BootCaT: Bootstrapping Corpora and Terms from the Web

Marco Baroni and Silvia Bernardini

SSLMIT, University of Bologna
Corso della Repubblica 136, 47100 Forlì, Italy
{baroni,silvia}@sslmit.unibo.it

Abstract

This paper introduces the BootCaT toolkit, a suite of perl programs implementing an iterative procedure to bootstrap specialized corpora and terms from the web. The procedure requires only a small set of seed terms as input. The seeds are used to build a corpus via automated Google queries, and more terms are extracted from this corpus. In turn, these new terms are used as seeds to build a larger corpus via automated queries, and so forth. The corpus and the unigram terms are then used to extract multi-word terms. We conducted an evaluation of the tools by applying them to the construction of English and Italian corpora and term lists from the domain of psychiatry. The results illustrate the potential usefulness of the tools.

1. Introduction

Despite certain obvious drawbacks (e.g., lack of control, sampling, documentation etc.), there is no doubt that the World Wide Web is a mine of language data of unprecedented richness and ease of access (Kilgarriff and Grefenstette, 2003). It is also the only viable source of "disposable" corpora (Varantola 2003) built ad hoc for a specific purpose (e.g., a translation task, the compilation of a terminological database, domain-specific machine learning). These corpora are essential resources for language professionals who routinely work with specialized languages, where new terms are introduced at a fast pace and standard reference corpora must be complemented by easy-to-construct, focused, up-to-date text collections.

While it is possible to construct a web-based corpus through manual queries and downloads, this process is extremely time-consuming. The time investment is particularly unjustified if the final result is meant to be a single-use corpus.

In this paper, we introduce the BootCaT toolkit,[1] a suite of perl programs implementing an iterative procedure to bootstrap specialized corpora and terms from the web, requiring only a small list of "seeds" (terms that are expected to be typical of the domain of interest) as input.

The basic idea is very simple: Build a corpus by automatically searching Google[2] for a small set of seed terms; extract new (single-word) terms from this corpus; use the latter to build a new corpus via a new set of automated Google queries; extract new terms/seeds from this corpus and so forth. The final corpus and unigram term list are then used to build a list of multi-word terms. These are sequences of words that must satisfy a set of constraints on their structure, frequency and distribution.

In developing the toolkit, we followed the old Unix adage that each program should do only one thing, but do it well. Thus, we designed a small, independent tool for each separate subtask of the algorithm. As a result, BootCaT is extremely modular: One can easily run a subset of the programs, look at intermediate output files, add new tools to the suite, or change one program without having to worry about the others.

The rest of the paper is structured as follows: In 2. we shortly review some related work. In 3. we describe the algorithm implemented in the BootCaT tools. In 4. we present an experimental evaluation of the tools. We conclude in 5. by suggesting possible directions for further development and evaluation of the toolkit.

2. Related work

The idea of building a corpus using automated search engine queries originates from Ghani et al. (2001), who apply it to the creation of minority language corpora.

In BootCaT, we compare frequencies in specialized and reference corpora to look for terms typical of the former. This is a fairly common idea in terminology extraction and corpus comparison work. See, for example, Rayson and Garside (2000).

The multi-word term extraction method we implement has some similarities with the ones proposed by Enguehard and Pantera (1995) and Pantel and Lin (2001). In particular, we share with the former the idea of looking for typical *connectors* (see 3.2. below) and with the latter the recursive method to search for multi-word terms.

However, as far as we know, we are the first to propose a full procedure for the automated extraction of specialized corpora and technical terms by web-mining.

3. The BootCaT procedure

The BootCaT procedure can be divided into two main phases: We first use an iterative algorithm to bootstrap corpora and unigram terms from the web. We then proceed to extract multi-word terms on the basis of the final corpus and unigram term list we extracted in the previous phase. Of course, one can stop after collecting the corpus and unigram list; and, *vice versa*, one can use our multi-word term extraction method on corpora that were not downloaded from the web.

Figure 1 summarizes the steps of the BootCaT procedure.

