Proceedings of the 12th Web as Corpus Workshop, pages 14–22
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

From Web Crawl to Clean Register-Annotated Corpora

Veronika Laippala, Samuel Rönnqvist, Saara Hellström, Juhani Luotolahti
Liina Repo, Anna Salmela, Valtteri Skantsi and Sampo Pyysalo
Turku NLP Group, University of Turku, Turku, Finland
{mavela,saanro,sherik,mjluot,tlkrep,annsalm,valtteri.skantsi,sampo.pyysalo}@utu.fi

Abstract

The web presents unprecedented opportunities for large-scale collection of text in many languages. However, two critical steps in the development of web corpora remain challenging: the identification of clean text from source HTML and the assignment of genre or register information to the documents. In this paper, we evaluate a multilingual approach to this end. Our starting points are the Swedish and French Common Crawl datasets gathered for the 2017 CoNLL shared task, particularly the URLs. We 1) fetch HTML pages based on the URLs and run boilerplate removal, 2) train a classifier to further clean out undesired text fragments, and 3) annotate text registers. We compare boilerplate removal against the CoNLL texts, and find an improvement. For the further cleaning of undesired material, the best results are achieved using Multilingual BERT with monolingual fine-tuning. However, our results are promising also in a cross-lingual setting, without fine-tuning on the target language. Finally, the register annotations show that most of the documents belong to a relatively small set of registers which are relatively similar in the two languages. A number of additional flags in the annotation are, however, necessary to reflect the wide range of linguistic variation associated with the documents.

Keywords: Register, genre, web-as-corpus, boilerplate removal, web data, web scraping

1. Introduction

Traditionally, linguistic corpora are collected in order to represent a language or a specific part of it (McEnery and Wilson, 1996; Biber et al., 1998; Kytö and Ludeling, 2008). Typically, in order to do so, corpora are composed of texts chosen to represent different genres or *registers*, that is, situationally defined text varieties such as news, blogs or discussion forum comments (Biber, 1988). Many web-based language resources diverge from this process by not being based on detailed compilation criteria (see, however, Schäfer (2016c)). Instead of the collection of coherent, high-quality text, the construction of web language resources commonly emphasizes gathering as much data as possible, for instance by using a dedicated crawl or extracting data from existing crawl-based datasets, such as Common Crawl[1]. As crawling and compilation pipelines are based on automatic processes, the resulting data can contain boilerplate texts, machine translations, and even text in languages other than that targeted in the corpus construction. Furthermore, there is typically no information on the kinds of registers that the web language resources represent. Although both linguistic and NLP efforts have achieved significant advances using web data (e.g. Mikolov et al. (2013), Bojanowski et al. (2017), Yang et al. (2019)), for a number of end uses, better structured web language resources with clean, full texts and register information would be essential to realizing their full potential.

Currently, a number of large web-crawled datasets are available. However, resources emphasizing the collection of clean texts, such as WaCky (Baroni et al., 2009) and COW (Schäfer, 2016b), represent only a limited number of languages. The ones with a more extensive selection of languages, such as OSCAR (Ortiz Suárez et al., 2019), have not gone through detailed text cleaning processes. Moreover, register information or further NLP processing steps such as syntactic analysis is typically not included at all.

In this paper, we present efforts toward the automatic creation of multilingual web-based language resources that consist of coherent, clean texts and include similar metadata to what traditional language resources have, in particular registers identified using a detailed, systematic register hierarchy. By *coherent texts*, we understand texts where each text part is linked to the others to form a full, meaningful whole (Halliday, 1976).

Our starting point is the Common Crawl dataset gathered for the 2017 CoNLL shared task (Ginter et al., 2017). Altogether, the dataset includes 56 languages, but in this paper, we focus on the Swedish and French collections. We 1) fetch pages from the URLs found in the collections and run boilerplate removal on the raw HTML, 2) train a classifier to further remove undesired text fragments that may remain, and 3) annotate text registers. The registers, such as *News report* or *Description with intent to sell*, are annotated using the taxonomy presented for English by Egbert et al. (2015) and also applied in Finnish by Laippala et al. (2019). To evaluate the need for boilerplate removal, we compare three versions of the data that have gone through different cleaning processes: 1) texts as included in the CoNLL collections, 2) raw texts after simple removal of markup from the fetched HTML pages, and 3) texts from the HTML pages cleaned of boilerplate and other unwanted elements using the web scraping tool Trafilatura[2]. The process is described in Figure 1.

We make all the resources introduced in this effort freely available under open licences at `https://github.com/TurkuNLP/WAC-XII`.

