

Cleaneval: a competition for cleaning web pages

Marco Baroni, Francis Chantree, Adam Kilgarriff, Serge Sharoff

Univ Trento, Lexical Computing Ltd, Lexical Computing Ltd, Leeds Univ.

Abstract

Cleaneval is a shared task and competitive evaluation on the topic of cleaning arbitrary web pages, with the goal of preparing web data for use as a corpus for linguistic and language technology research and development. The first exercise took place in 2007. We describe how it was set up, results, and lessons learnt.

Introduction

“Cleaning” web pages is increasingly a bottleneck for language technology (LT). More and more LT research and development uses the web as its data source. The following questions always arise:

- how do we detect and get rid of navigation bars, headers, footers and other textual data of no linguistic interest
- how do we identify paragraphs and other structural information
- how do we produce output in a standard form of regular text suitable for further linguistic processing?

It is a low-level, unglamorous task and yet it is increasingly 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 by each group using web data (and it has not been seen as interesting enough to publish on). Resources have not been pooled, and it has often not been done well. In Cleaneval we put cleaning centre-stage. The goals of the exercise are to identify good strategies (for each of the many different kinds of ‘dirt’ found in web pages) and to foster sharing of ideas and programs.

It takes the form of an open competition: who can do the best job of cleaning arbitrary web pages? It may seem odd to foster collective effort through competition, but the evidence from a number of LT competitions is that it works: the process of setting up the exercise precipitates discussion about the critical questions in the field, the ‘game’ aspect brings in additional participants including junior ones, and the whole exercise sets benchmarks which then become common reference points. It also supports progress, with the field as a whole benefiting from the leading technologies as identified in the competition, and because future rounds of the exercise can build on previous ones. For discussions of the approach and its benefits see e.g. Gaizauskas (1998), Belz and Kilgarriff (2006).

The stages of the process are:

- Announce the overall theme of the evaluation and invite people to participate
- Identify data; divide between development set and test set
- Employ people to produce sets of correct answers (the “gold standard”)
 - o Distribute development set (with correct answers)
- Develop scoring software

- Distribute test data (without correct answers)
- Participants process data, submit their system's answers
- Organisers score participants' systems
- Workshop

A first Cleaneval exercise was held in summer 2007, with the workshop in September (combined with the third Web-as-Corpus workshop of ACL's SIGWAC, under whose auspices Cleaneval was organised). We addressed two languages, English and Chinese: English, because it is the largest and most important on the web, and Chinese, firstly, as evidence that we were not blindly anglocentric, and secondly, to explore the issues that a language with a different character set and no word-break character presented. In this paper we describe the preparation of the data, scoring, and results. For descriptions of participating systems see Fairon et al (2007).

Data preparation

Data selection

The basic unit was the web page. For the exercise we used a random sample of pages from internet corpora that had already been developed for English and Chinese as described in Sharoff (2006). Thus the data samples carry the imprint of the choices made in the development of those corpora, for example the crawling method and exclusion of pages that were too long, or too short, or presented *prima facie* evidence of not containing usable text. While the method for data selection is open to challenge, cleaning techniques will always be applied to pages which are the output of corpus-collection strategies, and the corpus-collection strategies behind the corpora we used are documented and reasonably generic. For Cleaneval-1 we used only html pages.

We divided the data into development set and evaluation set, as shown, with data sizes, in Table 1.

	Dev EN	Dev ZH	Test EN	Test ZH
Files	57	60	684	653
KBytes	1892	1943	10701	9845

Table 1: Data sets

In most shared task exercises, the ratio of training to evaluation data has been higher. This is typically because the organisers want to support machine-learning methods. We imagined that these kinds of methods would not be the most suitable for this task, as there are so many different varieties of 'dirt' to be cleaned. However in the event several systems did use ML methods and, for the next exercise, it is likely that effort will be put into preparing a large training set.

Practical arrangements

We recruited 23 Masters students in Computer-Assisted Translation at the University of Leeds, including some Chinese native speakers who worked on the Chinese data. (All had a sufficient level of English to work on English.).

The taggers worked with two windows open, one showing the page as it was designed to be seen, in a browser, and the other showing a pre-cleaned version of the page, in a plain-text editor (by default NotePad++). The pre-cleaning used lynx: it removed html markup, javascript and other clearly unwanted material. It also converted all pages to UTF-8. (The script was made available to participants.)

At the workshop there were criticisms that this method meant that the link between the original markup and the retained material was lost. We accept the criticism. At the time we needed to set up arrangements that allowed the taggers to start work promptly, on dependable software: there was neither the time nor the money to set up an editor that retained the html markup in a way that did not make the editing much more cumbersome for the taggers.

The taggers were instructed as follows:

Your task is to "clean up" a set of webpages so that their contents can be easily used for further linguistic processing and analysis. In short, this implies

1. removing all HTML/Javascript code and "boilerplate" (headers, copyright notices, link lists, materials repeated across most pages of a site, etc.);
2. adding a basic encoding of the structure of the page using a minimal set of symbols to mark the beginning of headers, paragraphs and list elements.