[1] BootCaT stands for ***Boot****strapping* ***C****orpora* ***a****nd* ***T****erms*. The toolkit is freely available from:
`http://sslmit.unibo.it/~baroni/bootcat.html`

[2] `http://www.google.com/apis`

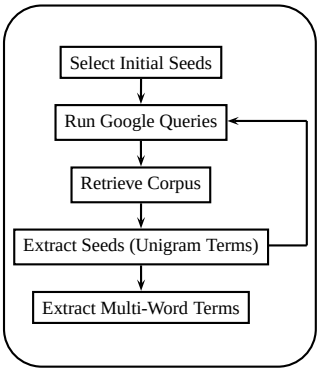

Figure 1: The BootCaT flow

3.1. Extraction of corpora and unigram terms

The bootstrapping process starts with a small list of seeds that are expected to be representative of the domain under investigation. For well-defined specialized domains, a small list of seeds (in the 5-to-15 range) is typically sufficient, and we obtained interesting results by starting with as few as two seeds (see 4.1.2. below).

The seed terms are randomly combined and each combination is used as a Google query string. The top *n* pages returned for each query are retrieved and formatted as text.

New unigram seeds are extracted from the corpus of retrieved pages by comparing the frequency of occurrence of each word in this set with its frequency of occurrence in a reference corpus. We compare frequencies using the log odds ratio measure (Everitt, 1992).

Random combinations of the newly extracted seed terms are then used for another round of Google queries and a new corpus is created by retrieving and formatting the top *n* pages found in this round.

The iterative term extraction/corpus downloading procedure is repeated as many times as desired (e.g., until the corpus reaches a certain size). In our experiments, we never found the need to repeat the process more than two or three times.

The user must control several important parameters, such as the number of queries issued for each iteration, the number of seeds used in a single query, the number of pages to be retrieved, etc.

3.2. Extraction of multi-word terms

The first step of this phase is to extract a list of single- and two-word *connectors* from the corpus, by looking for words and bigrams that frequently occur between two single-word terms (e.g., *of*, *of the*).

We then extract a list of stop words, i.e., words with a very high document frequency that were not identified as connectors.

At this point, we can look for multi-word terms, which we define for our current purposes as sequences of words that satisfy the following constraints:

- They contain at least one unigram term;
- they do not contain stop words;
- they may contain connectors, but these cannot occur at the edges nor be adjacent to each other;
- they have frequency above a certain threshold (dependent on length);
- they cannot be part of longer multi-word terms with frequency above $k * fq$, where k is a constant between 0 and 1 (but typically much closer to the upper end of the range) and fq is the frequency of the current term;[3]
- conversely, they cannot contain shorter multi-word terms with frequency above $(1/k) * fq$.

The multi-word terms are searched recursively. Starting with bigrams, we look left and right for an n+1gram term containing the current ngram and satisfying the constraints above, except the one banning edge connectors (otherwise, we would not find longer terms with inner connectors). For each seed bigram, the longest well-formed term containing it and without edge connectors is returned (this, of course, can equally be the bigram itself).

Again, the user must set various parameters, such as the minimum frequency for bigram terms and the value of the constant k (the minimum frequency threshold for longer terms will follow from these two parameters).

It would be interesting (and relatively straightforward) to add a filter that keeps only bigrams with a high mutual information (or other association measures) as possible starting points for the recursive multi-word term search procedure.

If the relevant resources are available, it would also be possible to filter out multi-word terms that do not match certain part-of-speech patterns.

4. Empirical assessment

Evaluating the performance of an unsupervised algorithm is hard, since typically we do not have pre-annotated data to be used for testing. In this case, the situation is further complicated by the inherent difficulty of estimating precision (is a certain web-page/term a hit or a miss?), and the impossibility of estimating recall (have we obtained an exhaustive list of all the web-pages/terms pertaining to a given domain?)

A few simple attempts at quantitative and qualitative evaluation have however proved encouraging. These were based on the results of an experiment in which we created a corpus and a term list intended to aid a trainee technical translator in translating an English psychiatric article (Fleisher et al., 2002) into Italian.

4.1. Corpus and term list construction

4.1.1. English

Our initial seeds were the six words in the abstract of the article to be translated that did not occur in the Brown corpus (Kučera and Francis, 1967), i.e., *dissociative, epilepsy, interventions, posttraumatic, pseudoseizures, ptsd*.

We iterated the corpus building/unigram term extraction procedure twice. In the first iteration, we queried Google

[3] In other words, a multi-word term cannot be part of a longer term with frequency close to its own.

for 15 randomly constructed seed triplets, retrieving a corpus of 181 pages (about 396,000 words). We calculated token-frequency-based log odds ratios by comparing this corpus to the Brown corpus, and we extracted 40 new seeds.

In the second iteration, we queried Google for 30 triplets randomly built from the new seeds and we retrieved a corpus that contained 516 new pages. We merged the two corpora obtaining a final corpus of about 1.5 million words.

