Annotated web as corpus

Paul Rayson
Computing Department,
Lancaster University, UK
p.rayson@lancs.ac.uk

James Walkerdine
Computing Department,
Lancaster University, UK
j.walkerdine@lancs.ac.uk

William H. Fletcher
United States Naval
Academy, USA
fletcher@usna.edu

Adam Kilgarriff
Lexical Computing Ltd., UK
adam@lexmasterclass.com

Abstract

This paper presents a proposal to facilitate the use of the *annotated web as corpus* by alleviating the annotation bottleneck for corpus data drawn from the web. We describe a framework for large-scale distributed corpus annotation using peer-to-peer (P2P) technology to meet this need. We also propose to annotate a large reference corpus in order to evaluate this framework. This will allow us to investigate the affordances offered by distributed techniques to ensure replicability of linguistic research based on web-derived corpora.

1 Introduction

Linguistic annotation of corpora contributes crucially to the study of language at several levels: morphology, syntax, semantics, and discourse. Its significance is reflected both in the growing interest in annotation software for word sense tagging (Edmonds and Kilgarriff, 2002) and in the long-standing use of part-of-speech taggers, parsers and morphological analysers for data from English and many other languages.

Linguists, lexicographers, social scientists and other researchers are using ever larger amounts of corpus data in their studies. In corpus linguistics the progression has been from the 1 million-word Brown and LOB corpora of the 1960s, to the 100 million-word British National Corpus of the 1990s. In lexicography this progression is paralleled, for example, by Collins Dictionaries' initial 10 million word corpus growing to their current corpus of around 600 million words. In addition, the requirement for *mega-* and even *giga-corpora*[1] extends to other applications, such as lexical frequency studies, neologism research, and statistical natural language processing where models of sparse data are built. The motivation for increasingly large data sets remains the same. Due to the Zipfian nature of word frequencies, around half the word types in a corpus occur only once, so tremendous increases in corpus size are required both to ensure inclusion of essential word and phrase types and to increase the chances of multiple occurrences of a given type.

In corpus linguistics building such mega-corpora is beyond the scope of individual researchers, and they are not easily accessible (Kennedy, 1998: 56) unless the web is used as a corpus (Kilgarriff and Grefenstette, 2003). Increasingly, corpus researchers are tapping the Web to overcome the sparse data problem (Keller et al., 2002). This topic generated intense interest at workshops held at the University of Heidelberg (October 2004), University of Bologna (January 2005), University of Birmingham (July 2005) and now in Trento in April 2006. In addition, the advantages of using linguistically annotated data over raw data are well documented (Mair, 2005; Granger and Rayson, 1998). As the size of a corpus increases, a near linear increase in computing power is required to annotate the text. Although processing power is steadily growing, it has already become impractical for a single computer to annotate a mega-corpus.

Creating a large-scale annotated corpus from the web requires a way to overcome the limitations on processing power. We propose distributed techniques to alleviate the limitations on the

[1] See, for example, those distributed by the Linguistic Data Consortium: http://www.ldc.upenn.edu/

volume of data that can be tagged by a single processor. The task of annotating the data will be shared by computers at collaborating institutions around the world, taking advantage of processing power and bandwidth that would otherwise go unused. Such large-scale parallel processing removes the workload bottleneck imposed by a server based structure. This allows for tagging a greater amount of textual data in a given amount of time while permitting other users to use the system simultaneously. Vast amounts of data can be analysed with distributed techniques. The feasibility of this approach has been demonstrated by the SETI@home project[2].

The framework we propose can incorporate other annotation or analysis systems, for example, lemmatisation, frequency profiling, or shallow parsing. To realise and evaluate the framework, it will be developed for a peer-to-peer (P2P) network and deployed along with an existing lexicographic toolset, the Sketch Engine. A P2P approach allows for a low cost implementation that draws upon available resources (existing user PCs). As a case study for evaluation, we plan to collect a large reference corpus from the web to be hosted on servers from Lexical Computing Ltd. We can evaluate annotation speed gains of our approach comparatively against the single server version by utilising processing power in computer labs at Lancaster University and the United States Naval Academy (USNA) and we will call for volunteers from the corpus community to be involved in the evaluation as well.

