NoWaC: a large web-based corpus for Norwegian

Emiliano Guevara
Tekstlab,
Institute for Linguistics and Nordic Studies,
University of Oslo
e.r.guevara@iln.uio.no

Abstract

In this paper we introduce the first version of *noWaC*, a large web-based corpus of Bokmål Norwegian currently containing about 700 million tokens. The corpus has been built by crawling, downloading and processing web documents in the *.no* top-level internet domain. The procedure used to collect the *noWaC* corpus is largely based on the techniques described by Ferraresi et al. (2008). In brief, first a set of "seed" URLs containing documents in the target language is collected by sending queries to commercial search engines (Google and Yahoo). The obtained seeds (overall 6900 URLs) are then used to start a crawling job using the Heritrix web-crawler limited to the *.no* domain. The downloaded documents are then processed in various ways in order to build a linguistic corpus (e.g. filtering by document size, language identification, duplicate and near duplicate detection, etc.).

1 Introduction and motivations

The development, training and testing of NLP tools requires suitable electronic sources of linguistic data (corpora, lexica, treebanks, ontological databases, etc.), which demand a great deal of work in order to be built and are, very often copyright protected. Furthermore, the ever growing importance of heavily data-intensive NLP techniques for strategic tasks such as machine translation and information retrieval, has created the additional requirement that these electronic resources be very large and general in scope.

Since most of the current work in NLP is carried out with data from the economically most impacting languages (and especially with English data), an amazing wealth of tools and resources is available for them. However, researchers interested in "smaller" languages (whether by the number of speakers or by their market relevance in the NLP industry) must struggle to transfer and adapt the available technologies because the suitable sources of data are lacking. Using the web as corpus is a promising option for the latter case, since it can provide with reasonably large and reliable amounts of data in a relatively short time and with a very low production cost.

In this paper we present the first version of *noWaC*, a large web-based corpus of Bokmål Norwegian, a language with a limited web presence, built by crawling the *.no* internet top level domain. The computational procedure used to collect the *noWaC* corpus is by and large based on the techniques described by Ferraresi et al. (2008). Our initiative was originally aimed at collecting a 1.5–2 billion word general-purpose corpus comparable to the corpora made available by the WaCky initiative (`http://wacky.sslmit.unibo.it`). However, carrying out this project on a language with a relatively small online presence such as Bokmål has lead to results which differ from previously reported similar projects. In its current, first version, *noWaC* contains about 700 million tokens.

Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop, pages 1–7,
Los Angeles, California, June 2010.

1.1 Norwegian: linguistic situation and available corpora

Norway is a country with a population of ca. 4.8 million inhabitants that has two official national written standards: *Bokmål* and *Nynorsk* (respectively, 'book language' and 'new Norwegian'). Of the two standards, Bokmål is the most widely used, being actively written by about 85% of the country's population (cf. `http://www.sprakrad.no/` for detailed up to date statistics). The two written standards are extremely similar, especially from the point of view of their orthography. In addition, Norway recognizes a number of regional minority languages (the largest of which, North Sami, has ca. 15,000 speakers).

While the written language is generally standardized, the spoken language in Norway is not, and using one's dialect in any occasion is tolerated and even encouraged. This tolerance is rapidly extending to informal writing, especially in modern means of communication and media such as internet forums, social networks, etc.

There is a fairly large number of corpora of the Norwegian language, both spoken and written (in both standards). However, most of them are of a limited size (under 50 million words, cf. `http://www.hf.uio.no/tekstlab/` for an overview). To our knowledge, the largest existing written corpus of Norwegian is the *Norsk Aviskorpus* (Hofland 2000, cf. `http://avis.uib.no/`), an expanding newspaper-based corpus currently containing 700 million words. However, the *Norsk Aviskorpus* is only available though a dedicated web interface for non commercial use, and advanced research tasks cannot be freely carried out on its contents.

Even though we have only worked on building a web corpus for Bokmål Norwegian, we intend to apply the same procedures to create web-corpora also for Nynorsk and North Sami, thus covering the whole spectrum of written languages in Norway.