We extracted the final unigram term list (1800 terms) from the corpus using a combination of log odds ratios based on token frequency and document frequency. This list and the corpus were used to extract multi-word terms of up to 5 words. We extracted 1507 such multi-word terms.

4.1.2. Italian

The two acronyms occurring in the English abstract (*eeg, ptsd*) constituted the initial seeds (we expect technical acronyms to be relatively stable across languages).

To compute the odds ratios, we used a reference corpus of Italian web pages of about 17 million words.

We iterated the bootstrap procedure twice, using parameter settings very similar to those we used for English. The main differences were that we queried Google for pairs rather than triplets (since there are less Italian pages on the web, we needed to boost up recall) and we filtered out the words also occurring in the English corpus before selecting the final unigram list. Of course, the automated Google queries used the Italian language option.

4.2. Evaluation

Evaluation of the corpora and term lists was carried out in three steps. First, 30 randomly selected web-pages from each corpus were (subjectively) evaluated according to their informativeness, relatedness to the target topic, and reliability. Second, 100 unigram terms and 100 multi-word terms in both English and Italian were also randomly selected and subjectively evaluated according to their well-formedness and relevance. Third, terms were manually collected from the English source text for which the corpus was originally constructed and from its translation. These lists were then compared with the BootCaT corpora and term lists.

4.2.1. Corpus quality

This procedure was meant to spot-check the quality and informativeness of the web-pages found by the algorithm, as against the standards a human might apply to corpus inclusion. The same procedure and the same criteria were applied to the English and to the Italian data. A web page was evaluated positively if it was found to be sufficiently informative, reliable, and consistent in topic and register with the task for which the corpus had been assembled.

Out of 30 web-pages randomly selected from the English corpus, 20 were found to be acceptable, and 10 to be unacceptable. The unacceptable ones were uninformative in 7 out of 10 cases (e.g., reference and staff lists) and unrelated in 3 out of 10 cases (e.g., a page on eye surgery).

Out of 30 web-pages randomly selected from the Italian corpus, 21 were acceptable and 9 were unacceptable. The unacceptable ones were either uninformative (6 cases, mainly reference lists or conference programs), or unreliable (e.g., a text on hypnosis from an ufological perspective) or unrelated (Amnesty International press releases).

Looking more in depth at the relevant pages, the Italian ones seem more informative, reliable and appropriately technical than the English ones. On the other hand, they are at times slightly off-topic, being concerned more with physiological than psychiatric topics. Further research on the grounds of these differences is needed.

4.2.2. Term lists: Precision

To estimate precision, we randomly extracted 100 multi-word terms from the English and Italian BootCaT lists and classified them according to their well-formedness and relevance.

For English, 10 were incomplete or badly-formed, 4 belonged to Internet jargon, 13 were proper nouns, 32 were general medical terms (e.g., *clinical symptoms*, *Environmental health*) and 41 were psychiatric terms (e.g., *abuse survivors*, *clinically significant distress*), giving a total of 73 "good" candidate terms out of 100.

For Italian, 3 were incomplete or badly-formed, 1 belonged to Internet jargon, 33 were proper nouns, 6 were general medical terms (e.g., *effetto analgesico*, *soggetti autistici*) and 26 were neuro-physiological terms (e.g., *oppioidi esogeni*, *patofisiologia della schizofrenia*), giving a total of 32 "good" candidate terms out of 100.

For the Italian multi-word term list a further category has to be mentioned, namely English chunks. As many as 31/100 terms (12 well-formed, 19 badly-formed) in our random set are made of English words. This is no doubt a consequence of the large percentage of English text found in Italian medical papers (e.g., in abstracts, references etc.), but also, possibly, an effect of the frequent use of English terms and phrases in medical Italian. For the purposes of this evaluation we have excluded all English term candidates from our list of "good" terms. It is not unlikely however that at least some of these terms should be moved up to the "good" terms list.

4.2.3. Term lists: Recall

Our evaluation of recall is more task-oriented, seeking to determine how many of the terms present in our source and target texts are also to be found in the BootCaT term lists and corpora.

For English, we found that 38 out of the 43 (88%) single-word terms and 11 of the 29 (38%) multi-word terms we extracted manually from the source text were also harvested by BootCaT on the web.

Of the 18 multi-word terms that were not in the BootCaT term list, 9 were present in the corpus it generated.

At least one content-rich component of each of the remaining terms was also found in the corpus, typically in contexts that could help understand the term as a whole. For instance, the term *sleep dysfunction* was not attested in the corpus, which however included 122 occurrences of *dysfunction*, with collocates such as *behavioral*, *family* and *sexual*.

