Matching Red Links with Wikidata Items

Kateryna Liubonko[1] and Diego Sáez-Trumper[2]

[1] Ukrainian Catholic University, Lviv, Ukraine
aloshkina@ucu.edu.ua
[2] Wikimedia Foundation
diego@wikimedia.org

Abstract. This work is dedicated to Ukrainian and English editions of the Wikipedia encyclopedia network. The problem to solve is matching red links of a Ukrainian Wikipedia with Wikidata items that is to make Wikipedia graph more complete. To that aim, we apply an ensemble methodology, including graph-properties, and information retrieval approaches.

Keywords: Wikipedia, red link, information retrieval, graph

1 Introduction

Wikipedia can be considered as a huge network of articles and links between them. Its specifics is that Wikipedia is being constantly created and thus this network is never complete and quite disordered. One of the mechanisms that helps Wikipedia to grow is creating red links. Red links are links to the pages that do not exist (either not yet created or have been deleted). Red links are loosely connected with the other nodes of the Wikipedia network having only incoming links from the articles where these are mentioned.

1.1 Problem Statement

In fact, there is a big amount of red links, which can be corresponded to full articles in the same or in the other edition of Wikipedia. It leads to many inconveniences such as giving a reader of a Wikipedia article misleading information on the gap in Wikipedia or not sufficient information that the whole Wikipedia network contains. The process of creating new articles through the stage of red links have to be really optimized and fostered. The last problem is the one that we tackle in our work. If managed appropriately red links may be better encapsulated in the Wikipedia network and faster transformed to full articles. We try to reach it by finding the correspondent Wikidata items for red links. Our project also considers red links of Ukrainian Wikipedia with correspondence to English Wikipedia articles in particular.

We approach this problem as a Named Entity Resolution task for the reason that the majority of Wikipedia articles are about Named Entities. Thus, it is an NLP problem in the context of a graph. The research is going to be carried out on data, which has changed through time so that the results are better proved. The first data stamp is from

In: Proceedings of the 1st Masters Symposium on Advances in Data Mining, Machine Learning, and Computer Vision (MS-AMLV 2019), Lviv, Ukraine, November 15-16, 2019, pp. 115–124

Ukrainian and English Wikipedia editions of September 2018 and the second is of September 2019.

2 Related Work

2.1 Wikimedia Projects

To the best of our knowledge, there are no scientific publications focusing on matching red links to Wikidata items. Several projects were held by Wikipedia community but with no published peer-reviewed papers. Community efforts include the Red Link Recovery Wiki Project for English Wikipedia [1]. The community had been contributing to the project until 2017. The main goal of this project was to reduce the number of irrelevant red links. Red Link Wiki Project is of our interest because there was developed a tool Red Link Recovery Live to suggest alternative targets for red links. Although the targets were in the same Wikipedia edition, the methods used there can be applied to our task as well. Some of the techniques to evaluate this similarity for Red Link Recovery Live are the following:

- A weighted Levenshtein distance
- Names with alternate spellings
- Matching with titles transliterated (from originally non-Latin entities) using alternative systems (e.g. Pinyin, Wade-Giles)
- Matching with titles spelled with alternative rules (e.g. anti personnel / anti-personnel / antipersonnel)

There was also a project proposal in Wikimedia community, called “Filling red links with Wikidata” [2], whose intention was to make red links a part of a Wikipedia graph. The aim is similar to ours but it is related to the particular moment of creating a red link. Its idea is to create placeholder articles filled with data from Wikidata. This project proposal had a wide perspective not only connecting red links to Wikidata items but also automatically creating Wikipedia pages. Nevertheless, it was not implemented. The discussion on that project involved many questions on how to maintain and edit these new ’semi-articles’.

Furthermore, the suggestion by one of the Wikimedia users and contributors Maarten Dammers to connect a red link to a Wikidata item appeared in [3]. The main question was about the technical implementation of this idea (create a new property on Wikidata to store the name of the future article, hover of the link to get a hovercard in user’s favorite backup language etc.). As we see these suggestions are also related to the process of connecting red links to the appropriate Wikidata items when creating them. This idea was also not implemented.

The projects described above are all in the domain of the English Wikipedia edition. For the Ukrainian edition, the only thing that was found related to the red links problem is gathering lists of red links and combining them into topics. There is also a powerful tool called PetScan [4] that helps obtain information on red links with a user interface.

It is developed by Wikimedia Toolforge [5], which is a hosting environment for developers working on services that provide value to the Wikimedia movement. There we could find more information for our work.