2. Related work

Web-based language resources are widely applied both in linguistic and NLP research. The WaCky Corpus Collection (Baroni et al., 2009) with more than a billion words in

[1] `https://commoncrawl.org`

[2] `https://github.com/adbar/trafilatura`

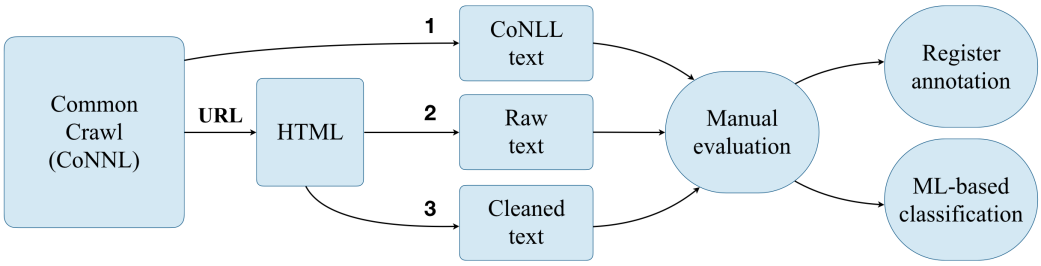

Figure 1: Text preprocessing and annotation process. Three versions of text are manually evaluated: 1) texts taken directly from the CoNLL version of Common Crawl that have undergone a cleaning process, 2) raw texts extracted from HTML based on the CoNLL URLs, and 3) texts extracted from CoNLL URLs by the boilerplate removal system (Trafilatura)

English, French, Italian and German was one of the earliest ones and perhaps mainly targeted at research questions in linguistics. Similarly, the COW corpora (Schäfer, 2016b) are linguistically processed and include billions of words in six European languages. CommonCrawl is a free and openly available web crawl maintained by the CommonCrawl foundation. The dataset is available at Amazon EC2-cloud as both plain text and HTML. The data totals petabytes in size. Lately the Common Crawl dataset has been used to gather text corpora for a number of NLP projects, such as the recently introduced massive multilingual corpus OSCAR (Ortiz Suárez et al., 2019).

An important part of processing web-based datasets for use in linguistic and NLP research is the extraction of the main body of the text and the removal of boilerplate text, such as lists, links and other unwanted material. These decrease the quality of the data as they brake the coherence of the texts by not including full sentences and by presenting individual, repetitive segments such as copyright indications. In the existing web-based language resources, the cleaning process is performed in different ways. The WaCkies (Baroni et al., 2009) use regular expressions and heuristic rules to remove boilerplates. The heuristics are based on the idea that HTML tags co-occur frequently with boilerplates, whereas the document parts with low HTML tag density belong often to main text body. Cow corpora (Schäfer et al., 2013) are processed based on a detailed pipeline with a tool classifying paragraphs as boilerplate or not (Schäfer, 2016a) and a another one classifying entire documents as coherent text or not (Schäfer et al., 2013). These are based on manually annotated data and a document-level unsupervised method to evaluate the text quality based on short and very frequent words. To create the monolingual OSCAR subcorpora, Ortiz Suárez et al. (2019) processed Common Crawl data using a pipeline based on the system by Grave et al. (2018), which included language detection using fastText (Joulin et al., 2016), basic heuristic cleaning, and hashing-based deduplication, but no boilerplate removal.

A number of readily available boilerplate removal packages exist. JusText[3] is a frequently applied boilerplate removal package in python. Trafilatura[4] is a recently developed web-scraping python library that preserves also some of the web page structure. According to an evaluation included in its documentation,[5] Trafilatura achieves an accuracy of 91% and outperforms a number of similar tools, including jusText.

Language	Documents	Tokens
French	18.2 million	10.5 billion
Swedish	19.4 million	7.7 billion

Table 1: Sizes of the deduplicated CoNLL 2017 Common Crawl-based datasets for French and Swedish

Thus, several well-developed web corpus resources and ready-made solutions for boilerplate removal and text cleaning exist. In contrast, the addition of register information to web-scale corpora is not yet common practice and involves many challenges. A first challenge has been the lack of annotated corpora that represent all the registers found online. Because of this, there has been no training data available to develop web register identification systems that could be applied to classify web-based language resources. Two large corpora with register annotations exist for English, the Leeds Web Genre Corpus (Asheghi et al., 2016) and the Corpus of Online Registers of English (CORE) (Egbert et al., 2015). A small collection of online registers has also been released for Finnish (Laippala et al., 2019). Second, another challenge with online registers is that online language use cannot necessarily be described in terms of discrete register categories. For instance, an online text might simultaneously have characteristics of a news article and a persuasive text. Thus, discrete register classification systems where each document belongs to exactly one register category do not necessarily suit web data sets very well. To solve this, the CORE corpus includes *hybrid* register categories that combine several register labels, such as *narrative+opinion* (see Biber and Egbert (2018)). Another solution is suggested by Sharoff (2018), who analyzes registers by describing texts based on proportions of dimensions, such as argumentative or hard news.