For Italian, our results were rather less encouraging. Only 2 out of 40 unigram terms and no multi-word term from the text was found in the candidate term lists.

However, 34 out of 40 (85%) of the single word terms and 9 out of 28 (32%) of the multi-word terms are attested in the corpus generated by BootCaT. As was the case with English, the contexts of term components are also revealing: For instance, while *funzioni neuroanatomiche* is not attested in the corpus, *funzione/funzioni* occurs 132 times, with collocates such as *psico-motorie*, *corticali*, *fisiologiche* and so forth.

The choice of English seeds in the search for Italian texts, and the use of a translated text as a benchmark for comparison (as opposed to an original text) may have had an effect on the poorer performance of the term extractor in Italian. Inherent characteristics of the two languages (e.g., the rich Italian morphology), as well as potential differences in the conventions regulating medical web pages in English vs. Italian may also have played a role. Further analysis is needed in this area.

5. Conclusion

While our current results indicate that the BootCaT tools are already mining usable data (corpora and terms), the tools can be improved in many respects, including the ones we briefly mention here.

We said that the suite is extremely modular. The downside of this is that users must possess relatively advanced Unix command line skills, in order to be able to combine the BootCaT programs as desired. In the next version of the toolkit, we will provide a small set of "meta-scripts" running the whole procedure at once or in few steps, in order to help potential users who are not experienced with the command line and/or who do not plan to exploit the possibilities offered by the modular design.

Moreover, at the moment the BootCaT tools simply ignore documents in non-textual formats such as PS, PDF and Microsoft Word. However, documents in such formats tend to be content-rich (e.g., scientific papers). Thus, future BootCaT versions should be able to handle them.

We also plan to provide an interface to the UCS toolkit (Evert, 2004). Using this interface, one will be able to experiment with different measures and combinations of measures, besides the log odds ratio, in the unigram term extraction phase.

Farther away, we would like to support more sophisticated web mining techniques (perhaps exploiting some of the "hacks" of Calishain and Dornfest (2003) and Hemenway and Calishain (2004)) and further analysis of the documents that are retrieved.

In the meantime, we also plan to investigate the reasons why the term-extractor worked better in English than in Italian, and fine-tune the whole procedure accordingly. Experiments with other languages should follow.

We believe that the ultimate criterion to assess the validity of the tools is the extent to which the intended users find them useful, and prefer them to manual procedures. In this perspective, we are experimenting with the BootCaT tools in the translation classroom and we are collecting reports from trainee translators and terminologists (Bernardini and Baroni, 2004).

While these reports may not provide us with standard quantitative evaluation measures, we believe that they will be extremely valuable for assessing the performance of the tools in realistic settings and that they will provide us with precious hints for future development.

6. References

S. Bernardini and M. Baroni. 2004. Web mining in the translation classroom. *Submitted*.

T. Calishain and R. Dornfest. 2003. *Google Hacks*. O'Reilly.

C. Enguehard and L. Pantera. 1995. Automatic natural acquisition of bilingual terminology. *Journal of Quantitative Linguistics*, 2:27–32.

B. Everitt. 1992. *The Analysis of Contingency Tables.* Chapman and Hall, 2nd edition.

S. Evert. 2004. *The Statistics of Word Cooccurrences: Bigrams and Collocations.* Ph.D. thesis (in progress), University of Stuttgart.

W. Fleisher, D. Staley, P. Krawetz, N. Pillay, J. Arnett, and J. Maher. 2002. A comparative study of trauma-related phenomena in subjects with pseudoseizures and subjects with epilepsy. *American Journal of Psychiatry*, 159:660–663.

R. Ghani, R. Jones, and D. Mladenic. 2001. Mining the web to create minority language corpora. *CIKM 2001*, 279–286.

K. Hemenway and T. Calishain. 2004. *Spidering Hacks*. O'Reilly.

A. Kilgarriff and G. Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29:333–347.

H. Kučera and N. Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press.

P. Pantel and D. Lin. 2001. A statistical corpus-based term extractor. *Proceedings of AI 2001*.

P. Rayson and R. Garside. 2000. Comparing corpora using frequency profiling. *Proceedings of Workshop on Comparing Corpora of ACL 2000*, 1-6.

K. Varantola. 2003. Translators and disposable corpora. In F. Zanettin, S. Bernardini and D. Stewart, editors, *Corpora in Translator Education*, pages 55–70, StJerome.