2.2 BabelNet

A relevant work concerning our task of Named Entity Resolution is made by a research community from the Sapienza University of Rome. They implemented the BabelNet knowledge base [6], which serves as a multilingual encyclopedic dictionary and a semantic network. BabelNet is initially constructed on Wikipedia concepts and WordNet[1] database. The main idea behind it is that encoding knowledge in a structured way helps to solve different NLP tasks even better than statistical techniques. Now it contains data from 47 sources (OmegaWiki[2], VerbNet[3], GeoNames[4], Semcor[5] automatic translations, etc.). Its power is different for different languages as each one has a particular amount of supporting sources. The most powerful is obviously English. Nowadays, BabelNet contains knowledge bases for 284 languages. These knowledge bases include not only lexicalized items but also images. To show its powers general statistics on the last version of BabelNet and its main constituents are presented in Table 1.

Table 1. General Statistics of BabelNet 4.0

Languages	284
Babel synsets	15 780 364
Babel senses	808 974 108
Babel concepts	6 113 467
Named Entities	9 666 897
Images	54 229 458
Sources	47

BabelNet had not been applied to the red links problem in Wikipedia before but we assumed it as powerful to solve our problem. The closest for our project task, where the power of BabelNet was tested, is Multilingual All-Words Sense Disambiguation and Entity Linking in SemEval-2015 Task 13 [7]. Thanks for its content and structure, BabelNet showed high results for finding the correct translations for polysemic words, especially it worked well for nouns and noun phrases, which make the majority of Wikipedia titles.

The core concept of BabelNet is a ’synset’ which is deciphered as a synonym set. It is a set of synonyms in multiple languages for a word meaning. For example for a word

[1] https://wordnet.princeton.edu/

[2] http://www.omegawiki.org/Meta:Main_Page

[3] https://verbs.colorado.edu/verbnet/

[4] https://www.geonames.org/

[5] https://www.semcor.net/

'play' in the meaning of a dramatic work intended for performance by actors on a stage there is a multilingual synset <play, Theaterstück, dramma, obra, . . . , pièce de théâtre>. On the other hand for a word 'play' in the meaning of a contest with rules to determine a winner the synset is <game, jeu, Spiel, ..., juego>. For this reason, BabelNet can tackle the ambiguity problem.

For solving our task, we also checked other works on Wikipedia links. Of the most interest for us were [8, 9]. In [8], server logs of Wikipedia are used to predict which links are needed to make Wikipedia graph more complete. This can even be applied further in our work to rank red links by their importance. In [9], the authors propose several approaches to embed Wikipedia concepts and entities and evaluate their performance based on Concept Analogy and Concept Similarity tasks. Their main contribution is implementing non-ambiguous word embedding for Wikipedia concepts (here each Wikipedia page is regarded as a concept). These experiments on embedding Wikipedia pages gave us the understanding of embedding possibilities in terms of Wikipedia and some ideas for further research.

To sum up for now our review of the field, we state that red links in Ukrainian Wikipedia edition have attracted little attention. Work on reducing red links had been carried by a WikiProject 'Red Link Recovery' from 2005 to 2017 but it concentrated on finding existing articles for red links in the same edition. Project proposals concerning red link problems were made within Wikimedia community. Powerful tools are provided by Wikimedia Foundation, that are useful for information retrieval on Wikipedia items. A potent knowledge base BabelNet was developed that may solve matching red links to existent pages in Wikipedia but the tool was not yet applied to this particular problem. The work previously done can be used in our Master project either partially or as the ideas for further work and applications of our model within Wikipedia.

3 Results to Date

3.1 Experimental Data Collection and Pre-processing

The specifics of this work is that no prepared data and ground truth is available from the beginning. Thus, we have created it on our own using Wikipedia XML dumps (https://dumps.wikimedia.org) – a langlinks SQL dump and a Wikipedia pages network. Wikipedia XML dump is a Wikipedia database backup of a certain version (time) and a certain edition (language). Langlinks SQL dump contains Wikipedia inter-language link records. For now, the dumps we processed contain Wikipedia data of the version dated the 20th of September, 2018.

3.2 Data Retrieval and Pre-processing of the Whole Dataset

Our goal is to obtain red links of Ukrainian Wikipedia edition and all the corresponding information that would help solve our matching problem. Data retrieval and some parts of pre-processing have already been done using the work of our team in the Mining Massive Datasets course project at Ukrainian Catholic University in Summer 2018 [10].

