Identifying Bilingual Topics in Wikipedia for Efficient Parallel Corpus Extraction and Building Domain-Specific Glossaries for the Japanese-English Language Pair

Bartholomäus Wloka
University of Vienna
Centre for Translation Studies
bartholomaeus.wloka@univie.at.at

Abstract

This paper presents an approach and its implementation as a software toolset for examining what portion of the multilingual content of Wikipedia is viable for harvesting bilingual data in order to build parallel corpora and domain-specific glossaries. An algorithm is presented which analyzes the link topology of topics and subtopics and the co-occurance in another language. This algorithm is implemented in the Python language and can be used to examine an arbitrary number of topics for Japanese-English as well as other language pairs with minor adjustements. The goal of the toolchain is ease of use and transparency as well as flexibility towards language combinations. The findings of a test with several thousands topics is presented as a showcase. The toolchain is open source under the Creative Commons license.

Keywords: automatic language resource harvesting, parallel corpora, data retrieval, wikipedia harvesting, multilingual comparison

1. Introduction

Wikipedia, as commonly known, is a conglomeration of articles written independently by people from all over the world. It is an extensive collection of knowledge represented in many languages. It is obvious that the content of these articles across these languages would vary in structure and semantics depending on the point of view of the particular country or region. However, there are also pages which are directly translated by groups of people who dedicate their time to make the articles consistent and to close gaps which might occur with certain topics across languages. Furthermore, some articles are written independently, but the information is often derived from articles in languages, where the content is more comprehensive. Often the English Wikipedia serves as a pivot for this purpose, given the status of English as lingua franca, and the fact that on Wikipedia English is represented more than any other language, as seen in Fig. 1. A probably even more representative quantifier for the overall activity in a given language, and perhaps an indicator for the quality of the content is the count of active contributors, i.e. contributors that have edited at least one article in the past month (Fig. 2). Here we see even more clearly how dominant the English Wikipedia is in terms of community contribution. Due to the activity of English, the English-Japanese language pair was chosen for this showcase.

The following chapters elaborate on related works (Chapter 2.), the approach of comparing multilingual content on Wikipedia articles in Chapter 3., the algorithm of comparison in Chapter 4., and its implementation in Chapter 5.. Chapter 6. presents a showcase of a comparison. Finally, a conclusion is presented in Chapter 7. as well as a description of future work plans in Chapter 8..

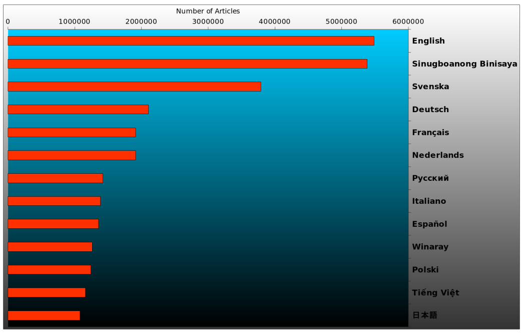


Figure 1: Count of Wikipedia articles by language

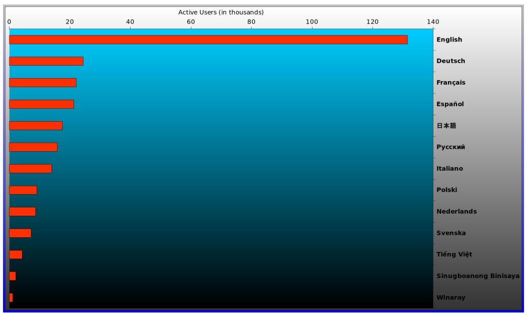


Figure 2: Count of active contributors for the largest Wikipedias

2. Related Work

Multilingual language resource extraction from Wikipedia is becoming increasingly popular, due to a consistent structure of Wikipedia articles. Comparable corpora have been built for many language pairs (Otero and López, 2010; Reese et al., 2010), especially for resource-scarce languages. However, the automatic generation of parallel, i.e.

sentence aligned corpora, across different languages is a very difficult task and requires evaluation and assessment of multilingual correspondence between articles (Paramita et al., 2012). (Ljubesic et al., 2016) shows that crawling text from Wikipedia results in acceptable results, measured by BLEU (Papineni et al., 2002). An approach to crawl depending on a domain is shown in (Labaka et al., 2016), by a breadth first search starting from a certain root-article and advancing via comparison with a domain specific dictionary and the assumption that occurrence of words from this dictionary signify that it belongs in this domain. These and other numerous works show promising results, however, it seems that their approach focuses mainly on few language combinations.This paper attempts to complement these results by a software toolchain which computes straight forward co-occurrence between articles, while not relying on machine translation or dictionary lookup, but on the structured architecture of topic IDs (Vrandecic and Krötsch, 2014), hence providing ease of use and flexibility regarding the language combination.