3. Data and annotation

In this section, we present the CoNLL data we use as source, the preprocessing steps we applied, and the annotation processes we performed. The overall workflow is presented in Figure 1.

[3] https://pypi.org/project/jusText/

[4] https://trafilatura.readthedocs.io/en/latest/

[5] https://trafilatura.readthedocs.io/en/latest/evaluation.html

Label	Text
1	Prinsessan Madeleine besökte Childhood-projekt i Florida och New York - Sveriges Kungahus 'Princess Madeleine visited Childhood projects in Florida and New York - The Swedish Royal Court'
0	Länk till sidan Anpassa webbplatsen 'Link to the site Customize the Web Site'
0	Länk till Startsidan 'Link to the home page'

Table 2: Example of text quality annotation for Swedish data. Lines marked with the label 1 are judged to be part of the main body of the text.

3.1. Source data

The source data for our study is gathered from the Common Crawl-based dataset prepared for the 2017 CoNLL shared task (Ginter et al., 2017). The Common Crawl data is available on the Amazon cloud, which was used for data collection and language detection. The Compact Language Detect 2 (CLD2) language detector[6] was applied in processing due to its speed and the availability of python bindings. For each processed plain text input file, the first 100 000 tokens per language were kept, and deduplication based on URLs was performed. The resulting dataset is composed of altogether 56 languages and nearly 100 billion words. The statistics of the French and Swedish collections used in this study are summarized in Table 1.

3.2. Text quality annotation

We evaluate the quality of texts by manually annotating three text versions that have gone through different cleaning processes: 1) text as they are included in the CoNLL data, 2) raw texts extracted from HTML source without boilerplate removal, and 3) texts extracted from HTML and processed with Trafilatura to remove boilerplate material. The raw texts are included in order to assess whether any good text content may have been lost and to provide an up-to-date point of reference for Trafilatura, as some of the online documents may have changed after the collection of the original source data in 2017.

The evaluation was done by 1) selecting from CoNLL data 40 documents (20 in Swedish and 20 in French) with active URLs and 2) manually annotating the quality of all three versions of these documents. The annotation was done on a line-by-line basis, coding which lines are part of the coherent texts and which are part of boilerplate.[7] To define boilerplate, we followed Schäfer (2016a), according to whom boilerplate is all material that "*remains after markup stripping, and which does not belong to one of those blocks of content on the web page that contain coherent text.*"

The annotation was performed by four annotators in total. Annotations were done individually, but difficult cases were discussed jointly with an annotation coordinator. Although many lines and text segments are easy to define as not belonging to the coherent text, the quality annotation was by no means a trivial task. Many lines could have been defined as either coherent text or boilerplate. Examples of undesired lines include links and lists of words or headlines that were not connected to body text, e.g., when serving as links to other pages. Automatically generated text was similarly excluded, e.g., headlines in a banner, phrases such as *visa mer* 'show more' and *fäll ihop* 'hide'.

Table 2 shows examples of lines annotated as belonging to the text and lines annotated as undesired material. The first line is a headline describing the text to come and its topic: the visit of the Princess Madeleine of Sweden. As this headline is not followed by other headlines, it is considered as belonging to the coherent text. The next two lines, in turn, are both links to other parts of the website. They do not belong to the coherent text and are thus annotated as undesired material to be rejected.

Narrative
News report / news blog, sports report, personal blog, historical article, fiction, travel blog, community blog, online article

Opinion
Review, opinion blog, religious blogs/sermon, advice

Informational Description
Description of a thing, encyclopedia article, research article, description of a person, information blog, FAQ, course material, legal terms / condition, report, job description

Discussion
Discussion forum, question-answer forum

How-to
How-to/instruction, recipe

Informational Persuasion
Description with intent to sell, news+opinion blog / editorial

Lyrical
Songs, poem

Spoken
Interview, formal speech, TV transcript

Table 3: Register classes in the taxonomy. Main register classes are shown in bold.

3.3. Register annotation

The register-annotated documents are sampled from the CoNLL data. The register annotation follows the register taxonomy presented for the English CORE corpus by Egbert et al. (2015) and for Finnish by Laippala et al. (2019). The advantage of this taxonomy is that it is developed in a data-driven manner and it covers the full range of registers and linguistic variation found online. Furthermore, as dis-

[6] https://github.com/CLD2Owners/cld2

[7] Lines correspond broadly to blocks of text uninterrupted by tags in the source HTML, such as titles or paragraphs.

cussed in Section 2., the annotation allows the assignment of multiple register labels for one document, which guarantees that the annotation covers the full range of language use in web documents. The taxonomy is hierarchical with eight main register classes with functional labels. These are divided into a number of sub-register categories that are perhaps more intuitive, such as *News report* and *Review*. The taxonomy is presented in Table 3.

