тезаурусом». При этом, чтобы оценить потенциал викисловаря, достаточно вспомнить успех «просто» тезауруса WordNet, который совсем не «многофункциональный» и не «много-

[1] См. данное определение по ссылке *http://ru.wiktionary.org*.

язычный», но при всех своих недостатках[2] активно применяется во многих приложениях. Машинно-читаемый словарь (MRD), построенный по данным викисловаря, востребован в следующих случаях:

— в поисковых системах;
— в системах сравнения онтологий (ontology matching);
— при распознавании запроса в запросно-ответных системах;
— при определение значения многозначного слова;
— при автоматическом создание тезаурусов;
— в машинном переводе;
— в компьютерных программах, помогающих в изучении иностранных языков.

Целью данной работы является создание такого машинно-читаемого словаря. Для этого потребуется уточнить, какие именно данные будут извлекаться из викисловаря. Также необходимо будет определить структуру для хранения извлеченных данных.

Среди ограниченного числа научных работ, посвященных такому новому направлению лексикографии как викисловарь, можно отметить работу [1], где описаны программные интерфейсы к электронной энциклопедии «Википедия» и к викисловарю (его английской и немецкой версиям) и статью [2], посвященную автоматическому улучшению сети синонимов в викисловаре.

2. Структура статьи в викисловаре и структура базы данных. Стандартная словарная статья викисловаря дает представление о слове с точки зрения таких разделов языкознания, как фонетика, орфография, морфология, синтаксис, семантика, этимология. На данный момент только часть этой информации извлекается и сохраняется в разработанной базе данных:

1) само слово (сохраняется в таблицу *page*, см. рисунок);

2) язык слова и его часть речи (таблицы *lang_pos*, *lang*, *part_of_speech*);

3) толкование (таблица *meaning*);

4) ссылки для ключевых слов в каком-либо из разделов словарной статьи: толкование, перевод, семантические отношения; т. е. ссылки в в «викифицированном»[3] тексте (таблицы *wiki_text*, *wiki_text_words*, *page_inflection* и *inflection*);

5) семантические отношения (таблицы *relation* и *relation_type*);

6) переводы (таблицы *translation* и *translation_entry*), при этом одна запись в таблице *translation* соответствует одному значению слова и

[2] См. *http://en.wikipedia.org/wiki/WordNet#Problems_and_Limitations*.
[3] См. *http://ru.wikipedia.org/wiki/Википедия:Викификация*.

одна запись в *translation_entry* соответствует переводу этого значения на один язык.

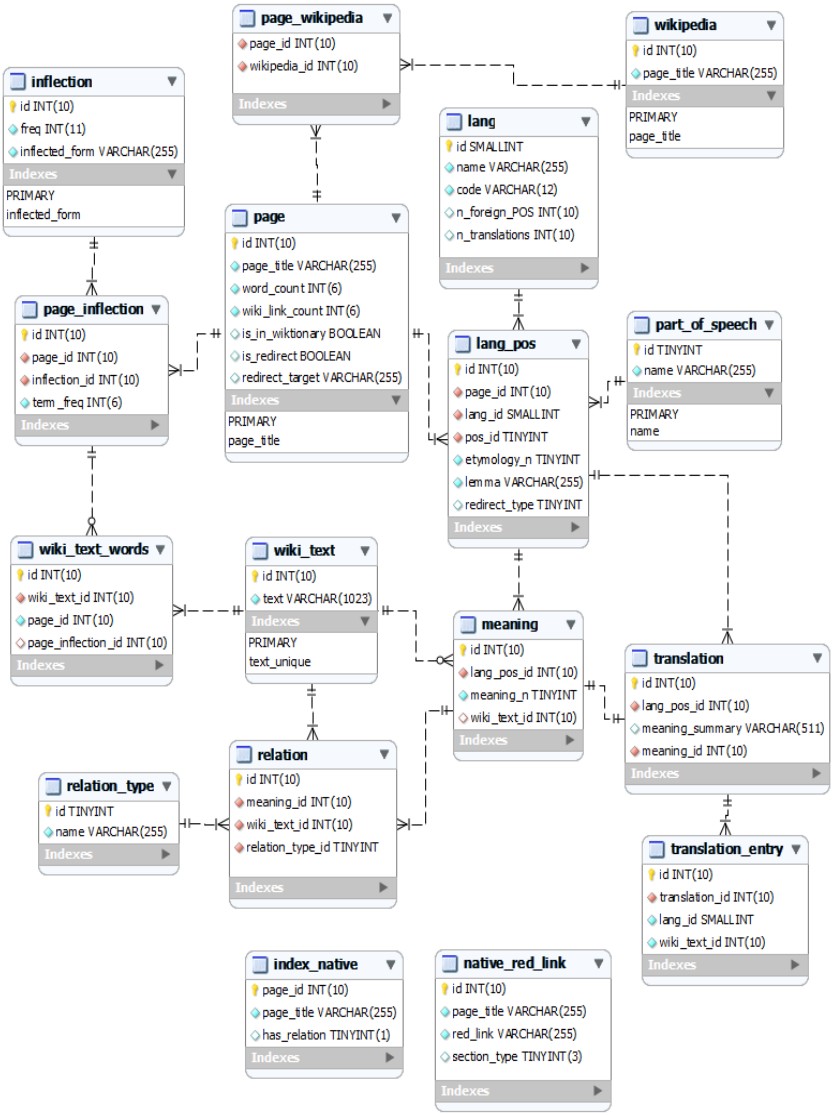

Рис. Таблицы и отношения в базе данных обработанного викисловаря.