1.2 Obtaining legal clearance

The legal status of openly accessible web-documents is not clear. In practice, when one visits a web page with a browsing program, an electronic exact copy of the remote document is created locally; this logically implies that any online document must be, at least to a certain extent, copyright-free if it is to be visited/viewed at all. This is a major difference with respect to other types of documents (e.g. printed materials, films, music records) which cannot be copied at all.

However, when building a web corpus, we do not only wish to visit (i.e. download) web documents, but we would like to process them in various ways, index them and, finally, make them available to other researchers and users in general. All of this would ideally require clearance from the copyright holders of each single document in the corpus, something which is simply impossible to realize for corpora that contain millions of different documents.[1]

In short, web corpora are, from the legal point of view, still a very dark spot in the field of computational linguistics. In most countries, there is simply no legal background to refer to, and the internet is a sort of no-man's land.

Norway is a special case: while the law explicitly protects online content as intellectual property, there is rather new piece of legislation in *Forskrift til åndsverkloven av 21.12 2001 nr. 1563, § 1-4* that allows universities and other research institutions to ask for permission from the Ministry of Culture and Church in order to use copyright protected documents for research purposes that do not cause conflict with the right holders' own use or their economic interests (cf. `http://www.lovdata.no/cgi-wift/ldles?ltdoc=/for/ff-20011221-1563.html`). We have been officially granted this permission for this project, and we can proudly say that *noWaC* is a totally legal and recognized initiative. The results of this work will be legally made available free of charge for research (i.e. non commercial) purposes. *NoWaC* will be distributed in association with the *WaCky* initiative and also directly from the University of Oslo.

[1] Search engines are in a clear contradiction to the copyright policies in most countries: they crawl, download and index billions of documents with no clearance whatsoever, and also redistribute whole copies of the cached documents.

2 Building a corpus of Bokmål by web-crawling

2.1 Methods and tools

In this project we decided to follow the methods used to build the *WaCky* corpora, and to use the related tools as much as possible (e.g. the *BootCaT tools*). In particular, we tried to reproduce the procedures described by Ferraresi et al. (2008) and Baroni et al. (2009). The methodology has already produced web-corpora ranging from 1.7 to 2.6 billion tokens (German, Italian, British English). However, most of the steps needed some adaptation, fine-tuning and some extra programming. In particular, given the relatively complex linguistic situation in Norway, a step dedicated to document language identification was added.

In short, the building and processing chain used for *noWaC* comprises the following steps:

1. Extraction of list of mid-frequency Bokmål words from Wikipedia and building query strings
2. Retrieval of seed URLs from search engines by sending automated queries, limited to the *.no* top-level domain
3. Crawling the web using the seed URLS, limited to the *.no* top-level domain
4. Removing HTML boilerplate and filtering documents by size
5. Removing duplicate and near-duplicate documents
6. Language identification and filtering
7. Tokenisation
8. POS-tagging

At the time of writing, the first version of *noWaC* is being POS-tagged and will be made available in the course of the next weeks.

2.2 Retrieving seed URLs from search engines

We started by obtaining the Wikipedia text dumps for Bokmål Norwegian and related languages (Nynorsk, Danish, Swedish and Icelandic) and selecting the 2000 most frequent words that are unique to Bokmål. We then sent queries of 2 randomly selected Bokmål words though search engine APIs (Google and Yahoo!). A maximum of ten seed URLs were saved for each query, and the retrieved URLs were collapsed in a single list of root URLs, deduplicated and filtered, only keeping those in the *.no* top level domain.

After one week of automated queries (limited to 1000 queries per day per search engine by the respective APIs) we had about 6900 filtered seed URLs.

2.3 Crawling

We used the *Heritrix* open-source, web-scale crawler (`http://crawler.archive.org/`) seeded with the 6900 URLs we obtained to traverse the internet *.no* domain and to download only HTML documents (all other document types were discarded from the archive). We instructed the crawler to use a multi-threaded breadth-first strategy, and to follow a very strict politeness policy, respecting all `robots.txt` exclusion directives while downloading pages at a moderate rate (90 second pause before retrying any URL) in order not to disrupt normal website activity.