This is the opening of the annotation guidelines (available at http://cleaneval.sigwac.org.uk/annotation_guidelines.html); the guidelines comprise two pages, and include examples. Given that this was the first exercise of its kind, and that time was limited, we chose to work on the basis that students would usually be able to follow these instructions in a systematic and replicable way, rather than legislating in the guidelines for many different cases.

All files in the development set were annotated by two people. The rationale was to test the reliability of our assumptions on what can be considered as boilerplate. There was some variation in how the annotators interpreted the guidelines, but the scores comparing two versions of manually cleaned files were on average 94%, much higher than the best systems taking part in the competition.

Taggers cleaned an average of 120 kB/hour for English, 50kB per hour for Chinese (and were paid a rate per kB of data cleaned). The tagging was completed in March 2007, funded by a small grant from Lexical Computing Ltd. The funding available defined the quantity of tagging that could take place. We prioritised "quantity of data" over double-tagging, and, as noted above, we only double-tagged the development sets.

Scoring

The scoring program needed to measure the similarity between two files. For the actual evaluation, one would be the participant's and the other, the gold standard version, though the program could be used for comparing any two versions. It needed to be able to deliver both "text only" scores, which ignored the markup of paragraphs, headers and lists, and "text with markup" scores, which took them into account.

The scoring was based on Levenshtein edit distance: the smallest number of 'insert word' and 'delete word' steps to get from the one text to the other. (This was satisfactory for English; for Chinese see below.) Prior to calculating edit distance, we normalised both files by lower-casing, deleting punctuation and other non-alphanumeric characters and normalising whitespace.

To check the validity of the scoring program (with a variety of parameter settings) we took several pairs of human annotations for the same input, and manually classified them according to whether the first file was better-cleaned, worse-cleaned, or about the same as the second.

We then ensured that the scorer that we used agreed with these human-expert judgements. Looked at another way, we could say that writing the scorer involved operationalising what we meant by “good cleaning”.

Participants and Results

For Chinese only one participating system, from University of Osnabrück, returned results in the format suitable for the script (other submissions had problems with either encodings or file format). The system performance was 18%, although it is likely that the low figure results from the scoring algorithm not aligning on appropriate units: this was not resolved by the time of the workshop.

For English there were nine participants, from four continents and from both academia and one company. The participants and their results are shown in Table 2. The results given “Text and Markup” (TM) , “text-only” (TO) and the Average (Ave).

Students	TM	TO	Ave	Non-students	TM	TO	Ave
Bauer et al, Osnabrück Uni, Germany	53.5	73.5	63.5	Gao & Abou-Assaleh, GenieKnows, Canada	63.9	83.4	73.6
Marek, Pecina & Sprousta; Charles Uni, Czech Republic	65.3	84.1	74.7	Girardi, IRST, Italy	65.6	82.5	74.0
Hofmann and Weerkamp, Uni Amsterdam, NL	65.5	83.0	74.2	Saralegi & Leturia, Elhuyar Foundation, Spain	65.3	83.4	74.3
Chaudhury, India	59.5	80.9	70.2	Evert, Osnabrück Uni, Germany	60.3	82.9	71.6
Conradie, North West Uni, South Africa	45.5	60.2	52.9				

Table 2: Participants and results.

The results are remarkably close. Except for two student outliers, all ‘average’ results are between 70% and 75%, with the four highest-scoring systems all between 74% and 75%. Adding the markup correctly was substantially harder than simply finding which text to retain, as shown by TO scores being around 20% higher than TM results.

Discussion, lessons learnt, and way forward

We have completed a first run of a Cleaneval exercise. It ran satisfactorily. Several points emerged in the discussions at the workshop. Counter to our expectations participants were largely interested in using machine learning techniques, and for this a larger training set is required. Also most systems used the HTML structure of the input page as an input to their algorithms, and it is desirable that the gold standard preparation is done in a way that adds extra markup into the original text to say where “good text” begins and ends.

There was a high level of interest on there being a further exercise; this should include more languages and should cover pdf as well as html documents. It should also address POS-tagging: POS taggers which have not been developed for web data typically perform badly on it, so it is interesting to add another (optional) task to the exercise which aims to support the development of POS-taggers which perform well on web data. A volunteer to run the next exercise was also discovered.

References

- Belz, A. and A. Kilgarriff 2006. [Shared task evaluations for HLT: lessons for NLG](#). *Proc. Intl Natural Language Generation Conference*. Sydney.
- Fairon, C., H. Naets, A. Kilgarriff and G-M. de Schryver, eds. 2007. Building and Exploring Web Corpora: Proceedings of the third Web-as-Corpus workshop, incorporating Cleaneval. Presses universitaires de Louvain, Louvain-la-Neuve, Belgium.
- Gaizauskas, R., ed. 1998. *Computer Speech and Language*, 12 (3) Special Issue on Evaluation of Speech and Language Technology.
- Sharoff, S. 2006. [Creating general-purpose corpora using automated search engine queries](#). *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna.