WebBootCaT: instant domain-specific corpora to support human translators

Marco Baroni[♣], Adam Kilgarriff[♢], Jan Pomikálek[♢♠], and Pavel Rychlý[♠]

[♣]SSLMIT, Bologna University (Italy)
[♢]Lexical Computing Ltd, Brighton (UK)
[♠]Masaryk University, Brno (Czech Republic)

Abstract

We present a web service to aid translators by 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.

1 Introduction

Where should a translator look if they want to find the terminology of a specialist area? 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will not cover it, specialist dictionaries, if they exist, will be hard to find and expensive and are likely to be out of date.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 translator and terminologist does as a matter of course. The question, then, is how to do it effectively and efficiently. [1]

Baroni and Bernardini ([2004]) responded to the challenge with the BootCaT tools. The basic method is

- Select a few "seed terms".
- 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 (POS-tagging, lemmatization), these functions are optional extras. Figure 1 shows the WebBootCat interface.

- In the first field, the user inputs the seed terms. They can be either single words,

[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]).

or multi-word terms enclosed in double-quotes.

- In the second field, 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 using the Google API.

- In the third field the user can specify the language. The system takes advantage of this information to use Google's language identifier and to perform language dependent steps on the downloaded pages (e.g. tokenization, POS-tagging, lemmatization). WebBootCaT can also be used for languages which are not supported by the 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.

- "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 (e.g. 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, e.g., "*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 on the number of seeds and URLs, and 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 under five minutes, but occasionally more than twenty.

2.1 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 time-outs 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 5 kB, for "very long", over 2 MB.

- 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).

- 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 tickbox was ticked, the corpus is lemmatized and POS-tagged.
- The corpus is loaded into the Sketch Engine.

The user then sees a page like Figure 2, which reports the completion of the process. (The same information is sent by email.) The figures shown are from an actual run, using the seed terms in Figure 1 and all default settings. Forty-one pages were retrieved and used, making a corpus of 143,883 words, in about two and a half minutes.

The corpus is available for download either as plain text, or as tokenized, 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 *translate* we see Figure 3. As the corpus is lemmatized, we see instances of various forms of the verb.

The reference numbers in the left columns show which concordance lines come from the same sources. If the user clicks on the number in the left hand column, they are shown the URL that the document is from, 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]

2.2 Finding key words and terms for the domain

We would like to compare this corpus to a reference corpus and to find the key terms of the domain. The reference corpora we use are developed from the web, using similar methods on a larger scale, to represent general language. Currently, we have five reference corpora available (for English, German, French, Italian and Spanish) of about 500 million words in average. If the user clicks the "extract terms" link on Fig. 2, they are shown (after a parameter-setting screen) a report of key terms as in Fig 4.

The comparison to the reference corpus is done by using frequency lists of words and multiword expressions. We use several statistical methods to find out which words and expressions appear in the corpus more frequently then in a general language represented by the reference corpus. These expressions are considered keywords.

Using the keywords extraction form, users can select from three statistical methods and set several parameters to filter out words which are very unlikely to be keywords, e.g. stoplist words (function words), expressions not containing any alphabetic character, very short words or unfrequent words.

3 Conclusion

We have presented WebBootCaT, a web service designed to help translators (and others, including MT system developers) find terms and other specialist-domain language on the web. The tool is fast and easy-to-use, and presents both candidate term lists and access to a specialist-language corpus in an advanced corpus query tool, the Sketch Engine. The tool will be publicly available from Spring 2006.

[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.

References

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

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

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

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

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

[2004] Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine. Proc Euralex. Lorient. (2004)

[1996] McEntry, T., Wilson, A.: Corpus Linguistics. Edinburgh University Press. (1996)

[2005] Sharoff, S.: Creating general-purpose corpora using automated search engine queries. In M. Baroni and S. Bernardini, editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna. (2005)

[2000] Varantola, K.: Translators and Disposable Corpora. Proc. CULT (Corpus use and learning to translate), Bertinoro, Italy. (2005)

