A Parallel English - Serbian - Bulgarian - Macedonian Lexicon of Named Entities

Aleksandar Petrovski
Faculty of Informatics
International Slavic University
Marshal Tito 77 Sv. Nikole, North Macedonia
a.petrovski.sise@gmail.com, aleksandar.petrovski@msu.edu.mk

Abstract

This paper describes the creation of a parallel multilingual lexicon of named entities from English to three South Slavic languages: Serbian, Bulgarian and Macedonian, with Wikipedia as a source. The basics of the proposed methodology are well known. This methodology provides a cheap opportunity to build multilingual lexicons, without having expertise in target languages.

Wikipedia's database dump can be freely downloaded in SQL and XML formats. The method presented here has been used to build a Python application that extracts the English – Serbian – Bulgarian – Macedonian parallel titles from Wikipedia and classifies them using the English Wikipedia category system. The extracted named entity sets have been classified into five classes: PERSON, ORGANIZATION, LOCATION, PRODUCT, and MISC (miscellaneous). It has been achieved using Wikipedia metadata. The quality of classification has been checked manually on 1,000 randomly chosen named entities. The following are the results obtained: 97% for precision and 90% for recall.

Keywords: parallel lexicons, named entities, Wikipedia

1 Introduction

Wikipedia is a free online encyclopedia, made and maintained as an open coordinated effort venture by a network of volunteer editors, utilizing a wiki – based editing system. Hosted and supported by the Wikimedia Foundation, since its start in 2001, the site has grown in both popularity and size. At the time of writing this paper (March 2022), Wikipedia contained over 58 million articles in 323 languages; its English version has over 6 million articles. The richness of information and texts continuously makes it an object of special research interest among the NLP (Natural Language Processing) community. By attracting approximately 6 billion visitors per month (Statista, 2021), it is the largest and most popular general reference work on the World Wide Web.

The term named entity (NE) refers to expressions describing objects, like persons, locations, and organizations. It was first introduced to the NLP community at the end of the 20th century. Named entities are often denoted by proper names. They can be abstract or have a physical existence. Some other expressions, describing money, percentage, time, and date might also be considered as named entities. Examples of named entities include *United States*, *Paris*, *Google*, *Mercedes Benz*, *Microsoft Windows*, or anything else that can be named.

The role of named entities has become more and more important in NLP. Their information is crucial in information extraction. As recent systems mostly rely on machine learning techniques, their performance is based on the size and quality of given training data. This data is expensive and cumbersome to create because experts usually annotate corpora manually to achieve high–quality data. As a result, these data sets often lack coverage, are not up to date, and are not available in many languages. To overcome this problem, semi – automatic methods for resource construction from other available sources were deployed.

Even though Wikipedia isn't made and maintained by linguists, metadata about articles, for instance, translations, disambiguations, or categorizations are accessible. Its structural features, size, and multilingual availability give a reasonable base to derive specialized resources, like multilingual lexicons (Bøhn and Nørvag,

2010). Researchers have found that around 74% of Wikipedia pages describe named entities (Nothman et al., 2008), a clear indication of Wikipedia's high coverage for named entities. Each Wikipedia article associated with a named entity is identified with its title, which is itself a named entity. That is a perfect opportunity to build parallel lexicons of named entities between them.