A key aspect of our case study research will be to investigate extending corpus collection to new document types. Most web-derived corpora have exploited raw text or HTML pages, so efforts have focussed on boilerplate removal and clean-up of these formats with tools like Hyppia-BTE, Tidy and Parcels[3] (Baroni and Sharoff, 2005). Other document formats such as Adobe PDF and MS-Word have been neglected due to the extra conversion and clean-up problems they entail. By excluding PDF documents, web-derived corpora are less representative of certain genres such as academic writing.

[2] http://setiathome.ssl.berkeley.edu/

[3] http://www.smi.ucd.ie/hyppia/, http://parcels.sourceforge.net and http://tidy.sourceforge.net.

2 Related Work

The vast majority of previous work on corpus annotation has utilised either manual coding or automated software tagging systems, or else a semi-automatic combination of the two approaches e.g. automated tagging followed by manual correction. In most cases a stand-alone system or client-server approach has been taken by annotation software using batch processing techniques to tag corpora. Only a handful of web-based or email services (CLAWS[4], Amalgam[5], Connexor[6]) are available, for example, in the application of part-of-speech tags to corpora. Existing tagging systems are 'small scale' and typically impose some limitation to prevent overload (e.g. restricted access or document size). Larger systems to support multiple document tagging processes would require resources that cannot be realistically provided by existing single-server systems. This corpus annotation bottleneck becomes even more problematic for voluminous data sets drawn from the web. The use of the web as a corpus for teaching and research on language has been proposed a number of times (Kilgarriff, 2001; Robb, 2003; Rundell, 2000; Fletcher, 2001, 2004b) and received a special issue of the journal *Computational Linguistics* (Kilgarriff and Grefenstette, 2003). Studies have used several different methods to mine web data. Turney (2001) extracts word co-occurrence probabilities from unlabelled text collected from a web crawler. Baroni and Bernardini (2004) built a corpus by iteratively searching Google for a small set of seed terms. Prototypes of Internet search engines for linguists, corpus linguists and lexicographers have been proposed: WebCorp (Kehoe and Renouf, 2002), KWiCFinder (Fletcher, 2004a) and the Linguist's Search Engine (Kilgarriff, 2003; Resnik and Elkiss, 2003).

A key concern in corpus linguistics and related disciplines is verifiability and replicability of the results of studies. Word frequency counts in internet search engines are inconsistent and unreliable (Veronis, 2005). Tools based on static corpora do not suffer from this problem, e.g. BNCweb[7], developed at the University of Zurich, and View[8] (Variation in English Words and Phrases, developed at Brigham Young University)

[4] http://www.comp.lancs.ac.uk/ucrel/claws/trial.html

[5] http://www.comp.leeds.ac.uk/amalgam/amalgam/amalghome.htm

[6] http://www.connexor.com

[7] http://homepage.mac.com/bncweb/home.html

[8] http://view.byu.edu/

are both based on the British National Corpus. Both BNCweb and View enable access to annotated corpora and facilitate searching on part-of-speech tags. In addition, PIE[9] (Phrases in English), developed at USNA, which performs searches on n-grams (based on words, parts-of-speech and characters), is currently restricted to the British National Corpus as well, although other static corpora are being added to its database. In contrast, little progress has been made toward annotating sizable sample corpora from the web.

"Real-time" linguistic analysis of web data at the syntactic level has been piloted by the Linguist's Search Engine (LSE). Using this tool, linguists can either perform syntactic searches via parse trees on a pre-analysed web collection of around three million sentences from the Internet Archive (www.archive.org) or build their own collections from AltaVista search engine results. The second method pushes the new collection onto a queue for the LSE annotator to analyse. A new collection does not become available for analysis until the LSE completes the annotation process, which may entail significant delay with multiple users of the LSE server. The Gsearch system (Corley et al., 2001) also selects sentences by syntactic criteria from large on-line text collections. Gsearch annotates corpora with a fast chart parser to obviate the need for corpora with pre-existing syntactic mark-up. In contrast, the Sketch Engine system to assist lexicographers to construct dictionary entries requires large pre-annotated corpora. A word sketch is an automatic one-page corpus-derived summary of a word's grammatical and collocational behaviour. Word Sketches were first used to prepare the Macmillan English Dictionary for Advanced Learners (2002, edited by Michael Rundell). They have also served as the starting point for high-accuracy Word Sense Disambiguation. More recently, the Sketch Engine was used to develop the new edition of the Oxford Thesaurus of English (2004, edited by Maurice Waite).

