Predicting the translations of red links

Olga Chernytska , Maksym Gontar , Kateryna Liubonko , and
Oleksandr Zaytsev

Ukrainian Catholic University
Faculty of Applied Sciences
Lviv, Ukraine
{chernytska,hontar,aloshkina,oleks}@ucu.edu.ua

Abstract

In this work we propose a technique for matching the red links of Wikipedia to the corresponding articles in another language. We build the graph of all Wikipedia pages and use it to measure similarity between the pages in different languages, one of which exists only as a red link. Such mathing would make the process of filling the gaps in Wikipedia more effective by allowing contributors to translate the existing page that corresponds to a red link rather than writing whole text from scratch.

1 Introduction

Wikipedia allows its users to create links to the pages that were not yet created. Such links become red and if you click on them, you get redirrected to an empty page and asked to write the missing article. Red links are an important source of information as they let us know what are the gaps in Wikipedia and how important it is to fill them. Every red link is the reference to a non-existing page and the number of such references can serve as a measure of interest that community of contributors has in a certain topic. Filling those gaps is a perfect way of making Wikipedia more complete and interconnected.

A major problem with red links is that because of the way they are stored in database they can not be linked to pages in other languages in a same way as existing pages are linked to their translations. This complicates the contribution process as it is much easier for people to translate and edit the existing page than to write a new one.

We propose a solution for finding the potential translation of the missing pages in other languages. It would allow Wikipedia to automatically match red links to existing pages and make contribution much easier.

2 Data collection and preparation

We have downloaded the full dumps of English[1] and Ukrainian Wikipedia[2] articles from 20/06/2018.

Then we have parsed those articles with regular expressions to get outgoing red and blue links: the link article, text and position in the current article text. There was red links in English Wikipedia and red links Ukrainian Wikipedia. There was blue links in English Wikipedia and blue links Ukrainian Wikipedia. This data is stored in enwiki-20180620-pages-links.csv and ukwiki-20180620-pages-links.csv files. Format of those files is following:
id - id of a page
link_id - id of a linked page
link_pos - position of link in a page markup text
link_pos_perc - relative position of link in a page markup text, range from 0 (at the beginning of page text) to 1 (at the end of page text)
link_val - title of a linked page
link_text - link text, if available
is_red_link - boolean, whether a link is red or not

Then we have downloaded Wiki interlanguage link records [3] and parsed out all interlingual links between En and Uk Wiki articles. There was 441928 pairs of Uk-En Wiki articles. This data is stored in 20180620-langlinks_uk_en.csv file. Format of this file is following:
id_uk - id of a page in Uk Wiki
id_en - id of a linked page in En Wiki

From dumps we collected data about pages aliases (redirects) in En and Uk Wiki. The alias page is the page user can come upon by searching the article not by it's original name, but by it's alias name, then user is redirected to the original page. Alias data is important for our task, since links may lead not to the original article, but to it's redirect page. This data is stored in set of files:

enwiki-20180620-id_alias_title_alias.csv, ukwiki-20180620-id_alias_title_alias.csv with format:
id_alias - id of alias page
title_alias - title of alias page

enwiki-20180620-id_alias_id_orig.csv, ukwiki-20180620-id_alias_id_orig.csv with

[1] `https://dumps.wikimedia.org/enwiki/20180620/`

[2] `https://dumps.wikimedia.org/ukwiki/20180620/`

[3] `https://dumps.wikimedia.org/ukwiki/20180620/ukwiki-20180620-langlinks.sql.gz`

format:
id_alias - id of alias article
id_orig - id of original article

enwiki-20180620-id_orig_title_alias.csv, ukwiki-20180620-id_orig_title_alias.csv
with format:
id_orig - id of original article
title_alias - title of alias pages

Also we composed a list of all pages in En and Uk Wiki. There were 5669865 in En and 581098 articles Uk Wiki. This data is stored in enwiki-20180620-id_name.csv, ukwiki-20180620-id_name.csv files. Format of those files is following:
id - id of a page
title - title of a page
length - length of a page markup text

Besides that, we did a statistical analysis for red links, we calculated how many times each red link was used and saved results in the enwiki-20180620-red_name_count.csv, ukwiki-20180620-red_name_count.csv files with format:
link_title - title of a red link
in_count - number of times it was used

Also we calculated how many red links there are with a certain number of use, and stored results into the enwiki-20180620-red_count_by_count.csv, ukwiki-20180620-red_count_by_count.csv files with format:
count - number of red link usage
in_count - number of this count case

For example, in En Wiki red link was used only once for 4354094, and twice for 811612 times.

For every eng red link in the matrix we calculate similarity (using similarity metrics selected on the previous step) to those ukr articles that do not have eng version. The most similar ukr article is the one that correeponds to this red link. Possibly, we will find several ukr articles with the same similarities, so add them as candidates for further preprocessing.

3 Baseline assumption

$$p(a)p(b)p(c) < p(a)p(b)$$