2 Related work

Building multilingual lexicons from Wikipedia has been a subject of research for more than 10 years. Schönhofen et al. (Schönhofen et al., 2007) exploited Wikipedia hyperlinkage for query term disambiguation. Tyers and Pienaar (Tyers and Pienaar, 2008) described a simple, fast, and computationally inexpensive method for extracting bilingual dictionary entries from Wikipedia (using the interwiki link system) and assessed the performance of this method with respect to four language pairs. Yu and Tsujii (Yu and Tsujii, 2009) proposed a method using the interlanguage link in Wikipedia to build an English-Chinese lexicon. Knopp (Knopp, 2010) showed how to use the Wikipedia category system to classify named entities. Bøhn and Nørvåg (Bøhn and Nørvag, 2010) described how to use Wikipedia contents to automatically generate a lexicon of named entities and synonyms that are all referring to the same entity. Halek et al. (Hálek et al., 2011) attempted to improve machine translation from English of named entities by using Wikipedia. In (Ivanova, 2012), the author evaluated a bilingual bidirectional English-Russian dictionary created from Wikipedia article titles. Higashinaka et al. (Higashinaka et al., 2012) aimed to create a lexicon of 200 extended named entity (ENE) types, which could enable fine-grained information extraction. Oussalah and Mohamed (Oussalah and Mohamed, 2014) demonstrated how to use info-boxes in order to identify and extract named entities from Wikipedia.

3 English, Serbian, Bulgarian, and Macedonian Wikipedias

The English Wikipedia is the English language edition of the Wikipedia online encyclopedia. English is the first language in which Wikipedia was written. It was started on 15 January 2001 (Wikimedia Foundation, 2001b), but versions of Wikipedia in other languages were quickly developed. There are three Wikipedias in concerned South Slavic languages among these versions: Serbian, Bulgarian, and Macedonian. They are all written in the Cyrillic alphabet, although there are few articles in Serbian Wikipedia written in Latin. The Serbian Wikipedia (Wikimedia Foundation, 2003c) was initiated in February 2003, the Macedonian (Wikimedia Foundation, 2003b) in September 2003, and the Bulgarian (Wikimedia Foundation, 2003a) in December 2003.

A list of all Wikipedias is published regularly on the Internet, along with several parameters for each language (Wikimedia Foundation, 2001a). Four parameters are considered: number of articles, the total number of pages (articles, user pages, images, talk pages, project pages, categories, and templates), number of active users (registered users who performed at least one change in the last thirty days), and depth (a rough indicator of the collaborative quality of Wikipedia, which shows how often articles are updated).

As shown in Table 1, as of 01 April 2022, the English Wikipedia contains 6,476,873 articles. It is by far the largest edition on Wikipedia. The Serbian Wikipedia contains 657,062 articles, the Bulgarian 280,535, and the Macedonian 126,265. According to the number of articles, the Serbian Wikipedia is the 21st largest edition of Wikipedia, the Bulgarian is 39th, and the Macedonian is 65th. The low value of the depth parameter for the Bulgarian Wikipedia is noticeable. It does not refer to low academic quality, which cannot be computed, but to Wikipedian quality, i.e. the depth of collaborativeness.

4 Method

Wikipedia's database dump can be freely downloaded in SQL and XML formats (Wikimedia Foundation, 2001c). The method presented here has been used to build a Python application that extracts the English – Serbian – Bulgarian – Macedonian parallel titles from Wikipedia and classifies them using the English Wikipedia category system.

The flowchart presented in Figure 1 shows

Parameter	en	sr	bg	mk
Number of articles	6,476,873	657,062	280,535	126,265
Total number of pages	55,506,698	3,959,090	625,235	516,913
Number of active users	127,578	868	756	258
Depth	1,111	156	27	88

Table 1: Parameters of the English, Serbian, Bulgarian, and Macedonian Wikipedias.

the process used for building the lexicon.