In the English CORE for which this taxonomy was developed, each document was annotated by four coders, and hybrid annotations resulted from consistent disagreements among the coders. In our study, we do not have the resources to have such an extensive annotation process. Instead, documents were first double-annotated, and when a certain level of agreement and confidence was found between the coders, the process was changed to single annotation. However, difficult cases were always discussed and resolved jointly. In our setting, during the annotation, annotators could select several register labels for a document when necessary to fully characterize it. This allows the direct annotation of hybrid documents even by a single annotator. Moreover, if the document could not be described by a specific sub-register label, annotators could select a more general, main register label only. The annotations were done using a custom annotation tool. The tool provides annotators with a wide selection of flags that can be toggled to identify additional aspects of the texts. The set of flags was developed during the annotation with the objective of marking text properties that may have an effect on the further analysis of the data. For instance, these include *untypical for the register* and *multiple texts*.

4. Classifiers for further cleaning

We next describe our approach to training and evaluating methods for further cleaning the texts after boilerplate removal. We experiment with two supervised machine learning methods:

BERT (Devlin et al., 2018) is a deep transfer learning approach based on the Transformer architecture (Vaswani et al., 2017). We apply the Multilingual BERT (mBERT) model released by Google[8], which has been pre-trained on a combination of Wikipedia texts in 104 languages, including French and Swedish. In addition to monolingual classification in the two languages, we apply mBERT also in multilingual and cross-lingual training setups. Following Devlin et al. (2018), we add a final classification layer to the pre-trained transformer stack, and fine-tune all model weights.

fastText (Joulin et al., 2016) is a text classification tool emphasizing computational efficiency, making it a popular choice for machine learning on web-scale data. We apply fastText as a baseline method using the supervised text classification facilities of the tool.

We train and evaluate BERT and fastText in the basic binary classification setting where each line is labelled as either 0 (rejected) or 1 (accepted). We divide both the French and the Swedish datasets into training, development, and test subsets on the document level, so that text drawn from a single document is only included in exactly one of the subsets. We perform a random stratified split so that the positive/negative distribution of each subset roughly matches that of the whole dataset (max. 2% point deviation). The test subsets were held out during method development and parameter selection.

For BERT, we perform a grid search on maximum sequence length, learning rate, batch size and number of training epochs, while evaluating on the development set. For fastText, we select the maximum number of word n-grams and the number of training epochs using grid search on the development data. We additionally evaluate the effect of initializing the word vectors for the method using pre-trained language-specific word vectors (Grave et al., 2018).

We evaluate classification performance primarily in terms of accuracy, i.e. the proportion of texts that are predicted to have the correct class. We additionally report precision and recall, summarizing performance across different classification thresholds with precision-recall curves.

[8] https://github.com/google-research/bert/blob/master/multilingual.md

French	Accept	Reject
Words	8374 (52%)	7789 (48%)
Lines	288 (23%)	956 (77%)

Swedish	Accept	Reject
Words	20694 (78%)	5961 (22%)
Lines	495 (31%)	1110 (69%)

Table 4: Text quality for CoNLL source text.

French	Accept	Reject
Words	8097 (36%)	14662 (64%)
Lines	408 (9%)	4227 (91%)

Swedish	Accept	Reject
Words	17228 (39%)	27324 (61%)
Lines	568 (11%)	4809 (89%)

Table 5: Text quality for raw text.

French	Accept	Reject
Words	6401 (79%)	1713 (21%)
Lines	306 (55%)	255 (45%)

Swedish	Accept	Reject
Words	12224 (94%)	794 (6%)
Lines	403 (84%)	77 (16%)

Table 6: Text quality for Trafilatura-processed text.

5. Results

In this section, we present the results of the evaluations. We start with the analysis of the text quality based on the manual annotations, then move on to the machine learning experiments to further clean the texts from undesired material, and finally analyze the register annotations.

