Deduplication in Large Web Corpora

Vladimír Benko
Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
vladimir.benko@juls.savba.sk

Abstract

Our paper tries to find answers to some questions related to deduplication process in large-scale web-crawled corpora. An experiment based on eight corpora from the Aranea family is introduced, and first results are presented.

1 Introduction

During past years, detection of duplicate data has been subject of increased research activity, motivated by efforts to save disk space in large-scale cloud-based storage systems (Mao et al., 2014), or to decrease size of index structures for web-based information system, such as search engines (Broder and Nelson, 1996; Zelenkov and Segalovich, 2007). In both cases, preference was given to algorithms capable of detecting duplicate data dynamically, i.e., such that evaluate each new document "as soon as it has arrived" (Waraporn et al., 2014).

In the context of corpus linguistics, the problem of duplicate data emerged relatively recently, mostly with the advent of the "Web as Corpus" research paradigm resulting in much larger corpora containing dramatically more duplicities. Due to the characteristics of a typical corpus processing pipeline, the detection of duplicates needs not to be performed for each document or text segment "on the fly", but rather the respective processing can be performed over the whole corpus (Pomikálek, 2011; Benko, 2013).

In both cases, it is obvious that detection of 100% duplicates is a relatively simple task, both from the theoretical and implementation perspective (Broder 1993), and the challenging part is the detection of near-duplicates (Pomikálek, op. cit.).

2 The Problem

Our paper will introduce a series of on-going experiments related to deduplication in large web-based corpora, in the framework of which we want to find answers to questions including (yet not limited to) as follows:

- How does the size of corpus influence the ratio of duplicate text segments of different level (documents, paragraphs and sentences).
- What are the optimal parameters of deduplication performed by the *Onion*[1] utility.
- What is the optimal method/metric for assessment the "quality" of deduplication.
- What is the nature of data that has been removed.

Our work is motivated mostly by the fact that we were able to find only very few papers devoted to these questions. In the framework of his PhD research, author of the *Onion* program based his evaluation of the deduplication process on counting the "surviving duplicate n-grams", and he worked with relatively small corpora only. The corpus used in Benko (2013) was larger, but still by at least one order of magnitude smaller than a typical web corpus. Moreover, that experiment had been performed on a traditional corpus with arguably different structure of duplicate phenomena in comparison with those in web corpora.

3 Deduplicating Aranea

For the first stage of our experiment, we decided to use data from our Aranea family of web corpora (Benko, 2014; Benko and Zakharov, 2016) that are not only sufficiently big but are also available in various source and intermediate formats suitable for the envisaged experiments.

As the deduplication of large corpora requires great amounts of computing resources (both RAM and processing time), corpus creators usually tend to optimize the process by opting for single pass and deduplicating on one type of text segment only, typically on paragraphs (Kilgarriff, 2014). In our case, however, we decided to perform the whole procedure in a progressive

[1] http://corpus.tools/wiki/Onion

manner, i.e., on the document, paragraph and sentence level, respectively. The advantage of such an approach is that the resulting corpora are available in several formats suitable for different types of use.

3.1 The Onion Pipeline

Onion is a mature, stable and extremely efficient tool optimized to detect and remove duplicate content for large-scale textual data files used in building language corpora. The way how it works is beyond the scope of this paper, and is described both in the already mentioned Pomikálek's dissertation, as well as in our previous work (Benko, 2013).

The program can basically work in two modes: by the default, the duplicates detected are simply deleted. Alternatively, duplicate text segments are only marked and the further decision what to do with them is left to an external utility – this was the functionality we used in the framework of our experiment.

3.2 "Onioning" the Paragraphs

As the input for our first experiment we used data of eight Aranea corpora, with four of them representing the "large" languages (English, French, German, and Russian), and the other four the "small" languages (Czech, Slovak, Swedish, and Latvian). Data of all these corpora had already been subject to standard pre-processing, such as filtration, tokenization, segmentation on sentences, and also document-level deduplication.