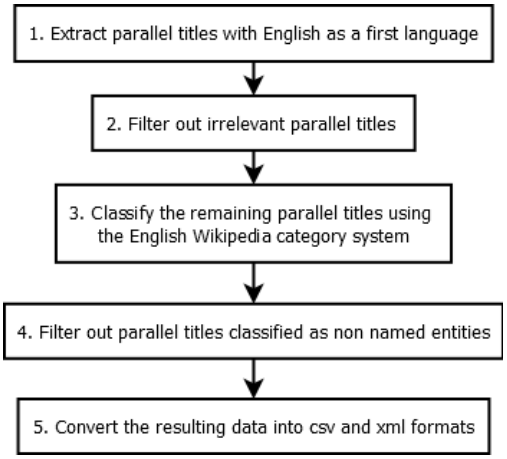

Figure 1: The process flowchart

1. *Extract parallel titles with English as a first language* – For building multilingual lexicons, two tables from the database are necessary: table of pages and table of interlanguage links. The page table is the "core of the wiki". It contains titles and other essential metadata for different Wikipedia namespaces. The interlanguage links table contains links between pages in different languages. Using these two tables, it is an easy programming task to create huge multilingual dictionaries without having any language expertise. In case of building multilingual lexicons with more than one language besides English, a new entry is created if there is a match between English and at least one of the other languages.

2. *Filter out irrelevant parallel titles* – The extracted parallel titles from the previous step contain a lot of noise. This step solves this issue. First, the program removes all the titles that don't belong to the main, template, or category namespaces. Second, there are titles containing some words or word stems that increase the noise and should be filtered out. The page table contains many entries that could not be a part of any lexicon, like usernames, nicknames, template names, etc. There are also titles, containing exclusively digits or blanks, which should be removed, too.

3. *Classify the remaining parallel titles using the English Wikipedia category system* – To classify the extracted named entities, one additional table from the database is required: a table of category links. The task of classifying named entities using category links is more complex. Wikipedia articles are generally members of categories. A category may have subcategories, each subcategory its subcategories, etc. The problem is that the graph could be cyclic, which may cause the program to go into an endless loop. Various authors propose different classes for named entities. Here, there are five classes: PERSON, ORGANIZATION, LOCATION, PRODUCT, and MISC (MISCELLANEOUS). Each named entity belongs to at least one of these classes. The classes comprise:

ORGANIZATION – political organizations, companies, schools, rock bands, sports teams;

PERSON – humans, gods, saints, fictional characters;

LOCATION – geographical terms, fictional places, cosmic terms;

PRODUCT – industrial products, software products, weapons, artworks, documents, concepts, standards, laws, formats, anthems, algorithms, journals, coats of arms, platforms, websites;

MISC – events, languages, peoples, tribes, alliances, orders, scientific discoveries, theories, titles, currencies, holidays, dynasties, positions, projects, historical periods, battles, competitions, alliances, deceases, programs, set of locations, awards, musical genres, missions, artistic directions, sets of organizations, networks.

4. *Filter out parallel titles classified as non – named entities* – Most Wikipedia titles are named entities, but not all of them. For example, certain natural terms – like biological species and substances which are very common

on Wikipedia are not included in the lexicon.

5. *Convert the resulting data into CSV and XML formats* – The lexicon comes in two formats: CSV and XML. An example of a lexicon entry in XML format is shown in Figure 2. The first four element names of each entry are *en*, *sr*, *bg*, and *mk* for English, Serbian, Bulgarian, and Macedonian, respectively. The text content of the elements is a translation of *Sofia* in respected languages. The fifth element contains the class, or classes, the entry belongs to. In this case, it is a LOCATION.

```
<entry>
    <en>Sofia</en>
    <sr>Софија</sr>
    <bg>София</bg>
    <mk>Софија</mk>
    <classes>
        <class>LOCATION</class>
    </classes>
</entry>
```

Figure 2: A lexicon entry in XML format

5 The lexicon

5.1 Statistics

The method presented in previous chapter has been used to build a Python application which extracts title sets independently on the languages. This program was applied to the Wikipedia database to extract the English - Serbian - Bulgarian - Macedonian sets of named entities. The result of the extraction after the first two steps from Figure 1 was 586,355 entries for English, 374,691 for Serbian, 258,940 for Bulgarian, and 149,633 for Macedonian. There are few titles in all South Slavic Wikipedias that are written in the original language (mostly English). In addition to that, there are few titles in Serbian Wikipedia in the Latin alphabet. The transliteration from Latin to Cyrillic and vice versa in Serbian is relatively straightforward.

