Web as corpus

Adam Kilgarriff
ITRI, University of Brighton

1. Introduction

The corpus resource for the 1990s was the BNC. Conceived in the 80s, completed in the mid 90s, it was hugely innovative and opened up myriad new research avenues for comparing different text types, sociolinguistics, empirical NLP, language teaching and lexicography.

But now the web is with us, giving access to colossal quantities of text, of any number of varieties, at the click of a button, for free. While the BNC and other fixed corpora remain of huge value, it is the web that presents the most provocative questions about the nature of language. It also presents a convenient tool for handling and examining text.

Compared to LOB, the BNC is an anarchic object, containing 'texts' from 25 to 250,000 words long, screeds of painfully formulaic entries from the Dictionary of National Biography, conversations monosyllabic and incoherent, sermons, pornography and the electronic discourse of the Leeds United Football Club Fan Club.

Compared to the web, the BNC is an English country garden. Whatever perversities the BNC has, the web has in spades. First, not all documents contain text, and many of those that do are not only text. Second, it changes all the time. Third, like Borges's Library of Babel, it contains duplicates, near duplicates, documents pointing to duplicates that may not be there, and documents that claim to be duplicates but are not. Next, the language has to be identified (and documents may contain mixes of language). Then comes the question of text type: to gain any perspective on the language we have at our disposal in the web, we must classify some of the millions of web pages, and we shall never do so manually, so corpus linguists, and also web search engines, need ways of telling what sort of text a document contains: chat or hate-mail; learned article or bus timetable.

These may sound like arguments for *not* studying the web: for scientific progress, we need to fix certain parameters so we can isolate the features we want to look at, and the web is not a good environment for that. This is true. For the web to be useful for language study, we must address its anarchy. If the web is a torrent and nothing more, it is not useful; for it to be useful, we must channel off manageable quantities to irrigate the pastures of scientific and technological progress.

2. The D3CI

We are developing the D3CI (Distributed Data Distributed Collection Initiative) a framework for distributed corpora. This will comprise a set of corpora contributed by anyone with an on-line corpus to offer, where each corpus comes in the form of a set of URLs. The "virtual multicorpus" website will then be a place to visit for anyone wishing to download a corpus of some known language-variety. Corpus measures (Kilgarriff 2001) will be used to identify the homogeneity of each submitted corpus. We shall check for duplicates (Bouyad-Agha and Kilgarriff 1999). We shall provide a program which will go and collect a set of web pages and deliver it to a user. We shall develop links with that part of the WWW community which is examining ways of using links between documents and other strategies to automatically identify interesting clusters of interconnected pages (e.g. Chakrabarti 2000). Our medium-term goal is to set up a suite of web-based corpora that can be used by linguists and language technologists to answer questions of the form: "my theory/algorithm/program works well on the text type I developed it for: I wonder how well it generalises to other text types."

The use of the web addresses the hobgoblin of corpus builders: copyright. If material is on the web, it has been published and can be downloaded without infringing copyright. If I wished to store that material, put it on a CD and distribute that CD, I would be infringing copyright. If I merely present a list of URLs and announce to the world that this URL set comprises a corpus (of a given text type which I also describe) then I am clearly not infringing copyright. There are also no administrative, CD-burning or postage costs associated with web-based corpora.

To the objection that web pages die, so a corpus defined as a set of URLs would be forever shrinking, we propose the following solution. Our virtual corpora are monitored by an agent, which periodically checks that all URLs are still live. On discovering that one no longer is, the agent, which has gathered a statistical profile of each of its pages, sets out to find a new page or pages to replace the deceased. First, it submits a web search, using the terms in the deceased as search terms. This gathers in a set of candidates. Then, using corpus similarity measures, it identifies which of the candidates do in fact have the same linguistic form as the deceased. It then adds them to the corpus. The virtual corpus will evolve.

Some may object, “but that is not suitable as use as a corpus because the texts that are there today are not identical to those that were there yesterday, so how can we compare results?” Results can be compared because the text type is the same. To demand more is to demand that tomorrow’s experiments on the water flow in the River Lune involve the same water molecules as yesterday’s.

3. Related Work

We are not the first to note the web’s usefulness for corpus research, despite its short history. Since the mid-nineties, the net has commonly been used by summarisation researchers as a source of documents to summarise. In this context, Radev and McKeown (1997) use internet-accessible newswire as a knowledge source for a language generation system.

