A Bilingual English-Ukrainian Lexicon of Named Entities Extracted from Wikipedia

Aleksandar Petrovski

Faculty of Informatics
International Slavic University
Marshal Tito 77 Sv. Nikole, North Macedonia
aleksandar.petrovski@msu.edu.mk

Abstract

This paper describes the creation of a bilingual English - Ukrainian lexicon of named entities, with Wikipedia as a source. The proposed methodology provides a cheap opportunity to build multilingual lexicons, without having expertise in target languages. The extracted named entity pairs have been classified into five classes: PERSON, ORGANIZATION, LOCATION, PRODUCT, and MISC (miscellaneous). It has been achieved using Wikipedia metadata. Using the presented methodology, a huge lexicon has been created, consisting of 624,168 pairs. The classification quality has been checked manually on 1,000 randomly selected named entities. The results obtained are 97% for precision and 90% for recall.

1. Introduction

The term named entity (NE) refers to expressions describing real world objects, like persons, locations, and organizations. It was first introduced to the Natural Language Processing (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 of America*, *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. One of these sources is Wikipedia.

The method presented here has been used to build a Python application which extracts the English - Ukrainian pairs from Wikipedia and classifies them using the English Wikipedia category system. Since both English and Ukrainian are among languages with most articles on Wikipedia, the result is a huge lexicon.

The goal of this paper is to present a method of extracting multilingual lexicons of classified named entities from Wikipedia. The method has been implemented to build a huge English - Ukrainian lexicon of named entities.

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. Wikipedia

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.

3.1. Wikipedia as a source

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.

Wikipedia is a very cheap resource of multilingual lexicons of named entities. Its database dump can be freely downloaded in sql and XML formats. But, taking into account the fact that Wikipedia articles have been written by millions of contributors, a question arises: What is the quality of these lexicons, and how reliable are they for using, e.g., in machine translation?

3.2. English and Ukrainian 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, 2022b), but versions of Wikipedia in other languages were quickly developed. Among these versions, there is one in Ukrainian language. The Ukrainian Wikipedia (Wikimedia Foundation, 2022c), written in the Cyrillic alphabet, was initiated in the year 2004.

A list of all Wikipedias is published regularly on the Internet, along with several parameters for each language (Wikimedia Foundation, 2022a). Four parameters are important: 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 quality of Wikipedia, which shows how often articles are updated).

As shown in Table 1, as of 26 March 2022, the English Wikipedia contains 6,473,638 articles and 55,472,454 pages. There are 127,722 active users. The depth value is 1,110. It is by far the largest edition of Wikipedia. The Ukrainian Wikipedia contains 1,144,596 articles and 3,992,549 pages. There are 2,702 active users. The depth value is 54. It is the 17th largest edition of Wikipedia, according to number of articles.

Parameter	en	uk
Number of articles	6,473,638	1,144,596
Total number of pages	55,472,454	3,992,549
Number of active users	127,722	2,702
Depth	1,110	54

Table 1: Parameters of the English and Ukrainian Wikipedias.

4. Method

The flowchart presented in Figure 1 shows the process used for building the lexicon.

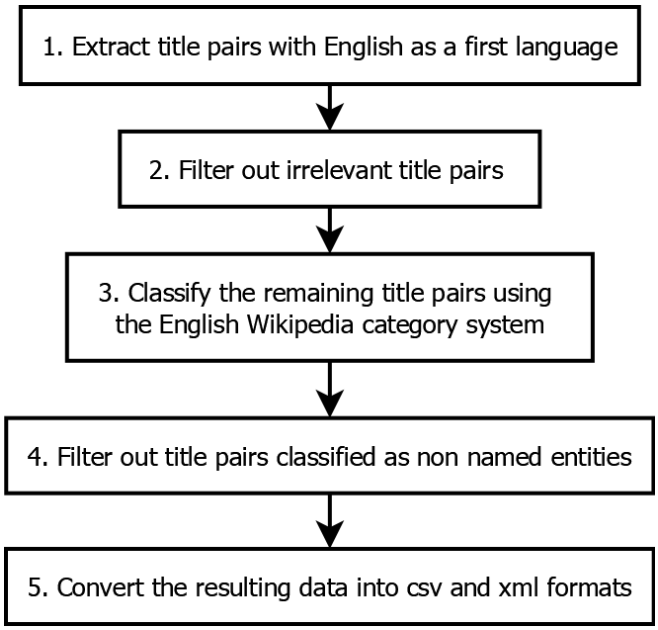

Figure 1: The process flowchart.