Table 2 shows the number of entries per language after filtering out non named entities. The number of named entities in English is equal to the number of entries in the lexicon. The entries' numbers in Serbian, Bulgarian, and Macedonian are lower, which is understandable taking into account the number of articles in these Wikipedias, given in Table 1.

Language	Number of titles
English	400,930
Serbian	257,542
Bulgarian	179,854
Macedonian	106,351

Table 2: Number of titles per language

Class	Number
PERSON	81,724
ORGANIZATION	23,127
LOCATION	161,524
PRODUCT	32,951
MISC	107,973
All	407,299

Table 3: Distribution of classes

5.2 Distribution of classes

The resulting parallel English – Serbian – Bulgarian – Macedonian lexicon consists of 400,930 named entities. Each one belongs to at least one class, some of them to more. The distribution of classes is presented in Table 3.

The total number of classes, 407,299 is slightly higher than the number of entries, since some named entities may belong to more classes. The lexical entry presented in Figure 3 is such an example. *Bulgarian Academy of Sciences* is classified as both ORGANIZATION (the academy as an organization) and LOCATION (the buildings where the organization is located).

The examples of lexicon entries presented in figures 2 and 3 contain titles in all languages considered. But, this is not always the case. For example, as it is presented in Figure 4, the English title *Mark Antony* has translations only in Serbian and Bulgarian. There is no Macedonian translation since there is not such an article in the Macedonian Wikipedia.

5.3 Evaluation of classification

To evaluate classification, two common metrics in information retrieval have been used: precision and recall. Precision refers to the percentage of classes that are correct. On the other hand, recall refers to the percentage of

```
<entry>
    <en>Bulgarian Academy of Sciences</en>
    <sr>Бугарска академија наука</sr>
    <bg>Българска академия на науките</bg>
    <mk>Бугарска академија на науките</mk>
    <classes>
        <class>ORGANIZATION</class>
        <class>LOCATION</class>
    </classes>
</entry>
```

Figure 3: A lexicon entry belonging to two classes

```
<entry>
    <en>Mark Antony</en>
    <sr>Марко Антоније</sr>
    <bg>Марк Антоний</bg>
    <mk></mk>
    <classes>
        <class>PERSON</class>
    </classes>
</entry>
```

Figure 4: A lexicon entry with a missing translation in Macedonian

total relevant classes correctly classified by the algorithm.

An alternative to having two measures is the F – measure, which combines precision and recall into a single performance measure. This metric is known as F1 – score, which is simply the harmonic mean of precision and recall.

In order to evaluate the classification, a random sample containing 1,000 entries has been extracted from the lexicon. The entries from the sample have been classified manually and then compared to the classification performed by the algorithm. The results are presented in Table 4.

The precision of classification is between 94% for ORGANIZATION and 99% for PERSON. The recall is slightly lower, from 83% for PRODUCT and MISC to 97% for PERSON. The overall results are 97% for precision and 90% for recall.

The higher values of precision show that the classification algorithm was adjusted to classify the named entities correctly, rather than to extract more named entities for the lexicon.

6 Utilization

Lexicons, like the one presented in this paper, can be used in machine translation (MT). Most statistical MT systems do not deal explicitly with named entities, simply relying on the model of selecting the correct translation, i.e., mistranslating them as generic nouns. It is also possible that, when not identified, named entities may be left out of the output translation, which also has implications for the readability of the text. Because most NEs are rare in texts, statistical MT systems are not capable of producing quality translations for them. Another problem with MT systems is that failure to recognize NEs often harms the morpho – syntactic and lexical context outside of NEs itself. If named entities are not immediately identified, certain morphological features of adjacent and syntactically related words, as well as word order, may be incorrect. However, developers of commercial MT systems often do not pay enough attention to the correct automatic identification of certain types of NE, e.g. names of organizations. This is partly due to the greater complexity of this problem (the set of proper nouns is open and very dynamic), and partly due to lack of time and other development resources. One solution to this problem is using a parallel lexicon of named entities. If the lexicon contains a translation of the named entity, the translation quality will probably be good.