3. Comparing Wikipedia Articles

The goal of multilingual Wikipedia topic comparison is to find out whether the corresponding articles content can be regarded as a good translation. However, it is not feasible to compare each sentence of every article, especially if many language combinations are to be examined. Due to an exponential computational complexity, dictionary lookups, and the fact that multiple millions of tokens have to be processed, these methods have limits regarding multilingual flexibility. Furthermore, the task of automatic translation quality assessment of entire texts is in itself a computationally intensive process, since it often includes *machine translation* and/or *word sense disambiguation* as well as the use of external language resources such as lexical resources, terminology-bases, etc. The method described in this paper suggests a comparison, which selects articles by identifying the subtopics, hence omitting the problem of huge amounts of data. It uses the Wikipedia *pageID* property to identify articles of the same topic in different languages, which helps to avoid the usage of dictionary lookups, thus eliminating the problem of ambiguity and eliminates the need for additional language resources.

The process begins with choosing a topic and selecting its article page. This Wikipedia page is analyzed for all of the topic links mentioned in this article. Next, all the topics which were found in the first step are analyzed in the same way resulting in a collection of tuples of topics and subtopics. This process is repeated for the second language. Finally, the two lists of tuples are compared to each other and the co-occurring topics/subtopics are counted. The articles with a high percentage of co-occurrences indicate potentially similar content and indicate candidates for bilingual language harvesting and parallel corpora creation.

In addition to identifying parallel corpora candidate pages the algorithm finds the equivalents terms for each topic, which results in a term glossary in a certain domain, depending on the starting point, i.e. the initial topic.

4. Comparison Algorithm

The algorithm presented in this paper starts with harvesting topic links within one article page. Wikipedia article pages are well structured and consistent, so the topic acquisition is easily achieved via extraction of all *href links* with the *title* tag. These tags are stored in a *tuple* (data structure which stores two objects of data), and the tuples themselves are stored in a list. Once all topics of the current page are collected, the first subtopic of the first article, i.e. the second element of the first tuple in the list is used as the main topic and all its subtopics are extracted in the same manner. This process is repeated until the list of tuples of the initial list is exhausted. A formal representation of the algorithm is noted below.

```
function extract(topic)
    for all subtopics in topic
        extract subtopic
        add topic , subtopic to tuple
        store tuple in list
    return list of tuples
call function extract(topic)
for all subtopics in tuple
    call function extract(subtopic)
```

This process is done for the starting topic in English and its equivalent topic in Japanese. These two lists are then compared for co-occurance. In order to do so, the Japanese topics are translated by finding the equivalent topic via the Wikipedia API using the pageID property. Each Wikipedia page has a json file associated with it, which contains the set of all available language representations of this topic. This is slightly different and in some cases more precise than a direct translation of the word, describing this concept, since this translation is focused on the concept, rather than a dictionary entry, which may present several options, or be very generic.

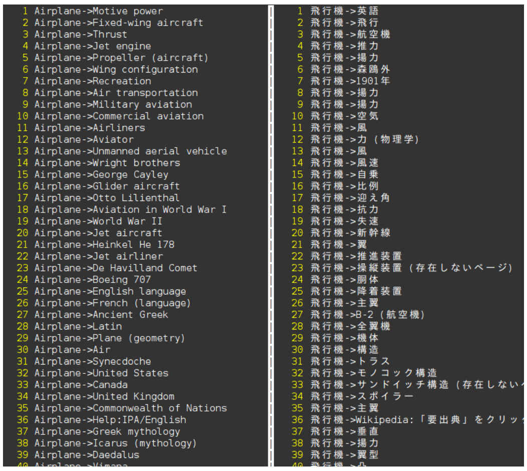
After getting the English topic equivalents for the Japanese list, the two lists are compared for co-occurance of topics. At the same time this results in a glossary for this list of topics.

5. Algorithm Implementation

The implementation of the algorithm described in Chapter 4. is done with the Python language. The modules used in the toolchain are: *BeautifulSoup*, for extraction of data from HTML, *requests* for HTTP access, *json* for reading pageID's from Wikipedia's API, and *re* for string comparison with regular expressions. The source code is open source under the Creative Commons License, and is available from the author upon request.