1. Extract title pairs 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 bilingual dictionaries without having any language expertise.

2. Filter out irrelevant title pairs

The extracted title pairs from the previous step contain a lot of noise. This step deals with it. First, the algorithm 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 user names, nicknames, template names, etc. There are also titles, containing exclusively digits or blanks, which should be removed too.

3. Classify the remaining title pairs using the English Wikipedia category system

In order 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 by means of category links is more complex. Wikipedia articles are generally members of categories. A category may have subcategories, each subcategory its own subcategories, etc. The problem is that the graph could be cyclic, which may cause the algorithm to go into an endless loop.

Various authors propose different classes for named entities. Here, there are five: PERSON, ORGANIZATION,

en	uk	PERS	ORG	LOC	PROD	MISC
Odessa	Одеса	0	0	1	0	0

Figure 2: A lexicon entry in CSV format.

LOCATION, PRODUCT, and MISC. Each named entity belongs to at least one of these classes. The classes comprise:

- ORGANIZATION- political organizations, companies, schools, rock bands, sport 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 title pairs 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.

The first row in the CSV file is a title row and tab is used as a field separator. The columns' titles are: en, uk, PERSON, ORGANIZATION, LOCATION, PRODUCT, and MISC. All other rows contain the data: English name, Ukrainian name, and five binary digits. These digits denote the class the named entity belongs to. For example, according to Figure 2, the named entity *Odessa* belongs to the class LOCATION, since the column LOC contains 1. All other classes contain 0's.

The structure of the XML file is similar. An equivalent of the entry from Figure 2 is shown in Figure 3. The columns' names en and uk from the CSV file are now names of elements and *class* denotes the classification.

In realizing the steps 2-3 of Figure 1, which refer to noise reduction and classification of named entities, the experience of creating a parallel lexicon of named entities from English to South Slavic languages (Slovenian, Croatian, Croatian, Bosnian, Ukrainian, Macedonian, and Bulgarian) (Petrovski, 2019) was of great benefit. That lexicon contains 26,155 entries, and the steps 2-3 were done manually.

This methodology has been used to create a multilingual English – Hebrew – Yiddish – Ladino lexicon of named entities. A tool that can be used to search it, can be found on the Internet (Petrovski, 2021).

```
<entry>
        <en> Odessa</en>
        <uk>Одеса</uk>
        <classes>
                <class>LOCATION</class>
        </classes>
</entry>
```

Figure 3: A lexicon entry in XML format.

5. Results

The method presented in previous chapter has been used to build a Python application which extracts title pairs independently on the languages. This application was applied to the Wikipedia database to extract the English - Ukrainian pairs of named entities. The result of the extraction after the first two steps from Figure 1 was 687,799 pairs. After filtering out non named entities, 624,168 pairs remained.

One part of the lexicon is presented in Figure 4.

