GridNLP

Ewan Klein Miles Osborne Lex Holt

Version 1.3, March 18, 2003

Abstract

The vision of Grid technology has motivated a huge effort in developing a Grid architecture which will support large-scale sharing of data and computing resources. Most attention so far has focused either on building infrastructure or on applications in disciplines such as physics, biology, medicine and astronomy. Over the last five years or so, modern Natural Language Processing has evolved into a large-scale endeavour. Consequently, the time is ripe for the international NLP community to participate in building Grid scale services and applications.

1 Introduction

Over the last few years, core activities in Natural Language Processing (NLP) have moved from 'applied research" to 'engineering'. In this paper, we argue that we are undergoing a further transition from small-scale to 'big engineering', and that this needs to be supported at the infrastructure level. NLP has reached a scale — in terms of data, models and computational requirements — that it should rank as a Grid level activity. Consequently, the international NLP community should now be taking steps to build a software and social framework which will support Grid-enabled experiments and collaborative research.

According to Foster et al. [2001], the crucial notion underlying the Grid is "coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations", where resources include computing power and software as well as data. Grid technology is intended to take the concept of the web one stage further by allowing seamless access and use of distributed computing resources as well as information. For example, a query to a Grid search engine should not only find the data but also the necessary data processing techniques and the computing power to carry them out before delivering the results. Within the context of serious NLP, a typical GridNLP application might involve a computational linguist in San Diego carrying out POS tagging using statistical models developed in Edinburgh, applied to a text corpus located in Oxford, and utilising computing power from Manchester. Using the Grid in this manner would enable the computational linguist in San Diego to apply very large models (those containing millions of parameters) to the billions of words of textual material that we have amassed over time. By using models developed in Edinburgh, the computational linguist would be sharing the

intellectual resources present there. All of this would only be possible if NLP was tackled as a Grid problem.

2 Why the Current Infrastructure isn't Good Enough

NLP has substantial and growing resource requirements in these areas:

- data sets,
- modelling,
- annotation,

We look at these in turn.

Data Sets

Most NLP researchers use statistical methods to make progress on some of the hardest problems in natural language processing, including part-of-speech tagging, word sense disambiguation, parsing, machine translation, information retrieval, question-answering and discourse analysis. The goal is to overcome the so-called knowledge-acquisition bottleneck by processing large quantities of data, and estimating various facts that any speaker of the language would be expected to know, e.g. word frequencies, word associations and typical predicate-argument relations.

In an influential paper, Banko and Brill [2001b] suggested that for at least some tasks, the benefits of examining a large enough corpus outweighs the contributions of different learning algorithms, since the learning curves of their four different learners are still rising appreciably at the one billion word mark. This suggests that significant progress in speech and language technologies will only come about when we are able to gather, manage and use extremely large datasets. Inferences about language made with 'small' datasets may not hold when things scale-up. And, as we are really only interested in language as a whole, it is vital that we are not mislead by conclusions that drawn from 'smaller' datasets.

It has often been observed that text is available like never before. It was not all that long ago that researchers referred to the Brown Corpus, consisting of a million words, as a 'large' corpus. Somewhat later, the British National Corpus, consisting of 100 million words, came to be regarded as 'large'. We are now moving to a point where a 'large' corpus is at least a billion words in size, and a very large corpus is the web, which Keller et al. [2002] estimated to be in the region of 50–100 billion words (of English?). The literature on using corpora of one billion words or more is already significant, and includes [Agirre and Martinez, 2000, Banko and Brill, 2001a,b, Grefenstette, 1999, Grefenstette and Nioche, 2000, Jones and Ghani, 2000, Keller et al., 2002, Keller and Lapata, 2003, Mihalcea and Moldovan, 1999, Volk, 2001, Chen and Goodman, 1996, Wu and Khudanpur, 2002, Curran and Osborne, 2002, Dumais et al., 2002]. Efficiently dealing with such large volumes of data cannot be tackled without high-speed disks that are

securely stored. Such needs are unlikely to be served by the average desktop machine typically used to store currently used datasets.