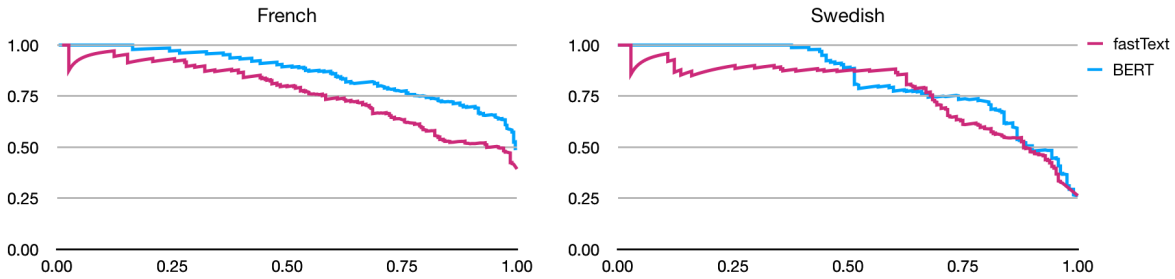

Figure 2: Precision-recall curves for the two machine learning methods. (x-axis: recall, y-axis: precision)

5.1. Text quality based on the manual annotations

The results on the manual evaluation of the text quality are presented in Table 4 for the CoNLL texts, in Table 5 for the raw text, and in Table 6 for the text processed with Trafilatura to remove boilerplate (see Section 3.2.). In the source CoNLL data, 48% of the words in French and 22% of the words in Swedish were evaluated as rejected, i.e., they appeared on lines that were not considered to belong to the coherent texts. On the line level, the proportions were even more drastic: in Swedish, 69% of the lines and in French 77% were marked as rejected. These findings suggest that the source texts may be too noisy to be used without further cleaning for many purposes and that the quality of the French CoNLL data is somewhat lower than that of the Swedish data. Moreover, the different distributions indicate that length is already a strong signal of the line belonging to the coherent text. This seems natural, as many of the short lines enumerating links are very short.

In the raw text versions extracted from HTML, the proportion of words evaluated as not belonging to the coherent texts was 64% in French and 61% in Swedish. On the line level, the rejected proportions were approximately 90% for both languages. Thus, despite its issues, the CoNLL data is clearly cleaner and of better quality than text extracted directly from HTML.

For Trafilatura, the proportions of rejected material were clearly lower than in the other settings. On the word level, the Swedish contained only 6% of rejections and the French 21%, while on the line level, the proportions were 16% and 45% (resp.). Text processed with Trafilatura is thus cleaner than the CoNLL data, and its use is motivated even if the CoNLL data has already gone through some cleaning. On the other hand, the Trafilatura cleaning process does also discard some parts of the raw text that were evaluated as belonging to the text. For Swedish, 5004 words – approximately 29% of accepted words in the raw text extracted from HTML – were deleted by Trafilatura. Similarly, in, French, 1696 accepted words, that is, 21%, were deleted. Thus, obtaining cleaner text in this way also has the downside of not acquiring all the text available. Whether this trade-off is acceptable is likely to depend on the purpose for which the text is processed.

5.2. Classifiers for further cleaning

The machine learning results are based on altogether 50+50 documents from the CoNLL data: 20+20 as described in Table 4 and an additional set of 30+30 documents we annotated in order to guarantee high system performance. Table 7 summarizes the key statistics of the training, development, and test division of the data.

French	Train	Dev	Test	Total
Lines	2437	673	696	3806
Words	44529	15415	11636	71580
Positives	908	253	274	1435
Negatives	1530	421	423	2374

Swedish	Train	Dev	Test	Total
Lines	2867	788	806	4461
Words	37855	9286	10007	57148
Positives	812	222	212	1246
Negatives	2056	567	595	3218

Table 7: Data statistics. Positives refer to the accepted lines annotated as part of the coherent texts, while negatives are the rejected lines annotated as undesired material.

We set machine learning method parameters in a monolingual setting by optimizing the hyperparameters for French and Swedish separately on the development subsets. For mBERT, we found the optimal hyperparameter settings to be largely in agreement across the two languages: both models use a maximum sequence length of 192, batch size of 16 and are trained for 6 epochs. The Swedish model was trained with a learning rate of $2.5e^{-6}$ and the French with $5.0e^{-6}$. For fastText, we selected word n-grams up to length three and training for 30 epochs, initializing the word vectors randomly as pre-initialized vectors did not show a clear benefit in evaluation on the development data. The final evaluation results on the test sets are shown in Table 8. Both fastText and mBERT clearly outperform the majority baseline, and mBERT achieves the best results for both languages, with a more notable advantage for the French data, reaching an accuracy of 85.62% for French and 81.64% for Swedish. Figure 2 shows the precision-recall curves for the two methods. We find that mBERT systematically outperforms fastText across the entire recall range for French, but dips below the precision of fastText for part of the scale for Swedish.

We continued to explore whether training the better-performing method, mBERT, on data combining annotations from both languages could further improve performance, evaluating on each language separately. The multilingual model was trained with the above settings, and the

	French	Swedish
mBERT	**81.64**	**85.62**
fastText	75.68	84.61
Majority	60.69	73.73

Table 8: Monolingual classification accuracy.

