Last Words

Googleology is bad science

Adam Kilgarriff
Lexical Computing Ltd. and
University of Sussex

The web is enormous, free, immediately available, and largely linguistic. As we discover, on ever more fronts, that language analysis and generation benefit from big data, so it becomes appealing to use the web as a data source. The question, then, is how.

The low-entry-cost way to use the web is via a commercial search engine. If the goal is to find frequencies or probabilities for some phenomenon of interest, we can use the hit count given in the search engine's hits page to make an estimate. People have been doing this for some time now. Early work using hit counts included Grefenstette (1999) who identified likely translations for compositional phrases and Turney (2001) who found synonyms; perhaps the most cited study is Keller and Lapata (2003) who established the validity of frequencies gathered in this way using experiments with human subjects. Leading recent work includes Nakov and Hearst (2005) who build models of noun compound bracketing.

The initial-entry cost for this kind of research is zero. Given a computer and a web connection, you input the query and get a hit count. But if the work is to proceed beyond the anecdotal a range of issues must be addressed

Firstly, the commercial search engines do not lemmatise or part-of-speech tag. To take a simple case: to estimate web frequencies for the verb-object pair *fulfil obligation*, Keller and Lapata make thirty-six queries (to cover the whole inflectional paradigm of both verb and noun and to allow for definite and indefinite articles to come between them) to each of Google and Altavista. It would be desirable to be able to search for *fulfil obligation* with a single search. If the research question concerns a language with more inflection, or a construction allowing more variability, the issues compound.

Secondly, the search syntax is limited. There are animated and intense discussions on the CORPORA mailing list, the chief forum for such matters, on the availability or otherwise of wild cards and 'near' operators with each of the search engines, and cries of horror when one of the companies makes changes. (From my reading of the CORPORA list, these changes seem mainly in the direction of offering less metalanguage.)

Thirdly, there are constraints on numbers of queries and numbers of hits per query. Google only allows automated querying via its API, limited to 1000 queries per user per day. If there are thirty-six Google queries per single 'linguistic' query, we can make just twenty-seven linguistic queries per day. Other search engines are currently less restrictive but that may arbitrarily change (particularly as corporate mergers are played out), and also Google has (probably) the largest index, and size is what we are going to the web for.

Fourthly, search hits are for pages, not for instances.

Working with commercial search engines makes us develop workarounds. We become experts in the syntax and constraints of Google, Yahoo, Altavista etc. We become googleologists. The argument that the commercial search engines provide low-cost

© 2006 Association for Computational Linguistics

access to the web fades, as we realise how much of our time is devoted to working with and against the constraints that the search engine imposes.

But science is hard work, and there are usually lots of foothill problems to be mastered before we get to the mountains that are our true goal. So this is all regular science.

Or so it may seem until we consider the arbitrariness of search engine counts. They depend on many specifics of the search engine's practice, including how it handles spam and duplicates (see entries "Yahoo's missing pages" (2005) and "Crazy duplicates" (2006) in Jean Véronis's blog.[1]) The engines will give you substantially different counts, even for repeats of the same query. In a small experiment, queries repeated the following day gave counts over 10% different 9 times in 30, and a factor of two different 6 times in 30. The reasons are that queries are sent to different computers, at different points in the update cycle, and with different data in their caches.

People wishing to use the URLs, rather than the counts, that search engines provide in their hits pages face another issue: the hits are sorted according to a complex and unknown algorithm (with full listings of all results usually not permitted) so we do not know what biases are being introduced. If we wish to investigate the biases, the area we become expert in is googleology not linguistics.

An academic-community alternative

An alternative is to work like the search engines, downloading and indexing substantial proportions of the web, but to do so transparently, giving reliable figures, and supporting language researchers' queries. In Baroni and Kilgarriff (2006) we report on a feasibility study: we prepared web corpora for German ('DeWaC') and Italian ('ItWaC') with around 1.5 billion words each, now loaded into a sophisticated corpus query tool and available for research use.[2] (Of course there are various other large web datasets which research groups have downloaded and are using for NLP.) By sharing good practice and resources and developing expertise, the prospects of the academic research community having resources to compare with Google, Microsoft etc., improves.