More recently, researchers have used collections of papers found on the web for a very wide range of purposes. Grefenstette and Nioche (2000) and Jones and Ghani (2000) explore the potential of the web as a source of language corpora for languages where electronic resources are in short supply, and Resnik (1999), as a source for bilingual parallel corpora. Fujii and Ishikawa (2000) use the web to generate encyclopaedia entries. Grefenstette (1999) presents prospects and experiments regarding the web as a source of lexical information; as the web provides thousands of contextualised instances of even fairly rare words, for many languages, it offers vast opportunities for automatic distillation of lexical entries from empirical evidence. Varantola (2000) pursues a similar theme, showing how translators, when confronted with a rare term, can find ample evidence of the term, its contexts, and associated vocabulary, through the simple use of a search engine. Specialised ‘lexicographic’ search engines have been produced (see http://www.webcorp.org.uk) though their relative merits compared to, e.g., google (which provides some linguistic context for each occurrence of the word, all at breathtaking speed) remains an open question. Mihalcea and Moldovan (1999) and Agirre and Martinez (2000) use the web as a lexical resource, and as a source of test data, for Word Sense Disambiguation. Jacquemin and Bush (2000) use it as a source for harvesting lists of named entities. There has recently been a Web track in the TREC Information Retrieval Competition (see http://pastime.anu.edu.au/WAR/webtrax.html).

A field such as this, with its newness and no entry costs, is immediately appealing to students and others, and the list above is of course incomplete. It does indicate how the use of the web as a corpus is taking off fast.

4. Conclusion

To conclude: the BNC was one of the greatest innovations for linguistics in the 1990s. Now the world has moved on. As corpus linguists, we are in the fortunate position of having a particular perspective and channel of attack for examining the web --perhaps the most extraordinary phenomenon of our time -- which also just happens to provides solutions to many of our practical problems and an endless stream of new data. We have presented a model which uses the web as a source of data, the web as a delivery medium, and in which the web, and its language, are objects to be explored.

The corpus of the new millennium is the web.

5. References

Agirre E and Martinez D. Exploring automatic word sense disambiguation with decision lists and the web. In *proceedings of COLING Workshop on Semantic Annotation and Intelligent Content*, Saarbruecken, Germany. August 2000.

Bouayad-Agha N and Kilgarriff A. Duplication in Corpora. In *proceedings of Second Computational Lingusitics in the UK Colloquium,* Essex, 1999.

Chakrabarti S. Invited talk. *Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*. Hong Kong. October 2000. (http://www.cse.iitb.ernet.in/~soumen)

Fujii A and Ishikawa T. Utilizing the world wide web as an encyclopaedia: Extracting term descriptions from semi-structured text. In *proceedings of the 38th Meeting of the ACL*, Hong Kong, October 2000, pp. 488-495.

Grefenstette G. The WWW as a Resource for Example-Based MT Tasks. Invited Talk, ASLIB *‘Translating and the Computer’ conference*, London. October 1999.

Grefenstette G and Nioche J. Estimation of English and non-English Language Use on the WWW. In *proceedings of RIAO (Recherche d'Informations Assistee par Ordinateur*), Paris, 2000.

Jacquemin C and Bush C. Combining Lexical and Formatting Clues for named entity acquisition from the web. In *proceedings of Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*. Hong Kong. October 2000, pp. 181-189.

Jones R and Ghani R. Automatically building a corpus for a minority language from the web. *38th Meeting of the ACL, Proceedings of the Student Research Workshop*. Hong Kong. October 2000, pp. 29-36.

Kilgarriff A. 2001 (in press) Comparing Corpora. *International Journal of Corpus Linguistics*.

Mihalcea R and Moldovan D. A method for word sense disambiguation of unrestricted text. In *proceedings of the 37th Meeting of ACL*. Maryalnd, USA, June 1999, pp. 152-158.

Radev D and McKeown K. Building a generation knowledge source using internet-accessible newswire. In *proceedings of the Fifth Applied Natural Language Processing* conference. Washington D. C.., April 1997, pp. 221-228.

Resnik P. Mining the web for bilingual text In *proceedings of the 37th Meeting of ACL*. Maryalnd, USA, June 1999, pp. 527-534.

Varantola K. Translators and disposable corpora. In *proceedings of CULT (Corpus Use and Learning to Translate)*. Bertinoro, Italy. November 2000.