The standard Onion pipeline has been modified to produce continuous logging of the results (tokens in duplicate vs. non-duplicate text segments) after a user-settable threshold is reached (100 M by default). The deduplication was performed on 5-grams with a threshold of 0.9 and smoothing switched off[2], i.e., a text segment was considered duplicate if it contained over 90% n-grams already encountered in the previous text.

The results of paragraph-level deduplication are shown in Figure 1.

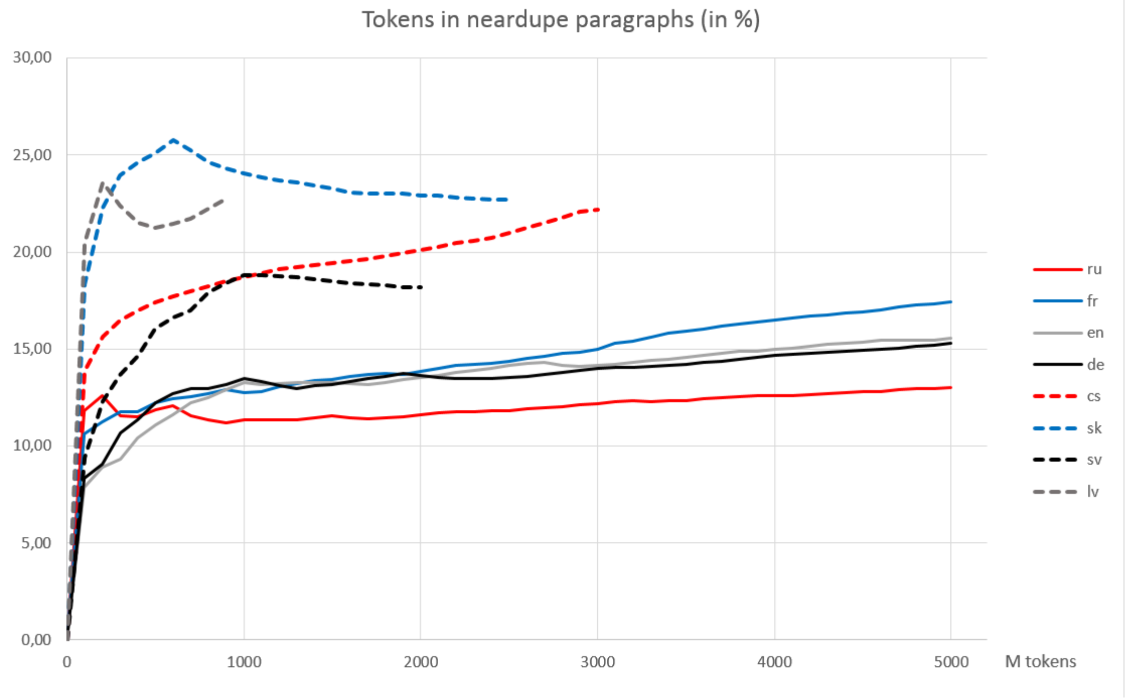

Figure 1: Deduplication on the paragraph level.

The result is somewhat surprising: it can be seen that the respective curves look very similar for the "large" languages, and after reaching the saturation, only small increase is observed. Although more data was available for these languages, we decided to cut the graph at the 5,000 Megatoken threshold to make the curves for "small" languages with less data more apparent.
The shape of curves for small languages is somewhat disparate, but we can observe that the ratios of duplicates are almost twice larger in comparison with "large" languages.

3.3 Deduping Sentences

Sentence-level deduplication is typically performed only in corpora that are to be analyzed by "reading", such as those used for lexicographic purposes. Duplicate sentences tend to negatively influence frequencies of

[2] In the smoothing mode, Onion also removes short non-duplicate segments between two duplicate ones.

lexical units and collocations, and impose additional burden for lexicographers compiling dictionary entries.

Lexicographers, however, belong to the "heaviest" users of our corpora (especially those containing the Slovak and Czech data), and sentence-level deduplication is therefore standard component of our processing pipeline.

The Figure 2. shows the result the process applied to the same eight corpora.