en	uk	PERSON	ORGANIZATION	LOCATION	PRODUCT	MISC
Kyiv	Київ	0	0	1	0	0
Kyiv Academic Puppet Theatre	Київський академічний театр ляльок	0	0	1	0	0
Kyiv Academic Theatre of Drama and Comedy on the left bank of Dnieper	Київський академічний театр драми і комедії на лівому березі Дніпра	0	0	1	0	0
Kyiv Academic Theatre of Ukrainian Folklore	Київський академічний театр українського фольклору «Берегиня»	0	0	1	0	0
Kyiv Academic Young Theatre	Київський національний академічний Молодий театр	0	0	1	0	0
Kyiv Ballet	Київський балет	0	1	0	0	0
Kyiv Boryspil Express	Kyiv Borys pil Express	0	1	1	0	0
Kyiv Camerata	Національний ансамбль солістів «Київська камерата»	0	1	0	0	0
Kyiv Central Bus Station	Київський центральний автовокзал	0	0	1	0	0
Kyiv Chaika Airfield	Чайка	0	0	1	0	0
Kyiv Chamber Choir	Київ	0	1	0	0	0
Kyiv Chamber Orchestra	Київський камерний оркестр	0	1	0	0	0
Kyiv Christian Academy	Київська християнська академія	0	1	0	0	0
Kyiv City Council	Київська міська рада	0	1	1	0	0
Kyiv City Duma building	Будинок Київської думи	0	0	1	0	0
Kyiv City State Administration	Київська міська державна адміністрація	0	0	1	0	0
Kyiv Conservatory	Національна музична академія України імені Петра Чайковського	0	1	1	0	0
Kyiv Conservatory alumni	Випускники Київської консерваторії	0	1	0	0	0
Kyiv Conservatory faculty	Викладачі Київської консерваторії	0	0	0	0	1
Kyiv Day	День Києва	0	0	1	0	0
Kyiv Day and Night	Київ вдень та вночі	0	0	0	1	0
Kyiv Fortress	Київська фортеця	0	0	1	0	0
Kyiv Funicular	Київський фунікулер	0	0	1	0	0
Kyiv Half Marathon	Київський півмарафон	0	1	0	0	0
Kyiv Higher Party School alumni	Випускники Вищої партійної школи при ЦК КПУ	0	1	0	0	0
Kyiv Hydroelectric Power Plant	Київська ГЕС	0	0	1	0	0
Kyiv Independence Day Parade	Парад на честь дня Незалежності України	0	0	0	0	1
Kyiv Institute of Business and Technology	Київський інститут бізнесу та технологій	0	1	0	0	0
Kyiv International Airport	Міжнародний аеропорт «Київ»	0	0	1	0	0
Kyiv International Film Festival "Molodist"	Молодість	0	1	0	0	0
Kyiv International Institute of Sociology	Київський міжнародний інститут соціології	0	1	0	0	0
Kyiv International School	Київська міжнародна школа	0	1	0	0	0

Figure 4: A part of the lexicon.

The distribution of classes is presented in Table 2.

Class	Number
PERSON	142,850
ORGANIZATION	39,348
LOCATION	237,229
PRODUCT	56,952
MISC	159,952
Total	636,331

Table 2: Distribution of classes.

The total number of classes, 636,331, is slightly higher than the number of entries, since some named entities may belong to more classes. The lexical entry presented in Figure 5 is such an example. *Kherson State University* is classified as both ORGANIZATION (the university as an educational organization) and LOCATION (the building where the organization is located).

```
<entry>
        <en>Kherson State University</en>
        <uk>Херсонський державний університет</uk>
        <classes>
                <class>ORGANIZATION</class>
                <class>LOCATION</class>
        </classes>
</entry>
```

Figure 5: A lexicon entry belonging to two classes.

It is expected that the most of Wikipedia titles are multiwords, i.e. they contain either a space or a hyphen. Table 3 shows the number of multiword NEs per class in the lexicon for both English and Ukrainian.

Class	en	uk
PERSON	132,219	131,354
ORGANIZATION	34,114	30,509
LOCATION	116,974	99,399
PRODUCT	45,781	43,378
MISC	146,498	141,665
Total	475,586	446,305

Table 3: Number of multiword NEs per class.

Table 4 shows the percentage of multiword NEs per class.

It can be seen that the percentage of multiwords is higher in the English than in the Ukrainian Wikipedia. This is most noticeable in the classes ORGANIZATION and LOCATION. Some examples from the lexicon where there is a multiword in English and a single word in Ukrainian are given in Table 5 for the class ORGANIZATION and Table 6 for the class LOCATION.

Contributors to the English Wikipedia add words to the base title, which define it in more detail, or it is simply a matter of adding a definite article, e.g. *Sacramento, California* - Сакраменто, *Malkiya Club* - Малкія, *The Acropolis* - Акрополіс.