Apart from the practical problems of dealing with a resource the size of the web, there is also the fact that, linguistically, it is very 'noisy'. One symptom of this can easily be ascertained by typing an arbitrary mispelled English word into Google. (For example, *requiments* retrieves 14,600 documents.) Similarly, many of the documents are ungrammatical, and many of the documents in English are written by non-native speakers, or translated by machine, or contain admixtures of other languages. From a statistical point of view, the sample of language represented by the web may be anomalous in various ways. For example, the string *the GNU General Public License is intended to guarantee your freedom to* recovers 15,900 documents from Google, and presumably many of these are identical copies of the GNU GPL.

In many NLP experiments, a linguistic corpus is viewed as an atemporal snapshot. Nevertheless, this perspective is not always appropriate. Curran and Osborne [2002], in analysing a billion word corpus assembled from the North American News Text Corpus and the Reuters Corpus, found that the unigram probability estimates of a number of words failed to converge to their eventual value as the corpus size approached the one billion word total. The reason is that the occurence of some words is bursty, particularly those that are topically related in the newspaper texts. More generally, as noted in [Kleinberg, 2002], it is important to develop appropriate modelling techniques for bursty behaviour in document streams, like email or news articles, that arrive continuously over time. These finding only serve to underline the pressing need to have access to very large datasets.

Finally, many datasets contain copyright material. The current practice is either to license the material to all-and-sundry, or else to totally deny access. However, this ignores the possibility of accessing the data at a Grid level and only consideraing aggregate properties. Such a possibility would increase the chances of the community sharing datasets, but still allowing the owners to manage access.

Modeling

Whilst data is valuable, it is useless unless we can make inferences over it. This means developing *models* of some phenomena. For example, if our data consisted of web pages containing job opportunities, then a model might predict who was the employeer, and what the job category was. Models typical grow at a rate that is proportional to the size of the data. However, since the parameters of these models can be associated with arbitrary subsets of the data, the actual size of the models can be enormous. For example, when developing language models (distributions over sentences), state-of-the-art systems use $5-grams$. This means that potentially, a model might contain on the order of n^5 different parameters, where n is the number of distinct words in a corpus.

Usually, researchers either take steps to prune the number of parameters, or else only use models that are relatively simple (for example, only contain trigrams). These steps are necessary because the sheer size of the models pose hard implementational problems that cannot easily be solved by the average NLP reseacher. Furthermore, some models (for example log-linear models) require iterative constraint satisfaction methods if they are to be estimated. Such approaches add

even more overhead to the size of the model. An inability to deal with large models will handicap progress in the field. Consequently, it is clear that very large models of language will only ever be feasible if we have large enough machines capable of processing them. Moore's Law is unlikely to help here as the volume of data and hence model size increases at a far faster rate than do individual (desktop) computers. This means using Grid computing.

Annotation

In many cases, NLP proceeds by adding various kinds of annotations to raw text,[1] and this can be cascaded; for example, to learn discourse relations from a corpus, you probably need to know about the syntactic structure of sentences in the corpus. More generally, a typical NLP task will involve serial processing by different components; say, tokenization followed by part-of-speech tagging followed by chunking. Each component will produce annotated output that is taken as input by the succeeding component.

Obviously there are resource factors involved in annotating large corpora: speed of processing, memory requirements and storage requirements. Substantial progress in addressing these has been made by the LT-XML processing paradigm. Although there are a large number of XML parsers and transducers available, the LT-XML tools tend to be significantly faster. Second, because the LT-XML tools adopt an event based processing model which streams data at input and output, the memory requirements are constrained, and there is no obligation to store intermediate annotation results.

Now, for some tasks, it may be necessary to store intermediate results. On a naive approach, the annotated copy of a large corpus would be an even larger corpus (excluding the option of applying compression techniques). However, by using stand-off annotation [Thompson and McKelvie, 1997, Ide, 2000], there is no need to copy the input text. Instead, we associate annotation labels with (sequences of) pointers to the original text. Standoff annotation has a further big advantage, namely that different annotators (human or automatic) can markup the same source data in parallel. This is a key component in supporting distributed NLP.