The final crawling job was stopped after 15 days. In this period of time, a total size of 1 terabyte was crawled, with approximately 90 million URLs being processed by the crawler. Circa 17 million HTML documents were downloaded, adding up to an overall archive size of 550 gigabytes. Only about 13.5 million documents were successfully retrieved pages (the rest consisting of various "page not found" replies and other server-side error messages).

The documents in the archive were filtered by size, keeping only those documents that were between 5Kb and 200Kb in size (following Ferraresi et al. 2008 and Baroni et al. 2009). This resulted in a reduced archive of 11.4 million documents for post-processing.

2.4 Post-processing: removing HTML boilerplate and de-duplication

At this point of the process, the archive contained raw HTML documents, still very far from being a linguistic corpus. We used the BootCaT toolkit (Baroni and Bernardini 2004, cf. `http://sslmit.unibo.it/~baroni/bootcat.html`) to perform the major operations to clean our archive.

First, every document was processed with the HTML boilerplate removal tool in order to select

only the linguistically interesting portions of text while removing all HTML, Javascript and CSS code and non-linguistic material (made mainly of HTML tags, visual formatting, tables, navigation links, etc.)

Then, the archive was processed with the duplicate and near-duplicate detecting script in the the BootCaT toolkit, based on a 5-gram model. This is a very drastic strategy leading to a huge reduction in the number of kept documents: any two documents sharing more than 1/25 5-grams were considered duplicates, and both documents were discarded. The overall number of documents in the archive went down from 11.40 to 1.17 million after duplicate removal.[2]

2.5 Language identification and filtering

The complex linguistic situation in Norway makes us expect that the Norwegian internet be at least a bilingual domain (Bokmål and Nynorsk). In addition, we also expect a number of other languages to be present to a lesser degree.

We used Damir Cavar's tri-gram algorithm for language identification (cf. `http://ling.unizd.hr/~dcavar/LID/`), training 16 language models on Wikipedia text from languages that are closely related to, or that have contact with Bokmål (Bokmål, Danish, Dutch, English, Faeroese, Finnish, French, German, Icelandic, Italian, Northern Sami, Nynorsk, Polish, Russian, Spanish and Swedish). The best models were trained on 1Mb of random Wikipedia lines and evaluated against a database of one hundred 5 Kb article excerpts for each language. The models performed very well, often approaching 100% accuracy; however, the extremely similar orthography of Bokmål and Nynorsk make them the most difficult pair of languages to spot for the system, one being often misclassified as the other. In any case, our results were relatively good: *Bokmål* Precision = 1.00, Recall = 0.89, F-measure = 0.94, *Nynorsk* Precision = 0.90, Recall = 1.00, F-measure = 0.95.

The language identifying filter was applied on a document basis, recognizing about 3 out of 4 documents as Bokmål:

- 72.25% Bokmål
- 16.00% Nynorsk
- 05.80% English
- 02.43% Danish
- 01.95% Swedish

This filter produced another sensible drop in the overall number of kept documents: from 1.17 to 0.85 million.

2.6 POS-tagging and lemmatization

At the time of writing *noWaC* is in the process of being POS-tagged. This is not at all an easy task, since the best and most widely used tagger for Norwegian (the Oslo-Bergen tagger, cf. Hagen et al. 2000) is available as a binary distribution which, besides not being open to modifications, is fairly slow and does not handle large text files. A number of statistical taggers have been trained, but we are still undecided about which system to use because the available training materials for Bokmål are rather limited (about 120,000 words). The tagging accuracy we have obtained so far is still not comparable to the state-of-the-art (94.32% with TnT, 94.40% with SVMT). In addition, we are also working on creating a large list of tagged lemmas to be used with *noWaC*. We estimate that a final POS-tagged and lemmatized version of the corpus will be available in the next few weeks (in any case, before the WAC6 workshop).

3 Comparing results

While it is still too early for us to carry out a fully fledged qualitative evaluation of *noWaC*, we are able to compare our results with previous published work, especially with the WaCky corpora we tried to emulate.