	Test	
Train	French	Swedish
French	**81.64**	76.61
Swedish	69.72	**85.62**
Fr + Sv	79.34	82.28

Table 9: Cross- and multilingual classification accuracy with mBERT. Monolingual results are repeated for reference.

learning rate of $2.5e^{-6}$ was found to perform best on the development set. Despite the increase in training data size, the multilingual model falls behind its monolingual counterparts by 2-3% points on the two languages (Table 9).

Finally, we assessed how well the monolingually trained classifiers perform in a zero-shot, cross-lingual learning setting, i.e., how well they can predict in a language not seen during fine-tuning. While we observed a 5% point drop for Swedish, the drop was 16% points for French (Table 9). Nevertheless, both models manage to outperform the majority baseline even in this setting. This is encouraging for the multilingual long-term objective of our project, as it shows that machine learning-based text cleaning is possible even without language-specific training data.

5.3. Large-scale identification of coherent text

Finally, we apply the developed classifiers to a large body of unannotated texts to further assess the ratio of clean text in the source data. In the French and Swedish CoNLL data, we randomly sample URLs from which we then extract the texts using Trafilatura. The process is continued until we reach 10,000 lines in each language.

We classify these lines using the French and Swedish monolingually tuned mBERT models described above, and observe the class proportions as summarized in Table 10. Both languages exhibit a similar distribution – about 27–29% of lines are accepted by the models – while in terms of number of words the ratios are close to the inverse. Somewhat less content is accepted for French than for Swedish, even though the class distribution in the training data was more skewed toward the negative class for Swedish. This supports our earlier finding that the French source data has a lower ratio of clean text than Swedish (Section 5.1.).

	French	Swedish
Lines	26.89%	29.48%
Words	70.91%	71.47%

Table 10: Proportion of accepted text in Trafilatura output based on mBERT predictions.

5.4. Register annotation results

The register-annotated datasets include 688 documents in French and 1085 in Swedish. The most frequent registers in these datasets as well as the frequencies of the additional flags are show in Tables 11 and 12, and the proportions of the registers in the two languages are illustrated in Figure 3. Although the rankings of the registers differ, the sets of the most frequent registers in the two languages are quite similar. In other words, similar registers seem to be the most frequent ones, and many of the registers described in the annotation scheme (Table 3) remain infrequent. Both languages include a large number of texts labeled as *Description with intent to sell*, *News* and *Personal blog*. Differences arise with *Machine translation, Personal opinion blog* and *Encyclopedia article*. The frequency of *Machine translation* is certainly a sign of its frequency on the Internet. For the other classes, the differences may reflect true language-specific distributions of registers. These will be further examined in future work with more extensive datasets.

Another interesting property in the annotations is that *Informational persuasion* is the only main register among the most frequent ones in both languages. Its frequency may reflect linguistic variation displayed within this register and the fact that documents within it are difficult to assign a specific category. Additionally, it is noteworthy that hybrid categories are relatively infrequent and do not show among the most frequent classes.

The additional flags show the range of linguistic variation and textual composition displayed by the documents. Many of the flags reflect textual properties that can affect the modeling of the documents. *Comments* can be particularly frequent in some registers. In the analyzed data, this is the case with Swedish *Opinion blog* and *Personal blog*. Linguistically, they may be more conversational than the bodies of the texts, which motivates the annotation of the flag.

Similarly, *foreign language* and *generated text* may be used in the text for instance in quotations. These are naturally very different from the language otherwise used in the documents. In our data, *foreign language* seems relatively infrequent, but *generated text* is flagged quite often. Its proportion can, however, decrease when the text cleaning process improves.

Multiple texts and *missing text*, again, are frequent properties of web documents. For instance, a document from a news site may include many headlines and beginnings of the actual news articles, which are then fully displayed on a page of their own. The structuring of these texts may show also in their linguistic characteristics. In our annotation results, these properties are flagged in both languages with frequencies ranging between 0% and 39%. Similar to *comments*, the frequency of these flags can correlate with specific register classes. For instance, 25% of the French and 39% of the Swedish annotations in the *News report* class were flagged as *multiple texts*, while the frequency of this flag was 0% for the *Discussion forum* class in both languages.