Class	en	uk
PERSON	93%	92%
ORGANIZATION	87%	78%
LOCATION	49%	42%
PRODUCT	80%	76%
MISC	92%	89%
All	75%	70%

Table 4: Percentage of multiword NEs per class.

en	uk
Malkiya Club	Малкія
Dnipro Kherson	Дніпро
Sharjah FC	Шарджа
Shin Bet	Шабак
Newtown A.F.C.	Ньютаун
The Day After Tomorrow	Післязавтра

Table 5: Examples of multiwords in English and single words in Ukrainian, class ORGANIZATION.

6. 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 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 7.

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.

7. Conclusion

Using the methodology presented in this paper, an English - Ukrainian lexicon of named entities has been created. Its size is 624,168 pairs. The named entities have been classified into five classes: PERSON, ORGANIZATION, LOCATION, PRODUCT, and MISC (miscellaneous). The quality of classification has been assessed: 97% for precision and 90% for recall.

en	uk
Malmö Airport	Мальме
Shintoku, Hokkaido	Сінтоку
Amarillo, Texas	Амарилло
Sacramento, California	Сакраменто
The Dakota	Дакота
The Acropolis	Акрополіс

Table 6: Examples of multiwords in English and single words in Ukrainian, class LOCATION.

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

Table 7: The results of the classification check.

The lexicon is available on (Petrovski, 2022) under CC-BY-NC-4.0 license (free for non commercial use).

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. It can be concluded that the identification of named entities in the source text is the first task of machine translators (Hálek et al., 2011). 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.

The European Commission called for language data in Ukrainian to/from all EU languages to train automatic translation systems (European Commission, 2022), (European Union's Horizon 2020 Research and Innovation Programme, 2020) supporting refugees and helpers in the Ukraine crisis. This lexicon was sent to ELRC (European Language Resource Coordination) Secretariat as a response.

8. References

Christian Bøhn and Kjetil Nørvag. 2010. Extracting Named Entities and Synonyms from Wikipedia. In *Proceedings of International Conference on Advanced Information Networking and Applications*, pages 1300–1307.

European Commission. 2022. Digital Europe Programme Language Technologies. https://language-tools.ec.europa.eu/.

European Union's Horizon 2020 Research and Innovation Programme. 2020. Bergamot Translations. https://translatelocally.com/web/.

Ryuichiro Higashinaka, Kugatsu Sadamitsu, Kuniko Saito, Toshiro Makino, and Yoshihiro Matsuo. 2012. Creating an Extended Named Entity Dictionary from Wikipedia. In *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. In *Proceedings of the Australian Language Technology Workshop*.

Mourad Oussalah and Muhidin Mohamed. 2014. Identifying and Extracting Named Entities from Wikipedia Database Using Entity Infoboxes. In *International Journal of Advanced Computer Science and Applications*, volume 5, pages 164–169.

Aleksandar Petrovski. 2019. EnToSSLNE - a Lexicon of Parallel Named Entities from English to South Slavic Languages. http://catalogue.elra.info/en-us/repository/browse/ELRA-M0051/.

Aleksandar Petrovski. 2021. Jewish Lexicons of Named Entities. https://www.jewishlex.org/.

Aleksandar Petrovski. 2022. A Bilingual English-Ukrainian Lexicon of Named Entities Extracted from Wikipedia. https://catalogue.elra.info/en-us/repository/browse/ELRA-M0104/.

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

Statista. 2021. Worldwide visits to Wikipedia.org from January to June 2021. https://www.statista.com/statistics/1259907/wikipedia-website-traffic/.

Francis M. Tyers and Jacques A. Pienaar. 2008. Extracting Bilingual Word Pairs from Wikipedia. In *Proceedings of*

the SALTMIL Workshop at the Language Resources and Evaluation Conference, LREC2008.

Wikimedia Foundation. 2022a. List of Wikipedias – Meta. https://meta.wikimedia.org/wiki/List_of_Wikipedias.

Wikimedia Foundation. 2022b. Wikipedia, the Free Encyclopedia. https://en.wikipedia.org/wiki/Main_Page.

Wikimedia Foundation. 2022c. Wikipedia, the Free Encyclopedia. https://uk.wikipedia.org/wiki/Main_Page.

Kun Yu and Jun'ichi Tsujii. 2009. Bilingual Dictionary Extraction from Wikipedia. In *Machine Translation Summit*, volume 12.