The outstanding characteristic of the input data is its size. The size of English Wikipedia is 28.0 GB in compressed format. It contains 5 719 743 full English articles. Whereas Ukrainian Wikipedia edition dump size, which we took as an input, is 2.1 GB. It contains 817 892 Ukrainian Wikipedia articles of full size. The special approach was required to process this data on one computer. Mostly we split it into chunks and processed them one by one.

At first, we retrieved the whole Ukrainian Wikipedia graph of pages, the whole English Wikipedia graph of pages and language links between Ukrainian and English Wikipedia. From these datasets, red links were obtained and other supporting datasets were formed. Eventually, we obtained 2 443 148 red links in Ukrainian Wikipedia among which 1 554 986 were unique titles.

For further matching red links to Wikidata items, English Wikipedia data was processed. Thus, from English Wikipedia XML dump and langlinks SQL dump we retrieved all non-translated English articles, the correspondences between Ukrainian and English articles, and all the incoming links to non-translated articles in English Wikipedia. The number of articles not translated to Ukrainian in English Wikipedia is 5 264 607 which means that only 8 % of English Wikipedia is translated into Ukrainian. Vice versa, the number of langlinks between Ukrainian and English Wikipedia is 599 636 which is 73 % of all Ukrainian Wikipedia articles. Moreover, we kept links between all Ukrainian and English Wikipedia editions to use it further in our model. For English Wikipedia it is 161 017 765 links between pages and for Ukrainian Wikipedia – 22 693 778 links.

In Fig. 1, we can see the frequency of occurring red links in Ukrainian articles that have langlinks to English Wiki. If put into numbers, there are 1 010 955 red links which occur only once. The most frequent link is "ацетилювання". It occurs in 941 articles. The tendency here is not linear.

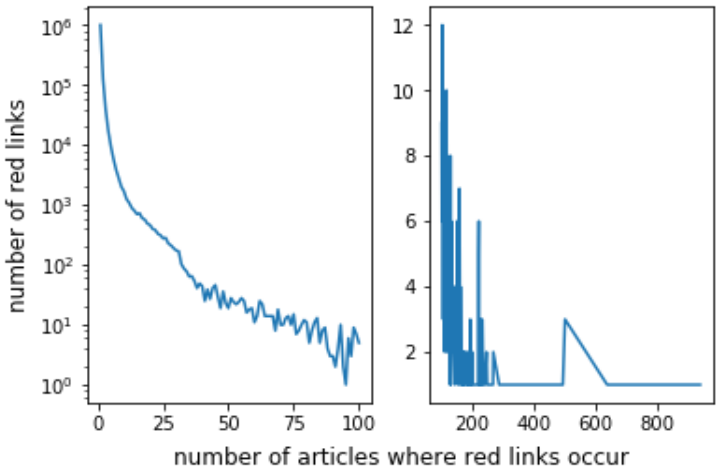

Fig. 1. Frequency of occurring red links in Ukrainian articles which have langlinks to English Wikipedia. *Left*: the number of red links in 0 to 100 articles (noted the log scale). *Right*: the number of red links in >100 articles.

3.3 Retrieving the Sample

For the reason that we cannot obtain the ground truth for such amount of red titles, we decided to work with samples in our project. Thus, we obtained a sample of 3 194 red titles which were in Ukrainian Wikipedia by the 20th of September, 2018. The sample was obtained by choosing red titles that occur in 20 or more articles that have corresponding articles in English Wikipedia (the langlinks mentioned above). Therefore, the chances to match a red link with an article from English Wikipedia are higher for this sample and, because of their popularity, these articles may be more needed in Wikipedia.

The characteristics of the obtained sample.

3 079 items of considered red links are Proper Names, which is 96 % of the sample. They include names of people, animal species (mostly moths), plant species, sport events, names of publishing houses, media sources, geographic locations and territories (mostly French regions), names of sport clubs, airports, administrative institutions, cinema awards, and a few other minor name categories.

Among these Proper Names, there are 969 red links for people's names, which is 30% of the sample. The biggest group of these names belong to tennis players.

Surprisingly, a great part of red links in Ukrainian Wikipedia (at least, as represented by our sample) are not in Ukrainian and many are spelled in other than Cyrillic script. The represented languages are English (e.g. 'John Wiley Sons'), Russian (e.g. 'Демографический энциклопедический словарь'), Latin (e.g. 'Idaea serpentata') and Japanese. Moreover, there are 989 red links spelled in Latin script, which is 31 % of the sample. Among these are red titles in English, Latin, and Ukrainian spelled in Latin script.