Data cleaning

The process involves crawling, downloading, 'cleaning' and de-duplicating the data, then linguistically annotating it and loading it into a corpus query tool. Expertise and tools are available for most of these steps, with the internet community providing crawlers and a de-duplication algorithm (Broder et al. 1997) and the NLP community providing corpus query tools, lemmatisers and POS-taggers for many languages. But in the middle there is a logjam. The questions:

- how do we detect and get rid of navigation bars, headers, footers, ...
- how do we identify paragraphs and other structural information
- how do we produce output in a standard form suitable for further processing?

1 http://aixtal.blogspot.com
2 http://www.sketchengine.co.uk

always arise. Cleaning is a low-level, unglamorous task, yet crucial: the better it is done, the better the outcomes. All further layers of linguistic processing depend on the cleanliness of the data.

To date, cleaning has been done in isolation (and it has not been seen as interesting enough to publish on). Resources have not been pooled, and it has been done cursorily if at all. Thus, a paper which describes work with a vast web corpus of 31 million pages devotes just one paragraph to the corpus development process, and mentions de-duplication and language-filtering but no other cleaning (Ravichandran, Pantel, and Hovy 2005, section 4). A paper using that same corpus notes, in a footnote, "as a preprocessing step we hand-edit the clusters to remove those containing non-English words, terms related to adult content, and other webpage-specific clusters" (Snow, Jurafsky, and Ng 2006). The development of open-source tools which identify and filter out each of the many sorts of 'dirt' found in web pages to give clean output will have many beneficiaries, and the CLEANEVAL project [3] has been set up to this end. There will of course be differences of opinion about what should be filtered out, and a full toolset will provide a range of options as well as provoking discussion on what we should include and exclude, to develop a low-noise, general-language corpus that is suitable for linguistic and language technology research by a wide range of researchers. (In the below, I call the data which meet these criteria "running text".)

How much non-duplicate running text do the commercial search engines index, and can the academic community compare?

While the anti-googleology arguments may be acknowledged, researchers often shake their heads and say "ah, but the commercial search engines index so much data". If the goal is to find frequencies of arbitrary <noun, preposition, verb> and <noun, preposition, noun> triples for PP-attachment disambiguation, then a very, very large dataset is needed to get many non-zero counts. Researchers will continue to use Google, Yahoo and Altavista unless the NLP community's resources are 'Google-scale'. The question this forces is "how much non-duplicate running text do Google and competitors index?"

For German and Italian, we addressed the question by comparing frequency counts for a sample of words in DeWaC and ItWaC with Google frequencies. Thirty words were randomly selected for each language. They were mid-frequency words which were not common words in English, French, German (for Italian), Italian (for German), Portugese or Spanish, with at least five characters (since longer words are less likely to clash with acronyms or words from other languages). For each of these words, Google was searched with a number of parameters:

- with and without "safe search" for excluding adult material
- with language set to German/Italian
- with the "all-in-text" box checked, so that documents were only included as hits if they contained the search term, and
- with and without the site filter set to .it domain only (for Italian), .de or .at domains only for German.

3 http://cleaneval.sigwac.org.uk

Word	max	min	raw	clean
besuchte	10,500	3,800	82	18
stirn	3,380	620	32	11
gerufen	7,140	3,720	67	27
verringert	6,860	3,460	52	16
bislang	24,400	11,600	239	90
brach	4,360	2,260	45	20

Table 1
Comparing Google and Dewac frequencies for a sample of words. 'max' and 'min' are the maximum and minimum from a set of six Google searches. 'raw' and 'clean' are counts for he numbers of documents that the word occurred in in DeWaC, before and after the cleaning, filtering and de-duplication. All numbers in thousands.