Parallelising or distributing processing has been suggested before. Clark and Curran's (2004) work is in parallelising an implementation of log-linear parsing on the Wall Street Journal Corpus, whereas we focus on part-of-speech tagging of a far larger and more varied web corpus, a technique more widely considered a prerequisite for corpus linguistics research. Curran (2003) suggested distributed processing in terms of web services but only to "allow components developed by different researchers in different locations to be composed to build larger systems" and not for parallel processing. Most significantly, previous investigations have not examined three essential questions: how to apply distributed techniques to vast quantities of corpus data derived from the web, how to ensure that web-derived corpora are representative, and how to provide verifiability and replicability. These core foci of our work represent crucial innovations lacking in prior research. In particular, representativeness and replicability are key research concerns to enhance the reliability of web data for corpora.

In the areas of Natural Language Processing (NLP) and computational linguistics, proposals have been made for using the computational Grid for data-intensive NLP and text-mining for e-Science (Carroll et al., 2005; Hughes et al, 2004). While such an approach promises much in terms of emerging infrastructure, we wish to exploit existing computing infrastructure that is more accessible to linguists via a P2P approach. In simple terms, P2P is a technology that takes advantage of the resources and services available at the edge of the Internet (Shirky, 2001). Better known for file-sharing and Instant Messenger applications, P2P has increasingly been applied in distributed computational systems. Examples include *SETI@home* (looking for radio evidence of extraterrestrial life), *ClimatePrediction.net* (studying climate change), *Predictor@home* (investigating protein-related diseases) and *Einstein@home* (searching for gravitational signals).

A key advantage of P2P systems is that they are lightweight and geared to personal computing where informal groups provide unused processing power to solve a common problem. Typically, P2P systems draw upon the resources that already exist on a network (e.g. home or work PCs), thus keeping the cost to resource ratio low. For example the fastest supercomputer cost over $110 million to develop and has a peak performance of 12.3 TFLOPS (trillions of floating-point operations per second). In contrast, a typical day for the SETI@home project involved a performance of over 20 TFLOPS, yet cost only $700,000 to develop; processing power is donated by user PCs. This high yield for low start-up cost makes it ideal for cheaply developing effective computational systems to realise, deploy and evaluate our framework. The deployment of computational based P2P systems is supported by archi-

[9] http://pie.usna.edu/

tectures such as BOINC[10], which provide a platform on which volunteer based distributed computing systems can be built. Lancaster's own P2P Application Framework (Walkerdine et al., submitted) also supports higher-level P2P application development and can be adapted to make use of the BOINC architecture.

3 Research hypothesis and aims

Our research hypothesis is that distributed computational techniques can alleviate the annotation bottleneck for processing corpus data from the web. This leads us to a number of research questions:

- How can corpus data from the web be divided into units for processing via distributed techniques?
- Which corpus annotation techniques are suitable for distributed processing?
- Can distributed techniques assist in corpus clean-up and conversion to allow inclusion of a wider variety of genres and to support more representative corpora?

In the early stages of our proposed research, we are focussing on grammatical word-class analysis (part-of-speech tagging) of web-derived corpora of English and aspects of corpus clean-up and conversion. Clarifying copyright issues and exploring models for legal dissemination of corpora compiled from web data are key objectives of this stage of the investigation as well.

4 Methodology

The initial focus of the work will be to develop the framework for distributed corpus annotation. Since existing solutions have been centralised in nature, we first must examine the consequences that a distributed approach has for corpus annotation and identify issues to address.

