Removing Spam from Web Corpora Through Supervised Learning and Semi-manual Classification of Web Sites

Vít Suchomel[1,2]

[1] Natural Language Processing Centre
Masaryk University, Brno, Czech Republic
xsuchom2@fi.muni.cz
https://nlp.fi.muni.cz/en/
[2] Lexical Computing, Brno, Czech Republic

Abstract. Internet spam is a major issue hindering the usefulness of web corpora. Unlike traditional text corpora collected from trustworthy sources, the content of web based corpora has to be cleaned.

In this paper, two experiments of non-text removal based on supervised learning are presented. First, an improvement of corpus based language analyses of selected words achieved by a supervised classifier is shown on an English web corpus. Then, a semi-manual approach of obtaining samples of non-text web pages in Estonian is introduced. This strategy makes the supervised learning process more efficient.

The result spam classifiers are tuned for high recall at the cost of precision to remove as much non-text as possible. The evaluation shows the classifiers reached the recall of 71 % and 97 % for English and Estonian web corpus, respectively.

A technique for avoiding spammed web sites by measuring the distance of web pages from trustworthy sites is studied too.

Keywords: Web corpora, Web spam, Supervised learning.

1 Introduction

It is known that boilerplate, duplicates, and spam skew corpus based analyses and therefore have to be dealt with. While the first two issues have been successfully addressed, e.g. by [8,10,17,12], spam might be still observed in web corpora as reported by [7,14].

While the traditional definition of web spam is *actions intended to mislead search engines into ranking some pages higher than they deserve* [4], the text corpus point of view is not concerned with intentions of spam producers or the justification of the search engine optimisation of a web page. A text corpus built for NLP or linguistics purpose should contain coherent and consistent, meaningful, natural and authentic sentences in the target language. Only texts created by spamming techniques breaking those properties should be detected and avoided.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2020, pp. 113–123, 2020. © Tribun EU 2020

The unwanted non-text is this: computer generated text, machine translated text, text altered by keyword stuffing or phrase stitching, text altered by replacing words with synonyms using a thesaurus, summaries automatically generated from databases (e.g. stock market reports, weather forecast, sport results – all of the same kind very similar), and finally any incoherent text. Varieties of spam removable by existing tools, e.g. duplicate content, link farms (quite a lot of links with scarce text), are only a minor problem.

Automatically generated content does not provide examples of authentic use of a natural language. Nonsense, incoherent or any unnatural texts such as the following short instance have to be removed from a good quality web corpus: *Edmonton Oilers rallied towards get over the Montreal Canadiens 4-3 upon Thursday.Ryan Nugent-Hopkins completed with 2 aims, together with the match-tying rating with 25 seconds remaining within just legislation.*[3]

Avoiding web spam by selecting trustworthy corpus sources such as Wikipedia, news sites, government and academic webs works well: [2] show it is possible to construct medium sized corpora from URL whitelists and web catalogues. [13] used a similar way of building a Czech web corpus. Also the BootCaT method [3] indirectly avoids spam by relying on a search engine to find non-spam data. Despite the avoiding methods being successful yet not perfect [14], it is doubtful a huge web collection can be obtained just from trustworthy sources.

Furthermore, language independent methods of combating spam might be of use. [9] reported web spamming was not only a matter of the English part of internet. Spam was found in their French, German, Japanese and Chinese documents as well. According to our experience in building web corpora in more than 50 languages, non-text content is still on the rise.

In this paper, two experiments of spam removal based on supervised learning are introduced: Section 2 shows the improvement of corpus based language analyses of selected words achieved by a supervised classifier applied to an English web corpus. Section 3 presents an experiment with an Estonian web corpus. A semi-manual approach of obtaining samples of non-text web pages made the supervised learning process more efficient. The result spam classifier reached a very high recall of 97 %. The usefulness of measuring the distance of web domains from initial web domains of a web crawl as a means to avoid low quality web sites was also studied. Results of this work and challenges for the future are summarised in Section 4.

2 Removing Spam from English Web Corpus Through Supervised Learning

This section describes training and evaluation of a supervised classifier to detect spam in web corpora.

[3] Source: `http://masterclasspolska.pl/forum/`, accessed in December 2015.