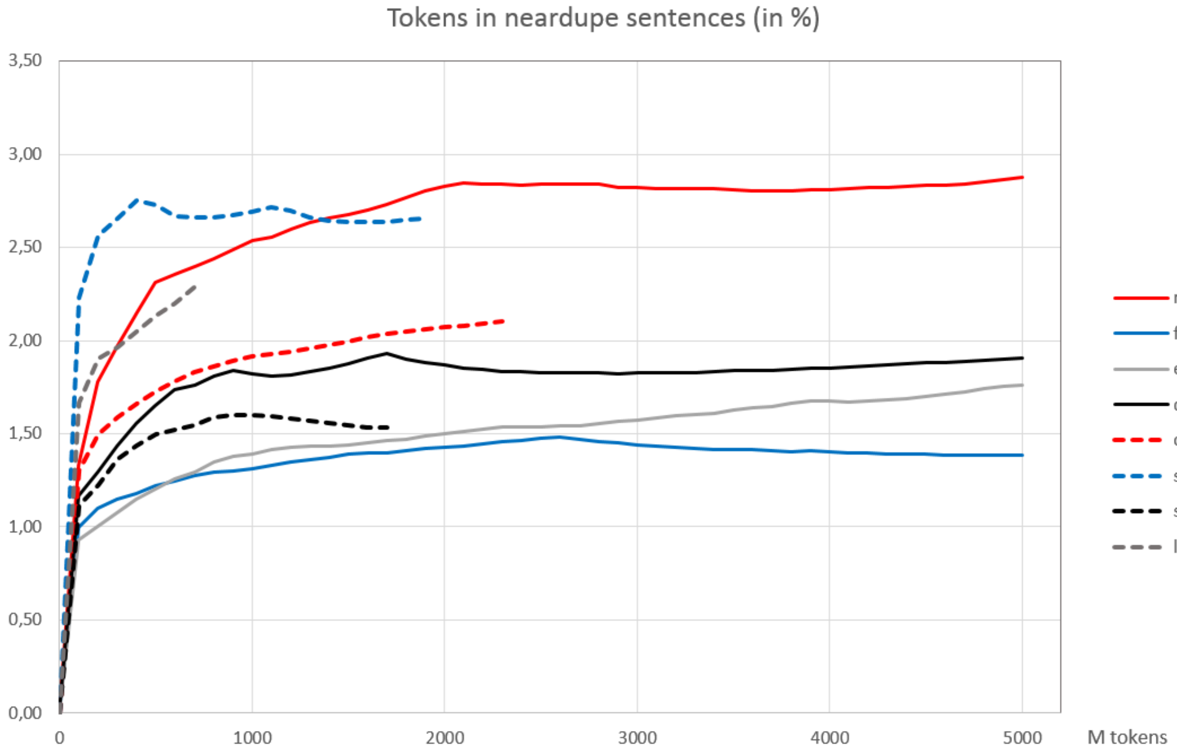

Figure 2: Deduplication on the sentence level.

Two phenomena can be observed in the figure. Firstly, the percentage of removed tokens is – not surprisingly – much smaller than in the paragraph-level deduplication. And secondly, the respective curves are much more similar, even for the "small" languages.

It might be quite interesting to observe that would happen if only sentence-level deduplication were performed – we'll probably make a new experiment targeted at this issue in the future.

4 Why Languages Differ

There are more ways how to examine the reasons of different "deduplication behavior" among languages involved of our experiment. Based on a suggestion of the anonymous reviewer of our paper, we decided to have a look at the number of Internet domains in the resulting corpora that could be used as a measure of data variety – the more different domains, the grater probability of differences in data.

To make the evaluations as simple as possible, we did not perform any new round of deduplication, and made use of the data already available: we produced frequency lists of Internet domains for all the *Minus* (100 Megaword) and *Maius* (1 Gigaword) versions of all the corpora involved. The results are shown in Table 1.

Language	**Domains**		**Ratio**
	Minus (100 MW)	**Maius** (1 GW)	
Russian	118,982	387,040	3.25
Czech	88,604	246,181	2.78
German	72,411	134,944	1.86
English	62,031	158,871	2.56
French	61,418	192,664	3.14
Slovak	49,738	126,024	2.53
Swedish	33,481	105,217	3.14
Latvian[3]	8,512	11,944	1.40

Table 1: Internet domains

[3] Only the *Parvus* class of corpus (530 MW) was available for Latvian.