Finally, the flag *untypical for the register* reflects linguistic variation within register categories, and is used when the document differs from a typical example of its register. Indicating this helps to further analyze the annotation

	Number of documents	Comments	Missing text	Foreign language	Generated text	Untypical for register	Multiple texts
Description with intent to sell	136	0%	10%	2%	13%	3%	10%
News report / news blog	75	1%	28%	0%	7%	7%	25%
Encyclopedia article	45	0%	18%	0%	22%	7%	2%
Description of a thing	45	0%	16%	2%	20%	0%	2%
Personal blog	33	3%	15%	3%	6%	9%	12%
Discussion forum	33	0%	0%	0%	33%	12%	0%
Reviews	32	3%	16%	0%	28%	9%	22%
How-to / instruction	25	0%	0%	0%	24%	12%	4%
Informational persuasion	25	0%	4%	0%	28%	0%	8%

Table 11: Annotation statistics for the French data

	Number of documents	Comments	Missing text	Foreign language	Generated text	Untypical for register	Multiple texts
Encyclopedia article	223	0%	18%	0%	83%	6%	0.5%
Personal blog	157	32%	8%	6%	31%	2%	9%
Description with intent to sell	136	4%	6%	0%	30%	6%	12%
News report / news blog	109	3%	28%	0%	17%	28%	39%
Opinion blog	45	24%	13%	2%	27%	4%	11%
MT /generated text	37	8%	3%	8%	11%	16%	22%
Description of a thing	27	0%	15%	0%	26%	0%	15%
Discussion forum	20	5%	0%	5%	35%	15%	0%
Informational persuasion	19	0%	11%	0%	32%	0%	16%

Table 12: Annotation statistics for the Swedish data

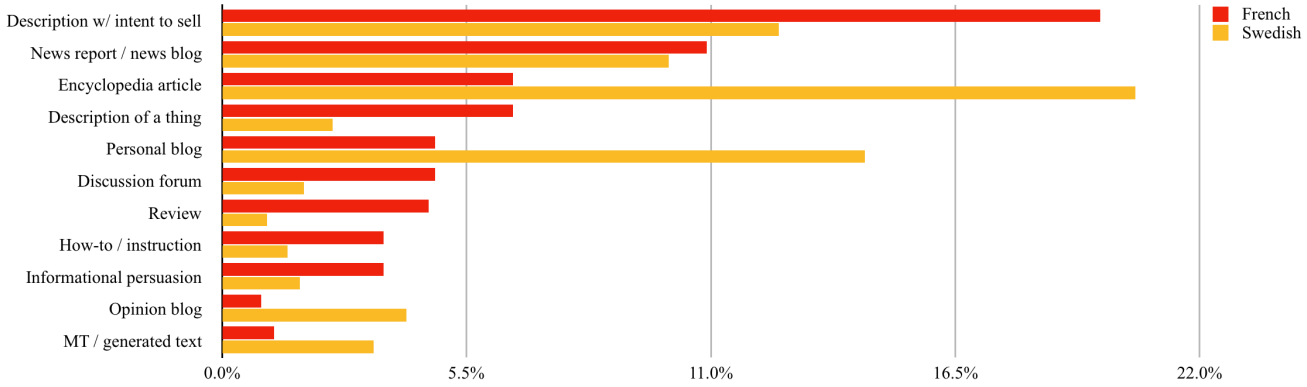

Figure 3: Proportions of registers in the two languages.

decisions if needed. In the annotations, this flag is marked for approximately 10% of the documents. In particular, the flag is frequent in the Swedish *News report* class with a proportion of 28%. This can be symptomatic of the range of linguistic variation within this register.

The register annotation and the different flags are illustrated in Table 13. The example text is annotated as belonging to the Review register. The text is taken from the middle of the original document which is a customer review in an online book store. The actual text is preceded and followed by automatically generated text that is frequent in this kinds of web documents: 'Add to cart' and 'More books on'. The text includes two separate reviews. The first one is present in its entirety, but the second review ends with . . . and continues on another page. These properties are described in the annotation by the additional flags.

6. Conclusions and future work

In this study, we have explored the challenges in deriving clean, register-annotated texts from the web. Our starting points were the Swedish and French Common Crawl datasets gathered for the 2017 CoNLL shared task (Ginter et al., 2017), and our approach consisted of three steps: the evaluation of the text quality in order to assess the benefit of boilerplate removal, the development of a classifier to further clean the texts, and the annotation of registers.