7 Conclusion

Using the methodology presented in this paper, a multilingual lexicon of named entities

Class	Precision	Recall	F1–score
PERSON	99%	97%	98%
ORGANIZATION	94%	87%	90%
LOCATION	98%	92%	95%
PRODUCT	96%	83%	89%
MISC	96%	83%	89%
All	97%	90%	93%

Table 4: The results of the classification check

from English to Serbian, Bulgarian, and Macedonian has been created. The named entities have been classified into five classes: PERSON, ORGANIZATION, LOCATION, PRODUCT, and MISC (miscellaneous). The number of lexical entries for these South Slavic languages varies and is dependent on the size of their Wikipedias, from 106,351 for Macedonian to 257,542 for Serbian. The quality of classification has been assessed: 97% for precision, and 90% for recall.

References

Christian Bøhn and Kjetil Nørvag. 2010. Extracting named entities and synonyms from wikipedia. *Proceedings of International Conference on Advanced Information Networking and Applications*, pages 1300–1307.

Ryuichiro Higashinaka, Kugatsu Sadamitsu, Kuniko Saito, Toshiro Makino, and Yoshihiro Matsuo. 2012. Creating an extended named entity dictionary from wikipedia. *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*, pages 1163–1178.

Ondrej Hálek, Rudolf Rosa, Aleš Tamchyna, and Ondrej Bojar. 2011. Named entities from wikipedia for machine translation. In *Proceedings of the Conference on Theory and Practice of Information Technologies*, pages 23–30.

Angelina Ivanova. 2012. Evaluation of a bilingual dictionary extracted from wikipedia. In *Computer Science*.

Johannes Knopp. 2010. *Classification of Named Entities in a Large Multilingual Resource Using the Wikipedia Category System*. University of Heidelberg, Master's thesis, Heidelberg, Germany.

Joel Nothman, James Curran, and Tara Murphy. 2008. Transforming wikipedia into named entity training data. *Proceedings of the Australian Language Technology Workshop*.

Mourad Oussalah and Muhidin Mohamed. 2014. Identifying and extracting named entities from wikipedia database using entity infoboxes. *International Journal of Advanced Computer Science and Applications*, 5:164–169.

Péter Schönhofen, András Benczúr, Istvan Biro, and Károly Csalogány. 2007. Cross-language retrieval with wikipedia. *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007ed Papers*, 5152:72–79.

Statista. 2021. Worldwide visits to wikipedia.org from january to june 2021. https://www.statista.com/statistics/1259907/wikipedia-website-traffic/, Last accessed on 2022-03-31.

Francis M. Tyers and Jacques A. Pienaar. 2008. Extracting bilingual word pairs from wikipedia. *Proceedings of the SALTMIL Workshop at the Language Resources and Evaluation Conference, LREC2008*.

Wikimedia Foundation. 2001a. List of wikipedias. https://meta.wikimedia.org/wiki/List_of_Wikipedias, Last accessed on 2022-03-31.

Wikimedia Foundation. 2001b. Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Main_Page, Last accessed on 2022-03-31.

Wikimedia Foundation. 2001c. Wikipedia:database download. https://en.wikipedia.org/wiki/Wikipedia:Database_download, Last accessed on 2022-03-31.

Wikimedia Foundation. 2003a. Bulgarian wikipedia. https://bg.wikipedia.org/wiki, Last accessed on 2022-03-31.

Wikimedia Foundation. 2003b. Macedonian wikipedia. https://mk.wikipedia.org/wiki, Last accessed on 2022-03-31.

Wikimedia Foundation. 2003c. Serbian wikipedia. https://sr.wikipedia.org/wiki, Last accessed on 2022-03-31.

Kun Yu and Jun'ichi Tsujii. 2009. Bilingual dictionary extraction from wikipedia. *Machine Translation Summit*, 12.