`langid.py` for better language modelling

Paul Cook[♡] **and Marco Lui**[♡♣]
♡ Department of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia
♣ NICTA Victoria Research Laboratory
{paulcook,mhlui}@unimelb.edu.au

Abstract

Large corpora are crucial resources for building many statistical language technology systems, and the Web is a readily-available source of vast amounts of linguistic data from which to construct such corpora. Nevertheless, little research has considered how to best build corpora from the Web. In this study we consider the importance of language identification in Web corpus construction. Beginning with a Web crawl consisting of documents identified as English using a standard language identification tool, we build corpora of varying sizes both with, and without, further filtering of non-English documents with a state-of-the-art language identifier. We show that the perplexity of a standard English corpus is lower under a language model trained from a Web corpus built with this extra language identification step, demonstrating the importance of state-of-the-art language identification in Web corpus construction.

1 The need for large corpora

Corpora are essential resources for building language technology (LT) systems for a variety of applications. For example, frequency estimates for n-grams — which can be used to build a language model, a key component of many contemporary LT systems — are typically derived from corpora. Furthermore, bigger corpora are typically better. Banko and Brill (2001) show that for a classification task central to many LT problems, performance increases as a variety of models are trained on increasingly large corpora.

The Web is a source of vast amounts of linguistic data, and the need for large corpora has motivated a wide range of research into techniques for building corpora of various types from the Web (e.g., Baroni and Bernardini, 2004; Ferraresi et al., 2008; Kilgarriff et al., 2010; Murphy and Stemle, 2011). In stark contrast to manual corpus construction, such automatic methods enable large corpora to be built quickly and inexpensively. Moreover, large Web crawls have recently been produced which are readily-available to the LT community (e.g., ClueWeb09[1] and CommonCrawl[2]) and can easily be exploited to build corpora much larger than those currently available (and indeed Pomikálek et al. (2012) have already done so); based on the findings of Banko and Brill, such corpora could be exploited to improve LT systems.

Despite the importance of large Web corpora, the issue of how to best derive a corpus from a Web crawl remains an open question. Once a large collection of documents is obtained (from, e.g., either a Web crawl or the results of issuing queries to a commercial search engine) they must be post-processed to remove non-linguistic document portions, for example, boilerplate text such as menus; filter unwanted content such as documents in languages other than that intended for the corpus, and spam; and finally remove deduplicate or near-duplicate documents or document portions to produce a corpus. Furthermore, this document post-processing can potentially have a tremendous impact on corpus quality (Kilgarriff, 2007). For example, if texts in languages other than the target language(s) are not reliably identified and removed, n-gram frequency estimates for the target language will be less accurate than they would otherwise be, potentially having a negative

[1] http://lemurproject.org/clueweb09/
[2] http://commoncrawl.org/

Paul Cook and Marco Lui. 2012. langid.py for better language modelling. In *Proceedings of Australasian Language Technology Association Workshop*, pages 107−112.

impact on LT systems trained on such a corpus. Similar problems are encountered with the presence of boilerplate text, and duplicate or near-duplicate documents or text segments.

Although document post-processing is clearly important to corpus construction, little work has studied it directly, with the notable exception of CleanEval (Baroni et al., 2008), a shared task on cleaning webpages by removing boilerplate and markup. Liu and Curran (2006) and Versley and Panchenko (2012) compare Web corpora with standard corpora in task-based evaluations, but do not specifically consider the impact of document post-processing. Web corpus construction projects have tended to rely on readily-available tools, or simple heuristics, to accomplish this post-processing. This is not a criticism of these projects — their goals were to build useful language resources, not specifically to study the impact of document post-processing on corpora. Nevertheless, because of the immediate opportunities for improving LT by building larger Web corpora, and the importance of post-processing on the quality of the resulting corpora, there appear to be potential opportunities to improve LT by improving Web corpus construction methods.