At the moment, "processed" corpora tend to have a brief life span. It's often not that easy to share the results of, eg, one experimental parsing of the BNC within an institute, let alone in the wider research community. But grid middleware is supposed to facilitate transparent, location-neutral access to large datasets, with clear mechanisms for global identification/naming (and access control, etc.). Thus this might make feasible new research angles: eg, I want to try my parser on three versions of some corpus, each one tagged with a different tagger (from three different institutes).

When we consider a task that involves annotating arbitrary text documents on the web, there is clearly an issue about repeatibility. We have no guarantee that the contents of the document will be persistent over time, and therefore no guarantee that a given annotation will remain valid. In general, this is just an inescapable feature of the web, and cannot be solved as such. However, an approach like standoff annotation which uses, say, token offsets from the beginning of the document at least allows us to locate regions of change in the source document.

[1] For an overview of NLP approaches to annotation, see `http://www.ldc.upenn.edu/annotation/`

We also note that annotating data is extremely time-counsuming. However, advances in progress on almost any language problem correlate with increases in the amount of labelled material made available to the reseacher. The usual approach to annotation is employ a few people. This method does not scale very well. A much more ambitious way to annotate material would be to treat it as a distributed, collaborative effort. This approach would require geographically distinct groups of people to work together. As such, it requires a suitable network—a Grid.

3 Challenges for Grid computing

The previous points addressed problems with current approaches to NLP. It is also possible to turn things around, and show that language raises interesting challenges for the Grid community to tackle. For example, as mentioned before, a typical NLP task will involve serial processing by multiple components. By contrast, most current Grid applications are much simpler, typically consisting of a data source, a single processor, and some kind of presentation front end. For this reason, NLP tasks offer immediate scope for testing ideas about the about the configuration and re-configuration of Grid components. There is clearly a need to develop good methods for composing different NLP pipelines in a robust, large scale, fault tolerant way.

We also mentioned the case of intermediate 'processed' corpora. It's critical, of course, that each intermediate corpus has accurate provenance data that tells you about its dependencies: its 'prior' corpus, and the tool used to produce it. In general, there's an interesting dependency graph here, across these various datasets, and it's not clear that any existing grid technology has nice ways to abstract and manipulate this.

4 Virtual Organisations

A virtual organisation (VO) is a collection of resource providers and consumers who observe a set of rules which dictate the conditions for sharing resources. As mentioned earlier, 'resources' here covers not just data but also computers and software. So far, the only instances of such virtual organisations within the NLP community are very restricted. For example, exception of LDC and ELRA adopt a narrow interpretation of resources as corpora (written and spoken) (and possibly lexicons?), where there is a single service provider involved in bilateral relationships with a large number of consumers.

We can see the beginning of an *ad hoc* VO in events like the 6-week summer workshops held by the Center for Language and Speech Processing at Johns Hopkins University. In such cases, a group of collaborating researchers may draw on a variety of resources across multiple sites in order to accomplish a shared task.

5 Grid Activities

What kind of steps have to be taken by an organisation to contribute a simple resource as a Grid service within a VO? We would expect to build on the Globus Toolkit (cf. `http://www.globus.`

org/), a set of software components that can be used to build Grid applications. Then the main tasks would be the following.

Gatekeeper: Each resource contributor in the VO would set up a machine as a gatekeeper; that is, as their public interface to the local computing resource. (The Globus Toolkit would be installed on the gatekeeper.)

Certificates: Access to the VO wouldl be controlled by certificates issued by a Certification Authority (CA). Initially, we could use the UK e-Science CA at Rutherford Appleton Labs. Subsequently, it might be appropriate to set up a specialised NLP CA. Each gatekeeper would be configured to accept certificates from members of the VO.

Job Brokering: Each gatekeeper will host a Resource Broker. This has the task of finding appropriate resources, arranging for job to be locally executed, and tracking their progress of jobs. Typically, it would be necessary to write software to ensure that local resources are made available to the Resource Broker.