We have manually annotated a collection of 1630 web pages from various web sources from years 2006 to 2015.[4] To cover the main topics of spam texts observed in our previously built corpora, we included 107 spam pages promoting medication, financial services, commercial essay writing and other subjects.

Both phrase level and sentence level incoherent texts (mostly keyword insertions, n-grams of words stitched together or seemingly authentic sentences not conveying any connecting message) were represented. Another 39 spam documents coming from random web documents identified by annotators were included. There were 146 positive instances of spam documents altogether.

Table 1. Comparison of the 2015 English web corpus before and after spam removal using the classifier. Corpus sizes and relative frequencies (number of occurrences per million words) of selected words are shown. By reducing the corpus to 55 % of the former token count, phrases strongly indicating spam documents such as *cialis 20 mg*, *payday loan*, *essay writing* or *slot machine* were almost removed while innocent phrases not attracting spammers from the same domains such as *oral administration*, *interest rate*, *pass the exam* or *play games* were reduced proportionally to the whole corpus.

	Original corpus	**Clean corpus**	**Kept**
Document count	58,438,034	37,810,139	64.7 %
Token count	33,144,241,513	18,371,812,861	55.4 %

Phrase	**Original hits/M**	**Clean hits/M**	**Kept**
viagra	229.71	3.42	0.8 %
cialis 20 mg	2.74	0.02	0.4 %
aspirin	5.63	1.52	14.8 %
oral administration	0.26	0.23	48.8 %
loan	166.32	48.34	16.1 %
payday loan	24.19	1.09	2.5 %
cheap	295.31	64.30	12.1 %
interest rate	14.73	9.80	36.7 %
essay	348.89	33.95	5.4 %
essay writing	7.72	0.32	2.3 %
pass the exam	0.34	0.36	59.4 %
slot machine	3.50	0.99	15.8 %
playing cards	1.01	0.67	36.8 %
play games	3.55	3.68	53.9 %

The classifier was trained using FastText [5] and applied to a large English web corpus from 2015. The expected performance of the classifier was evaluated using a 30-fold cross-validation on the web page collection. Since our aim was to remove as much spam from the corpus as possible, regardless false positives,

[4] The collection is a part of another experiment co-authored by us.

the classifier top label probability threshold was set to prioritize recall over precision.

The achieved precision and recall were 71.5 % and 70.5 % respectively. Applying this classifier to an English web corpus from 2015 resulted in removing 35 % of corpus documents still leaving enough data for the corpus use.

An inspection of the cleaned corpus revealed the relative count of usual spam related keywords dropped significantly as expected while general words not necessarily associated with spam were affected less as can be seen in Table 1.

Table 2. Top collocate objects of verb 'buy' before and after spam removal. Corpus frequency of the verb: 14,267,996 (original), 2,699,951 (cleaned) – 81 % reduction by cleaning (i.e. more than the average reduction of a word in the corpus). The highest scoring lemmas are displayed. Frequency denotes the number of occurrences of the lemma as a collocate of the headword in the corpus. The score represents the typicality value (calculated by collocation metric LogDice [11] here) indicating how strong the collocation is.

Original lemma	**frequency**	**score**		**Clean lemma**	**frequency**	**score**
viagra	569,944	10.68		ticket	52,529	9.80
ciali	242,476	9.56		house	28,313	8.59
essay	212,077	9.17		product	37,126	8.49
paper	180,180	8.93		food	24,940	8.22
levitra	98,830	8.33		car	20,053	8.18
uk	93,491	8.22		book	27,088	8.09
ticket	85,994	8.08		property	17,210	7.88
product	105,263	8.00		land	15,857	7.83
cialis	71,359	7.85		share	12,083	7.67

To show the impact of the method on data used in real applications, Word Sketches of selected verb, nouns and adjectives in the original corpus and the cleaned corpus were compared. A Word Sketch is a table like report providing a collocation and grammatical summary of the word's behaviour that is essential for lexicography e.g. to derive the typical context and word senses of headwords in a dictionary. [6,1]. To create a good entry in a dictionary, one has to know strong collocates of the headword. We will show better collocates are provided by the cleaned corpus than the original version in the case of selected headwords.

Table 2 shows that top collocates of verb 'buy' in relation 'objects of verb' were improved a lot by applying the cleaning method to the corpus. It is true that e.g. 'buy viagra' or 'buy essay' are valid phrases, however looking at random concordance lines of these collocations, vast majority come from computer generated un-grammatical sentences.