The data also has some innate characteristics that created obstacles for retrieval and pre-processing steps and which we had to take into account while building our model.

The first is double redirections – redirection pages that redirect to more redirections. For example a page 'Католицизм' redirected to 'Католицтво' which in turn redirected to 'Католицька церква' (The only full article here). Fortunately, these double redirections are constantly checked and cleaned by Wikipedia users or bots. By the time of writing, the redirections mentioned above were already removed and all of them redirected directly to the full article 'Католицька церква'.

The second type of noise in data is typos in red link titles. For example 'Панчакутек Юпанкі' is really 'Пачакутек Юпанкі', 'Сувалцьке воєводство' must be 'Сувальське воєводство'. It also goes for other mistakes in writing red links (e.g. 'Негрська раса' instead of 'Негроїдна раса'). The dangerous thing in this context is that articles for these red links really exist in the Ukrainian Wikipedia, but are not recognized because of the typos. Such 'false' red titles were revealed during the creation of the ground truth. Still, for our current model of matching red titles to Wikidata items they have no bad impact and do not influence the model. Nevertheless, this fact should be taken into account in further research.

3.4 Candidate Pairs Generation

This step is based on the work of our team in the Mining Massive Datasets course project at Ukrainian Catholic University in Summer, 2018 [10].

For each red link of our sample, we have retrieved a set of articles from English Wikipedia, which is more probable to contain an entity a red link refers to. Thus, it is called a candidate set and this phase in a pipeline is called a Candidate Entity Generation. Our approach on candidate generation is based on common links comparison. The chosen similarity measure was Jaccard score [11].

We could calculate Jaccard score similarity between red links and each of non-translated to Ukrainian English articles according to this formula:

$$S_{AB} = \frac{A \cap B}{A \cup B}, \quad (1)$$

where A is a set of incoming links for English non-translated articles and B is a set of incoming links for Ukrainian red links. Thus we obtain an array of tuples (score, candidate) for each red link. From these arrays, which correspond to red links, a table of red link candidate pairs is built. A part of this table is presented in Fig. 2.

	red_link	candidate
0	Pachetra sagittigera	Phytometra viridaria
1	Pachetra sagittigera	Conistra rubiginea
2	Pachetra sagittigera	Tholera decimalis
3	Pachetra sagittigera	Pachetra sagittigera
4	Pachetra sagittigera	Hoplodrina octogenaria
5	Pachetra sagittigera	Tiliacea aurago
6	Pachetra sagittigera	Apamea illyria
7	Pachetra sagittigera	Apamea lithoxylaea
8	Pachetra sagittigera	Apamea lateritia
9	Pachetra sagittigera	Actinotia polyodon

Fig. 2. Generated candidate pairs. Part.

The size of this set is 2 964 382 red link candidate pairs for 3 194 red links.

3.5 Creating Ground Truth

In the process of creating ground truth for the sample, we faced several other specific features of the dataset that made the evaluation more difficult. These are the following:

- Different names for one concept or person (e.g. 'Білозубкові' and 'Білозубки'). It also leads to the articles that already exist in the considered Wikipedia edition.
- Ambiguity. It is hard to find the right correspondence to a red title just by the name (e.g. 'Austin', 'Guilford', 'Йонас Свенссон'). In this context, it is often useful to point to a disambiguation page. Evidently, more information than just a title is required for matching.
- Red links which, by the time of checking for ground truth, already became full articles in the considered edition.
- Correspondences that were found by the time of checking for ground truth became deleted articles.

3.6 Metrics

Now having the set of candidate pairs and the ground truth we approached our problem as a ranking task, which are going to estimate with $F1$ score metric.

Thus, we concluded that methods to deal with a huge amount of data should be applied and one way is to take representative samples. Furthermore, due to Wikipedia nature and structure, people's mistakes, nature of the language itself and inner nature of the relations between Ukrainian and English Wikipedia editions the considered, data has some specific described characteristics that should be taken into account when building a model. The results above are yet going to be compared with the obtained in a near future data and statistics for Ukrainian red links of September 2019.

4 Goal

The main practical goal of the project is to bring more order and congruence to Wikipedia data by filling the red link gaps in its network. The concurrent aim is to understand the nature of red links (in Ukrainian and English Wikipedia editions in particular), the way they appear, and to sketch techniques for further creating red links in more consistent ways. We also aim to contribute to previous Named Entity Resolution models developed on Wikipedia data.