A key concern will be handling web pages within the framework, as it is essential to minimise the amount of data communicated between peers. Unlike the other distributed analytical systems mentioned above, the size of text document and analysis time is largely proportional for corpora annotation. This places limitations on work unit size and distribution strategies. In particular, three areas will be investigated:

- *Mechanisms for crawling/discovery of a web corpus domain* - how to identify pages to include in a web corpus. Also investigate appropriate criteria for handling pages which are created or modified dynamically.
- *Mechanisms to generate work units for distributed computation* - how to split the corpus into work units and reduce the communication / computation time ratio that is crucial for such systems to be effective.
- *Mechanisms to support the distribution of work units and collection of results* - how to handle load balancing. What data should be sent to peers and how is the processed information handled and manipulated? What mechanisms should be in place to ensure correctness of results? How can abuse be prevented and security concerns of collaborating institutions be addressed? BOINC already provides a good platform for this, and these aspects will be investigated within the project.

Analysis of existing distributed computation systems will help to inform the design of the framework and tackle some of these issues. Finally, the framework will also cater for three common strategies for corpus annotation:

- *Site based corpus annotation* - in which the user can specify a web site to annotate
- *Domain based corpus annotation* - in which the user specifies a content domain (with the use of keywords) to annotate
- *Crawler based corpus annotation* - more general web based corpus annotation in which crawlers are used to locate web pages

From a computational linguistic view, the framework will also need to take into account the granularity of the unit (for example, POS tagging requires sentence-units, but anaphoric annotation needs paragraphs or larger). Secondly, we need to investigate techniques for identifying identical documents, virtually identical documents and highly repetitive documents, such as those pioneered by Fletcher (2004b) and shingling techniques described by Chakrabarti (2002).

The second stage of our work will involve implementing the framework within a P2P environment. We have already developed a prototype of an object-oriented application environment to support P2P system development using JXTA (Sun's P2P API). We have designed this environment so that specific application functionality

[10] BOINC, Berkeley Open Infrastructure for Network Computing. http://boinc.berkeley.edu.

can be captured within *plug-ins* that can then integrate with the environment and utilise its functionality. This system has been successfully tested with the development of plug-ins supporting instant messaging, distributed video encoding (Hughes and Walkerdine, 2005), distributed virtual worlds (Hughes et al., 2005) and digital library management (Walkerdine and Rayson, 2004). It is our intention to implement our distributed corpus annotation framework as a plug-in. This will involve implementing new functionality and integrating this with our existing annotation tools (such as CLAWS[11]). The development environment is also flexible enough to utilise the BOINC platform, and such support will be built into it.

Using the P2P Application Framework as a basis for the development secures several advantages. First, it reduces development time by allowing the developer to reuse existing functionality; secondly, it already supports essential aspects such as system security; and thirdly, it has already been used successfully to deploy comparable P2P applications. A lightweight version of the application framework will be bundled with the corpus annotation plug-in, and this will then be made publicly available for download in open-source and executable formats. We envisage our end-users will come from a variety of disciplines such as language engineering and linguistics. For the less-technical users, the prototype will be packaged as a screensaver or instant messaging client to facilitate deployment.

5 Evaluation

We will evaluate the framework and prototype developed by applying it as a pre-processor step for the Sketch Engine system. The Sketch Engine requires a large well-balanced corpus which has been part-of-speech tagged and shallow parsed to find subjects, objects, heads, and modifiers. We will use the existing non-distributed processing tools on the Sketch Engine as a baseline for a comparative evaluation of the AWAC framework instantiation by utilising processing power and bandwidth in learning labs at Lancaster University and USNA during off hours.

We will explore techniques to make the resulting annotated web corpus data available in static form to enable replication and verification of corpus studies based on such data. The initial solution will be to store the resulting reference corpus in the Sketch Engine. We will also investigate whether the distributed environment underlying our approach offers a solution to the problem of reproducibility in web-based corpus studies based in general. Current practise elsewhere includes the distribution of URL lists, but given the dynamic nature of the web, this is not sufficiently robust. Other solutions such as complete caching of the corpora are not typically adopted due to legal concerns over copyright and redistribution of web data, issues considered at length by Fletcher (2004a). Other requirements for reference corpora such as retrieval and storage of metadata for web pages are beyond the scope of what we propose here.