3. Реализация. Функционально можно выделить две части разработанной компьютерной программы на языке Java: это парсер и API. Парсер обрабатывает базу данных (БД) русского викисловаря в формате MySQL, извлекает поля словарной статьи и сохраняет данные в БД обработанного викисловаря (рис.). API позволяет записывать, считывать данные и выполнять поиск в БД обработанного викисловаря. Программа и БД доступны на сайте проекта *http://code.google.com/p/wikokit*.

4. Заключение. Создан машинно-читаемый словарь на основе данных викисловаря. Разработана структура базы данных машинно-читаемого словаря. Реализована программа разбора русского викисловаря и программа для поиска и представления данных машинно-читаемого словаря.

Литература

1. *Zesch T., Mueller C., Gurevych I.* Extracting lexical semantic knowledge from Wikipedia and Wiktionary // Proc. of the Conf. on Language Resources and Evaluation (LREC), 2008.
2. *Navarro E., Sajous F., Gaume B., Prevot L. et al.* Wiktionary and NLP: Improving synonymy networks // Proc. of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources. Suntec, Singapore, 19–27 August 2009.

Крижановский Андрей Анатольевич — канд. техн. наук; старший научный сотрудник лаборатории интегрированных систем автоматизации Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: автоматическая обработка текста, корпусная лингвистика, вики. Число научных публикаций — 60. andrew.krizhanovsky@gmail.com, whinger.narod.ru; 14-я линия, д.39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-8071, факс +7(812)328-0685.

Krizhanovsky Andrew Anatoliyevich — Ph.D.; senior researcher at computer-aided integrated systems laboratory of Institution of the Russian Academy of Sciences St.Petersburg Institute for Informatics and Automation RAS (SPIIRAS). Research Interest: information retrieval, corpus linguistics, wiki. The number of scientific publications — 60. andrew.krizhanovsky@gmail.com, whinger.narod.ru; 14th Line V.O., 39, St.Petersburg, 199178, Russia; office phone +7(812)328-8071, fax +7(812)328-0685.

Поддержка исследований. Работа выполнена при финансовой поддержке РФФИ (проект № 08-07-00264), Президиума РАН (проект № 213).

Рекомендовано лабораторией интегрированных систем автоматизации, заведующий лабораторией А.В. Смирнов, д-р техн. наук, проф.
Статья поступила в редакцию 10.12.2009.

РЕФЕРАТ

Крижановский А.А. **Построение машинно-читаемого словаря на основе русского викисловаря.**

В виду большого количества словарных статей и разностороннему описанию слов (фонетика, орфография, морфология, синтаксис, семантика, этимология) викисловарь является важным лингвистическим ресурсом, например для таких задач, как: информационный поиск, сравнение онтологий, определение значения многозначных слов, проверка орфографии, автоматическое создание тезаурусов, машинный перевод и др.

В статье представлены практические вопросы извлечения данных из викисловаря, представляющего собой тезаурус и многофункциональный многоязычный словарь (только в русском викисловаре представлено более 300 языков).

Для хранения лексикографической информации, извлеченной из русского викисловаря, разработаны (1) структура базы данных машинно-читаемого словаря, (2) интерфейсы к этой базе данных.

Разработанный графический интерфейс позволяет выводить на экран карточки словарных статей. В работе рассказывается о создании машинно-читаемого словаря на основе данных русского викисловаря.

Необходимо отметить, что в данной работе не рассматривались другии языковые версии викисловарей, а только русский викисловарь, при этом только небольшая часть лексикографической информации была извлечена из текстов русского викисловаря: толкование, ссылки для ключевых слов, семантические отношения, перевод. Извлечение из викисловаря таких частей словарной статьи, как: произношение (фонетическая транскрипция, аудиофайл), разбиение на слоги, этимология, цитаты (примеры употреблений), параллельные тексты (цитаты с переводами), иллюстрация (фото или видео к значению слова) – здесь не рассматривается, поскольку это первый шаг в создании парсера викисловаря с открытым исходным кодом.

SUMMARY

Krizhanovsky A.A. **Constructing machine-readable dictionary based on Russian Wiktionary.**

Due to a big number of articles and many-sided word's description the Wiktionary is an important linguistics resource, e.g. for such tasks as information search, ontology alignment, word sense disambiguation, spell checking, machine translation, etc.

In this paper the practical questions of data extraction from Wiktionary are elaborated. Wiktionary is a multilingual, web-based project to create a free content dictionary, available in over 151 languages (and in Russian Wiktionary there are more than 300 languages).

In order to store the lexicographic data extracted from Russian Wiktionary (1) a database structure (tables and relations) was designed, (2) an application programming interface to this database was developed.

The structure of the developed database corresponds to the parts of the Wiktionary article. The application programming interface allows reading, writing and searching for data in this database.

The graphical user interface was implemented, which allows present the word-cards to the user. The paper is devoted of the creation of a machine-readable dictionary based on data from Russian Wiktionary.

It should be noted that (1) other language editions of Wiktionary are out of the scope of this paper, (2) only a small part of lexicographic information from Russian Wiktionary texts has been extracted and stored into machine readable dictionary. An extraction from Wiktionary of a pronunciation (phonetic transcription, a sound sample), a hyphenation, an etymology, a quotation (example sentence), a parallel text (examples with translations), a figure (which illustrates a word meaning) were not considered because this is a first step towards the creation of an open-source Wiktionary parser software.