Semi-Manual Annotation of Topics and Genres in Web Corpora

The Cheap and Fast Way

Vít Suchomel[†‡], Jan Kraus[‡]

†Natural Language Processing Centre
Faculty of Informatics, Masaryk University, Brno, Czechia

‡Lexical Computing
Brno, Czechia
name.surname@sketchengine.eu

Abstract. In this paper we present a cheap and fast semi-manual approach to annotation of topics and genres in web corpora.
The main feature of our method is assigning the same topic or genre label to all web pages coming from websites most represented in the corpus. We assume that web pages within a site share the topic of the whole domain. According to the evaluation of texts coming from sites that were manually assigned a topic label, our hypothesis holds in 92 % of cases. In other words, the noise in these semi-manually labelled web pages is just 8 %. That is clean enough to train a classifier of texts from websites not seen in the process.
The procedure of fast manual topic and genre labelling of web domains is described in this paper. Recommendations for training a topic or genre classifier using semi-manually labelled texts from large websites follow.

Keywords: web corpus, text corpus, topic, genre, text annotation

1 Introduction and Motivation

According to [1], text corpora built from texts from the internet are used for language modelling, information retrieval, question answering, automatic population of ontologies, translating terms and language teaching. Text collections large enough to find evidence of scarce language phenomena in natural language context have to be compiled from the largest, free and easy-to-use data source – the web.

Understanding the sources of a web corpus and its content is important for users of web corpora. However, the internet is not organized by linguistic properties of the text or text types so one has to add the desired metadata by manual or automated ways. This paper is based on our experience with adding information about text topic and genre to web pages in the TenTen family of web corpora [2] for corpus manager Sketch Engine [3].

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022, pp. 141–148, 2022. © Tribun EU 2022

Once the information about text types is inserted, Sketch Engine allows the users to see the size of data within a corpus, as document count, sentence count, or token count by topic or by genre. For example, genre *news* or topic *religion* usually belong to the most represented text types in corpora in languages with a small presence on the web.

The users can also limit their search queries or other corpus based analyses to a subcorpus made from a single text type or from a combination of text types. For example, terminologists may require working just with documents labelled by topic *nature & environment*. Another example could be a language model for a typing prediction software. The model would benefit from texts labelled by genre *discussion*. On ther other hand, genre *legal* should be avoided or reduced in this case of corpus use.

This paper is a follow-up to our recent work on semi-manual methods of computer generated text removal and annotation of topics and genres in web corpora. We proposed a semi-manual approach consisting of manually checking the largest sources of data and training a non-text classifier, using this data, for the rest of the corpus in [4, p. 85]. Our assumption in [5] was that all pages in a web domain shared the same properties with regards to text quality. We noted such hypothesis could lead to mistakes and noisy training data for a text quality classifier while there were two clear advantages of the approach: Millions of training samples for the classifier and a low cost of manually annotating the whole websites. [6] applied the approach to document metadata, namely the text topic.

An extension of the method to both topics and genres is described in this paper. Our aim is to reliably annotate a large part of a web corpus with only a small human effort, thus cheaply. The procedure of fast manual topic and genre labelling of web domains is documented in detail in the following chapters.

2 Determining a Set of Topics and Genres Feasible to Recognize

We understand the topic of a text as its subject, recognizable mostly by lexical properties of the text, i.e. its words. The genre is determined by both syntactic and lexical features of the text, i.e. defined by the style of writing.

While Dewey Decimal Classification provides a wide tree of subjects or text topics (by design ten main categories, each with ten subcategories and each with ten third level labels) and while there are 24 main genres with 31 subgenres in BNC 1994 or 8 main genres with 37 subgenres in BNC 2014 [7], web corpora are not constructed in a deliberate way and the internet is not populated by texts selected in order to belong to a pre-designed topic or genre hierarchy.

When determining topic or genre of web texts, we have to deal with large grey zones between class definitions. There are texts belonging to multiple classes, e.g. a post about a recent release of a football computer game in a personal site. – Is it a news, a blog, or both? Is the topic sports, games, or

both? How much text can form a separate topic to recognize in a multi-topic document?