To improve the representative nature of web-derived corpora, we will research techniques to enable the importing of additional document types such as PDF. We will reuse and extend techniques implemented in the collection, encoding and annotation of the PERC Corpus of Professional English[12]. A majority of this corpus has been collected by conversion of on-line academic journal articles from PDF to XML with a combination of semi-automatic tools and techniques (including Adobe Acrobat version 6). Basic issues such as character encoding, table/figure extraction and maintaining text flow around embedded images need to be dealt with before annotation processing can begin. We will comparatively evaluate our techniques against others such as pdf2txt, and Multivalent PDF ExtractText[13]. Part of the evaluation will be to collect and annotate a sample corpus. We aim to collect a corpus from the web that is comparable to the BNC in content and annotation. This corpus will be tagged using the P2P framework. It will form a test-bed for the framework and we will utilise the non-distributed annotation system on the Sketch Engine as a baseline for comparison and evaluation. To evaluate text conversion and clean-up routines for PDF documents, we will use a 5-million-word gold-standard sub-corpus extracted

[11] http://www.comp.lancs.ac.uk/ucrel/claws/

[12] The Corpus of Professional English (CPE) is a major research project of PERC (the Professional English Research Consortium) currently underway that, when finished, will consist of a 100-million-word computerised database of English used by professionals in science, engineering, technology and other fields. Lancaster University and Shogakukan Inc. are PERC Member Institutions. For more details, see http://www.perc21.org/

[13] http://multivalent.sourceforge.net/

from the PERC Corpus of Professional English[14].

6 Conclusion

Future work includes an analysis of the balance between computational and bandwidth requirements. It is essential in distributing the corpus annotation to achieve small amounts of data transmission in return for large computational gains for each work-unit.

In this paper, we have discussed the requirement for annotation of web-derived corpus data. Currently, a bottleneck exists in the tagging of web-derived corpus data due to the voluminous amount of corpus processing involved. Our proposal is to construct a framework for large-scale distributed corpus annotation using existing peer-to-peer technology. We have presented the challenges that lie ahead for such an approach. Work is now underway to address the clean-up of PDF data for inclusion into corpora downloaded from the web.

Acknowledgements

We wish to thank the anonymous reviewers who commented our paper. We are grateful to Shogakukan Inc. (Tokyo, Japan) for supporting research at Lancaster University into the process of conversion and clean-up of PDF to text, and to the Professional English Research Consortium for the provision of the gold-standard corpus for our evaluation.

References

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of LREC2004*, Lisbon, pp. 1313-1316.

Baroni, M. and Sharoff, S. (2005). Creating specialized and general corpora using automated search engine queries. *Web as Corpus Workshop*, Birmingham University, UK, 14th July 2005.

Carroll, J., R. Evans and E. Klein (2005) Supporting text mining for e-Science: the challenges for Grid-enabled natural language processing. In *Workshop on Text Mining, e-Research And Grid-enabled Language Technology at the Fourth UK e-Science Programme All Hands Meeting (AHM2005)*, Nottingham, UK.

[14] This corpus has already been manually re-typed at Shogakukan Inc. from PDF originals downloaded from the web.

Chakrabarti, S. (2002) *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann.

Clark, S. and Curran, J. R.. (2004). Parsing the wsj using ccg and log-linear models. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL '04)*.

Corley, S., Corley, M., Keller, F., Crocker, M., & Trewin, S. (2001). Finding Syntactic Structure in Unparsed Corpora: The Gsearch Corpus Query System. *Computers and the Humanities*, 35, 81-94.

Curran, J.R. (2003). Blueprint for a High Performance NLP Infrastructure. In *Proc. of Workshop on Software Engineering and Architecture of Language Technology Systems (SEALTS)* Edmonton, Canada, 2003, pp. 40 – 45.

Edmonds, P and Kilgarriff, A. (2002). Introduction to the special issue on evaluating word sense disambiguation systems. *Journal of Natural Language Engineering*, 8 (2), pp. 279-291.

Fletcher, W. H. (2001). Concordancing the Web with KWiCFinder. *Third North American Symposium on Corpus Linguistics and Language Teaching*, Boston, MA, 23-25 March 2001.

