Extracting Weighted Language Lexicons from Wikipedia

Gregory Grefenstette
Inria Saclay/TAO, Rue Noetzlin - Bât 660
91190 Gif sur Yvette, France
E-mail: gregory.grefenstette@inria.fr

Abstract

Language models are used in applications as diverse as speech recognition, optical character recognition and information retrieval. They are used to predict word appearance, and to weight the importance of words in these applications. One basic element of language models is the list of words in a language. Another is the unigram frequency of each word. But this basic information is not available for most languages in the world. Since the multilingual Wikipedia project encourages the production of encyclopedic-like articles in many world languages, we can find there an ever-growing source of text from which to extract these two language modelling elements: word list and frequency. Here we present a simple technique for converting this Wikipedia text into lexicons of weighted unigrams for the more than 270 languages present currently present in Wikipedia. The lexicons produced, and the source code for producing them in a Linux-based system are here made available for free on the Web.

Keywords: lexicon, Wikipedia, language model, under-resourced languages

1. Introduction

Most natural language processing applications involve language-specific processing tools, and one essential element in these tools is a lexicon of words for the language processed. In the best cases these lexicons contain morphological, grammatical and usage information for each word (Sérasset, 2012) (Hathout *et al.,* 2014). In the worst cases, which is the case for the majority of the 6000+ languages spoken in the world, no computational lexicons are available[1]. Thus, a useful first step for under-resourced would be to provide at least a list of known words in that language. In addition to having a word list for the language, providing unigram probabilities, or relative frequencies of the known words, allows some natural language applications, such as optical character reading, spelling correction (Li, et al., 2011) machine translation (Tiers & Pienaar, 2008), or information retrieval, to make better decisions.

There are a few automatic ways to produce lexicons for languages. One way is to crawl the web and using a language identifier to classify pages into languages. Then, once boilerplate is removed, one can extract a lexicon from these pages (See, for example Erjavec *et al.,* 2008). In order to start a targeted crawling (Medelyan et al., 2006) , one starts from a list of URLs known to be in the desired language and a list of words to filter new pages discovered during the crawl. The Crúbadán Project (Scannell, 2007) uses such a crawler to continuously gather text for under-resourced languages. So far, the Crúbadán project has created lexicons for 2124 writing systems. For each language, anyone can download[2] a list of URLs that have been gleaned and hand edited for the language, and a list of character tri-grams that can be used to create a language identifier (Grefenstette, 1995). The Crúbadán pages also make available a weighted lexicon, such as we present here, from this crawled text as well as the word bigrams found during the web crawl, with their frequencies. This is, by far, the most advanced project for creating language resources for developed and under-developed languages.

Here we present the tool for crawling Wikipedia, just a subset of the sources crawled, by the Crúbadán Project, and making a weighted lexicon from this crawl. Although the Crúbadán Project releases the lexicons it produces under a Creative Commons Attribution licence, we are not aware of the release of its underlying tools, such as we present here.

2. Extracting Lexicons from Wikipedia

Wikipedia is a crowd-sourced "free-content" encyclopedia hosted by the non-profit Wikimedia Foundation. It currently exists in more than 270 languages[3], with the number of pages ranging from almost 5 million articles (English Wikipedia) to a few hundred pages (Gothic, Twi, Tsonga, Bislama, …)[4]. As of March 2016, 281 Wikipedias were considered 'active'[5] in the sense that at least one user user modified the Wikipedia in the preceding 30 days. Different language versions of Wikipedia are represented by a two or three character language code (ISO 639-1and ISO 639-3), forming part of the language specific URL, such as fr.wikipedia.org for the French version.

The contents of each language specific Wikipedia can be downloaded from dumps.wikimedia.org, the latest versions at dumps.wikimedia.org under the directories
/arcwiki/latest/<ISO>wiki-latest-pages-articles.xml.bz2
where <ISO> is the one or two character language code (e.g. aa, ab, ace, af, ak, …, en, fr, … zh, zu).