6. Showcase

The algorithm described in Chapter 4. is used to examine three topics and all of their subtopics up to the second level. The starting articles are *cat*, *language*, and *airplane*. The corresponding Japanese articles are ネコ,言語, and 飛行機. The output of the results and intermediate results for this starting topic *airplane* are described in this chapter. The article *airplane* yielded 406 topic entries, while their subtopics resulted in a combined total of 62,634 topics. A sample of the output for the first 40 topics found in English and Japanese are shown in Fig. 3.



```
 1 Airplane->Motive power
 2 Airplane->Fixed-wing aircraft
 3 Airplane->Thrust
 4 Airplane->Jet engine
 5 Airplane->Propeller (aircraft)
 6 Airplane->Wing configuration
 7 Airplane->Recreation
 8 Airplane->Air transportation
 9 Airplane->Military aviation
10 Airplane->Commercial aviation
11 Airplane->Airliners
12 Airplane->Aviator
13 Airplane->Unmanned aerial vehicle
14 Airplane->Wright brothers
15 Airplane->George Cayley
16 Airplane->Glider aircraft
17 Airplane->Otto Lilienthal
18 Airplane->Aviation in World War I
19 Airplane->World War II
20 Airplane->Jet aircraft
21 Airplane->Heinkel He 178
22 Airplane->Jet airliner
23 Airplane->De Havilland Comet
24 Airplane->Boeing 707
25 Airplane->English language
26 Airplane->French (language)
27 Airplane->Ancient Greek
28 Airplane->Latin
29 Airplane->Plane (geometry)
30 Airplane->Air
31 Airplane->Synecdoche
32 Airplane->United States
33 Airplane->Canada
34 Airplane->United Kingdom
35 Airplane->Commonwealth of Nations
36 Airplane->Help:IPA/English
37 Airplane->Greek mythology
38 Airplane->Icarus (mythology)
39 Airplane->Daedalus
40 Airplane->Vimana
```

```
 1 飛行機->英語
 2 飛行機->飛行
 3 飛行機->航空機
 4 飛行機->推力
 5 飛行機->揚力
 6 飛行機->森鴎外
 7 飛行機->1901年
 8 飛行機->揚力
 9 飛行機->揚力
10 飛行機->空気
11 飛行機->風
12 飛行機->力 (物理学)
13 飛行機->風
14 飛行機->風速
15 飛行機->自乗
16 飛行機->比例
17 飛行機->迎え角
18 飛行機->抗力
19 飛行機->失速
20 飛行機->新幹線
21 飛行機->翼
22 飛行機->推進装置
23 飛行機->操縦装置 (存在しないページ)
24 飛行機->胴体
25 飛行機->降着装置
26 飛行機->主翼
27 飛行機->B-2 (航空機)
28 飛行機->全翼機
29 飛行機->機体
30 飛行機->構造
31 飛行機->トラス
32 飛行機->モノコック構造
33 飛行機->サンドイッチ構造 (存在しない
34 飛行機->スポイラー
35 飛行機->主翼
36 飛行機->Wikipedia:「要出典」をクリッ
37 飛行機->垂直
38 飛行機->揚力
39 飛行機->翼型
40 飛行機->凸
```

Figure 3: Output from collecting topics

In the next step, the equivalent English topics for each Japanese topic is found, as described in Chapter 4.. Furthermore, the number of common topics is found in these lists. A high occurance of co-occuring topics indicates a high overlap of information and indicates a potential topic collection for parallel corpus harvesting.