To address the issue of grey zones and find a set of topics and genres feasible to recognize, we merged the definition of categories with the process of manually labelling texts. We started with a large English web corpus[1], the list of topics in web directory Curlie.org (formerly DMOZ.org)[2] and with Sharoff's Functional text dimensions for large web corpora [8]. We merged categories that caused problems to decide in which of them real web documents belong or where there was a small number of such texts, even though their definition in an annotation manual seemed clear. The real texts were just neither white nor black but grey. We did not want to keep labels with a low annotator agreement.

On the other hand, we introduced topic labels that were easy to recognize. We added more words in category names too to help understand which content belongs there. This approach is comparable to text types in Estonian National Corpus – [9, p. 215–216], in fact we were inspired by some of their topics but since we wanted to keep a high content variety within each class, labels assigned to documents from a low number of websites were discardeed – that is why we kept less categories than ENC in the end.

The list of topics and genres recognized in the English web corpus can be found in Table 1 and in Table 2, respectively. Classes with a low number of source websites were disregarded. When we applied the same label selection process to smaller corpora (in other languages than English), more classes were discarded to keep a variety of sources within each class.

3 Fast Manual Topic and Genre Labelling of Web Domains

The procedure of manual topic and genre labelling of the content of whole websites follows.

First, web domains represented in the corpus are ranked by the count of tokens they contributed. Top ranking sites, i.e. the largest sources of text, are examined thoroughly while the time spent by examining smaller sources drops with the domain size. 3,000 largest domains in the English web corpus were inspected. These sources cover 40 % of corpus tokens. For other languages, the count depends on the corpus size – usually between 300 and 1,500 largest domains, covering at least 60 % and even up to 90 % of corpora – by checking just a small part of websites to keep the process efficient.

Documents from a single domain sharing frequent prefixes of paths can be examined separately, independently on the rest of documents within the domain, to adapt to websites with multiple topics. This technique works for sites with path prefixes such as "/sports/", "/culture/", etc.

[1] The corpus was enTenTen21, obtained from the web in 2021, comprising of 65 billion tokens at this stage of processing in which sources of bad texts are identified and text types are determined.

[2] `https://curlie.org/`

Table 1: Topics recongized in a large English web corpus. Out of top 3,000 websites that were inspected, 887 were assigned a topic. Note four categories marked by the red colour that were not represented by enough websites so their labels were discarded in the final revision of the data.

Topic	Websites	Tokens
arts	12	169 655 242
beauty & women	6	45 899 006
cars & bikes	49	268 201 168
construction & real estate	1	4 610 212
culture & entertainment	123	695 609 769
economy, finance & business	62	387 271 125
education	15	79 155 574
food & drinks	2	9 774 572
gambling & casinos	1	7 839 308
games	52	324 004 431
health	59	426 786 724
history	24	176 510 675
hobbies	18	111 828 110
home, family & children	7	47 126 547
lifestyle	0	0
nature & environment	6	64 495 602
pets & animals	9	33 432 198
politics & government	27	243 239 797
reference/encyclopedias	10	4 210 237 110
religion	71	424 919 420
science	51	594 461 579
sex	10	209 398 259
sports	103	647 268 352
technology & IT	138	887 566 212
travel & tourism	31	162 020 069
Total	**887**	**10 231 311 061**

Table 2: Genres recongized in a large English web corpus. Out of top 3,000 websites that were inspected, 611 were assigned a genre.

Genre	Websites	Tokens
blog	99	748 208 188
discussion	194	1 327 118 539
fiction	55	1 009 319 746
legal	37	507 984 084
news	226	1 284 058 175
Total	**611**	**4 876 688 732**

Second, an annotator records the topic and/or the genre in an inspection table. The table is generated by a script from the list of largest sources of the corpus provided by the corpus manager. Each row of the table is dedicated to inspecting one website. To increase the efficiency of the process, the quality of the site content is checked in this phase, together with determining text types.