Comparison of modifiers of noun 'house' in Table 3 reveals that the Word Sketch of a seemingly problem-free headword such as 'house' can be polluted by a false collocate – 'geisha'. Checking random concordance lines for co-occurrences of 'house' and 'geisha', almost none of them are natural English sentences. While 'geisha' is the fifth strongest collocate in the original corpus, it is not present among top 100 collocates in the cleaned version.

Table 3. Top collocate modifiers of noun 'house' before and after spam removal. Corpus frequency of the noun: 10,873,053 (original), 3,675,144 (cleaned) – 66 % reduction.

Original lemma	frequency	score		Clean lemma	frequency	score
white	280,842	10.58		publishing	20,314	8.63
opera	58,182	8.53		open	39,684	8.47
auction	41,438	8.05		guest	13,574	7.94
publishing	41,855	8.02		opera	9,847	7.67
geisha	38,331	7.95		old	32,855	7.64
open	37,627	7.78		haunted	9,013	7.58
old	73,454	7.52		auction	8,240	7.40
guest	28,655	7.44		manor	7,225	7.28
country	26,092	7.07		bedroom	7,717	7.26

The last comparison in Table 4 showing nouns modified by adjective 'green' is an example of cases not changed much by the cleaning.It is worthy of noting that apart from other words in this evaluation, the relative number of hits of adjective 'green' in the corpus was decreased less than the whole corpus. Although the classifier deliberately prefers recall over precision, the presence of non-spam words in the corpus was reduced less than the count of 'spam susceptible' words.

Table 4. Top collocate nouns modified by adjective 'green' before and after spam removal. Corpus frequency of the adjective: 2,626,241 (original), 1,585,328 (cleaned) – 40 % reduction (less than the average in the corpus).

Original lemma	frequency	score		Clean lemma	frequency	score
tea	86,031	10.04		tea	45,214	9.94
light	54,991	8.74		light	33,069	8.86
bean	28,724	8.63		space	51,830	8.72
egg	26,150	8.45		roof	17,916	8.72
space	55,412	8.19		bean	15,398	8.52
vegetable	20,906	8.16		economy	24,181	8.21
roof	18,910	8.1		energy	18,101	7.8
leave	16,712	7.74		infrastructure	13,331	7.69
economy	25,261	7.72		leave	9,754	7.69

3 Removing Spam from Estonian Web Corpus Through Semi-manual Classification of Web Sites

Unlike the spam classification of English web pages described in the previous chapter, where human annotators identified a small set of spam documents representing various non-text types, the annotators classified whole web domains this time. An Estonian web corpus crawled in 2019 was used in this experiment. Similarly to our previous result, supervised learning using FastText was employed to classify the corpus.

Our assumption in this setup is that all pages in a web domain are either good – consisting of nice human produced text – or bad – i.e. machine generated non-text or other poor quality content. Although this supposition might not hold for all cases and can lead to noisy training data for the classifier, it has two advantages: Much more training samples are obtained and the cost to determine if a web domain tends to provide good text or non-text is not high. In this case, that work was done by Kristina Koppel from the Institute of Estonian Language at University of Tartu in several days.

Furthermore, it is efficient to check the most represented domains in the corpus. Thus a lot of non-text can be eliminated while obtaining a lot of training web pages at the same time. Spam documents coming from less represented web domains can be traced by the classifier once it is built.

A list of 1,000 Estonian web sites with the highest count of documents or the highest count of tokens in the corpus was used in the process of manual quality checking. There were also path prefixes covering at least 10 % of all paths within each site available to provide information about the structure of the domain. If the site was known to the human annotator, it was marked as good without further checks. If the site name looked suspiciously (e.g. a concatenation of unrelated words, mixed letters and numbers, or a foreign TLD), the annotator checked the site content on the web or its concordance lines in Sketch Engine.

Site name rules were formulated by observation of bad web domains. E.g. all hosts starting with `ee.`, `est.`, or `et.` under generic TLDs `.com`, `.net`, `.org`[5] were marked as non-text since there was machine translated content usually observed in these cases.

77 % of web pages in the corpus were semi-manually classified this way. 16 % of these documents were marked as computer generated non-text, mostly machine translated. 6 % of these documents were marked as bad for other reasons, generally poor quality content.