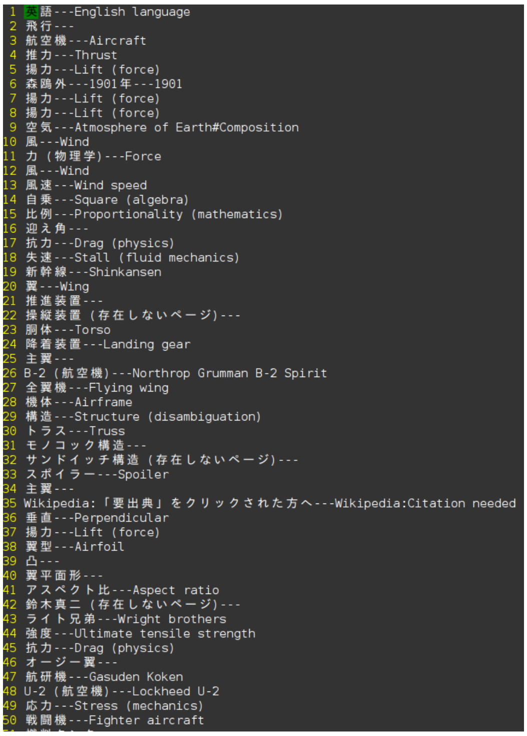
Figure 4 shows the first 50 entries of the Japanese topics with their English Wikipedia concept counterparts. Apart from being the basis of comparison of topic overlap, this collection is a glossary in a domain that stems from the initially chosen topic. This way, a glossary of any spesific topic domain can be compiled quickly and dynamically.

7. Summary

This paper presents a method for extracting Wikipedia articles and all its subtopics up to the second link level for the English-Japanese language pair and is extensible to other language pairs. A showcase of a topic search is presented as an example. Since this is a work in progress, there are no exact numbers yet on the precise topic overlap, although first samples indicate promising results. In the process of analyzing the topic co-occurance several domain specific terminology glossaries have been produced.

8. Future Work

It is planned to analyze large amounts of topic collections to identify parallel corpus harvesting candidates across Wikipedia. Further it is planned to identify and process article sections for parallel corpus extraction. Additionally, the toolchain will be expanded with a graphical user interface, which will make it easy and intuitive to use. A graphical implementation of the step to build glossaries will be tested at the Centre for Translation studies in a class room setting.



```
1  英語---English language
2  飛行---
3  航空機---Aircraft
4  推力---Thrust
5  揚力---Lift (force)
6  森鴎外---1901年---1901
7  揚力---Lift (force)
8  揚力---Lift (force)
9  空気---Atmosphere of Earth#Composition
10 風---Wind
11 力 (物理学)---Force
12 風---Wind
13 風速---Wind speed
14 自乗---Square (algebra)
15 比例---Proportionality (mathematics)
16 迎え角---
17 抗力---Drag (physics)
18 失速---Stall (fluid mechanics)
19 新幹線---Shinkansen
20 翼---Wing
21 推進装置---
22 操縦装置 (存在しないページ)---
23 胴体---Torso
24 降着装置---Landing gear
25 主翼---
26 B-2 (航空機)---Northrop Grumman B-2 Spirit
27 全翼機---Flying wing
28 機体---Airframe
29 構造---Structure (disambiguation)
30 トラス---Truss
31 モノコック構造---
32 サンドイッチ構造 (存在しないページ)---
33 スポイラー---Spoiler
34 主翼---
35 Wikipedia:「要出典」をクリックされた方へ---Wikipedia:Citation needed
36 垂直---Perpendicular
37 揚力---Lift (force)
38 翼型---Airfoil
39 凸---
40 翼平面形---
41 アスペクト比---Aspect ratio
42 鈴木真二 (存在しないページ)---
43 ライト兄弟---Wright brothers
44 強度---Ultimate tensile strength
45 抗力---Drag (physics)
46 オージー翼---
47 航研機---Gasuden Koken
48 U-2 (航空機)---Lockheed U-2
49 応力---Stress (mechanics)
50 戦闘機---Fighter aircraft
```

Figure 4: Output from collecting topics

9. Bibliographical References

Labaka, G., Alegria, I., and Sarasola, K. (2016). Domain adaptation in mt using titles in wikipedia as a parallel corpus: Resources and evaluation. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Ljubesic, N., Espla-Gomis, M., Toral, A., Rojas, S. O., and Klubicka, F. (2016). Producing monolingual and parallel web corpora at the same time - spiderling and bitextor's love affair. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Otero, P. G. and López, I. G. (2010). Wikipedia as multilingual source of comparable corpora. In *In Proceedings of the LREC Workshop on BUCC*, pages 30–37.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Paramita, M. L., Clough, P., Aker, A., and Gaizauskas, R. (2012). Correlation between similarity measures for inter-language linked wikipedia articles. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 23–25, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Reese, S., Boleda, G., Cuadros, M., Padró, L., and Rigau, G. (2010). Wikicorpus: A word-sense disambiguated multilingual wikipedia corpus. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 19–21, Valletta, Malta, may. European Language Resources Association (ELRA).

Vrandecic, D. and Krötsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, pages 78–85.