WikiTailor

Technical Manual
Version 1.0

Cristina España-Bonet, Alberto Barrón-Cedeño and Josu Boldoba
WORKING DOCUMENT

February 18, 2016

Abstract

WIKITAILOR is a Java toolkit designed to extract and analise corpora from Wikipedia in any language and domain[1]. This document describes the prerequisites and the usage of the toolkit, the main characteristics of the systems involved in the extraction of the corpora, and some methods to evaluate the extractions.

Contents

1 Introduction

2 Prerequisites and Installation

This software is distributed as a stand-alone jar[2], so nothing else but a Java Virtual Machine is needed. The source code is also available[3] in which case, it can be compiled using the included Ant buildfile (see Section 2.3). In both cases a Wikipedia dump of the desired languages must be stored in a database to retrieve the articles off-line. Sections 2.1 and 2.2 describe how to download the data and create the database in a format compatible with WIKITAILOR.

2.1 Downloading the Dumps

In order to preprocess the Wikipedia articles a dump of the desired language(s) must be downloaded from the Wikimedia webpage:

```
http://dumps.wikimedia.org/[LAN]wiki/[DATE]
```

where, and from here on, `LAN` is a two-character language identifier with the ISO 639 codes[4] and `DATE` is the date of the dump in the `YEARMONTHDAY` format.

The dump consists of several sql tables. WIKITAILOR needs the general purpose tables:

```
[LAN]wiki-[DATE]-pages-articles.xml.bz2
[LAN]wiki-[DATE]-pagelinks.sql.gz
[LAN]wiki-[DATE]-categorylinks.sql.gz,
```

and three additional tables to relate articles in different languages and follow redirections:

```
[LAN]wiki-[DATE]-langlinks.sql.gz
[LAN]wiki-[DATE]-page.sql.gz
[LAN]wiki-[DATE]-redirect.sql.gz
```

The first group can be preprocessed and upload into a MySQL database using the JWPL DataMachine; the second group can be uploaded directly.

2.2 The JWPL DataMachine and the Database

JWPL (Java Wikipedia Library) is a Java-based application programming interface that allows to access all information in Wikipedia [?]. WIKITAILOR uses its DataMachine to create a MySQL database with a preprocessed Wikipedia dump and as a external library to parse the MediaWiki syntax.

The full process to deal with the dumps can be summarised in the following steps:

1. Download the JWPL DataMachine and the tables.sql file from
   ```
   https://dkpro.github.io/dkpro-jwpl/
   ```

2. Database setup. Create a MySQL database (use UTF-8 encoding) and afterwards the tables:
   ```
   CREATE DATABASE IF NOT EXISTS [DBname] CHARACTER SET utf8 \
   COLLATE utf8_general_ci

   mysql -u [USER] -p [DBname] < tables.sql
   ```

---

[2]`http://cristinae.github.io/WikiTailor/dwnld/wikiTailor-v1.0.0.with-dependencies.tar.gz`
[3]`https://github.com/cristinae/WikiTailor`
[4]`http://www.loc.gov/standards/iso639-2/php/code_list.php`. ISO 639-1 is used when available.

where a `DBname` in the format `[DBrootname]_wiki_[LAN]_[YEAR]` is expected from WIKITAILOR.

3. JWPL preprocessing. Run the transformation:

```
java -jar -Xmx4g datamachine.jar [LANGUAGE] \
[CATEGORY:MAIN_CATEGORY_NAME] \
[CATEGORY:DISAMBIGUATION_CATEGORY_NAME] [SOURCE_DIRECTORY]
```

where:
`LANGUAGE` is a string matching one in languages.txt in this release,
`MAIN_CATEGORY_NAME` is the name of the main (top) category of the Wikipedia category hierarchy[5],
`DISAMBIGUATION_CATEGORY_NAME` is the name of the category that contains the disambiguation categories[6] and,
`SOURCE_DIRECTORY` is the path to the directory containing the data dumps.

This should create a lot of new data files in a "output" subfolder of each input folder.

Mind that the names of the main category or the category marking disambiguation pages may change over time. E.g. the English category for disambiguation pages was called "Disambiguation" for a long time, while now it is "All_disambiguation_pages".

Examples:
```
java -jar -Xmx4g \
de.tudarmstadt.ukp.wikipedia.datamachine-1.0-jar-with-dependencies.jar\
english Category:Contents Category:All_disambiguation_pages ./en/
```

```
java -jar -Xmx2g \
de.tudarmstadt.ukp.wikipedia.datamachine-1.0-jar-with-dependencies.jar\
catalan Categoria:Principal Categoria:Pàgines_de_desambiguació ./ca/
```

```
java -jar -Xmx2g \
de.tudarmstadt.ukp.wikipedia.datamachine-1.0-jar-with-dependencies.jar\
arabic تصنيف:محتويات ويكيبيديا_ تصنيف:صفحات _توضيح ./ar/
```

4. Import the generated data files into de database.

```
mysqlimport -u[USER] -h [HOST] -p -local \
-default-character-set=utf8 [DBname] *.txt
```

A MySQL database can host the tables related to the interlanguage links, the langlinks. Just modify the names of the tables in the downloaded files to fit the WIKITAILOR nomenclature and upload them into the database.

