Linguistic Search Engine

Adam Kilgarriff
adam@itri.brighton.ac.uk

Synopsis

We propose to build a linguistic search engine, similar in overall design to Google or Altavista but meeting the specifications and requirements of researchers into language.

Language scientists and technologists are increasingly turning to the web as a source of language data, because other resources are not large enough, because they do not contain the types of language the researcher is interested in, or simply because it is free and instantly available. The default means of access to the web is through a search engine such as Google. While the web search engines are dazzlingly efficient pieces of technology and excellent at the task they set themselves, for the linguist they are frustrating. The search engine results do not present enough instances (Google sets a limit of 2000) or enough context for each instance (Google generally provides a ca 10-word fragment), they are selected according to criteria which are, from a linguistic perspective, distorting, and they do not allow searches to be specified according to linguistic functions such as lemmatization and word class.

Each of these goals could straightforwardly be resolved, but approaches to Google to date have gone unanswered. However this suggests a better solution: rather than depend upon existing search engines, it would be far better to set up a linguistic search engine, dedicated to linguists' interests. Then the kinds of processing and querying would be designed explicitly to meet linguists' desiderata, without any conflict of interest or 'poor relation' role. Once this is set up, large numbers of possibilities open out. All those processes of linguistic enrichment and `linguistic data mining' which have been applied with impressive effect to smaller corpora could be applied to the web so that web searches could be specified in terms of linguistically interesting units such as lemmas, word classes, and constituents (e.g. noun phrase) rather than strings. Thesauruses and lexicons could be developed directly from the web. The way would be open for further anatomizing of web text types and domains, both a topic of interest in itself and one where strategies would be needed so that web-based lexical resources could be developed for specific text types or domains, or so that the biases of the web could be countered to provide 'general languages' and 'sublanguage' resources from the web. All of this can potentially be done for all of the many languages for which there is ample data on the web.

The web, teeming as it is with language data, of all manner of varieties and languages, in vast quantity and freely available, is potentially a fabulous linguists' playground. The Linguistic Search Engine will bring that dream closer to reality.

Historical background

As linguistics matures, so the methods it uses turn towards the empirical. It is no longer enough to introspect to gather linguistic insight. Data is required.

For linguistics, the maturation process is intimately linked to the development of the computer. For chemistry or biology, the computer is merely a place to store and process information gleaned about the object of study. For linguistics the object of study itself (in one of its two primary forms, the other being acoustic) is found on computers. Text is an information object, and a computer's hard disk is as valid a place to go for its realization as the printed page or anywhere else.

The one-million word, computer-readable Brown corpus opened the chapter in the early 1960s. Noting the singular needs of lexicography for big data, in the late 1970s Sinclair and Atkins inaugurated the COBUILD project, which raised the threshold of viable corpus size from one million to ten million words (Sinclair 1987). Ten years on, Atkins again took the lead in the development of the British National

Corpus (BNC 1995), which raised horizons tenfold once again, with its 100M words, and was in addition widely available at low cost and covered a wide spectrum of varieties of contemporary British English.

Is 100M large enough?

100M words is large enough for many empirical strategies for learning about language, either for linguists and lexicographers (Baker et al 1998, Kilgarriff and Rundell 2002) or for technologies that need quantitative information about the behaviour of words as input (most notably parsers: see e.g. Briscoe and Carroll 1997, Korhonen et al 2000). However for some purposes it is not large enough. This is an outcome of the Zipfian nature of word frequencies. While 100M is a vast number, and the BNC contains ample information on the dominant meanings and usage-patterns for the 10,000 words that make up the core of English, the bulk of the lexical stock occurs less than 50 times in it, which is not enough to draw statistically stable conclusions about the word. For rarer words, rare meanings of common words, and combinations of words, we frequently find no evidence at all. Researchers are obliged to look to larger data sources (Keller et al 2003, Kilgarriff and Grefenstette 2003). They find that probabilistic models of the language based on very large quantities of data, even if that data is noisy, are better than ones based on estimates (using sophisticated smoothing techniques) from smaller, cleaner datasets.