In this paper we consider the importance of language identification — which has already been shown to benefit other LT tasks (e.g., Alex et al., 2007) — in Web corpus construction. We build corpora of varying sizes from a readily-available Web crawl (the English portion of ClueWeb09) using a standard corpus construction methodology. This dataset contains only documents classified as English according to a commonly-used language identification tool (TEXTCAT).[3] We then produce versions of these corpora from which non-English documents according to a state-of-the-art language identification tool (`langid.py`, Lui and Baldwin, 2012) are filtered. In this preliminary work, we measure the impact of language identification in a task-based evaluation. Specifically, we train language models on the Web corpora, and demonstrate that, for corpora built from equal amounts of crawl data, the perplexity of a standard (manually-constructed) corpus is lower under a language model trained on a corpus filtered using `langid.py`, than a model trained on a corpus without this filtering.

[3] `http://odur.let.rug.nl/vannoord/TextCat/`

2 Materials and methods

This section describes the language identification tools, corpus construction methods, and language modelling approach used in this study.

2.1 Language identification

The initial language identification for ClueWeb09 was performed using TEXTCAT, an implementation of the language identification method of Cavnar and Trenkle (1994),[4] which is based on the relative frequencies of byte n-grams. The reported language identification precision is over 99.7% across all 10 languages in ClueWeb09. However, the method of Cavnar and Trenkle has been shown to perform poorly when applied to test data outside the domain of the training data (Lui and Baldwin, 2011), as was the case for ClueWeb09 where the training data was drawn from newswire and European parliament corpora.

`langid.py` is an implementation of the method described in Lui and Baldwin (2011), which improves on the method of Cavnar and Trenkle (1994) in a number of ways; both classifiers are based on relative frequencies of byte n-grams, but `langid.py` uses a Naive Bayes classifier and cross-domain feature selection, allowing it to ignore non-linguistic content such as HTML, without the need to explicitly model such content. Lui and Baldwin (2012) show that `langid.py` significantly and systematically outperforms TEXTCAT on a number of domains, and we therefore use it in this study.

2.2 Corpora

We build corpora from subsets of the English portion of ClueWeb09, a Web crawl consisting of roughly 500 million webpages crawled from January–February 2009 that has been used in a number of shared tasks (e.g., Clarke et al., 2011). We build corpora of two types: corpora based on subsets of all documents in this crawl (which include only documents classified as English by TEXTCAT, but a small proportion of non-English documents according to `langid.py`) and corpora based on subsets of only those documents identified as English using `langid.py`.

Similar to Ferraresi et al. (2008), we select

[4] `http://boston.lti.cs.cmu.edu/clueweb09/wiki/tiki-index.php?page=Language+Identification+for+ClueWeb09`

documents of MIME type text/html with size between 5K and 200K bytes. Also following Ferraresi et al. we extract the textual portions of the selected HTML documents using the body text extraction algorithm (BTE, Finn et al., 2001) which heuristically removes removes boilerplate based on the frequency of HTML tags.[5] We use Pomikálek's (2011) implementation of BTE. We remove duplicate and near-duplicate paragraphs using onion (Pomikálek, 2011) — the same tool used by Pomikálek et al. (2012) — with its default settings. In this configuration onion makes a single pass through a corpus, and eliminates any paragraph which shares more than 50% of its 7-grams with the portion of the corpus analysed so far. Finally we tokenise and sentence split our corpora using tools provided by the Stanford Natural Language Processing Group.[6]

ClueWeb09 is broken into a number of files, each containing approximately 100M of compressed crawl data; we apply the above method to build corpora from the first 1, 5, 10, 50, and 100 files in English ClueWeb09.[7] The sizes, in tokens, of the resulting corpora are shown in Table 1.

2.3 Language modelling

We construct language models using SRILM (Stolcke, 2002), a commonly-used, off-the-shelf toolkit for building and applying statistical language models. For each corpus built from ClueWeb09, we build an open-vocabulary language model using the default settings of SRILM, which correspond to an order 3 language model with Good-Turing smoothing. All language models were built using the `make-big-lm` script provided with SRILM.

We evaluate our language models by measuring the perplexity of the written portion of the British National Corpus (BNC, Burnard, 2000), a corpus of roughly 87 million words of British English from the late twentieth century, spanning a variety of genres and topics. Perplexity is a standard evaluation metric for language models, with lower perplexity indicating the model better fits the test data. Perplexities were calculated using the `ngram` program from SRILM, and are normalized counting all input tokens, including end-of-sentence tags.