1. Create the database:

---

[5]A complete list with the labels for the main category can be obtained in `https://www.wikidata.org/wiki/Q1281`.

[6]A complete list with the labels for the disambiguation pages can be obtained in `https://www.wikidata.org/wiki/Q1982926`.

```
CREATE DATABASE IF NOT EXISTS [DBrootname_pairs] \
CHARACTER SET utf8 COLLATE utf8_general_ci
```

2. Within the sql file, change the name of the table `langlinks` into `wiki[LAN]_[YEAR]_langlinks`. Do the equivalent modification for `page` and `redirect`.

3. Upload the tables:

```
mysql -u[USER] -p [DBrootname_pairs] < [LAN]wiki-[DATE]-langlinks.modified.sql
mysql -u[USER] -p [DBrootname_pairs] < [LAN]wiki-[DATE]-page.modified.sql
mysql -u[USER] -p [DBrootname_pairs] < [LAN]wiki-[DATE]-redirect.modified.sql
```

2.3 WikiTailor Installation

This subsection provides the basic instructions for installing WikiTailor, building the source code and obtaining the external libraries. All these steps work on a linux distribution. Be sure that you have Apache Ant and Java (version $>$ Java 7) installed, otherwise use your favourite package manager or do:

```
sudo apt-get install ant
sudo apt-get install openjdk-7-jdk
```

2.3.1 Building WikiTailor

First of all download the last version of the code. Clone WikiTailor from the repository:

```
git clone https://github.com/cristinae/WikiTailor
```

`cd` into the WikiTailor folder and run ant to build the source:

```
ant deploy-src
```

Run the basic tests to check everything is fine:

```
ant testNoDB
```

If no errors appear, you are ready to use WikiTailor. Clean the installation, `cd` into the distribution folder, and see the following section:

```
ant clean
cd WT-v$version
```

2.3.2 External dependencies

WikiTailor uses the following java libraries, all of them under open source licenses:

- apfloat-1.8.1

- apache-commons-cli-1.2

- apache-commons-collections-4.4.0

- icu4j-4.8.1.1

- apache-commons-io-2.4

- apache-commons-lang3-3.2.1

- apache-log4j-2.0

- apache-lucene

- jama-1.0.3

- junit-4.11

- jwpl-0.9.2

- apache-commons-math3-3.4.1

- opencsv-2.3

- apache-opennlp-1.5.3

- snowball-stemmer

- gnu-trove-3.0.3

For the current version, we have included the external libraries in the git repository itself, so when you build the source code the needed libraries are already added. Ahora hay el paquete light y el completo, solo se deberia compilar uno por default aunque demos los dos como binarios. Cual?

3   Description and Usage

```
user@machine:~/WT-v1.0.0/java -jar wikiTailor.v1.0.0.light.jar -h

usage: java -jar wikiTailor.v1.0.0.light.jar [-c <arg> | -n <arg>] [-d <arg>] [-e <arg>]
       [-h] -i <FILE> -l <arg> [-m <arg>]  [-o <arg>] [-s <arg>] [-t <arg>] -y <arg>


where the arguments are:
 -c,--category <arg>      Name of the category (with '_' instead of ' '; you can use -n
                          instead)
 -d,--depth <arg>         Depth obtained in a previous execution
                          (default: 0)
 -e,--end <arg>           Last step for the process
                          (default: 7)
 -h,--help                This help
 -i,--ini <FILE>          Global config file for WikiTailor
 -l,--language <arg>      Language of interest (e.g., en, es, ca)
 -m,--model <arg>         Percentage of in-domain categories
                          (default: 0.5)
 -n,--numcategory <arg>   Numerical identifier of the category (you can use -c instead)
 -o,--outpath <arg>       Save the output into this directory
                          (default: current)
 -s,--start <arg>         Initial step for the process
                          (default: 1)
 -t,--top <arg>           Number of vocabulary terms within the 10%
                          (default: 100, all: -1)
 -y,--year <arg>          Wikipedia year edition (2013, 2015, 2016)

Ex: java -jar wikiTailor.v1.0.0.light.jar -l en -y 2015 -i wikiTailor.ini -c Science
```

3.1 Extraction of Content Articles

3.2 xxx

3.2.1 Including a new Language

/home/cristinae/pln/workspace/lump2/lump2-aq-textextraction/HOWTO.txt

4 Evaluating the Extractions

5 Additional Utilities

6 Towards WikiParalel