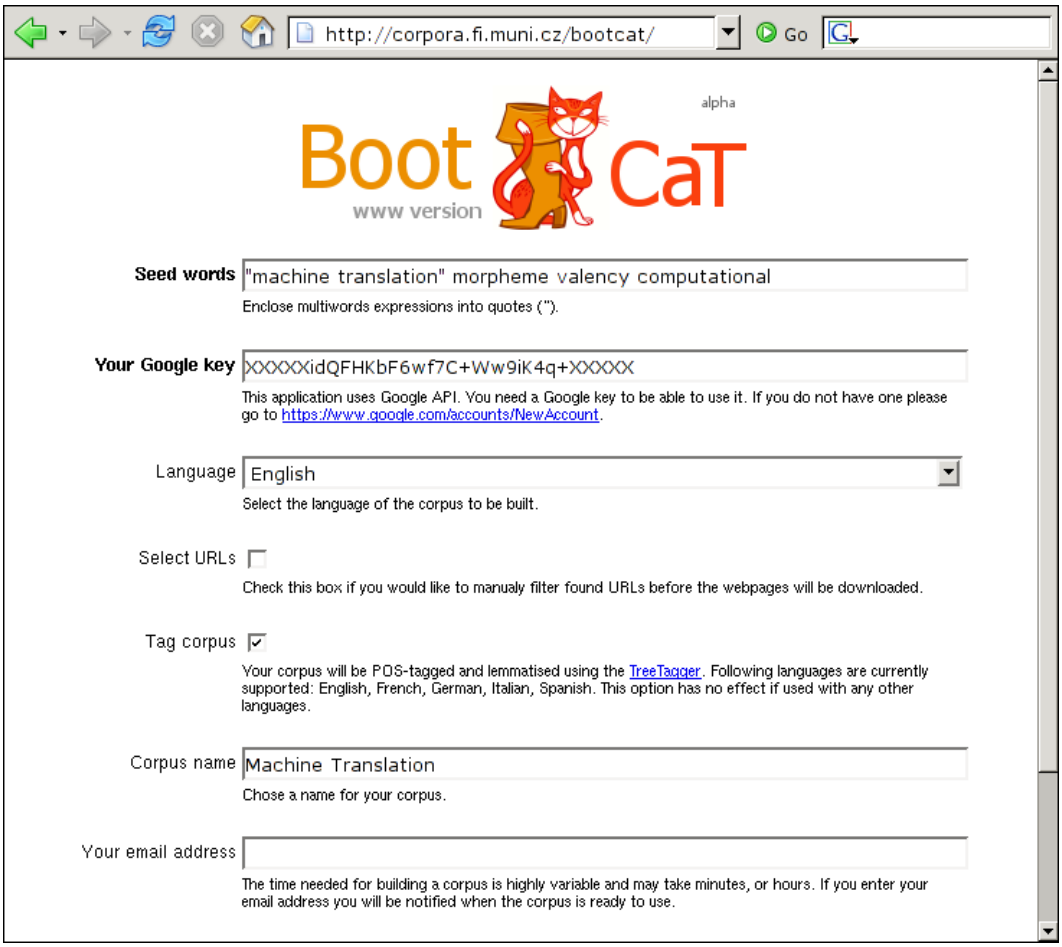

Figure 1: WebBootCaT search page

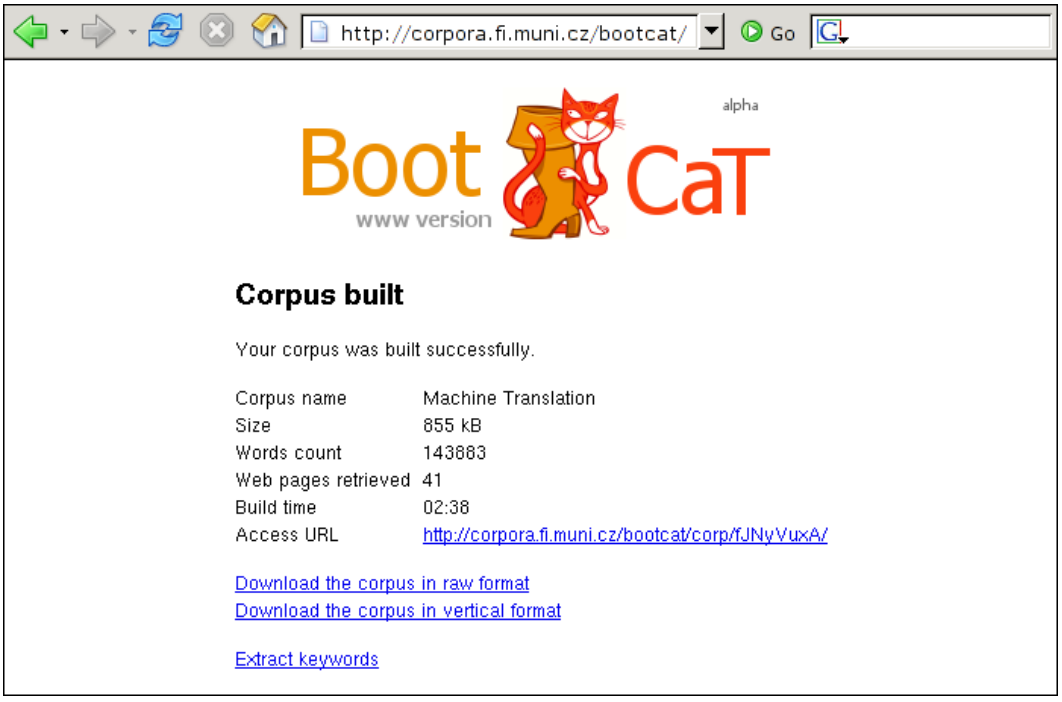

Figure 2: Report of completion of corpus build