The results are interesting but really need deeper analysis to be able to interpret the differences among the respective languages. It must be noted that several factors might have influenced the actual numbers – one of them being the number of crawling sessions that was varying from one or two for some languages to several dozens for the "featured" languages (Slovak, Czech and Russian).

5 What Data Has Been Removed

Our deduplication pipeline does not simply remove the duplicate content but rather splits the original file into two parts, i.e. retains the removed segments for possible further analysis. Due to the huge sizes of the respective files, this task is far from being easy. Here we show just a simple first step: finding the most frequent duplicate paragraphs and sentences.

As the Onion-based deduplication is performed on tokenized and tagged data, this procedure involves a reverse process, i.e., removing the annotation (lemma, tag and possible other attributes), "untokenizing" (converting vertical data to original one-paragraph-per-line format) and performing the respective frequency lists by means of standard *sort* and *uniq* utilities. The beginning of the resulting paragraph list is shown in Table 2.

Rank	Freq	Paragraph text
1	58,943	<p><s>Your email address will not be published.</s><s>Required fields are marked *</s></p>
2	55,739	<p><s>We've sent an email with instructions to create a new password.</s><s>Your existing password has not been changed.</s></p>
3	52,223	<p><s>It looks like you're already registered</s></p>
4	44,816	<p><s>Save changes Preview Cancel</s></p>
5	26,758	<p><s>Your password has been changed</s></p>
6	26,757	<p><s>Password has been successfully updated.</s></p>
7	26,619	<p><s>Conference Presentation Video</s></p>
8	26,149	<p><s>Email address is required.</s></p>
9	26,113	<p><s>Enter your email and we'll send you a link to reset your password.</s></p>
10	26,112	<p><s>You're almost there.We've just sent a confirmation email to .</s><s>Check it out to confirm your registration.</s></p>
11	26,112	<p><s>We have sent a confirmation email to .</s><s>Please check your email and click on the link to activate your account.</s></p>
12	26,112	<p><s>We are unable to process your request at this time.</s><s>Please try again later.</s></p>
13	26,112	<p><s>Thank you for registering</s></p>
14	26,112	<p><s>Please fill in the remaining fields below to complete your registration</s></p>
15	26,112	<p><s>It looks like you're already registered.</s></p>
16	26,112	<p><s>is already registered with .</s><s>You will be able to use the same account on .</s><s>Alternatively, you can create a new account with another email address.</s></p>
17	26,112	<p><s>Congratulations, you've just sealed the deal!</s><s>Sign in to your profile now to get started.</s></p>
18	26,112	<p><s>By registering you are agreeing to the Terms and Conditions of the website.</s></p>
10	26,111	<p><s>We didn't recognise that password reset code.</s><s>Enter your email address to get a new one.</s></p>
20	26,111	<p><s>We are unable to send your welcome email at this time.</s><s>Please try again later by clicking the resend welcome email link from your profile page.</s></p>

Table 2: Most frequent duplicate paragraphs (English)

As it can be seen, the most frequent dupes are surprisingly quite long and apparently come from very similar texts – at least their frequencies suggest so.

The Table 3 shows similar list resulting from the sentence-level deduplication. The situation here is different – the most frequent "sentences" are in fact short text fragments, and some of them even raise questions about appropriateness of the sentence segmentation policy.

Rank	Freq	Sentence text
1	532,867	<s>1.</s>
2	477,841	<s>2.</s>
3	407,229	<s>3.</s>
4	315,925	<s>4.</s>
5	247,789	<s>5.</s>
6	181,323	<s>6.</s>
7	145,202	<s>7.</s>
8	117,650	<s>8.</s>