These dumps are produced every three weeks, and older dumps are also available, so applying the Wikipedia

[1] https://www.ethnologue.com/world
[2] http://crubadan.org/writingsystems
[3] https://en.wikipedia.org/wiki/List_of_Wikipedias
[4] See https://stats.wikimedia.org for exact numbers
[5] Our scripts only found 274 Wikipedia with text in them

processing files described and released in this article might be used for diachronic study of word use.

The dump contains articles in an XML format such as seen in Figure 1 for the article on *anarchism*.

```
<page>
<title>Anarchism</title>
<ns>0</ns>
 <id>12</id>
  <revision>
  <id>683845221</id>
  <parentid>683465996</parentid>
  <timestamp>2015-10-02T21:47:22Z</timestamp>
  <contributor>
   <username>Laplacemat</username>
    <id>18133810</id>
    </contributor>
      <comment>/* Further reading */ replaced Woodcock wi
      <model>wikitext</model>
      <format>text/x-wiki</format>
      <text xml:space="preserve">{{Redirect2|Anarchist|Anar
     {{pp-move-indef}}
     {{Use British English|date=January 2014}}
     {{Anarchism sidebar}}
'''Anarchism''' is a [[political philosophy]] that advocates
[[stateless society|stateless societies]], often defined as
[[self-governance|self-governed]], voluntary
institutions,&lt;ref&gt;&quot;ANARCHISM, a social
philosophy that rejects authoritarian government and
maintains that voluntary institutions are best suited to
```

Figure 1. Extract of Wikipedia dump for the article "Anarchy".

To extract the textual part, we provide three programs (dewiki, UnSGML, and sentencize)[6] to be run over each file. These programs are released with Creative Commons Attribution licenses with this article ‘see section 4). The *awk*-based programs remove non-content HTML markup (e.g. the text, seen in Figure 1, *<model>wikitext</model>* is removed). They convert SGML character encodings[7] into UTF-8 (e.g., *&quot;* is changed in a quotation mark). Then, they produce a tokenized output of Wikipedia with one sentence per line, such as seen in Figure 2.

```
TITLE=Anarchism .
' Anarchism ' is a political philosophy that advocates
        stateless societies , often defined as self-governed ,
        voluntary institutions , but that several authors have
        defined as more specific institutions based on
        non-hierarchical free associations .
 Anarchism holds the state to be undesirable ,
        unnecessary , or harmful .
While anti-statism is central , anarchism entails opposing
        authority or hierarchical organisation in the conduct
        of human relations , including , but not limited to ,
        the state system .
```

Figure 2. Cleaned version of Wikipedia text.

When you process Wikipedia versions in languages other than English, you find that there are pages or sections of pages written in English (or Spanish, or Dutch). For example, at the time of writing this article, the Romanian page https://ro.wikipedia.org/wiki/Enschede contains the English sentence: “The industrialisation stimulated a large increase in population, which at first was rather chaotic.” Sometimes entire articles are written English on a non-English wiki, either by copying, or spamming, or poor incomplete translation. Often song, book, or movie titles are given in English in these wiki. For these reasons, as a mitigating heuristic, for non-English Wikipedia, we include a fourth program (detectEnglish) which eliminates any line that contains three consecutive words in English.

The scripts we provide then split the output on whitespaces, and the individual tokens are sorted and tabulated, producing an output such as:

```
3549 anarchism
10 anarchisme
6 anarchismo
4 anarchisms
10330 anarchist
105 anarchiste
35 anarchistes
305 anarchistic
2 anarchistically
2 anarchistischen
4096 anarchists
54 anarcho
2 anarchocommunism
6 anarchopunk
4 anarchos
4 anarchosyndicalism
22 anarchosyndicalist
3 anarchosyndicalists
2 anarchs
2767 anarchy
```

The first value is the count of the word in the language version of Wikipedia, followed by the word itself. The reader will note that non-English words (*anarchisme, anarchismo*, …) also appear in this lexicon, due to the fact that non-English titles and terms also appear in the English Wikipedia. If needed, one might filter out these terms by comparing lexicons for different languages. This final cleaning step is not addressed here, but it is an interesting problem. As a palliative, on can remove words appearing under a given threshold, or impose a character set range on the words retained. We leave these solutions for future research.