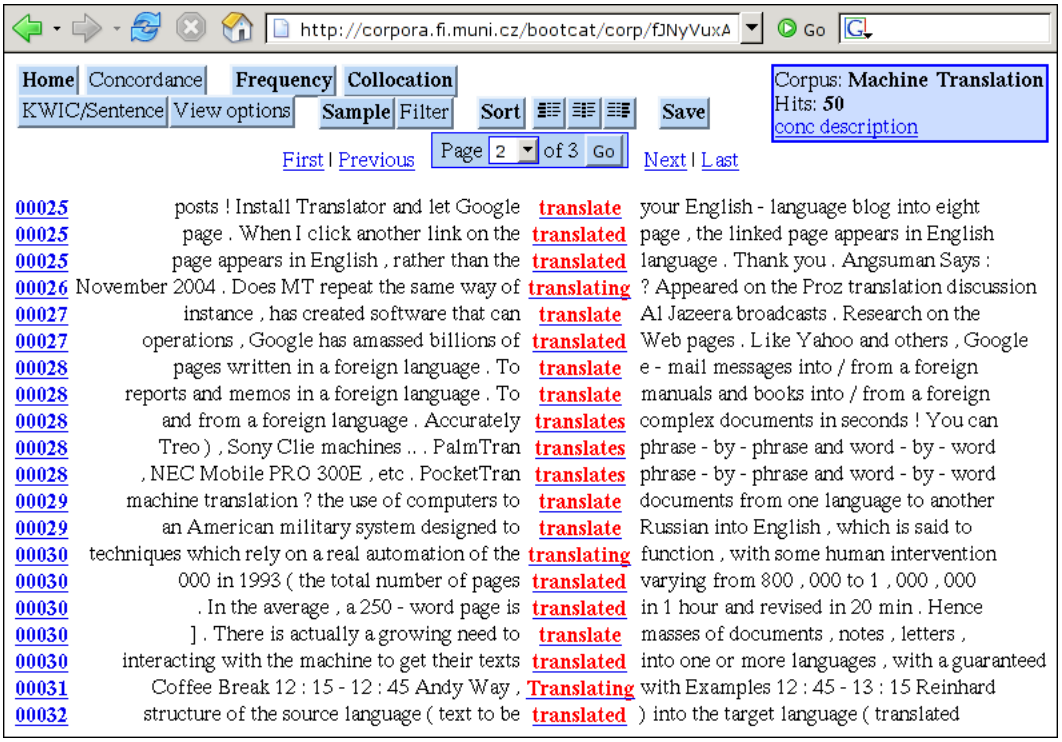

Figure 3: Instant corpus in the Sketch Engine

Figure 4: Key terms for the domain