There are the following columns in the table:

1. The hostname (e.g. "www.bbc.com") – Names with suspicious or long words, generic or foreign TLDs, language code in TLD are checked for generated content.
2. A link to the landing page of the site (e.g. "https://www.bbc.com/") – The page is checked in a web browser for low quality text, no text, hijacked/unrelated content, selectors with too many language mutations (high chance of machine translated (MT) content, MT scripts in the source code. A dead site is suspicious too (a high quality content does not get shut down often).
3. A link to 100 random triples of consecutive sentences in context displayed in Sketch Engine – 3 to 10 sentence triples are inspected, the rest is briefly seen and consulted more in case of doubtful content in the sentences that were read well or in case of a dubious hostname or a suspicious live site. Machine generated text, clusters of nonsense characters, unrelated phrases stitched together are indicators of bad content and lead to the removal of the whole source from the corpus. Each sentence triple can be tracked to the original web page within the domain (if the page still exists) to see the live content in a web browser.
4. Topic and genre – The annotator should be able to decide if the content shows lexical or syntactic features typical for a recognized text type. No label is given if the person is not sure. No class is given instead of assigning multiple labels to pages sites with many text types.
5. The size of the site in tokens – used to estimate how much time should be spent inspecting the source.

The procedure does not require an expert linguist or computer scientist. Any person with a bit of a sense of language and a common web browsing skill is sufficient. The inspection table is a guide easy to follow. In our experience, the annotator does not even need to understand the language after some exposure to

the task in a language they are familiar with. Photos in live pages tell much about the topic, e.g. sports or health, and the structure of the page can indicate the genre, e.g. a discussion forum, without understanding a word. Browser plugins connected to Google Translate or DeepL translating the page content help in other cases.

Depending on the rank of the website, a human annotator can spend between several minutes to as less as 20 seconds with each item to inspect. Not assigning any labels is encouraged to reduce noise in the annotations. Altogether, the procedure is quite efficient because it is fast and simple:

- A large part of a corpus is covered just by checking a small amount of random sentences from the most contributing sources and checking the live sites,
- all documents from a website are labelled with the same text type as the whole site,
- and no expert skill is required.

4 All for One, and One for All? The Evaluation

[6] compared manually assigned topic labels to 960 documents (a label assigned separately to each document, regardless the site labels) with the labels of their source sites. The document labels reportedly matched the respective domain labels in 92 % of cases. An estimated noise of just 8 % in the annotated data enabled training a topic classifier using the semi-manually obtained labels.

5 Conclusions

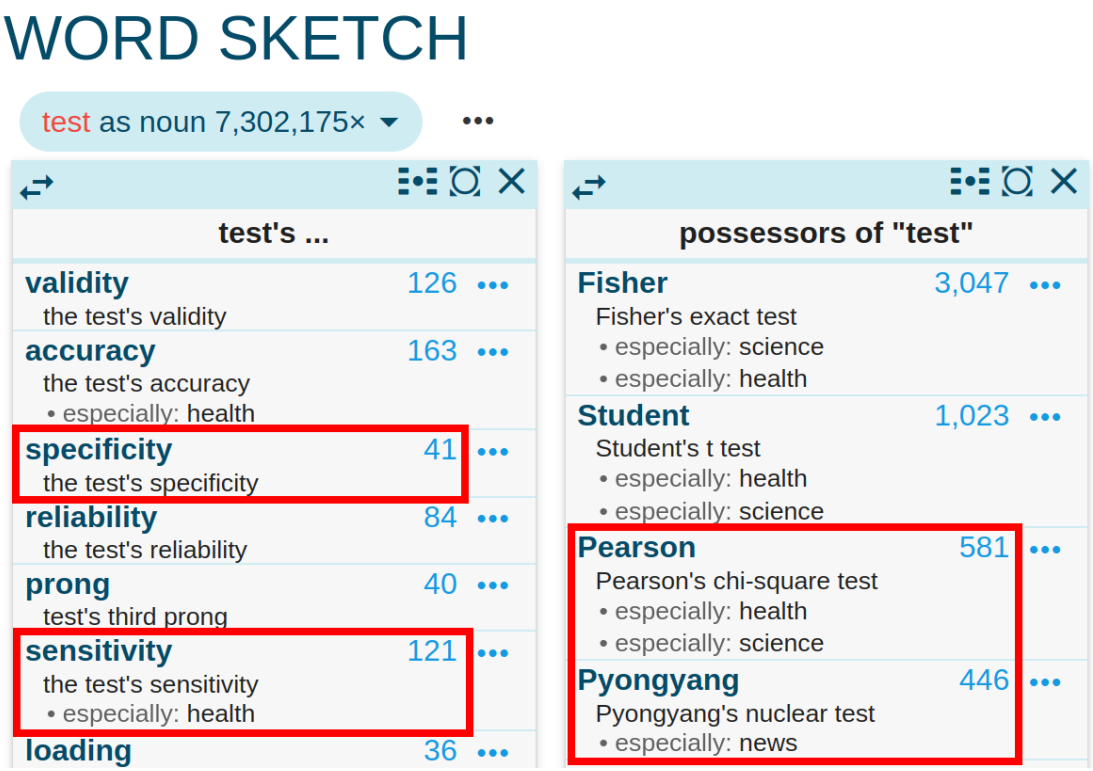
In this paper we presented a semi-manual procedure to annotate topics and genres in large web corpora. Unlike manually inspecting each sample and training a classifier on 100 % clean data – which is the usual approach – our method relies on seeing just random sentences and a few live web pages to represent up to tens of thousands of texts. The scheme was designed to decrease the time an annotator needs to check one website.