3. Characteristics of Lexicons

The same set of programs is run over 274 Wikipedia dumps. Abusively, all programs are tokenized using the same program developed for English, so languages that do not use ASCII whitespace and punctuation marks as word separators will have to be further segmented. For languages with a lot of Wikipedia pages (English, French, Spanish, Dutch), there are a lot of foreign language words present in the extracted lexicons. Place names are often given with their original language names; national anthems can be listed with their original language lyrics, as well as translations; book, song and

[6] http://pages.saclay.inria.fr/gregory.grefenstette/LREC2016/WikiLexMaker.tar

[7] https://www.w3.org/TR/html4/sgml/entities.html

movie titles can be given in their original language. Entire quotes from other works may be given in their original language with a translation. This means that lexicons derived from these “rich” Wikipedia versions can contain a lot of *noise*, that is, words that one would not consider as being part of the language. For these reasons, we find over 4 million tokens in the English lexicon, over 1.5 million in the French and Spanish lexicons.

One *quick-and-dirty* way to eliminate these words is to impose lower-bound thresholds, for example, only taking words appearing with an initial lowercase letter (to eliminate proper names) and appearing 200 times or more over the Wikipedia. Doing so, for the English lexicon, one reduces the lexicon from 4.3 million upper and lower case entries to 56,000 lowercase entries. Similar results give 47,000 tokens in Spanish and 42,000 tokens in French. This is enough information to apply lexical expansion techniques if we have a basic morphological dictionary for some of the words (Grefenstette *et al.,* 2002). Such a technique will not work for German, nor for languages which do not use upper and lower-case to distinguish between common and proper nouns.

The programs presented here produce over a million potential lexical items with their frequency in the corresponding Wikipedia version for the following languages: English, German, Japanese, Russian, French, Polish, Swedish, Dutch, Spanish, Chinese, Italian, Hungarian, Finnish, Ukranian, Czech, Portuguese, and Norwegian. At the bottom of the list, there are less than 1000 items for the following ISO coded languages: pih, st, tw, tum, sg, ng, fj, ik, arc, ks, iu, mh, dz, cho, ii, kj, ti.

3.1 Sample comparison

We can compare one lexicon produced by the programs described here and the lexicons available at the Crúbadán Project. For example, for the Akan language of Ghana, the Crúbadán Project has crawled 176 documents, and produced a lexicon of 547909 words. When we compare their lexicon[8] with the one produced by our scripts, we find that we have discovered a number of words not found in the Crúbadán version but appearing more than once in ak.wikipedia.org : …, wɔn, wo, wɔ, wƆ, Wɔ, wɔampene, wɔaware, wobeko, wobetumi, wɔbɔ, wɔbɔɔ, wode, Wode, wɔde, Wɔde,wodi, wɔdi, wɔfa, wɔfrɛ, wɔka, wɔkyeree, wom, won, wɔn, wɔnfa,wɔnho, wɔnhwɛ, wɔnn, wonni, wɔnnye, wontu, wɔnyɛ, wooee, Wörld,wɔsan, woso, wɔtaa, wotuu, wowo, wɔwɔ, wɔwoo, Wɔwoo, wɔyɛ, woyii, wu, wui, wuxuu, … These missing words may be due to the fact that Crúbadán crawled Wikipedia at an earlier date, but it is illustrates the fact that possessing one’s own programs for crawling Wikipedia, in addition to the data provided by the Crúbadán Project, allows for richer lexicons.

4. Programs and Scripts released

We release two Linux scripts[9] for creating the lexicons:

- getdumps.sh, which fetches the latest version of Wikipedia dumps for 274 languages
- convert.sh, which converts Wikipedia dumps into tokenized versions, one sentence per line, and then converts that sentencized version into a weighted lexicon.

