New Model Corpus

ABSTRACT
We introduce the New Model Corpus (NMC), a balanced 100m-word corpus of general English, freely available to all. It is a response to the fact that the British National Corpus has served as a hugely valuable stimulus to progress in many aspects of language technology and linguistics, but that it is now old, both in terms of the language it contains and its design model. The NMC draws all material from the web (making it a much faster and cheaper operation than the BNC to create). We aim to make the NMC a collaborative project where all NLP groups are encouraged to apply their NLP tools to it and return the resulting markup to us: we shall then integrate the markup to give a publicly-available, multiply-marked-up version of the resource (all available by web API for further exploration). The NMC is also a manageable-sized sample and model for a much larger web corpus we are building. The NMC will be publicly launched at or shortly before ICON 2009.

Introduction

Innovative resources that are large-scale, well-designed and widely available have played a central role in the development of language technologies over the last two decades. The Penn Treebank, WordNet and the British National Corpus (BNC) have been central to progress in the field. Looking particularly at the BNC, we note that it has made many things possible but also that, in 2009, it is dated: it is a pre-web resource, it contains no blogs and no hits for *blog*, it is of modest size (by 21st century standards; by 20th-century standards it was huge).

The New Model Corpus (NMC) is a 100m word corpus of English, with all data drawn from the web. It is freely available, WordNet-style, to all.

It is a model in two senses of the word. Firstly it is a model in the sense of having a design that others might adopt (cf Atkins et al (1993) for the BNC's design model). Secondly it is a model in the way that a child's model train is a model. It is a 1:100 scale model of BiWeC (Big Web Corpus), a ten billion word corpus that we are currently developing. We are using NMC as a prototype, to work out all desired characteristics and issues before scaling up.

A collaborative project

We are inviting others to contribute markup. This could be markup at several levels:

- the document level: if, for example, you have a document-domain classifier, or classifier by regional variety of English, you could run it over NMC and contribute a domain or region label to each constituent document.
- the lexical level: if you had a new POS-tagger or a WSD system, you could run that over the data and contribute new POS-tags or sense-labels for all or some words.
- the linguistic-structure level: if you have a named-entity tagger or noun phrase chunker or discourse-element chunker, then you can run that and propose structural units, or add attribute-value pairs to existing or new structural units

We shall distribute the corpus as plain text (with a variable amount of associated markup). We shall use a standoff markup model.

The resulting multiply-annotated corpora will then support research into the interaction between the different kinds of annotation, as well as giving researchers a large, well-balanced corpus, in a leading corpus query tool (anon), to do research on (and debugging of) their annotation.

The multiply-annotated corpus will be made available in three ways: to download, as a web service with a user interface for people, or as a web service via our corpus system's API. The first will be free, the second and third will be free to people engaged in the project, as suppliers of annotation or similar.

BiWeC will be a collaborative project with slightly different characteristics. We shall review the annotations on NMC (both the ones done by our own 'home team', and those undertaken by others) and identify what is useful, of good accuracy, scaleable, and makes a coherent overall set of annotations. We and collaborators will then mark up BiWeC with them.

This aspect of the project was developed following a reading of "Wikinomics" (Tapscott and Williams 2007).

Data sampling
Sample sizes will usually be small, with a default maximum of 1000 words taken from each web page. One reason for doing this is to reduce the risk of copyright-holders of web pages objecting to this use of their data (though, in the age of Google, this risk is small, and we are willing to bear it).

The corpus is comprised as follows:

- General crawl 50m
- Targeted crawls
    - Fiction 7m
    - Blog 7m
    - Newspaper RSS feeds 7m
    - Speech 10m
        - (Film transcripts, chat show transcripts)
    - Domain-specific 19m
        - Business, Law., Medical

TOTAL **100m**

Notes
We have state-of-the-art methods for both de-duplication (anon) and data cleaning (e.g., following the model of the CLEANEVAL initiative (Chantree et al 2008). We believe the corpus to be a model web corpus in both these regards.
General crawl: methods as described in Ferraresi et al (2008).
Fiction: all material taken from the Gutenburg project. Our experience with finding contemporary fiction on the web is that the skew toward science fiction and science fantasy is marked, and we suspect those genres will already be well covered in the general crawl. In contrast to all other parts of the corpus (with the possible exception

of film-transcripts) this will be older rather than current language, as the website hosts mainly out-of-copyright texts. We believe there is a case for including this material in a largely contemporary corpus since these texts are still being read: they are part of language input in 2009 even though they are not part of language output. Default maximum sample size per text for this component is 7000 words.
Blog: Not more than 5000 words from any single blog.
Newspaper: Not more than 100,000 words from any single newspaper.
Film transcripts: drawn from http://www.opensubtitles.com

References

S Atkins, J Clear, N Ostler (1992) Corpus Design Criteria. Literary and linguistic computing 7(1):1-16.

Marco Baroni, Francis Chantree, Adam Kilgarriff and Serge Sharoff 2008 CleanEval: a competition for cleaning web pages. Proc LREC. Marrakech, Morocco.

Marco Baroni, Adam Kilgarriff, Jan Pomikalek, Pavel Rychly 2006 WebBootCaT: a web tool for instant corpora *Proc. Euralex.* Torino, Italy.

Ferraresi, A., E. Zanchetta, S. Bernardini and M. Baroni 2008. Introducing and evaluating UKWaC, a very large web-derived corpus of English . Proceedings, 4th WAC workshop, LREC, Marrakech, Morocco.

Don Tapscott and Anthony Williams 2007. Wikinomics: How mass collaboration changes everything. Atlantic Books, London.