A binary classifier was trained using FastText on good and non-text web pages. URL of a page, plaintext word forms and 3 to 6 tuples of plaintext characters were the features supplied to FastText. 10 fold cross-validation was carried out to estimate the classifier's performance. Documents from the same web site were put in the same fold to make sure there was not the same content or the same URL prefix in multiple folds. Since the ratio of good to non-text samples in the data was approximately 77:16, the baseline accuracy (putting

[5] Excluding `et.wikipedia.org`.

all samples in the larger class) was 0.826. Despite the rather high baseline, the classifier performed well. FastText reported fold-averaged precision around 0.94 and recall from 0.93 to 0.76 based on the label probability threshold.

The final classifier was applied to the part of the corpus that had not been checked by the human annotator. 100 positive, i.e. non-text, and 100 negative, i.e. good, web pages were randomly selected for inspection. Kristina Koppel and Margit Langemets from the same institution checked the URL and plaintext[6] of each page. Three minimal probabilities of the top label were tested. The result precision and recall can be seen in Figure 1.

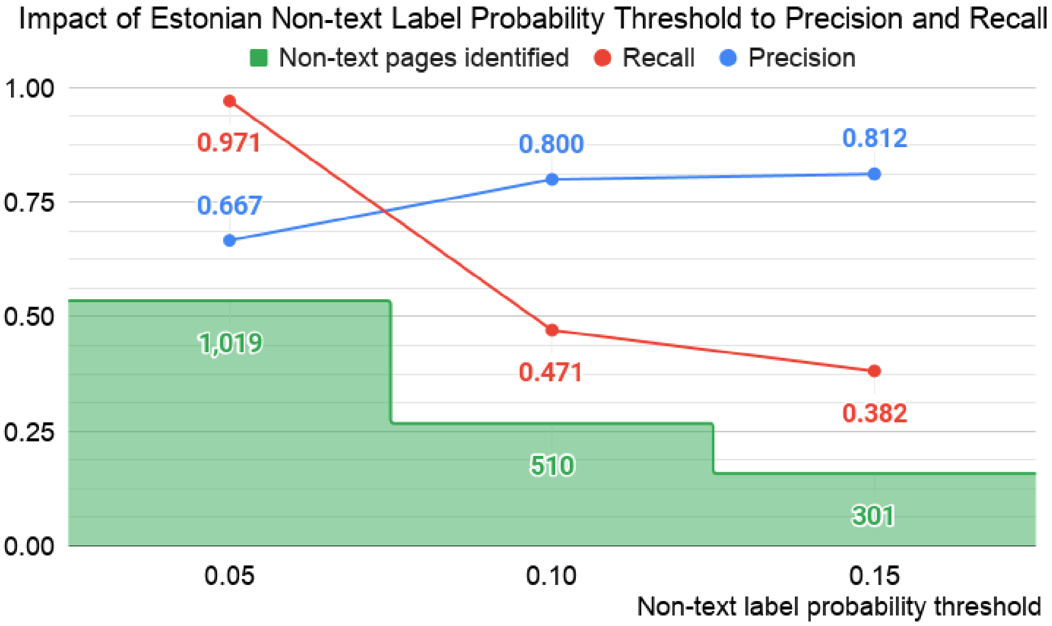

Fig. 1. Evaluation of the final binary spam classifier on documents not previously checked by a human annotator in Estonian web corpus. Precision and recall were estimated for minimal probabilities of the non-text label from 0.05 to 0.15. Since we aim for a high recall, the performance with the non-text label threshold set to 0.05 is satisfying. A higher threshold leads to an undesirable drop of recall.

It can be observed the recall dropped a lot with an increasing threshold. Therefore, the final top label probability applied to the corpus was set just to 0.05 to keep the recall high. We do not mind false positives as long as most of non-text is removed. We consider this setup and result as both time efficient and well performing. It will be applied to web corpora in other languages in cooperation with native speaker experts in the future.

Since web crawler SpiderLing [15], used to obtain the data, measures the distance of web domains from the initial domains, the value can be used to

[6] Texts were cropped to first 2,000 characters to speed up the process.

estimate the quality of the content of a web site. If our hypothesis was true, domains close to the seeds should be crawled more than domains far from the seeds.

