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

The goal of this project is to build the graph of all Wikipedia pages and learn from it to fing pages that correspond to the non-existing pages referenced by red links in other languages. This would allow to create the so-called "red pages" by translating the existing ones.

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. It is much easier for contributors to translate articles from other language than to write a new article from scratch.

We propose a solution of finding the potential translation of the missing page in other language. This would allow Wikipedia to automatically link red links to their translations in other languages and make the contribution process 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 to get outgoing red and blue links: the link article, text and position in the current article text. Also we have downloaded Wiki interlanguage link records[3] and parsed out all interlingual links between English and Ukrainian Wikipedia articles.

[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

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

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