3.1 NoWaC and the WaCky corpora

As we stated above, we tried to follow the WaCky methodology as closely as possible, in the hopes that we could obtain a very large corpus (we aimed at collecting above 1 billion tokens). However, even though our crawling job produced a much bigger initial archive than those reported for German, Italian and British English in Baroni et al. (2009), and

[2] As pointed out by an anonymous reviewer, this drastic reduction in number of documents may be due to faults in the boilerplate removal phase, leading to 5-grams of HTML or similar code counting as real text. We are aware of this issue, and the future versions of *noWaC* will be revised to this effect.

even though after document size filtering was applied our archive contained roughly twice as many documents, our final figures (number of tokens and number of documents) only amount to about half the size reported for the WaCky corpora (cf. table 1).

In particular, we observe that the most significant drop in size and in number of documents took place during the detection of duplicate and near-duplicate documents (drastically dropping from 11.4 million documents to 1.17 million documents after duplicate filtering). This indicates that, even if a huge number of documents in Bokmål Norwegian are present in the internet, a large portion of them must be machine generated content containing repeated n-grams that the duplicate removal tool successfully identifies and discards.[3]

These figures, although unexpected by us, may actually have a reasonable explanation. If we consider that Bokmål Norwegian has about 4.8 million potential content authors (assuming that every Norwegian inhabitant is able to produce web documents in Bokmål), and given that our final corpus contains 0.85 million documents, this means that we have so far sampled roughly one document every five potential writers: as good as it may sound, it is a highly unrealistic projection, and a great deal of noise and possibly also machine generated content must still be present in the corpus. The duplicate removal tools are only helping us understand that a speaker community can only produce a limited amount of linguistically relevant online content. We leave the interesting task of estimating the size of this content and its growth rate for further research. The Norwegian case, being a relatively small but highly developed information society, might prove to be a good starting point.

3.2 Scaling noWaC: how much Bokmål is there? How much did we get?

The question arises immediately. We want to know how representative our corpus is, in spite of the fact that we now know that it must still contain a great deal of noise and that a great deal of documents were plausibly not produced by human speakers.

To this effect, we applied the scaling factors methodology used by Kilgarrif (2007) to estimate the size of the Italian and German internet on the basis of the WaCky corpora. The method consists in comparing document hits for a sample of mid-frequency words in Google and in our corpus before and after duplicate removal. The method assumes that Google does indeed apply duplicate removal to some extent, though less drastically than we have. Cf. table 2 for some example figures.

From this document hit comparison, two scaling factors are extracted. The *scaling ratio* tells us how much smaller our corpus is compared to the Google index for Norwegian (including duplicates and non-running-text). The *duplicate ratio* gives us an idea of how much duplicated material was found in our archive.

Since we do not know exactly how much duplicate detection Google performs, we will multiply the duplicate ratio by a weight of 0.1, 0.25 and 0.5 (these weights, in turn, assume that Google discards 10 times less, 4 times less and half what our duplicate removal has done – the latter hypothesis is used by Kilgarriff 2007).

- Scaling ratio (average):
  Google frq. / noWaC raw frq. = 24.9
- Duplicate ratio (average):
  noWaC raw frq. / dedup. frq. = 7.8

We can then multiply the number of tokens in our final cleaned corpus by the scaling ratio and by the duplicate ratio (weighted) in order to obtain a rough estimate of how much Norwegian text is contained in the Google index. We can also estimate how much of this amount is present in *noWaC*. Cf. table 3.

Using exactly the same procedure as Kilgarrif (2007) leads us to conclude that *noWaC* should contain over **15%** of the Bokmål text indexed by Google. A much more restrictive estimate gives us about **3%**. More precise estimates are extremely difficult to make, and these results should be taken only as rough approximations. In any case, *noWaC* certainly is a reasonably representative web-corpus containing between 3% and 15% of all the currently indexed online Bokmål (Kilgarriff reports an estimate of 3% for German and 7% for Italian in the WaCky corpora).