# files	− `langid.py`		+ `langid.py`	
	# tokens	PPL	# tokens	PPL
1	16M	457.1	15M	457.5
5	81M	384.2	77M	381.0
10	156M	361.8	148M	359.4
50	795M	297.1	760M	294.9
100	1.6B	277.1	1.5B	275.4

Table 1: Number of tokens in each corpus built from increasing numbers of ClueWeb09 files, with and without document filtering using `langid.py`. The perplexity (PPL) of the BNC under a language model trained on the corresponding corpus is also shown.

3 Experimental setup and results

We train language models on each corpus derived from ClueWeb09, and then measure the perplexity of the written portion of the BNC (as described in Section 2.3). Results are shown in Table 1.

We begin by noting that for all corpus sizes considered with the exception of the smallest, the perplexity of the BNC is lower under a language model from the corpus filtered using `langid.py` than under a language model trained on a corpus built from the same original data but without this extra language identification step. This suggests that state-of-the-art language identification can indeed enable the construction of better corpora — at least for training language models for the BNC.

To assess whether the observed differences are significant, for each corpus size (i.e., number of ClueWeb09 files) we measure the perplexity of each BNC document under the language model from the corpus with, and without, filtering with `langid.py`. For a given corpus size this then gives us independent paired measurements, which we compare using a Wilcoxon rank sum test. For each corpus size the difference with and without `langid.py` filtering is highly significant ($p < 10^{-23}$ in each case).

Further analysing the case of the smallest cor-

[5] We do not consider JusText (Pomikálek, 2011), a recent alternative to BTE, because it incorporates rudimentary language identification in the form of stopword frequency; our specific goal is to study the effects of state-of-the-art language identification in corpus construction. We leave studying the interaction between various steps of corpus construction — including text extraction and language identification — for future work. Furthermore, BTE has been widely used in previous corpus construction projects (e.g., Baroni and Bernardini, 2004; Sharoff, 2006; Ferraresi et al., 2008).

[6] `http://nlp.stanford.edu/software/tokenizer.shtml`

[7] We use the files in en0000, the first section of ClueWeb09.

pus size considered, the perplexity is quite high in both cases, suggesting that the language model is under-fitting due to insufficient training data. It seems that in such cases — which correspond to corpora far smaller than one would typically build from a Web crawl — there is little to be gained from improved language identification (at least for the task of building trigram language models considered here).

With the exception of the smallest corpus, as corpus size increases, the absolute reduction in perplexity with and without `langid.py` decreases. In future work we plan to build much larger corpora to further examine this trend.

In addition to the BNC, we considered a number of other corpora for evaluation, including the Brown Corpus (Francis and Kucera, 1964) and a sample of texts provided with NLTK (Bird et al., 2009) from Project Gutenberg,[8] and found the results to be consistent with those on the BNC.

4 Discussion

In addition to demonstrating the importance of language identification in Web corpus construction, the results in Table 1 confirm Banko and Brill's (2001) findings about corpus size; in particular, for corpora built using the same method (i.e., with or without `langid.py`) bigger is better. However, for each corpus size (i.e., each number of files from ClueWeb09) the corpus filtered with `langid.py` is roughly 5% smaller — and yet produces a better language model — than the corresponding corpus not filtered in this way. Furthermore, because of their smaller size, the corpora filtered with `langid.py` have lower storage and processing costs.

Based on these findings, it appears we can improve corpora in two ways: by getting more data, and by better processing the data we have. Although it is certainly possible to build a larger Web crawl than ClueWeb09, doing so comes at a substantial cost in terms of bandwidth, processing, and storage (although Suchomel and Pomikálek (2012) have recently considered how to more-efficiently crawl the Web for linguistic data). Resources which are readily-available at relatively-low cost (such as ClueWeb09) are likely to serve as the basis for many corpus construction efforts, and it is therefore important to determine how to best exploit such a fixed resource in building corpora.

The largest language-filtered corpus built in this study consists of roughly 1.5B tokens. Although we eventually intend to build much larger corpora, this corpus size is on par with that of the ukWaC (Ferraresi et al., 2008) — a corpus that has been widely used in computational linguistics and as the basis for lexicographical analysis (e.g., Atkins, 2010). Our findings are therefore helpful in that they demonstrate the possibility for improving Web corpora of a size already shown to be of practical use. Nevertheless, in future work we intend to explore the impact of language identification on much larger corpora by building corpora from roughly an order of magnitude more data.