First, we manually evaluated three versions of the data that had gone through different cleaning processes: CoNLL versions, raw text versions derived from HTML by stripping markup and cleaned versions extracted from HTML using the boilerplate removal system Trafilatura. The evaluation of the text quality showed that the use of boilerplate removal improves the text quality clearly, although the process also incorrectly rejects some parts belonging to the main text body. In our project, the trade-off – loosing a small proportion of coherent text while improving overall

Original Swedish	Translation
Lägg i varukorg	'Add to cart'
jag tyckte boken var fin med vackra bilder, väntade mig dock mer lantlig känsla, vet ej varför fick bara det intrycket med titeln men alla hem var moderna med stads känsla, inredda med vintage och antikviteter	'i thought the book was nice with beautiful pictures, however, I expected a more rustic feeling, donât know why just got the impression from the title but all the homes were modern with a city-like feeling, decorated with vintage and antiquities '
Vartenda uppslag är fantastiskt! En ren njutning som...	'Every page is fantastic! Pure pleasure that . . . '
Fler böcker inom	'More books on'

Table 13: Swedish text example with English translations on the right. Register: Review; Additional flags: Generated text, Part of the text is missing.

quality – is acceptable, as it does not reduce the size of the data substantially.

To facilitate further cleanup of the resulting texts, as a second step, we trained classifiers for distinguishing coherent text content from other, undesirable material. Monolingually fine-tuned Multilingual BERT models achieved the best results for both French and Swedish. Additionally, we tested multi- and cross-lingual settings to investigate to what extent the cleaning could be realized with a joint model or in a language not seen during training. Combining the languages during training in the multilingual setup performed well, but did not outperform the monolingual classifiers. The cross-lingual, zero-shot setting did perform above baseline, which indicates that further cleaning of the texts can be done (to some extent) in multilingual settings without the time-expensive annotation of data in each of the languages under study. This is very encouraging for our project.

Finally, we examined the register annotations and the proportions of different registers in the two languages. This analysis showed that most of the documents belong to a relatively small set of the most frequent registers, although the annotation scheme does cover a wide range of registers and their combinations. Additionally, the sets of the most frequent registers are relatively similar in the two languages. This finding is also very encouraging for our future plans. Specifically, we intend to extend to a larger set of languages already covered in the CoNLL data. We will also experiment with the possibility of combining the line-wise estimates of text quality at the document level. Finally, we will continue the register annotations with the objective of being able to automatically attach detailed register information to all the data.

We release the materials and methods introduced in this study under open licenses at `https://github.com/TurkuNLP/WAC-XII`.

Acknowledgements

The work was funded the Foundation of Emil Aaltonen. Computational resources for this work were provided by CSC – Finnish IT Center for Science.

7. Bibliographical References

Asheghi, N. R., Sharoff, S., and Markert, K. (2016). Crowdsourcing for web genre annotation. *Language Resources and Evaluation*, 50(3):603–641.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, September.

Biber, D. and Egbert, J. (2018). *Register variation online*. Cambridge University Press, Cambridge.

Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press, Cambridge.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press, Cambridge.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135â146, Dec.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Egbert, J., Biber, D., and Davies, M. (2015). Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.

Ginter, F., Hajič, J., Luotolahti, J., Straka, M., and Zeman, D. (2017). CoNLL 2017 shared task - automatically annotated raw texts and word embeddings. LINDAT/CLARIN digital library at ÚFAL.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Halliday, M. A. K. (1976). *Cohesion in English*. English language series ; 9. Longman, London.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *CoRR*, abs/1607.01759.

Kytö, M. and Ludeling, A. (2008). Collection strategies and design decisions. In *Corpus Linguistics: An International Handbook*, chapter 9.

Laippala, V., Kyllönen, R., Egbert, J., Biber, D., and Pyysalo, S. (2019). Toward multilingual identification of online registers. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 292–

297, Turku, Finland, September–October. Linköping University Electronic Press.

McEnery, T. and Wilson, A. (1996). *Corpus Linguistics*. Edinburgh University Press, Edinburgh.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Ortiz Suárez, P. J., Sagot, B., and Romary, L. (2019). Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In Piotr Bański, et al., editors, *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom, July. Leibniz-Institut für Deutsche Sprache.

Schäfer, R., Barbaresi, A., and Bildhauer, F. (2013). The good, the bad, and the hazy: Design decisions in web corpus construction. In Stefan Evert, et al., editors, *Proceedings of the 8th Web as Corpus Workshop (WAC-8)*, pages 7–15, Lancaster. SIGWAC.

Schäfer, R. (2016a). Accurate and efficient general-purpose boilerplate detection for crawled web corpora. *Language Resources and Evaluation*. online first.

Schäfer, R. (2016b). Commoncow: Massively huge web corpora from commoncrawl data and a method to distribute them freely under restrictive eu copyright laws. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4500–4504, PortoroÅ$\frac{3}{4}$, Slovenia. European Language Resources Association (ELRA).

Schäfer, R., (2016c). *Proceedings of the 10th Web as Corpus Workshop*, chapter On Bias-free Crawling and Representative Web Corpora, pages 99–105. Association for Computational Linguistics.

Sharoff, S. (2018). Functional text dimensions for the annotation of web corpora. *Corpora*, 1(13):65–95.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding.