WebBootCaT: a web tool for instant corpora

Marco Baroni
SSLMIT
Bologna University, Italy

Adam Kilgarriff
Lexical Computing Ltd,
Brighton, UK

Jan Pomikálek
Masaryk University,Czech Republic
and Lexical Computing Ltd

Pavel Rychlý
Masaryk University
Brno, Czech Republic

Abstract
We present a web service for quickly producing corpora for specialist areas, in any of a range of languages, from the web. The underlying BootCaT tools have already been extensively used: here, we present a version which is easy for non-technical people to use as all they need do is fill in a web form. The corpus, once produced, can be either downloaded or loaded into the Sketch Engine, a corpus query tool, for further exploration. Reference corpora are used to identify the key terms in the specialist domain. The service is freely available to all on a trial basis.

1 Introduction

If I am investigating the terminology of lexicography in Italian, where should I go? Regular dictionaries will not cover it, specialist dictionaries, if they exist, will be hard to find and expensive and are likely to be out of date. The obvious answer is the web. In 2006, this is probably what every working lexicographer, terminologist and translator does as a matter of course. The question, then, is how to do it effectively and efficiently.[1]

Baroni and Bernardini (2003) responded to the challenge with the BootCaT tools. The basic method is
• Select a few 'seed terms'

---

[1] For early accounts of using the web in this way see Varantola (2000) and Jones and Ghani (2000). For an overview of the use of the web as a source of linguistic data see Kilgarriff and Grefenstette (2003).

• Send queries with the seed terms to Google
• Collect the pages that the Google hits page points to.

This is then a first-pass specialist corpus. The vocabulary in this corpus can be compared with a reference corpus and terms can be automatically extracted. The process can also be iterated with the new terms as seeds to give a 'purer' specialist corpus.

The software is freely available for download and has been widely used, both to produce specialist corpus for technical term extraction (see, e.g., Fantinuoli 2006) and to produce large general-language corpora (Sharoff 2005, Baroni and Kilgarriff 2006). However, the software must be downloaded and installed, and this presents a barrier for people without computer systems skills.

2 WebBootCaT

In this paper we present a web-service version of the BootCat tools, WebBootCaT. The user no longer needs to download or install software, as they use a copy of the software which is already installed on our webserver. Our webserver also holds the corpus and loads it into a corpus query tool, the Sketch Engine (Kilgarriff et al 2004) for further investigation and analysis. For languages where further linguistic processing tools are available, these functions are optional extras.

Figure 1. WebBootCaT search page

• In the first box, the user inputs the seed terms. They can be either single words, or multi-word terms enclosed in double-quotes.

• In the second box, they input a Google API key. Google and other search engines are at risk of being swamped by automated queries, sent without human intervention. So Google's terms of use forbid automated querying, unless it is done officially. The mechanism for doing it officially is to fill in a web form to obtain a Google account and an "API key". Automated queries must then be accompanied by the key, with a limit of 1000 per day. To use WebBootCaT, the user must obtain a key and enter it in the box. (WebBootCaT typically uses just twenty queries a time, so the 1000 limit will only constrain a very active user.)

• In the third box the user specifies the language. The list of languages offered is a subset of those offered by Google, and the language is specified as part of the Google query, so the

pages returned are of the language specified according to Google. (We find that similar languages – e.g., Spanish and Portuguese – are occasionally confused.)

– Rather than using Google's language identifier, it is possible to use common words of the language as seeds and to gather pages of the right language that way. The words chosen should not be common words of any other language, and not words of English at all.

– Languages with different character sets have not yet been tested. Preliminary testing for Chinese and Japanese shows that the core functionality is in place.

• "Select URLs" tickbox: The software uses Google to identify a set of pages, and the user has the option of checking these pages and deciding which should be excluded before proceeding. If the user ticks this box, they are shown a list of URLs, each with a tickbox beside it, and they leave the ones they want ticked before proceeding to the corpus gathering.

• "Tag corpus" tickbox: Currently this is available for English, French, German, Italian, Spanish; further languages will be added as lemmatizers and POS-taggers for more languages are prepared. The lemmatizer gives the lemma for the word, so, e.g., a corpus instance of the English form *invading* will be associating with the lemma *invade*. A POS-tag is a label associated with a word saying what its word class is. As discussed elsewhere (eg McEnery and Wilson 1996) more can be done with a corpus if the data is enriched with markup such as POS tags. It makes it straightforward to search for, eg, "*promise* as a noun preceded by an adjective".