Fletcher, W. H. (2004a). Facilitating the compilation and dissemination of ad-hoc Web corpora. In G. Aston, S. Bernardini and D. Stewart (eds.), *Corpora and Language Learners*, pp. 271 – 300, John Benjamins, Amsterdam.

Fletcher, W. H. (2004b). Making the Web More Useful as a Source for Linguistic Corpora. In Ulla Connor and Thomas A. Upton (eds.) *Applied Corpus Linguistics. A Multidimensional Perspective*. Rodopi, Amsterdam, pp. 191 – 205.

Granger, S., and Rayson, P. (1998). Automatic profiling of learner texts. In S. Granger (ed.) *Learner English on Computer*. Longman, London and New York, pp. 119-131.

Hughes, B, Bird, S., Haejoong, K., and Klein, E. (2004). Experiments with data-intensive NLP on a computational grid. Proceedings of the International Workshop on Human Language Technology. http://eprints.unimelb.edu.au/archive/00000503/.

Hughes, D., Gilleade, K., Walkerdine, J. and Mariani, J., Exploiting P2P in the Creation of Game Worlds. In the proceedings of ACM GDTW 2005, Liverpool, UK, 8th-9th November, 2005.

Hughes, D. and Walkerdine, J. (2005), Distributed Video Encoding Over A Peer-to-Peer Network. In the *proceedings of PREP 2005*, Lancaster, UK, 30th March - 1st April, 2005

Kehoe, A. and Renouf, A. (2002) WebCorp: Applying the Web to Linguistics and Linguistics to the Web.

World Wide Web 2002 Conference, Honolulu, Hawaii.

Keller, F., Lapata, M. and Ourioupina, O. (2002). Using the Web to Overcome Data Sparseness. *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Philadelphia, July 2002*, pp. 230-237.

Kennedy, G. (1998). *An introduction to corpus linguistics*. Longman, London.

Kilgarriff, A. (2001). Web as corpus. In *Proceedings of Corpus Linguistics 2001*, Lancaster University, 29 March - 2 April 2001, pp. 342 – 344.

Kilgarriff, A. (2003). Linguistic Search Engine. In *proceedings of Workshop on Shallow Processing of Large Corpora (SProLaC 2003)*, Lancaster University, 28 - 31 March 2003, pp. 53 – 58.

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

Mair, C. (2005). The corpus-based study of language change in progress: The extra value of tagged corpora. *Presentation at the AAACL/ICAME Conference*, Ann Arbor, May 2005.

Resnik, P. and Elkiss, A. (2003) The Linguist's Search Engine: Getting Started Guide. *Technical Report: LAMP-TR-108/CS-TR-4541/UMIACS-TR-2003-109, University of Maryland*, College Park, November 2003.

Robb, T. (2003) Google as a Corpus Tool? In *ETJ Journal*, Volume 4, number 1, Spring 2003.

Rundell, M. (2000). "The biggest corpus of all", *Humanising Language Teaching*. 2:3; May 2000.

Shirky, C. (2001) Listening to Napster, in Peer-to-Peer: Harnessing the power of Disruptive Technologies, O'Reilly.

Turney, P. (2001). Word Sense Disambiguation by Web Mining for Word Co-occurrence Probabilities. In *proceedings of SENSEVAL-3*, Barcelona, Spain, July 2004 pp. 239-242.

Veronis, J. (2005). Web: Google's missing pages: mystery solved? http://aixtal.blogspot.com/2005/02/web-googles-missing-pages-mystery.html (accessed April 28, 2005).

Walkerdine, J., Gilleade, K., Hughes, D., Rayson, P., Simms, J., Mariani, J., and Sommerville, I. A Framework for P2P Application Development. *Paper submitted to Software Practice and Experience*.

Walkerdine, J. and Rayson, P. (2004) P2P-4-DL: Digital Library over Peer-to-Peer. In Caronni G., Weiler N., Shahmehri N. (eds.) *Proceedings of Fourth IEEE International Conference on Peer-to-Peer Computing (PSP2004)* 25-27 August 2004, Zurich, Switzerland. IEEE Computer Society Press, pp. 264-265.