Therefore the main questions we want to answer in our project are:

- What is the nature of Wikipedia red links: quantitative and qualitative characteristics?
- What is a picture of Ukrainian Wiki red links in terms of English edition?
- What methods are more efficient to fill red link gaps in a Wiki Network?
- How to prevent further incoordination of red links in Wikimedia?

5 Methodology

In order to answer the above stated questions, we are working on Wikipedia dumps of September 2018 and September 2019 from Ukrainian and English Wikipedia editions. The modelling stage is preceded by a fine-grained analysis of red links in Ukrainian

Wikipedia with regard to English edition. The main features of analysis is that it is provided on samples because of the size of Wikipedia data and on data, which changed in a year to make more solid conclusions. Statistical inference from our data helps us choose the metrics, techniques, and tools for modelling. Having preliminary results of September 2018 Wikipedia data we have chosen to apply a supervised learning binary classification model and try a multi-factor approach considering graph, editing and embeddings information. In general, our further work consists on updating analysis and results on Wikipedia September 2019 dump, applying the chosen techniques to the latest data with improvements of models with regard to newly received knowledge.

6 Time Plan

The Master thesis pipeline is built in an iterative mode so that each monthly stage concludes with a prepared paper text and code.

September

1. Review of the modelling and writing on results of September 2018 data
2. Write and submit abstract and article to MS-AMLV-2019

October

1. Read related papers on Named Entity Recognition task
2. Retrieve data and create sample of September 2019 Ukrainian and English Wikipedia dumps
3. Write new sections on relevant to this work results in the field and on retrieval result of updated data
4. Clean the code and submit it to Github

November

1. Refine the final version of the paper to MS-AMLV-2019, submit and prepare an oral report
2. Read related papers on embedding techniques for Named Entity Recognition
3. Process data and make statistics on samples
4. Make comparative analysis of Ukrainian red links of Wikipedia September2018 edition and a year after data. Write new sections on the received results and conclusions; refine the paper text based on recommendations from MS-AMLV-2019.
5. Clean the code and submit it to Github

December

1. Apply chosen independent and ensemble methods to samples and check the results of models
2. Write the final sections to the Master thesis paper and refine the whole work
3. Clean the code and make final submissions to Github

7 Conclusions

With this Master project proposal, we are the first to state the problem of matching red links with items in another Wikipedia edition. We are also the first to begin solving this problem in the context of Ukrainian red links. For that, we created a dataset of Ukrainian red links and candidate pages from English Wikipedia. Then we applied BabelNet knowledge base to solve red links for Ukrainian Wikipedia. In the context of our project, BabelNet was regarded as a baseline. Next, we made a thorough data analysis. This helped us define the methodology to solve the problem as an Entity Resolution task. Finally, we presented a time plan for further research.

References

1. Wikipedia: WikiProject Red Link Recovery/RLRL. https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Red_Link_Recovery/RLRL
2. Filling red links with Wikidata, Wikimedia Meta-Wiki. https://meta.wikimedia.org/wiki/Filling_red_links_with_Wikidata .
3. Wiki-research-l Digest, Vol. 157, Issue 19. https://lists.wikimedia.org/pipermail/wiki-research-l/2018-September/006439.html
4. PetScan tool for Wikimedia. https://petscan.wmflabs.org/
5. Wikimedia Toolforge for developers. https://tools.wmflabs.org/
6. BabelNet 4.0 and Live Version. https://babelnet.org/
7. Moro, A., Navigli, R.: SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In: 9th International Workshop on Semantic Evaluation, pp. 288–297 (2015)
8. Paranjape, A., West, R., Zia, L., Leskovec, J.: Improving website hyperlink structure using server logs. In: 9th ACM International Conference on Web Search and Data Mining, pp. 615–624. ACM (2016)
9. Sherkat, E., Milios, E.E.: Vector embedding of Wikipedia concepts and entities. In: Frasincar, F., Ittoo, A., Nguyen, L., Métais E. (eds.) Natural Language Processing and Information Systems. NLDB 2017. LNCS, vol. 10260, pp. 418–428. Springer, Cham (2017)
10. Final project for the Mining Massive Datasets course at the Ukrainian Catholic University, 2018. https://github.com/olekscode/Power2TheWiki
11. Kosub, S: A note on the triangle inequality for the Jaccard distance. Pattern Recognition Letters 120, 36–38 (2019)