• The last two boxes are for a name for the corpus, and the user's email, so they can be notified when the corpus is ready.

The user then clicks "BUILD CORPUS" and sees a progress-monitoring screen. The length of time taken for processing is highly variable, depending both on the speed of the web connection and the target websites, both variable according to time of day etc. Our experience to date is that it usually takes minutes, but occasionally more than an hour.

For the time being we have imposed a 3MB limit (typically equivalent to a corpus of 500,000 words) for trial users. If the limit is reached, collection of new pages ceases but the program continues with the pages already gathered. Higher space allocations will be made available by agreement.

3 Processing details

The work going on behind the scenes includes the following.

• Producing permutations of the seed terms to send to Google. The default settings are that ten queries are sent to Google, each containing a randomly-selected triple of the seed terms.

• Each Google query returns up to 100 hits. We take, by default, the top ten for each query (and filter out duplicates).

• The collection of pages, including timeouts where no responses is received in good time for one of the URLs

• As widely noted, very short and very long web pages rarely contain useful samples of language, with running text. We filter these out. Default value for "very short" is less than 5KB, for "very long", over 2MB.

- Duplicate and near-duplicate web pages are deleted
- The remaining pages are further processed to filter out HTML, javascript, navigation bars, and other kinds of unwanted material (see Baroni and Kilgarriff 2006 for more details).
- A header is added which gives the URL the page was obtained from.
- The text is tokenized, to give a stream of words, punctuation characters etc. For languages written with spaces between words, most cases are straightforward but for languages such as Chinese and Japanese, this is a complex further stage.
- If the language is one for which a lemmatizer and POS-tagger is available, and the tick-box was ticked, the corpus is lemmatized and POS-tagged.
- The corpus is loaded into the Sketch Engine.

The user then sees a page like Fig. 2., which reports the completion of the process. (The same information is sent by email.)

Figure 2. Report of completion of corpus build

The figures shown are from an actual run, using the seed terms in Fig 1 and all default settings. Twenty-seven pages were retrieved and used, making a corpus of 155,700 words, in one and a half minutes.

127

The corpus is available for download (either as plain text, or as tokenised, part-of-speech-tagged, "vertical" (one-word-per-line) text. It can also be viewed in Sketch Engine. If we click on the "access URL" above we can search, in this corpus, for the lemma *collocazione* (and then sport by right context) we see Fig. 3.
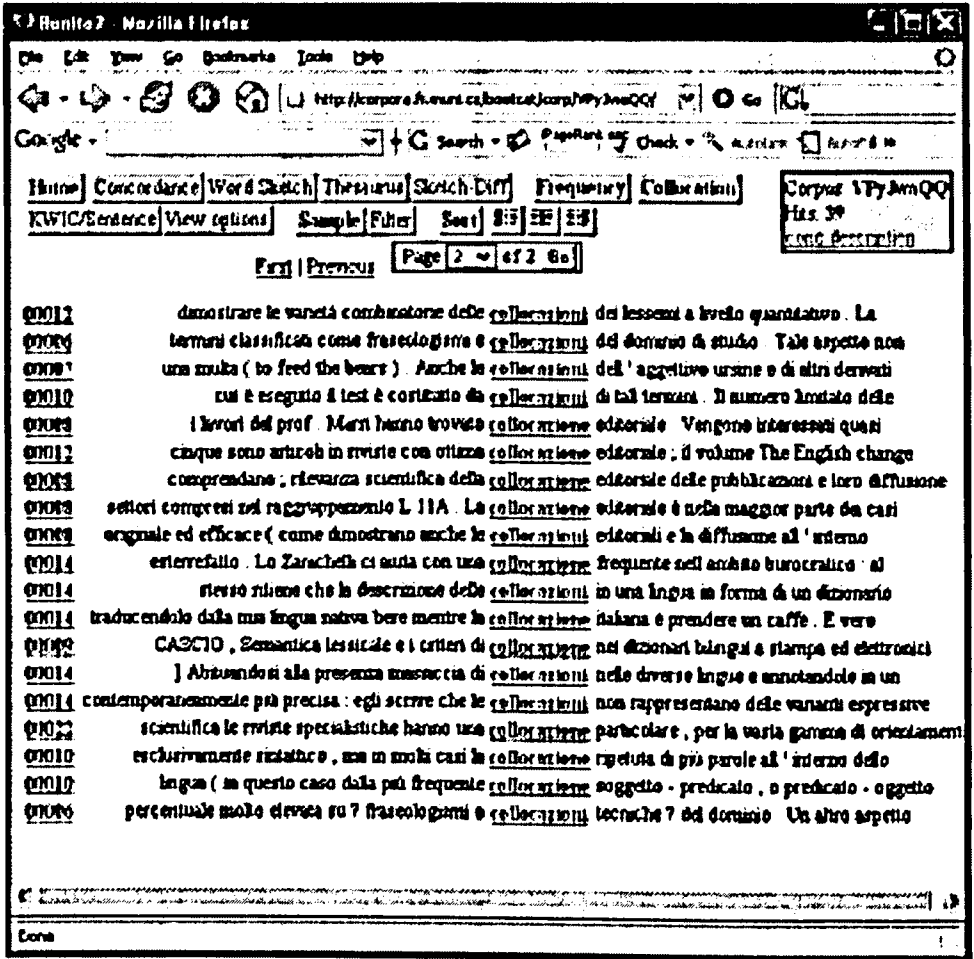