	DeWaC/ItWaC	Scaling 1	Scaling 2	% of Google	Estimate
German	1.41 bn	83.5	2.65	3.1	45 bn
Italian	1.67 bn	33.0	2.25	6.8	25 bn

Table 2
Scaling up from DeWaC/ItWaC size to estimate non-duplicate German/Italian running text indexed by Google. Scaling 1 compares Google frequencies with 'raw' DeWac/ItWaC frequencies. Scaling 2 compares 'raw' and 'filtered' DeWaC/ItWaC.

Results were not always consistent, with additional filters sometimes producing an increased hit count so for each word we took the midpoint of the maximum and minimum of the results and compared this number to the DeWaC/ItWaC document frequencies. Here there were two numbers to consider: the count before filtering and cleaning, and the count after. A sample of the results is shown in Table 1.

It would have been convenient to use the Google API but it gave much lower counts than browser queries: a substantial number were one eighteenth as large. Altavista, which has a reputation for NLP-friendliness, was also explored, but since Altavista's index is known to be smaller than Google's, and the goal was to compare with the biggest index available, Altavista results were not going to answer the critical question.

The goal is to use the figures to assess the quantity of duplicate-free, Google-indexed running text for German and Italian. The Google counts are best compared with DeWaC/ItWaC 'raw' counts, and a first scaling factor will give an indication of the size of the Google-indexed German/Italian web inclusive of non-running-text and duplicates. Taking the mid point between maximum and minimum and averaging across words, the ratio for German is 83.5:1 and for Italian, 33:1. A further scaling factor should then be applied, based on the raw:clean ratio, to assess how much of the material is duplicated or not running text. However we do not know to what extent Google applies de-duplication and other rubbish-filtering strategies before calculating counts, and DeWaC/ItWaC filtering and cleaning errs towards rejecting doubtful material. The mean ratio raw:clean is 5.3 for German, 4.5 for Italian: for a best estimate, we halve the figures. Best estimates for the Google-indexed, non-duplicative running text are then 45 billion words for German and 25 billion words for Italian, as summarised in Table 2.

Clearly this is highly approximate, and the notion of running text needs articulation. The point here is that a pilot project (of half a person year's effort) was able to provide

a corpus which was several percent of Google-scale, for two languages. It provides grounds for optimism that the web can be used, without reliance on commercial search engines and, at least for languages other than English, without sacrificing too much in terms of scale.

In sum

The most talked-about presentation of ACL 2005 was Franz-Josef Och's, in which he presented statistical MT results based on a 250 billion word English corpus. His results led the field. He was in a privileged position to have access to a corpus of that size. He works at Google.

With enormous data, you get better results. There are two possible responses for the academic NLP community. The first is to accept defeat: "we will never have resources on the scale of Google, Microsoft and Yahoo, so we should accept that our systems will not really compete, that they will be proofs-of-concept or deal with niche problems, but will be out of the mainstream of high-performance language technology system development." The second is to say: we too need to make resources on this scale available, and they should be available to researchers in universities as well as behind corporate firewalls: and we can do it, because resources of the right scale are available, for free, on the web, and between us we have the skills and the talent.

References

Baroni, Marco and Adam Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of European ACL*, Trento, Italy.

Broder, Andrei Z., Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. 1997. Syntactic clustering of the web. *Computer Networks*, 29(8-13):1157–1166.

Grefenstette, Gregory. 1999. The WWW as a resource for example-based MT tasks. In *ASLIB Translating and the Computer Conference*, London.

Keller, Frank and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Nakov, Preslav and Marti Hearst. 2005. Search engine statistics beyond the n-gram: Application to noun compound bracketing. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 17–24, Ann Arbor, Michigan.

Ravichandran, Deepak, Patrick Pantel, and Eduard Hovy. 2005. Randomized algorithms and NLP: Using locality sensitive hash functions for high speed noun clustering. In *Proceedings of ACL*, Ann Arbour, Michigan, USA.

Snow, Rion, Daniel Jurafsky, and Andrew Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of ACL*, Sydney.

Turney, Peter D. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *European Conference on Machine Learning*, pages 491–502.