According to the evaluation, the noise in the labelled texts is small enough to allow using the data for training a text type classifier. The classifier can be applied to documents coming from websites that were not inspected manually thus covering the whole corpus. We recommend to balance the training samples in order to keep a wide diversity of the set, e.g. by limiting the count of documents from a single web domain. We also suggest disregarding classes consisting of samples from less than 10 websites. Alternatively, one can find additional sites in the ranked list just to boost the size and variety of under-represented text types.

An example of a language analysis benefiting from text type labels produced by our method can be seen in Figure 1. Word Sketch, the report shown in the screenshot, is used by publishing houses to produce dictionaries.

The main contribution of our work is showing how to reliably annotate a large part of a web corpus with only a small human effort.

Fig. 1: Word Sketch of noun test in a large English web corpus from 2020. Frequent phrases and frequent text types are shown. The counts of co-occurrences of "test" with collocates are displayed too. Note the phrase "the test's sensitivity" is specific to topic health while the phrase "the test's specificity" is not more specific to topic health than to topic science. Indeed, health researchers seem to be more interested in the sensitivity of tests than general researchers. Also note that "Pearson chi-square tests" occur usually in texts on health or science while "Pyongyang's nuclear tests" can usually be found in the news. The information about text types in Word Sketches is appreciated by lexicographers.



Acknowledgements This work has been partly supported by the Ministry of Education of CR within the Lindat Clarin Center. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

References

1. Kilgarriff, A., Grefenstette, G.: Introduction to the special issue on the web as corpus. Computational linguistics **29**(3) (2003) 333–347
2. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The TenTen Corpus Family. International Conference on Corpus Linguistics, Lancaster (2013)
3. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: ten years on. Lexicography **1** (2014)

4. Suchomel, V.: Better Web Corpora For Corpus Linguistics And NLP. PhD thesis, Masaryk University (2020)
5. Suchomel, V., Kraus, J.: Website properties in relation to the quality of text extracted for web corpora. In: The Fifteenth Workshop on Recent Advances in Slavonic Natural Language Processing. (2021) 167–175
6. Papčo, R.: Topic classification for web corpora: Method comparison and crosslingual transfer. Master's thesis, Masaryk University (2022) Supervisor: V. Suchomel.
7. Brezina, V., Hawtin, A., McEnery, T.: The written british national corpus 2014 – design and comparability. Text & Talk **41**(5-6) (2021) 595–615
8. Sharoff, S.: Functional text dimensions for the annotation of web corpora. Corpora **13**(1) (2018) 65–95
9. Koppel, K., Kallas, J.: Eesti keele ühendkorpuste sari 2013–2021: mahukaim eesti-keelsete digitekstide kogu. Eesti Rakenduslingvistika Ühingu aastaraamat **18** (2022) 207–228