The Linux commands used by these shells are: *wget, sleep, bzcat, bzip2, egrep, gawk, echo, date, tr, cat, sort,* and *uniq*. The *gawk* programs used by these scripts are provided in the *tar* file:

- dewiki.awk, removes Wikipedia and metadata from the Wikipedia dump, retaining Wikipedia text, titles, categories, and redirects
- UnSGML.awk, translates SMGL-marked accents and special characeters into UTF-8
- sentencize.awk, tokenizes and rewrites input text with one sentence per output line, with spaces around each word.
- detectEnglish.awk, removes a line that contains three English words in a row (English lexicon included in the program source).

5. Future work

There are a number of ways that the lexicons can be made cleaner. The most obvious way would be to restrict the lexical entries to those within the UTF-8 code ranges for each individual language. Code charts exist for each language (see http://www.unicode.org/charts) and one could attach to each ISO codes the permissible character ranges from these charts. One could also use a known corpus of text from a given language and then calculate an edit-distance from the words extracted from this corpus to words found in the lexicon, keeping only the closest words. One might collect a set of “sure” lexical items (using the top 10% more rfrequent items, and then iteratively add lexical tiels found between two or four of these “sure” items. It is a difficult and ill-defined problem to decide whether a word should be part of a language lexicon or not (Zampieri, *et al.* 2015). Machine learning techniques might help.[10]

It would also be interesting to add bigrams from Wikipedia (as the Crúbadán project provides) and to produce the most similar words using a word embedding tool such as *word2vec* (Levy, *et al.*, 2015)*,* which goes part way to fulfilling the project of having a very large lexicon (Grefenstette, 2002) for the world’s languages.

6. Conclusion

This paper has presented four NLP programs and two scripts for converting Wikipedia into frequency weighted

[8] http://crubadan.org/files/ak.zip

[9] http://pages.saclay.inria.fr/gregory.grefenstette/LREC2016/WikiLexMaker.tar

[10] See “Large-Scale Language Classification” at http://nlp.stanford.edu/courses/cs224n/2015/reports/24.pdf

lexicons for all the languages of Wikipedia. The raw lexicons these scripts produced are also available online[11].

7. Acknowledgements

This research has been funded by a research grant from Inria/Saclay.

8. Bibliographical References

Erjavec, I. S., Erjavec, T., & Kilgarriff, A. (2008). A web corpus and word sketches for Japanese. Information and Media Technologies, 3(3), pp. 529-551.

Grefenstette, Gregory. (1995) Comparing two language identification schemes. In Proceedings of the Third International Conference on the Statistical Analysis of Textual Data (JADT'95), Rome, pages 263–268.

Grefenstette, G., Qu, Y., & Evans, D. (2002). Expanding lexicons by inducing paradigms and validating attested forms. In LREC 2002, Las Palmas.

Grefenstette, G. (2002). Multilingual corpus-based extraction and the very large lexicon. Language and Computers, 43(1), 137-149.

Hathout, N., Sajous, F., & Calderone, B. (2014). GLÀFF, a large versatile French lexicon. In Conference on Language Resources and Evaluation (LREC) pp. 1007-1012.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics, 3, 211-225.

Li, Y., Duan, H., & Zhai, C. (2011, July). Cloudspeller: Spelling correction for search queries by using a unified hidden markov model with web-scale resources. In Spelling Alteration for Web Search Workshop, pp. 10-14.

Medelyan, O., Schulz, S., Paetzold, J., Poprat, M., & Markó, K. (2006). Language specific and topic focused web crawling. In Proceedings of the Language Resources Conference LREC, pp. 267-269.

Scannell, K. P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop, Vol. 4, pp. 5-15.

Sérasset, G. (2012). Dbnary: Wiktionary as a LMF based Multilingual RDF network. In Language Resources and Evaluation Conference, LREC 2012.

Tyers, F. M., & Pienaar, J. A. (2008). Extracting bilingual word pairs from Wikipedia. Collaboration: interoperability between people in the creation of language resources for less-resourced languages, 19, pp. 19-22.

Zampieri, M., Tan, L., Ljubešic, N., Tiedemann, J., & Nakov, P. (2015, September). Overview of the dsl shared task 2015. In Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects

[11] Lexicons are found using the Wikipedia language code, for example, English is found at http://pages.saclay.inria.fr/gregory.grefenstette/LREC2016/lex/en.lex