Figure 3. Instant corpus in the Sketch Engine

As the corpus is lemmatized, we see instances of the singular and the plural. The sorting brings together the five instances where the following word is a form of *editoriale*, thereby promptly bringing our attention to a candidate term (though in this case we note that it is not a term from this domain.) The "collocation" button supports further exploration.

The reference numbers in the left column show that the five instances of *collocazione editoriale* come from two different documents. If the user clicks on the number in the left hand column, they are shown the URL that the document is form, and if they click that, they will be taken to the original page.[2]

The corpus can of course be explored in many further ways now that it is in Sketch Engine, a corpus query tool with a wide range of functions.[3]

4 Finding key words and terms for the domain

We would like to compare this corpus to a reference corpus, to find the key words of the domain.

Thee reference corpus we use for Italian is developed from the web, using similar methods on a larger scale, to represent general Italian. If the user clicks the "extract terms" button on the web page which reports the success of the corpus build, they are shown (after a parameter-setting screen) a report of key terms as in Fig. 4.

---

[2] It is of course possible that the original web page will no longer be "live".

[3] See http://www.sketchengine.co.uk for user guide and a trial account.

Figure 4. Key words report

This is a first version and currently reports only single-word items: a version that reports also multi-word terms will be available in March 2006.

References

Baroni, M, Bernardini, S. (2004), 'BootCaT: Bootstrapping corpora and terms from the web', Proceedings of LREC 2004, Lisbon: ELDA, pp. 1313-1316.

Baroni, M., Kilgarriff, A. (2006), 'Large linguistically processed web corpora for multiple languages', *Proc. EACL.* Trento.

Fantinuoli, C. (2006), 'Specialized corpora from the Web and term extraction for simultaneous interpreters', in Baroni, M. and Bernardini, S. (eds.), *WaCky! Working papers on the Web as Corpus.* Gedit, Bologna.

Jones, R., Ghani, R. (2000), 'Automatically building a corpus for a minority language from the Web', Proc. Students' session, 38th ACL, Hong Kong.

Kilgarriff, A, Grefenstette, G. (2003), 'Web as Corpus: Introduction to the Special Issue', *Computational Linguistics* 29 (3), pp. 333-347.

Kilgarriff, A, Rychlý, P., Smrz, P., Tugwell, D. (2004), 'The Sketch Engine', Proc Euralex. Lorient.

Sharoff, S. (2005), 'Creating general-purpose corpora using automated search engine queries', in Baroni M., Bernardini, S, (eds.), *WaCky! Working papers on the Web as Corpus.* Gedit, Bologna.

Varantola, K. (2000), 'Translators and Disposable Corpora', *Proc. CULT (Corpus use and learning to translate),* Bertinoro, Italy.