[3] Although we are aware that the process of duplicate removal in *noWaC* must be refined further, constituting in itself an interesting research area.

	deWaC	*itWaC*	*ukWaC*	***noWaC***
N. of seed pairs	1,653	1,000	2,000	1,000
N. of seed URLs	8,626	5,231	6,528	6,891
Raw crawl size	398GB	379GB	351GB	**550GB**
Size after document size filter	20GB	19GB	19GB	22GB
N. of docs after document size filter	4.86M	4.43M	5.69M	11.4M
Size after near-duplicate filter	13GB	10GB	12GB	**5GB**
N. of docs after near-duplicate filter	1.75M	1.87M	2.69M	1.17M
N. of docs after lang-ID	–	–	–	**0.85M**
N. of tokens	1.27 Bn	1.58 Bn	1.91 Bn	**0.69 Bn**
N. of types	9.3M	3.6M	3.8M	6.0M

Table 1: Figure comparison of noWaC and the published WaCky corpora (German, Italian and British English data from Baroni et al. 2009)

Word	Google frq.	noWaC raw frq.	noWaC dedup. frq.
bilavgifter	33700	1637	314
mekanikk	82900	3266	661
musikkpris	16700	570	171

Table 2: Sample of Google and noWaC document frequencies before and after duplicate removal.

noWaC	Scaling ratio	Dup. ratio (weight)	Google estimate	% in noWaC
0.69 bn	24.9	0.78 (0.10)	21.8 bn	3.15%
		1.97 (0.25)	8.7 bn	7.89%
		3.94 (0.50)	4.3 bn	15.79%

Table 3: Estimating the size of the Bokmål Norwegian internet as indexed by Google in three different settings (method from Kilgarriff 2007)

4 Concluding remarks

Building large web-corpora for languages with a relatively small internet presence and with a limited speaker population presents problems and challenges that have not been found in previous work. In particular, the amount of data that can be collected with similar efforts is considerably smaller. In our experience, following as closely as possible the WaCky corpora methodology yielded a corpus that is roughly between one half and one third the size of the published comparable Italian, German and English corpora.

In any case, the experience has been very successful so far, and the first version of the *noWaC* corpus is about the same size than the largest currently available corpus of Norwegian (i.e. Norske Aviskorpus, 700 million tokens), and it has been created in just a minimal fraction of the time it took to build it.

Furthermore, the scaling experiments showed that *noWaC* is a very representative web-corpus containing a significant portion of all the online content in Bokmål Norwegian, in spite of our extremely drastic cleaning and filtering strategies.

There is clearly a great margin for improvement in almost every processing step we applied in this work. And there is clearly a lot to be done in order to qualitatively assess the created corpus. In the future, we intend to pursue this activity by carrying out an even greater crawling job in order to obtain a larger corpus, possibly containing over 1 billion tokens. Moreover, we shall reproduce this corpus creation process with the remaining two largest written languages of Norway, Nynorsk and North Sami. All of these resources will soon be publicly and freely available both for the general public and for the research community.

Acknowledgements

Building *noWaC* has been possible thanks to NOTUR advanced user support and assistance from the Research Computing Services group (Vitenskapelig Databehandling) at USIT, University of Oslo. Many thanks are due to Eros Zanchetta (U. of Bologna), Adriano Ferraresi (U. of Bologna) and Marco Baroni (U. of Trento) and two anonymous reviewers for their helpful comments and help.

References

M. Baroni and S. Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, pages 1313–1316, Lisbon. ELDA.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, 09.

David Crystal. 2001. *Language and the Internet*. Cambridge University Press, Cambridge.

A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the WAC4 Workshop at LREC 2008*.

R. Ghani, R. Jones, and D. Mladenic. 2001. Mining the web to create minority language corpora. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 279–286.

Johannessen J.B. Nøklestad A. Hagen, K. 2000. A constraint-based tagger for norwegian. *Odense Working Papers in Language and Communication*, 19(I).

Knut Hofland. 2000. A self-expanding corpus based on newspapers on the web. In *Proceedings of the Second International Language Resources and Evaluation Conference*, Paris. European Language Resources Association.

Adam Kilgarriff and Marco Baroni, editors. 2006. *Proceedings of the 2nd International Workshop on the Web as Corpus (EACL 2006 SIGWAC Workshop)*. Association for Computational Linguistics, East Stroudsburg, PA.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–348.

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

S. Sharoff. 2005. Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, (11):435–462.