To prove or reject the hypothesis, the classification of spam from the previous experiment was put into a relation with the domain distance of the respective good or bad documents. Both semi-manual and machine classification web pages were included. The binary classification of texts – good and bad labels – aggregated by the distance of the respective web sites from seed domains is displayed on Figure 2. The evaluation does not support the hypothesis much, at least in the case of the Estonian web.

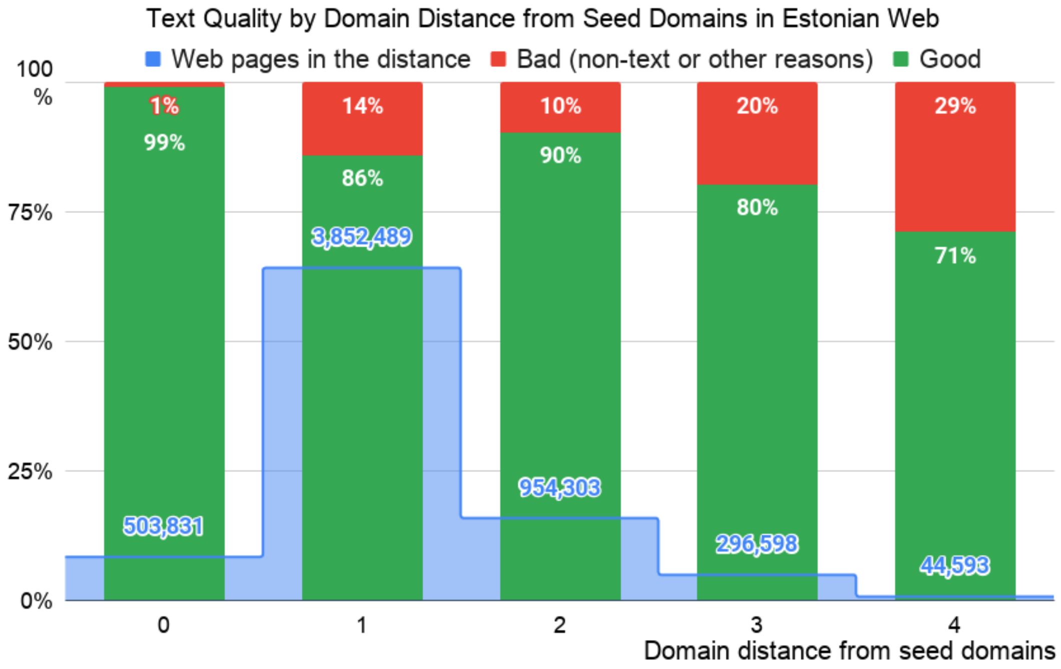

Fig. 2. Evaluation of the relation of the distance of web domain from the initial domains to the presence of non-text on the sites. Web pages of distances 0 to 4 classified semi-manually or by the spam classifier were taken into account. Two thirds of the pages were in distance 1. The percentage of good and bad documents within the same domain distance is shown. The presence of non-text in the data is notable from distance 1.

To sum up the findings of our experiments with Estonian web corpus:

1. A non-text classifier with a very high recall (at the cost of precision) can be trained on human annotated good and bad web sites.
2. The annotation process can be quite efficient: Checking web domains most represented in the corpus produces sufficient samples to classify the rest.
3. It is beneficial to start the crawling from trustworthy, quality content sites. However, there is non-text on web sites linked from the initial sites. The

domain distance is related to the presence of non-text but the correlation is not strong enough to make it an important feature in spam removal.

4 Conclusion and Future Challenges

Two experiments of spam removal based on supervised learning using FastText were presented in this paper.

A classifier trained on manually identified spam documents was applied to a recent English web corpus. The classifier was set to prefer recall at the cost of greatly reducing the size of the result corpus. Although the evaluation of the classifier on the training set reports a far from perfect recall of 71 %, it was able to notably decrease the presence of spam related words in the corpus.

An extrinsic evaluation was carried out by comparing the original data and the cleaned version in a lexicography oriented application: Relative corpus frequencies of words and Word Sketches of grammatical relations that could be used to make a dictionary entry for selected verb, noun and adjective were compared in the experiment.

Another experiment with a smaller Estonian corpus was carried out. An efficient human annotation lead to using more than two thirds of the corpus as training data for the spam classifier. The evaluation of the classifier shows a very high recall of 97 % was reached.