In an effort to better understand the differences between the language identifiers, we examined 100 documents from English ClueWeb09 classified as non-English by `langid.py`. We found that 33 were entirely non-English, 30 contained some text in English as well as another language, 27 were in fact English, and 10 contained no linguistic content. The prevalence of multilingual documents suggests that language identification at the sub-document (e.g., paragraph) level, or language identification methods capable of detecting mixtures of languages could lead to further improvements.

5 Conclusions

In this paper we have considered the impact of language identification on corpus construction, and shown that state-of-the art language identification leads to better language models. The ultimate goal of this research is to determine how to best derive a linguistic corpus from a Web crawl. In future work, we intend to consider other aspects of the corpus construction process, including webpage cleaning (e.g., removing boilerplate text) and deduplication. In this preliminary study we only considered language modelling for evaluation; in the future, we plan to carry out a more-comprehensive evaluation, including classification and rankings tasks (e.g., Banko and Brill, 2001; Liu and Curran, 2006; Versley and Panchenko, 2012) in addition to language modelling. To encourage further research on this problem, code to replicate the corpora created for, and experiments carried out in, this paper will be made publicly available upon publication.

[8] `http://www.gutenberg.org/`

Acknowledgments

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

Beatrice Alex, Amit Dubey, and Frank Keller. 2007. Using foreign inclusion detection to improve parsing performance. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 151–160. Prague, Czech Republic.

B. T. Sue Atkins. 2010. The DANTE Database: Its contribution to English lexical research, and in particular to complementing the FrameNet data. In Gilles-Maurice De Schryver, editor, *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks*. Menha Publishers, Kampala, Uganda.

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, pages 26–33. Toulouse, France.

Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the Web. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal.

Marco Baroni, Francis Chantree, Adam Kilgarriff, and Serge Sharoff. 2008. Cleaneval: A competition for cleaning Web pages. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pages 638–643. Marrakech, Morocco.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc., Sebastopol, CA.

Lou Burnard. 2000. *The British National Corpus Users Reference Guide*. Oxford University Computing Services.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR-94)*, pages 161–175. Las Vegas, USA.

Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Ellen M. Voorhees. 2011. Overview of the TREC 2011 Web Track. In *Proceedings of the Twentieth Text REtrieval Conference (TREC 2011), NIST Special Publication: SP 500-295*.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop: Can we beat Google*, pages 47–54. Marrakech, Morocco.

Aidan Finn, Nicholas Kushmerick, and Barry Smyth. 2001. Fact or fiction: Content classification for digital libraries. In *Proceedings of the Second DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries*. Dublin, Ireland.

W. Nelson Francis and Henry Kucera. 1964. *Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Brown University.

Adam Kilgarriff. 2007. Googleology is bad science. *Computational Linguistics*, 33(1):147–151.

Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and Avinesh PVS. 2010. A corpus factory for many languages. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 904–910. Valletta, Malta.

Vinci Liu and James Curran. 2006. Web text corpus for natural language processing. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pages 233–240. Trento, Italy.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 553–561. Chiang Mai, Thailand.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012) Demo Session*, pages 25–30. Jeju, Republic of Korea.

Brian Murphy and Egon Stemle. 2011. PaddyWaC: A minimally-supervised Web-corpus of Hiberno-English. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 22–29. Edinburgh, Scotland.

Jan Pomikálek. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. thesis, Masaryk University.

Jan Pomikálek, Miloš Jakubíček, and Pavel Rychlý. 2012. Building a 70 billion word corpus of English from ClueWeb. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 502–506. Istanbul, Turkey.

Serge Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, editors, *Wacky! Working papers on the Web as Corpus*, pages 63–98. GEDIT, Bologna, Italy.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901–904. Denver, USA.

Vit Suchomel and Jan Pomikálek. 2012. Efficient Web crawling for large text corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 39–43. Lyon, France.

Yannick Versley and Yana Panchenko. 2012. Not just bigger: Towards better-quality Web corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 44–52.