"General English" is a theoretically difficult notion, and a language can be seen as a modest core of lexis and constructions, plus a wide array of different sublanguages, as used in each of a myriad of human activities. This presents a challenge to corpus dedvelopers: should sublanguages be included? The three possible positions are:

1. no, none should
2. some but not all should
3. yes, all should.

The problem with 1. is that, with all sublanguages removed, the residual core gives an impoverished view of language (quite apart from demarcation issues, and the problem of determining what is left). The problem with 2. is that it is arbitrary. The BNC happens to include cake recipes and research papers on gastro-uterine diseases, but not car manuals or astronomy texts. 3. has not, until recently, been a viable option.

The web

Each step in the ascent of corpora was underwritten by computers becoming orders of magnitude larger (in disk and memory capacity), faster and cheaper. In the last decade, the greatest development in computing has been the web. It contains, at a conservative estimate, 100 billion words of English (1000 times more than the BNC) as well as lesser, but still very large, quantities of many other languages (Grefenstette and Nioche 2000), and conveniently supplies a mechanism for delivering the data it holds. It is large enough for extensive exploration of word combinations, for corpus linguistics for many 'smaller' languages, and to contain very, very many text types and sublanguages. Much of the data is classified, explicitly (with keywords or in Yahoo or similar topic hierarchies) or implicitly (eg through the pages a page links to, which may be classified, or simply through the vocabulary used) and this can support further uses of the data (e.g. Agirre and Martinez 2000).

Existing search engines

More and more people are starting to use the web for language research, for everything from spell-checking to finding academic papers. But the obvious way to use it – through existing search engines – is limiting and frustrating for linguistic research (see above) and can easily lead to wasted effort, distorted results and bad science. Because it is so large, downloading large parts of it is non-trivial. Resnik and Smith (2003) address the issue by using the Internet Archive, a snapshot of the web made available for research, and this is an interesting avenue to pursue, but involves a further organization with its own concerns, priorities and technical and legal constraints.

Our proposal may be contrasted with the approach taken by Webcorp (http://www.webcorp.org.uk). While their goal is similar to ours, they proceed though a meta-search engine which takes user input, converts it into a query for Google and other engines, and then packages the data returned by the engines as a concordance. They remain dependent on other search engines and whatever distortions they introduce. Anecdotal evidence is that the distortions are considerable, with anecdotal frequencies given for "pages containing the words x and y" varying wildly; the search engines' priority is to respond fast, speed mattering far more to most search engine users than the accuracy of frequency counts. Existing search engines cannot be depended on for reliable lexical statistics.

Much web search engine technology has been developed with reference to HLT. The prototype for Altavista was developed in a joint project between Oxford University Press, whose goal was to explore methods for corpus lexicography, and DEC, who used the lexicographic corpus as a very large database requiring fast access for a range of queries. Language identification algorithms, now widely used in web search engines, were developed as HLT technology. This project offers the prospect of a 'homecoming' for web search, with it now feeding the hand that fostered it.

Components

The components of the LSE are

1. web crawler
2. filter/classifier (o)
3. linguistic processor (o)
4. database
5. statistical summariser (o)
6. user interface.

The three items marked 'o' are optional, in that a search engine could meet linguists' needs, to the extent of giving linguistics-oriented web access, without them. However they are central to linguists getting more out of the web, and are key to this proposal.

1. Web Crawler

Setting up a web crawler to crawl the whole web will involve establishing a very high bandwidth web connection, and negotiating for it to bypass all caches.

For regular search engines, it is critical to be up-to-date, so they crawl the web very frequently and this places high performance demands on their hardware. For the LSE, it is not critical that the language data is this year's rather than last, so the web can be crawled slowly: once or twice a year will be adequate. Similarly, regular search engines take great pains to index as much of the web as they possibly can, so people do not fail to find the pages they want. For us, this will not be critical.

2. Filter/classifier

Many web pages do not contain text of the kinds the linguist is interested in, and will be filtered out. They included both obvious cases like images, animations and sound, and less obvious ones like timetables, price lists, pages of pointers and lists of people, places, products, departments etc. (Prior to rejection of the non-obvious cases, procedures will be required to convert documents in html, word, pdf, postscript, rtf and plain ASCII into a standard XML format.) As a general rule, pages and parts of pages that are not "in sentences" will be rejected. A page or text segment can be identified as "in sentences" using a simple regular expression sentence splitter and heuristics (for western languages) such as

> Inter-quartile range for word length falls within the range 3-25 characters
> Inter-quartile range for sentence length falls within the range 4-40 words
> Most words in a sentence comprise only the characters [a-zA-Z].

We shall use a language classifier to identify the language. LSE will be Unicode compliant, so will be usable for languages with Far-Eastern and other non-Roman scripts.

We would also like to implement classifiers for domain, text-type and sublanguage. (There is substantial overlap between these concepts and none are very tightly defined so it is not currently apparent whether this will be one, two or three classification systems.) This is a substantial research task which will involve the identification of relevant inventories of domains, text-types and sublanguages, of feature sets relevant to distinguishing them, and of the potential of training-data-driven as against self-organizing approaches to the classification task. The research here will build on the TypTex and TypWeb projects (Folch et al 2000), amongst others in the extensive document-classification literature. We consider it likely that we shall arrive at a taxonomy of several hundred web text types, each with its own classifier.

After the web pages, as harvested, have passed through the filters and classifiers, the output will be the text chunks of web pages, labeled for language, source URL, and potentially also (at least for English) for one or more of domain, text-type and sublanguage.

3 Linguistic processor

For as many languages as possible, we shall collaborate with language technology experts for that language (and also with experts such as the Xerox group who have produced technologies and applied them to many languages: Beesley and Kartunnen 2003). The language experts will provide tools to lemmatize, part-of-speech-tag and parse the text of that language. This may seem a tall order. It is not, for two reasons. Firstly, the people using the technology will be forgiving. They will be interested to get a first view of, for example, the objects of a verb, or complementation possibilities for a noun: we shall also give them the option of seeing the raw data (see User Interface below). It is not hard to give a starting point for linguistic analysis which is much better than Google, even if it is error-strewn: moderate-accuracy shallow parsing is all that is required. Secondly, speed. Linguistic processing will be in a pipeline applied to all incoming texts. No linguistic processing will be required at run-time. If ten million words can pass through the pipeline per day, then all the textual web will be processed in (very roughly) a year.

We speak on this theme with some confidence because we have processed the BNC in these ways, and have found it very fruitful, all on machines from several years back and at speeds in excess of ten million words a day, in the course of producing word sketches (Kilgarriff and Rundell 2002, op cit; http://wasps.itri.bton.ac.uk).

Other projects (e.g. COLLATE, co-ordinated by DFKI, Saarbrucken) have adopted a similar strategy of gathering language experts for the processing of different languages.

4. Database

Once linguistically processed, the data needs storing in a way that supports fast access. We envisage strategies borrowing from both the search engine world and the corpus world (eg Google's Brin and Page (1998) and the Stuttgart Corpus Tools (Schulze and Christ 1994)). Again, we note that the greater part of the demands and costs for Google stem from the sheer scale (ability to handle a million simultaneous searches) and the commercial obligation for the system to run without downtime. The LSE will, at least for the course of any initial research project, be a research prototype and will not offer guarantees regarding response speed or downtime.

5. Statistical summariser (o)

One great benefit of vast quantities of data is that statistics can be applied to find repeated patterns. Then the points of linguistic interest will emerge as 'signal' above the 'noise' of one-off or uninteresting lexical and grammatical choices, and the linguist no longer needs to trawl through large quantities of data to find significant patterns. A word sketch is a one-page, automatic, corpus-derived summary of a word's grammatical and lexical behaviour (http://wasps.itri.bton.ac.uk). We shall offer word-sketch-like output for common vocabulary for a range of languages, using lexical salience statistics as applied to the output of the shallow parser.

6. User interface

As well as word sketches, we shall offer users access to the full dataset in the form of concordances. Our model will be the Stuttgart Corpus Tools and particularly the search syntax of their corpus query language (now widely used in Corpus Linguistics). This allows for searching on raw word forms, lemma, word class, and combinations of these, both for individual words and for regular expression patterns of words. We shall provide full textual context and also source URLs, so that users wishing to see the entire context for a concordance line can do so (unless the URL happens to have died in the meantime). The recurring theme is that LSE will provide a point of entry to the language resource of the web. Those wishing to investigate a particular point in more detail may need to use LSE simply as step one, for gathering data, then discard the LSE analysis and replace it with their own. This humble role is one which LSE is very happy to play.

Stuttgart tools currently only provide a command line interface. LSE will develop a web-form interface as an alternative, giving most of the power that regular expressions offer to novice users.

Copyright and privacy

LSE will support document deletion. This is important as a response to copyright concerns. With the copyright status of web pages being wildly indeterminate (and it not even being clear that it is legal to cache a web page) it is important for projects to be responsive to individuals asking that web pages be removed from a resource. In addition to observing the ‘no-robots’ conventions when trawling the web, we shall build in procedures for deleting documents where the owner has asked that the document be deleted.

Grid technology

LSE will require very substantial connectivity and computing power. We plan to exploit grid technologies, undertaking a number of tasks in parallel across several machines, probably at multiple sites (particularly for the linguistic processing).

The LSE overall involves a leap of over three orders of magnitude from the BNC to the web. The stepping-up of the scale of the project will take place in stages, with the first full trawl planned to start in the beginning of year three of the project.

We see the LSE as a launching point for a range of further research questions involving language and grid computing. One such is thesaurus creation. We have developed a thesaurus in the context of the WASPS project (http://wasps.itri.bton.ac.uk, see also Lin (1998)) which involved a complex similarity computation for (in principle) all pairs of 40,000 nouns: 0.8 billion similarities. The computation took three weeks. In the course of LSE we plan to explore strategies for thesaurus generation and look at a range of ways in which thesauruses might be used, many of which will be computationally intensive.

Summary

We have presented the plan for a Linguistic Search Engine, which would make the linguistic resource of the web available, in a form designed to meet the needs of linguistic research, to anyone with web access. We believe that, while ambitious, it is perfectly feasible within the parameters of a substantial five-year research project.

References

Sinclair, J (editor) 1987 *Looking Up: An Account of the COBUILD Project in Lexical Computing.* Collins.

BNC 1995. *British National Corpus, User Reference Manual*, Oxford University Computing Service.

Baker, C. F., C. J. Fillmore, J. B. Lowe 1998. *The Berkeley FrameNet Project*. COLING-ACL, Montreal. Pp 8690.

Kilgarriff, A and M. Rundell 2002. *Lexical profiling software and its lexicographical applications - a case study.* EURALEX 2002, Copenhagen.

Briscoe E. J. and J. Carroll 1997. *Automatic Extraction of Subcategorization from Corpora*. Proc 5th ANLP, Washington D. C. Pp 356-363.

Korhonen A., G. Gorrell and D. McCarthy 2000. *Statistical Filtering and Subcategorization Frame Acquisition.* Proc WVLC and EMNLP, Hong Kong. Pp 199-206.

Keller, F., M. Lapata and O. Ourioupina 2003. *Using the Web to Obtain Frequencies for Unseen Bigrams.* Computational Linguistics, submitted.

Kilgarriff, A and G. Grefenstette 2003. *Web as Corpus: Introduction to the Special Issue, Computational Linguistics.* Forthcoming.

Grefenstette G. and J. Nioche 2000. *Estimation of English and non-English Language Use on the WWW.* Proc RIAO, Paris. Pp 237-246.

Agirre E. and D. Martinez 2000. *Exploring Automatic Word Sense Disambiguation with Decision Lists and the Web.* COLING.

Resnik, P. and N. Smith 2003. *The web as a parallel corpus.* Computational Linguistics. Forthcoming.

Folch, Heiden, Habert, Fleury, Illouz, Lafon, Nioche and Prevost 2000. *TypTex: Inductive typological text classification by multivariate statistical analysis for NLP systems tuning/evaluation*. Proc LREC, Athens. Pp 141-148.

Beesley K. and L. Karttunen 2003. *Finite State Morphology*. CSLI, Stanford, Ca.

Brin S. and L. Page 1998. *The anatomy of a large hypertextual web search engine*. Web publication, Stanford University.

Schulze B. and O. Christ 1994. *The IMS Corpus Workbench*. IMS, University of Stuttgart.

Lin D 1998 *Automatic retrieval and clustering of similar words.* COLING/ACL, Montreal. Pp 768-774.