This of course is just a beginning. Ideally, we would like to reach a situation where an NLP task could be specified declaratively, without necessarily having to list all the resources employed, and configuration software would assemble available resources to carry out the task. In order to pursue this avenue, a number of challenges would have to be addressed. The most important involve the question of how NL resources (in the sense of processing components and static knowledge resources) should be described as services. This general issue is receiving increasing attention within the Semantic Web community, and a framework like DAML-S—Semantic Markup for Web Services [Coalition, 2002]—will be an important contribution.

6 Summary

The time has come for NLP to move away from being seen as an amateur activity of small groups of people working in isolation, bounded by their abilities and the local availability of computing power. Instead, we need to realise the size of the problems that face us and take action: large problems are only tackled by a concerted, co-ordinated effort of researchers. The time of developing small systems is over. The future of research in NLP must be seen as a Grid activity.

References

Eneko Agirre and David Martinez. Exploring automatic word sense disambiguation with decision lists and the web. In *Proceedings of the 18th*, Saarbrücken/Luxembourg/Nancy, 2000.

Michele Banko and Eric Brill. Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing. In James Allan, editor, *Proceedings of the 1st International Conference on Human Language Technology Research*, San Francisco, 2001a. Morgan Kaufmann.

Michele Banko and Eric Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Toulouse, 2001b.

Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In Aravind Joshi and Martha Palmer, editors, *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310–318, San Francisco, 1996. Morgan Kaufmann Publishers. URL `citeseer.nj.nec.com/stanley98empirical.html`.

The DAML Services Coalition. DAML-S: Semantic markup for web services. Technical report, DARPA, 2002. URL `http://www.daml.org/services/daml-s/0.7/`.

James R. Curran and Miles Osborne. A very very large corpus doesn't always yield reliable estimates. In Dan Roth and Antal van den Bosch, editors, *Joint Proceedings of CoNLL02 and Workshop on Very Large Corpora*, pages 126–131, Taipei, Taiwan, 2002.

S. Dumais, M. Banko, E. Brill, J. Lin, and A. Ng. Web question answering: Is more always better? In *Proceedings of SIGIR'02*, pages 291–298, August 2002.

Ian Foster, Carl Kesselman, and Steven Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. *International Journal of Supercomputer Applications*, 15(3), 2001.

Gregory Grefenstette. The World Wide Web as a resource for example-based machine translation tasks. In *Proceedings of the ASLIB Conference on Translating and the Computer*, London, 1999.

Gregory Grefenstette and Jean Nioche. Estimation of English and non-English language use on the WWW. In *Proceedings of the RIAO Conference on Content-Based Multimedia Information Access*, pages 237–246, Paris, 2000.

Nancy Ide. The XML framework and its implications for corpus access and use. In *Proceedings of Data Architectures and Software Support for Large Corpora*, pages 28–32, Paris, 2000. European Language Resources Association.

Rosie Jones and Rayid Ghani. Automatically building a corpus for a minority language from the web. In *Proceedings of the Student Research Workshop at the 38th*, pages 29–36, Hong Kong, 2000.

Frank Keller and Maria Lapata. Using the web to obtain frequencies for unseen bigrams. to appear, 2003.

Frank Keller, Maria Lapata, and Olga Ourioupina. Using the web to overcome data sparseness. In Jan Hajič and Yuji Matsumoto, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Philadelphia, 2002.

Jon Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SiGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

Rada Mihalcea and Dan Moldovan. A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th*, pages 152–158, University of Maryland, College Park, 1999.

Henry S. Thompson and David McKelvie. Hyperlink semantics for standoff markup of read-only documents. In *Proceedings of SGML Europe '97*, Barcelona, May 1997. URL `http://www.ltg.ed.ac.uk/ht/sgmleu97.html`.

Martin Volk. Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie, and Shereen Khoja, editors, *Proceedings of the Corpus Linguistics Conference*, pages 601–606, Lancaster, 2001.

Jun Wu and Sanjeev Khudanpur. Building A topic-dependent maximum entropy model for very large corpora. In *Proceedings of ICASSP2002*, Orlando, USA, May 2002. URL `citeseer.nj.nec.com/493729.html`.