9	98,438	<s>9.</s>
10	92,226	<s>Why?</s>
11	91,129	<s>.</s>
12	85,738	<s>10.</s>
13	72,879	<s>Read more</s>
14	60,538	<s>Read More</s>
15	60,327	<s>Yes.</s>
16	59,769	<s>More</s>
17	58,953	<s>11.</s>
18	54,932	<s>Abstract</s>
10	52,645	<s>a.</s>
20	51,510	<s>12.</s>
21	49,238	<s>1</s>
22	46,641	<s>–</s>
23	46,616	<s>b.</s>
24	43,727	<s>13.</s>
25	42,615	<s>You are here</s>
26	42,460	<s>3</s>
27	40,405	<s>2</s>
28	39,024	<s>14.</s>
29	37,228	<s>Description</s>
30	37,005	<s>Comments</s>
32	36,643	<s>MR.</s>
32	36,521	<s>Pages</s>

Table 3: Most frequent duplicate sentences (English)

The optimal strategy for analyzing the files containing duplicate data is yet to be developed and may also depend on the expected use of the resulting corpus. For lexicographic use, for example, one of the promising options may be looking for lexical units present in duplicate data, yet missing in the deduplicated corpus, with the amount of them being used as a measure of the "quality" of deduplication.

6 Conclusion and Further Work

It is probably too early to make any final conclusions before this experiment is performed with more data and more parameters for the *Onion* program, perhaps also with finer logging thresholds to see the shape of the curve before the saturation.

What can be, however, said after this first stage of our experiment is that the amount of data removed during deduplication depends on many factors associated not only with the respective language itself, but also with the size of "searchable web" for the respective language.

Acknowledgments

This work has been, in part, funded by the Slovak KEGA and VEGA Grant Agencies, Project No. K-16-022-00, and 2/0017/17, respectively.

References

Vladimir Benko. 2013. *Data Deduplication in Slovak Corpora.* In Slovko 2013: Natural Language Processing, Corpus Linguistics, E-learning. RAM-Verlag: Lüdenscheid, 2013, pp. 27-39.

Vladimir Benko. 2014. *Aranea: Yet Another Family of (Comparable) Web Corpora.* In Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, 2014. pp. 257-264. ISBN: 978-3-319-10815-5 (Print), 978-3-319-10816-2 (Online).

Vladimír Benko, and Victor P. Zakharov. 2016. *Very Large Russian Corpora: New Opportunities and New Challenges.* In Kompjuternaja lingvistika i intellektual'nye tekhnologii: Po materialam mezhdunarodnoy konferentsii «Dialog» (2016), vypusk 15 (22). Moskva: Rossijskiy gosudarstvennyy gumanitarnyy universitet, 2016, pp. 79–93.

Andrei Z. Broder. 1993. *Some applications of Rabin's fingerprinting method.* In: Sequences II: Methods in Communications, Security, and Computer Science. Springer-Verlag.
http://xmail.eye-catcher.com/rabin_apps.pdf.

Adam Kilgarriff. 2014. *Personal communication.*

Bo Mao, Hong Jiang, Suzhen Wu, Yinjin Fu, Lei Tian. 2014. *Read-Performance Optimization for Deduplication-Based Storage Systems in the Cloud.* In ACM Transactions on Storage (TOS). Volume 10 Issue 2, March 2014.
https://doi.org/10.1145/2512348

Jan Pomikálek. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora.* Ph.D. thesis. Masaryk University in Brno, Faculty of Informatics.
http://is.muni.cz/th/45523/fi_d/phdthesis.pdf.

Yuriy G. Zelenkov, and Ilya V. Segalovich. 2007. *Sravnitel'nyj analiz metodov opredeleniya nechetkikh dublikatov dlya Web-dokumentov* (*Comparative analysis of near-duplicate detection methods of Web documents*). In Trudy 9-oj Vserossijskoj nauchnoj konferencii «Elektronnye biblioteki: perspektvinye metody i tekhnologii» RDCL 2007. Perelsavl'-Zalesskij: «Universitet goroda Pereslavlya», 2007. pp. 166–174.

Waraporn Leesakul, Paul Townend, and Jie Xu. 2014. *Dynamic Data Deduplication in Cloud Storage.* In IEEE 8th International Symposium on Service Oriented System Engineering. 7-11 April 2014, Oxford, UK.
https://doi.org/10.1109/SOSE.2014.46