We understand the process can take more time for large internet languages such as English, Spanish, Russian or Chinese. We admit the number of sites in our Estonian experiment is small in comparison to these languages. Nevertheless we believe this is a good way to go for all languages. After all, Google needed human intervention to identify certain types of spam too.[7]

Although promising results were shown, we still consider computer generated non-text the main factor decreasing the quality of web corpora.

Computer generated text is on the rise. Although starting the crawl from a set of trustworthy seed domains, measuring domain distance from seed domains and not deviating too deep from the seed domains using hostname heuristics are ways to avoid spam, a lot of generated non-text will still be downloaded.

Machine translation is a specific subcase. Although there might exist a solution – watermarking the output of statistical machine translation – suggested by [16], we are not aware of the actual spread of this technique.

Strategies of non-text detection using language models will just compete with the same language models generating non-text. Nevertheless, the web will remain the largest source of text corpora.

Acknowledgments. This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2018101.

[7] Document 'Fighting Spam' accessed at `http://www.google.com/insidesearch/howsearchworks/fighting-spam.html` in January 2015.

References

1. Baisa, V., Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Medveď, M., Měchura, M., Rychlý, P., Suchomel, V.: Automating dictionary production: a tagalog-english-korean dictionary from scratch. In: Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal. pp. 805–818 (2019)
2. Baisa, V., Suchomel, V.: Skell: Web interface for english language learning. In: Proceedings of 8th Workshop on Recent Advances in Slavonic Natural Languages Processing. pp. 63–70. Brno (2014)
3. Baroni, M., Bernardini, S.: Bootcat: Bootstrapping corpora and terms from the web. In: Proceedings of International Conference on Language Resources and Evaluation (2004)
4. Gyongyi, Z., Garcia-Molina, H.: Web spam taxonomy. In: First international workshop on adversarial information retrieval on the web (AIRWeb 2005) (2005)
5. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
6. Kilgarriff, A., Baisa, V., Busta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The sketch engine: Ten years on. Lexicography **1**(1), 7–36 (2014)
7. Kilgarriff, A., Suchomel, V.: Web spam. In: Stefan Evert, Egon Stemle, P.R. (ed.) Proceedings of the 8th Web as Corpus Workshop (WAC-8) @Corpus Linguistics 2013. pp. 46–52 (2013)
8. Marek, M., Pecina, P., Spousta, M.: Web page cleaning with conditional random fields. In: Building and Exploring Web Corpora: Proceedings of the Fifth Web as Corpus Workshop, Incorporationg CleanEval (WAC3), Belgium. pp. 155–162 (2007)
9. Ntoulas, A., Najork, M., Manasse, M., Fetterly, D.: Detecting spam web pages through content analysis. In: Proceedings of the 15th international conference on World Wide Web. pp. 83–92. ACM (2006)
10. Pomikálek, J.: Removing boilerplate and duplicate content from web corpora. Ph.D. thesis, Masaryk University (2011)
11. Rychlý, P.: A lexicographer-friendly association score. In: Proceedings of 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing. pp. 6–9 (2008)
12. Schäfer, R., Bildhauer, F.: Web Corpus Construction, vol. 6. Morgan & Claypool Publishers (2013)
13. Spoustová, J., Spousta, M.: A high-quality web corpus of czech. In: Proceedings of Eighth International Conference on Language Resources and Evaluation. pp. 311–315 (2012)
14. Suchomel, V.: Removing spam from web corpora through supervised learning using fasttext. In: Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. pp. 56–60. Birmingham (2017)
15. Suchomel, V., Pomikálek, J.: Efficient web crawling for large text corpora. In: Adam Kilgarriff, S.S. (ed.) Proceedings of the seventh Web as Corpus Workshop (WAC7). pp. 39–43. Lyon (2012)
16. Venugopal, A., Uszkoreit, J., Talbot, D., Och, F.J., Ganitkevitch, J.: Watermarking the outputs of structured prediction with an application in statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1363–1372. Association for Computational Linguistics (2011)

17. Versley, Y., Panchenko, Y.: Not just bigger: Towards better-quality web corpora. In: Proceedings of the seventh Web as Corpus Workshop (WAC7). pp. 44–52 (2012)